Computer Science and Information Systems

Published by ComSIS Consortium

Special Issue on Advances in Information Technology, Distributed and Model Driven Systems

Volume 14, Number 3
Special Issue,
September 2017

Computer Science and Information Systems

Vol 14, No 3, September 2017

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Editorial Office:
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Informatics
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
Phone: +381 21 458 888; **Fax:** +381 21 6350 458
www.comsis.org; Email: comsis@uns.ac.rs

Volume 14, Number 3, 2017
Novi Sad

Computer Science and Information Systems

Special Issue on Advances in Information Technology, Distributed and Model Driven Systems

ISSN: 1820-0214 (Print) 2406-1018 (Online)

The ComSIS journal is sponsored by:

Ministry of Education, Science and Technological Development of the Republic of Serbia
http://www.mpn.gov.rs/

Computer Science and Information Systems

AIMS AND SCOPE

Computer Science and Information Systems (ComSIS) is an international refereed journal, published in Serbia. The objective of ComSIS is to communicate important research and development results in the areas of computer science, software engineering, and information systems.

We publish original papers of lasting value covering both theoretical foundations of computer science and commercial, industrial, or educational aspects that provide new insights into design and implementation of software and information systems. In addition to wide-scope regular issues, ComSIS also includes special issues covering specific topics in all areas of computer science and information systems.

ComSIS publishes invited and regular papers in English. Papers that pass a strict reviewing procedure are accepted for publishing. ComSIS is published semiannually.

Indexing Information

ComSIS is covered or selected for coverage in the following:

- Science Citation Index (also known as SciSearch) and Journal Citation Reports / Science Edition by Thomson Reuters, with 2016 two-year impact factor 0.837,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Summon (Serials Solutions),
- EBSCO bibliographic databases,
- IET bibliographic database Inspec,
- FIZ Karlsruhe bibliographic database io-port,
- Index of Information Systems Journals (Deakin University, Australia),
- Directory of Open Access Journals (DOAJ),
- Google Scholar,
- Journal Bibliometric Report of the Center for Evaluation in Education and Science (CEON/CEES) in cooperation with the National Library of Serbia, for the Serbian Ministry of Education and Science,
- Serbian Citation Index (SCIndeks),
- doiSerbia.

Information for Contributors

The Editors will be pleased to receive contributions from all parts of the world. An electronic version (MS Word or LaTeX), or three hard-copies of the manuscript written in English, intended for publication and prepared as described in "Manuscript Requirements" (which may be downloaded from http://www.comsis.org), along with a cover letter containing the corresponding author's details should be sent to official journal e-mail.

Criteria for Acceptance

Criteria for acceptance will be appropriateness to the field of Journal, as described in the Aims and Scope, taking into account the merit of the content and presentation. The number of pages of submitted articles is limited to 20 (using the appropriate Word or LaTeX template).

Manuscripts will be refereed in the manner customary with scientific journals before being accepted for publication.

Copyright and Use Agreement

Computer Science and Information Systems

Volume 14, Number 3, Special Issue, September 2017

CONTENTS

Editorial
Guest Editorial: Advances in Distributed Computing and Data Analysis
Guest Editorial: Advances in Information Technology
Guest Editorial: Model Driven Approaches in System Development

Special Section: Advances in Distributed Computing and Data Analysis

579 **Imbalanced Data Classification Based on Hybrid Resampling and Twin Support Vector Machine**
Lu Cao, Hong Shen

597 **Promising Techniques for Anomaly Detection on Network Traffic**
Hui Tian, Jingtian Liu, Meimei Ding

611 **BHyberCube: a MapReduce Aware Heterogeneous Architecture for Data Center**
Tao Jiang, Huaxi Gu, Kun Wang, Xiaoshan Yu, Yunfeng Lu

629 **Click-Boosted Graph Ranking for Image Retrieval**
Jun Wu, Yu He, Xiaohong Qin, Na Zhao, Yingpeng Sang

643 **A Weighted Mutual Information Biclustering Algorithm for Gene Expression Data**
Yidong Li, Wenhua Liu, Yankun Jia, Hairong Dong

661 **An Optimization Scheme for Routing and Scheduling of Concurrent User Requests in Wireless Mesh Networks**
Zhanmao Cao, Chase Q. Wu, Mark L. Berry

Special Section: Advances in Information Technology

685 **Construction of Affective Education in Mobile Learning: The Study Based on Learner's Interest and Emotion Recognition**
Haijian Chen, Yonghui Dai, Yanjie Feng, Bo Jiang, Jun Xiao, Ben You

703 **A Retrieval Algorithm of Encrypted Speech based on Syllable-level Perceptual Hashing**
Shaofang He, Huan Zhao

719 **A Novel Link Quality Prediction Algorithm for Wireless Sensor Networks**
Chenhao Jia, Linlan Liu, Xiaole Gu, Manlan Liu

735 **Connected Model for Opportunistic Sensor Network Based on Katz Centrality**
Jian Shu, Lei Xu, Shandong Jiang, Lingchong Meng

751 **An Improved Artificial Bee Colony Algorithm with Elite-Guided Search Equations**
Zhenxin Du, Dezhi Han, Guangzhong Liu, Kun Bi, Jianxin Jia

769 **A DDoS Attack Detection System Based on Spark Framework**
Dezhi Han, Kun Bi, Han Liu, Jianxin Jia

789 **A kernel based true online Sarsa(λ) for continuous space control problems**
Fei Zhu, Haijun Zhu, Yuchen Fu, Donghuo Chen, Xiaoke Zhou

805 **Social evaluation of innovative drugs: A method based on big data analytics**
Genghui Dai, Xinshuang Fu, Weihui Dai, Shengqi Lu

823 **Sentiment information Extraction of comparative sentences based on CRF model**
Wei Wang, Guodong Xin, Bailing Wang, Junheng Huang, Yang Liu

839 **Distinguishing Flooding Distributed Denial of Service from Flash Crowds Using Four Data Mining Approaches**
Bin Kong, Kun Yang, Degang Sun, Meimei Li, Zhixin Shi

857 **Building a Lightweight Testbed Using Devices in Personal Area Networks**
Qiaozhi Xu, Junxing Zhang

Special Section: Model Driven Approaches in System Development

875 **Supporting the platform extensibility for the model-driven development of agent systems by the interoperability between domain-specific modeling languages of multi-agent systems**
Geylani Kardas, Emine Bircan, Moharram Challenger

913 **Towards OntoUML for Software Engineering: Transformation of Kinds and Subkinds into Relational Databases**
Zdeněk Rybola, Robert Perg

939 **Development of Custom Notation for XML-based Language: a Model-Driven Approach**
Sergej Chodarev, Jaroslav Porubän

EDITORIAL

This, the third issue of Volume 14 of the Computer Science and Information Systems journal, consists of three special sections:

1. **Advances in Distributed Computing and Data Analysis**,
2. **Advances in Information Technology**, and
3. **Model Driven Approaches in System Development**.

We thank all guest editors, authors and reviewers for the hard work and enthusiasm which were invested into preparing the current issue of our journal.

GUEST EDITORIAL

Special Section: Advances in Distributed Computing and Data Analysis

The Special Section on Advances in Distributed Computing and Data Analysis was inspired by the conference held in 2016, the 17th International Conference on Parallel and Distributed Computing, Applications and Technologies. According to the review results during conference preparation, ten papers related to the scope of this section were selected for possible inclusion. After two rounds of rigorous review process, we finally accepted six papers where each one has over 40% extension to their conference version. These papers present interesting algorithms and promising techniques in the field of Distributed Computing and Data Analysis.

In the first paper "Imbalanced Data Classification Based on Hybrid Re-sampling and Twin Support Vector Machine", a combined technique with twin support vector machine (TWSVM) was proposed to identify the minority class in imbalanced datasets. It employed over-sampling and under-sampling to balance the training data. The classification accuracy of the whole dataset can thus be improved. The efficiency of dealing with imbalanced data classification was also improved in their experiment.

The paper "Promising Techniques for Anomaly Detection on Network Traffic" discussed anomaly detection techniques based on analysis of global traffic. They introduced Principle Component Analysis-based and Diffusion Wavelets-based analysis techniques in details. After compared with various anomaly detection methods, these techniques show their outperformance in global traffic analysis for anomaly detection.

Tao Jiang et. al's paper "BHyberCube: a MapReduce Aware Heterogeneous Architecture for Data Center" proposed a new heterogeneous network, BHyberCube network (BHC), for the distributed data processing application, MapReduce. They addressed the heterogeneous nodes and scalability issues by considering the implementation of MapReduce in the existing topologies. Their simulations of BHC in multi-job injection and different probability of worker servers' communications scenarios showed that the BHC could be a viable interconnection topology in today's data center for MapReduce.

In paper "Click-Boosted Graph Ranking for Image Retrieval", Jun Wu et. al. proposed a novel click-boosted graph ranking framework for image retrieval, which addressed the limited effectiveness of the well-known semantic gap for image data. This framework consisted of two coupled components. The first one was a click predictor based on matrix factorization with visual regularization, which was used to alleviate the sparseness of the click through data. The second component was a soft-label graph ranker that conducts the image ranking using the enriched click-through data noise-tolerantly. The proposed method was effective for the tasks of click predicting and image ranking.

The paper "A Weighted Mutual Information Biclustering Algorithm for Gene Expression Data" presented a novel biclustering algorithm, which is called Weighted Mutual Information Biclustering algorithm (WMIB), to discover local characteristics of gene expression data. Traditional clustering methods were difficult to deal with this high dimensional data, whose a subset of genes were co-regulated under a subset of conditions. Their algorithm applied the weighted mutual information as new similarity measure which can simultaneously detect complex linear and nonlinear relationships between genes. In experiments on yeast gene expression data, their algorithm generated larger biclusters with lower mean square residues.

In last paper of this special section, "An Optimization Scheme for Routing and Scheduling of Concurrent User Requests in Wireless Mesh Networks", Z. Cao et. al. constructed analytical network models and formulated multi-pair data transfers as a rigorous optimization problem. They proposed an optimization scheme for cooperative routing and scheduling together with channel assignment to establish a network path for each request through the selection of appropriate link patterns. Their performance superiority was illustrated in experiments on various types of mesh networks.

PDCAT is an annual international conference covering the theory, design, analysis, evaluation and application of parallel and distributed computing systems. It started from Hong Kong in 2000, followed with the great successes in Taipei, China, Kanazawa, Japan, Chengdu, China, Singapore, Dalian, China, Adelaide, Australia, Dunedin, New Zealand, Hiroshima, Japan, Wuhan, China, Gwangju, Korea, Beijing, Jeju, Korea, and then Guangzhou, China in 2016. The PDCAT 2016 had the support of Sun Yat-Sen University and IEEE Computer Society Technical Committee on Parallel Processing. The conference aims to strengthen the drive towards a close and promoted networks of different areas on the latest research problems, innovations, trends, and needs in parallel computing.

We sincerely thank to the program committee members for their support in selecting paper and especially the reviewers for their valuable comments to improve selected papers. We also thank all authors for their contribution to this special section. Special thanks are given to Prof. Mirjana Ivanović, the Editor in Chief of ComSIS, for providing us the opportunity to publish this special section, valuable comments in improving quality of selected papers, and support in the whole process.

Guest Editor
Hui Tian
University of Adelaide
Beijing Jiaotong University

GUEST EDITORIAL

Special Section: Advances in Information Technology

The special section on recent advances in information technology has attracted a wide range of articles on technology theory, applications from many aspects, and design methods of information technology. Reviewing the papers in this special section, it is clear that many diverse fields such as computer science, cloud computing, wireless sensor networks, prediction, image annotation, and storage have been involved. The articles about recent advances in information technology tackled significant recent developments in the fields mentioned above, both of a foundational and applicable character.

Also, we can easily find that most contributors regard "information technology" as synonymous with tools such as the computer, mobile phone, and tablet and such issues as instructional design, mobile learning, social networking, and open source. Through the topic's development, research designs are appropriate for studying the potential of information technology applications under controlled situations.

In this section, eleven papers have been selected for publication. All selected papers followed the same standard (peer-reviewed by at least three independent reviewers) as applied to regular submissions. They have been selected based on their quality and their relation to the scope of the special section.

In the paper entitled "Construction of Affective Education in Mobile Learning: The Study Based on Learner's Interest and Emotion Recognition" Haijian Chen et al. propose the framework of affective education based on learner's interest and emotion recognition. Learner's voice, text and behavior log data are firstly preprocessed, then association rule analysis, SO-PMI (Semantic Orientation-Pointwise Mutual Information) and ANN-DL (Artificial Neural Network with Deep Learning) methods are applied to learner's interest mining and emotion recognition.

In the paper entitled "A Retrieval Algorithm of Encrypted Speech based on Syllable-level Perceptual Hashing" Shaofang He et al. propose a syllable-level perceptual hashing-based retrieval method. Different from the existing methods, the posterior probability features based on acoustic segment models of syllable are used to generate a perceptual hashing sequence, which is then embedded into encrypted speech as a digital watermark.

The paper entitled "A Novel Link Quality Prediction Algorithm for Wireless Sensor Networks" by Chenhao Jia et al. proposes a cloud reasoning-based link quality prediction algorithm based on multiple parameters, which classifies link quality parameters according to the cloud model. This algorithm overcomes the subjectivity of link quality classification, as different link quality parameters can represent different aspects of link quality.

In the paper entitled "Connected Model for Opportunistic Sensor Network Based on Katz Centrality" Jian Shu et al. consider the central characteristics of the sink node, the connectivity of OSNs is modeled by time graph, according to the characteristics of OSNs. The experimental results show that the proposed network connectivity model can reflect the connectivity of the whole network in different scenarios.

Regarding the paper entitled "An Improved Artificial Bee Colony Algorithm with Elite-Guided Search Equations" by Zhenxin Du et al.: In order to increase the exploitation ability of the ABC elite and seek a better balance between the abilities of exploration and exploitation, an improved ABC elite (the IABC elite) algorithm is put forward in this paper, combining two novel search equation and a new parameter with ABC elite.

The paper entitled "A DDoS Attack Detection System Based on Spark Framework" by Dezhi Han et al. presents a DDoS detection system based on Spark, to ensure accuracy in detection. In the meanwhile, the time for detecting DDoS attacks is reduced and the detection efficiency is improved significantly with the advantage of Spark technology.

The paper entitled "A Kernel Based True Online Sarsa(Λ) for Continuous Space Control Problems" by Fei Zhu et al. presents TOSarsa(Λ) algorithm with the dual heuristic dynamic programming algorithm to improve policy learning speed of policy search algorithms by replacing approximating using a neural network method with approximating using the kernel method.

The paper entitled "Social Evaluation of Innovative Drugs: A Method Based on Big Data Analytics" by Genghui Dai et al. presents a Hadoop platform and explored the social evaluation method of innovative drugs based on big data analytics. It aimed to provide the supplementary information for a comprehensive review on innovative drugs, as well as to make up the defects of a regular post-marketing evaluation.

The paper entitled "Sentiment Information Extraction of Comparative Sentences Based on CRF Model" by Wei Wang et al. introduces the conditional random fields model to extraction of Chinese comparative information and focuses on the task of element extraction from comparative sentences. The conditional random fields model is employed to extract comparative elements, which fuses various lexical, syntactic and heuristic features.

The paper entitled "Distinguishing Flooding Distributed Denial of Service from Flash Crowds Using Four Data Mining Approaches" by Bin Kong et al. proposes a new method that employs data mining approaches to discriminate between DDoS attacks and FCs. Experiments are conducted to evaluate the proposed method based on two real-world datasets.

The paper entitled "Building a Lightweight Testbed Using Devices in Personal Area Networks" by Qiaozhi Xu et al. proposes the design and implementation of the prototype of PANBED, building a small-scale personal testbed for users utilizing devices in their own personal area networks (PANs). The experiment results show that PANBED allows users to set up different network scenes to test applications easily using a home router, PCs, mobile phones and other devices.

In particular, we would like to acknowledge the program committee members of Ninth International Symposium on Information Processing (ISIP 2016) and 2016 IEEE International Workshop on Trust and Security in Wireless Sensor Networks (Trust WSN 2016), in conjunction with "The 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom 2016). This section contains revised and expanded versions of selected quality papers presented at the Ninth International Symposium on Information Processing (ISIP 2016). ISIP 2016 took place on August 20-21, 2016, in Changsha, China, and was cosponsored by Shanghai Institute of Electronics, China; Jishou University, China;

Peoples' Friendship University of Russia, Russia; South China University of Technology, China; Feng Chia University, Taiwan; Henan Polytechnic University, China; Nanchang Hangkong University, China; and Jiangxi University of Science and Technology, China. In closing, we would like to take this opportunity to thank the authors for the efforts they put in the preparation of the manuscripts and in keeping the deadlines set by editorial requirements. We hope that you will enjoy reading papers from this special section as much as we did putting it together.

We would also like to thank Prof. Mirjana Ivanović, the editor-in chief of ComSIS, for her support during the preparation of this special section in the journal.

Guest Editors
Fei Yu, Peoples' Friendship University of Russia, Moscow, Russia
Chin-Chen Chang, Feng Chia University, Taichung, Taiwan
Degang Sun, Chinese Academy of Sciences, Beijing, China
Iftikhar Ahmad, King Saud University, Riyadh, Saudi Arabia
Jun Zhang, Deakin University, Burwood, Australia
Jose Maria de Fuentes, Universidad Carlos III de Madrid, Madrid, Spain

GUEST EDITORIAL

Special Section: Model Driven Approaches in System Development

The Special Section on Model Driven Approaches in System Development was inspired by the event with the same title and acronym MDASD 2016, organized during 2016 in the scope of the Federated Conference on Computer Science and Information Systems (FedCSIS) in Gdansk, Poland. After a call to the prospective authors to submit their papers, and a rigorous reviewing procedure, the same as for regularly submitted papers, we finally accepted 3 papers presenting both theoretical and practical contributions in the field of *Model Driven Software Engineering*.

The conventional approach currently followed in the development of domain-specific modeling languages (DSMLs) for multi-agent systems (MASs) requires the definition and implementation of new model-to-model and model-to-text transformations from scratch in order to make the DSMLs functional for each different agent execution platforms. In their paper *Supporting the Platform Extensibility for the Model-Driven Development of Agent Systems by the Interoperability Between Domain-Specific Modeling Languages of Multi-Agent Systems*, Geylani Kardas, Emine Bircan, and Moharram Challenger present an alternative approach which considers the construction of the interoperability between MAS DSMLs for a more efficient way of platform support extension. The feasibility of using this new interoperability approach instead of the conventional approach is exhibited by discussing and evaluating the model-driven engineering required for the application of both approaches. Use of the approaches is also exemplified with a case study which covers the model-driven development of an agent-based stock exchange system. In comparison to the conventional approach, evaluation results show that the

interoperability approach requires both less development time and effort considering design and implementation of all required transformations.

OntoUML is an ontologically well-founded conceptual modelling language that distinguishes various types of classifiers and relations providing precise meaning to the modelled entities. The authors of the paper *Towards OntoUML for Software Engineering: Transformation of Kinds and Subkinds into Relational Databases*, Zdeněk Rybola and Robert Pergl, advocate that OntoUML has been overlooked so far as a conceptual modelling language for the platform independent model of application data. They outline the transformation of Rigid Sortal Types – Kinds and Subkinds and discuss the details of various variants of the transformation of these types and the rigid generalization sets. The result is a complete method for preserving high-level ontological constraints during the transformations, specifically special multiplicities and generalization set meta-properties in a relational database using views, CHECK constraints and triggers.

Sergej Chodarev and Jaroslav Porubän in their paper *Development of Custom Notation for XML-based Language: a Model-Driven Approach* present an approach for design and development of the custom notation for existing XML-based language together with a translator between the new notation and XML. The approach supports iterative design of the language concrete syntax, allowing its modification based on users feedback. The translator is developed using a model-driven approach. It is based on explicit representation of language abstract syntax as a metamodel, that can be augmented with mappings to both XML and the custom notation. The authors give recommendations for application of the approach and demonstrate them on a case study of a language for definition of graphs.

Guest Editor
Ivan Luković
University of Novi Sad, Serbia

Computer Science and Information Systems 14(3):579–595 https://doi.org/10.2298/CSIS161221017L

Imbalanced Data Classification Based on Hybrid Re-sampling and Twin Support Vector Machine

Lu Cao[1,3] and Hong Shen[1,2]

[1] School of data science and computer science,
Sun Yat-sen University, Guangzhou, China
caolu20001742@163.com
hongsh01@gmail.com
[2] School of Computer Science,
University of Adelaide, Australia
[3] School of Information Engineering,
Wuyi University, Jiangmen, China

Abstract. Imbalanced datasets exist widely in real life. The identification of the minority class in imbalanced datasets tends to be the focus of classification. As a variant of enhanced support vector machine (SVM), the twin support vector machine (TWSVM) provides an effective technique for data classification. TWSVM is based on a relative balance in the training sample dataset and distribution to improve the classification accuracy of the whole dataset, however, it is not effective in dealing with imbalanced data classification problems. In this paper, we propose to combine a re-sampling technique, which utilizes over-sampling and under-sampling to balance the training data, with TWSVM to deal with imbalanced data classification. Experimental results show that our proposed approach outperforms other state-of-art methods.

Keywords: over-sampling, under-sampling, imbalanced dataset, TWSVM, classification.

1. Introduction

Support vector machine (SVM) proposed by V. Vapnik et al. in 1960s is a machine learning technique based on statistical theory. SVM has excellent learning performance in the case of small samples and has been widely used in many fields such as pattern recognition, text classification and regression analysis[1-2]. SVM has a solid theoretical foundation, which is mainly embodied in three aspects: the maximum interval principle, the duality theory and the introduction of kernel function. The theory of maximum interval transforms the original problem of support vector machine to the solution of a convex quadratic programming problem. The kernel function is introduced according to the dual theory, which is used to solve the nonlinear problem. However, the high cost for training data required by SVM makes SVM not applicable for classification tasks on large datasets. A new learning method known as twin support vector machine (TWSVM), which extends a pair of parallel hyper-planes in SVM to the complex non-parallel-plane, was proposed in [3].

Compared with the traditional SVM, TWSVM has two important properties: (1) TWSVM can overcome some of the traditional SVM difficulties in dealing with data distribution, such as cross data. (2) TWSVM solves the quadratic programming problem in the quarter size of the original SVM and the constraint condition of the two programming problem does not contain all the sample points, which makes the training speed of TWSVM is remarkly less than that of traditional SVM. After TWSVM was proposed, researchers have paid close attention to how to further improve the TWSVM, thus a lot of methods have emerged in [4-7]. Although TWSVM has many advantages, it has drawbacks in dealing with imbalanced datasets directly. Imbalanced datasets exist widely in real life, such as cancer diagnosis [8], fraud detection [9] and insurance risk management [10]. The number of instances is much larger than that of the other samples, known as majority and minority class respectively. The recognition of the minority class in imbalanced datasets is greatly important to detect. Such as in the intrusion detection, the number of intrusion events must be far less than the number of normal events, but if an intrusion behavior is judged as a normal event, it may suffer serious losses. TWSVM, as a learning machine designed to optimize the performance of the whole dataset like other traditional classifiers, has low performance for minority class.

In this paper, we propose to combine a hybrid re-sampling technique, which utilizes over-sampling and under-sampling to balance the training data, with TWSVM to improve the recognization rate of the minority class samples in imbalanced datasets. The paper contains three technical components: (1) We present a hybrid re-sampling method to balance the training data by inserting synthetic points into minority classes with the over-sampling technique SMOTE (synthetic minority over-sampling technique) and simultaneously deleting samples carrying little information or noise from majority classes with the under-sampling technique OSS (one side selection). (2) As a new application of TWSVM, we show how to combine the above re-sampling technique with TWSVM to solve the imbalanced datasets classification problem. (3) We conduct extensive experiments to show the effectiveness of the proposed method in comparison with other state-of-art methods in term of *F-measure* and *G-mean*.

The rest of this paper is organized as follows. Section 2 presents the existing imbalanced datasets classification methods. Section 3 describes TWSVM theory and Section 4 introduces our approach to solve the imbalanced datasets classification problem. Section 5 compares the performance of the proposed approach with the existing methods. Finally in section 6, we conclude this paper and indicate our future work.

2. Related Work

A lot of research works have been carried out in the domestic and foreign scholars on the problem of imbalanced classification [11-13]. At present, the existing class imbalance classification methods can be simply categorized into two groups: data level strategy and algorithm level strategy. The data level approaches balance the training dataset of the classifier by re-sampling techniques, while the algorithmic approaches aim

to bias the learning process to enlarge the minority class domination. The two approaches are independent of each other and can be combined.

The data level approach is to resample imbalanced datasets, including under-sampling and over-sampling. The idea of re-sampling is to increase or decrease samples of balance datasets, in order to reduce adverse effects brought by the imbalanced datasets for classifiers. The simplest re-sampling method is to increase or decrease samples randomly, but the effect is not ideal [14]. People are inclined to a heuristic method. Synthetic minority over-sampling technique (SMOTE) is the most common over-sampling method, which adds new synthetic samples to the minority class by randomly interpolating pairs of the closest neighbors in the minority class [15]. SMOTE algorithm generates samples regardless of the majority class and is inclined to increase the minority samples close to the borderline and over-fitting. Han et al. presented an improved strategy of SMOTE, called borderline-SMOTE [16] to solve the problem of over-fitting by generating the minority samples near the classification hyper-plane instead of all minority sample points. Whether the original SMOTE algorithm or its improved algorithm, the generated samples are not consistent with the underlying true distribution of minority class, which could inevitably introduce noise into the training sample set and distort the spatial distribution of data. In [17], Adaptive Synthetic Sampling (ADASYN) algorithm is proposed to overcome the limitation of SMOTE by generate synthetic samples for minority class according to the distribution situation. Gao et al. introduce a novel over-sampling approach, which bases on kernel density method of the minority class to get probability density function estimation to solve two-class imbalanced classification problems [18]. The samples produced by this method can meet the probability density of the minority samples, but this approach is limited by the specific classifier. Zhang et al. presented a RandomWalk Over-Sampling approach (RWO-Sampling) to balance different class samples by creating synthetic samples through randomly walking from the real data. This method keeps the minority data distribution unchanged, but it is stated by the central limit theorem and some conditions must be satisfied [19]. In [20], an over-sampling technique MDO (Mahalanobis Distance-based Over-sampling), which can reduce the risk of overlapping between different class regions, is presented to generate synthetic samples by preserving the covariance structure of the minority class instances according to the probability contours. Two probabilistic over-sampling methods, RACOG (Rapidly Converging Gibbs) and wRACOG (wrapper-based Rapidly Converging Gibbs) are proposed in [21]. Both of these two methods generate new minority samples by using the joint probability distribution of data attributes and Gibbs sampling. RACOG generate new samples based on Markov chain, while wRACOG selects the samples which are most likely to be misclassified in probability.

Under-sampling method reduces the data samples of majority class. Random under-sampling (RUS) is the non-heuristic approach to delete some of the majority samples randomly to rebalance the dataset [22]. This method is simple and easy to implement. Because of the reduction of samples, the under-random sampling technique can reduce the training time. However, the representative information samples are inclined to be lost in this method. Therefore, it is the focus of the future research to retain the samples with large information and eliminate the samples with less information. One Side Selection (OSS) [23] is a typical under-sampling strategy, which divides majority samples into four groups according to Tomek Links technology. And it deletes noise

samples and borderline samples to balance the data samples of minority class. At the same time, researchers begin to try to use clustering method to find the information samples. Yen and Lee [24] propose cluster-based under-sampling approaches, which firstly divide all the training samples into some clusters, then select the representative data as training data in the cluster to improve the classification accuracy for minority class. An adiversified sensitivity-based under-sampling approach is presented in [25] by clustering and sampling iteratively. Majority samples are clustered to obtain the distribution information and improve the diversity of sampling in this method, then a random sensitive strategy is used to select samples from each cluster, finally, a relatively balanced dataset is obtained by iteratively clustering and resampling. In [26], a one-sided dynamic under-sampling (ODU) technique which adopts all samples in the training process, and dynamically determines whether a majority sample should be used for the classifier learning is proposed to solve multi class imbalance problems. For each training sample, ODU algorithm calculates the probability that it may be selected. When the probability is greater than a random number, the sample is considered to be representative. Otherwise the sample is not used in the training process. Lin et al. [27] introduces a dynamic sampling method (DyS) for multilayer perceptrons to solve multi-class imbalance classification. This approach dynamically selects informative samples according to the probability estimated to train the multilayer perceptron. In general, the most important thing in under-sampling is how to select the sample points which are useful for classification.

In addition to data preprocessing techniques, algorithmic level methods are also very popular to handle the imbalanced classification problem. Algorithm level is mainly to improve and enhance the existing algorithm Cost-sensitive learning by setting different misclassification cost to the majority and minority datasets is an effective solution [28]. Castro et al. presents a new cost-sensitive algorithm to improve the discrimination ability of multi-layer perceptrons by learning the Levenberg-Marquadt's rule for class imbalanced problem [29]. With the development of ensemble learning technology, more and more researches introduce the ensemble learning technology to the classification of imbalanced data. People are trying to combine the re-sampling technology and integration technology to come out the imbalanced data classification problem [30-31]. The two algorithms EasyEnsemble and BalanceCascade proposed in [30] are the typical ensemble classification algorithm based on the Boosting and Bagging techniques for undersampling data processing. Chen et al. proposes a Ranked Minority Over-sampling in Boosting (RAMOBoost) algorithm, which adaptively ranks minority class samples at each learning iteration according to a sampling probability distribution [31]. This approach can adaptively shift the decision boundary toward majority and minority samples which are difficult to learn by using a hypothesis assessed procedure. [32] are very comprehensive to summarize the existing boosting and bagging algorithms for imbalanced datasets classification. In addition, Shao et al. firstly introduce an efficient weighted Lagrangian twin support vector machine (WLTSVM) by using different training points to overcome the bias phenomenon in imbalanced classification [33].

3. Comparison of SVM and TWSVM

Consider a binary classification problem in the n dimensional, training dataset $T=\{(x_1,y_1),\cdots,(x_m,y_m)\}$, where m represents the number of samples, x_i is a sample in the input space X, $y_i \in \{-1,1\}$ is the label in the output Y.

The basic idea of support vector machine is to find an optimal hyper-plane, which ensures thd maximizes the area of both sides of the hyper-plane. As for standard linear se accuracy of the classification anupport vector classification (SVC), the separating hyper-plane can be defined as:

$$f(x)=w^{\mathrm{T}}x+b=0 \tag{1}$$

By introducing the regularization term and the slack variable ξ, the optimization problem can be described as follows:

$$\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2+C\sum_{n=1}^{N}\xi_i \\ &\text{s.t.} \quad y_i(w^{\mathrm{T}}x_i+b)\geq 1-\xi_i,\xi_i\geq 0 \end{aligned} \tag{2}$$

The problem of maximizing the interval is transformed into a convex quadratic programming problem by using the dual method in convex optimization. In order to cope with the non-linear problems, SVMs use nonlinear kernels to map low dimensional feature space to high dimensional space.

TWSVM constructs two non-parallel hyper-planes, each of which is close to a class of samples and is far from the other class. Two hyper-planes are denoted as:

$$\begin{aligned} f_+(x)&=w_+^{\mathrm{T}}x+b_+=0 \\ f_-(x)&=w_-^{\mathrm{T}}x+b_-=0 \end{aligned} \tag{3}$$

Two non-parallel hyper-planes can be obtained by solving two optimization problems, and the two optimization problems can be described as:

$$\begin{aligned} &\min_{w_+,b_+,\xi_-} \ \frac{1}{2}\|X_+w_++e_+b_+\|^2+c_1e_-^{\mathrm{T}}\xi_-, \\ &\text{s.t.} \quad -(X_-w_++e_-b_+)+\xi_-\geq e_-, \qquad \xi_-\geq 0 \end{aligned} \tag{4}$$

$$\begin{aligned} &\min_{w_-,b_-,\xi_+} \ \frac{1}{2}\|X_-w_-+e_-b_-\|^2+c_2e_+^{\mathrm{T}}\xi_+, \\ &\text{s.t.} \quad (X_+w_-+e_+b_-)+\xi_+\geq e_+, \qquad \xi_+\geq 0 \end{aligned} \tag{5}$$

where $c_1>0$, $c_2>0$, X_+ and X_- are two types of samples, e_+ and e_- are column vectors, ξ_+ and ξ_- are slack variables. The dual problems of the equation (4) and (5) can be expressed as:

$$\begin{aligned} &\max_{\alpha} \ e_-^{\mathrm{T}}\alpha-\frac{1}{2}\alpha^{\mathrm{T}}\overline{X}_-\left(\overline{X}_+^{\mathrm{T}}\overline{X}_+\right)^{-1}\overline{X}_-^{\mathrm{T}}\alpha, \\ &\text{s.t.} \quad 0\leq\alpha\leq c_1e_-, \end{aligned} \tag{6}$$

$$\begin{aligned} &\max_{\gamma} \ e_+^{\mathrm{T}}\gamma-\frac{1}{2}\gamma\overline{X}_+\left(\overline{X}_-^{\mathrm{T}}\overline{X}_-\right)^{-1}\overline{X}_+^{\mathrm{T}}\gamma, \\ &\text{s.t.} \quad 0\leq\gamma\leq c_2e_+, \end{aligned} \tag{7}$$

where $\overline{X}_+=[X_+\ e_+]$, $\overline{X}_-=[X_-\ e_-]$, α and γ are Lagrange multipliers. The inverse of the matrix is solved in (6) and (7). In order to avoid the possible ill-conditioning matrix,

TWSVM introduces factor εI to make the matrix inverse solvable. By solving (6) and (7), the two hyper-planes can be obtained as:

$$\begin{aligned} z_+ &= -\left(\overline{X}_+^{\mathrm{T}}\overline{X}_+ + \varepsilon I\right)^{-1}\overline{X}_-^{\mathrm{T}}\alpha \\ z_- &= -\left(\overline{X}_-^{\mathrm{T}}\overline{X}_- + \varepsilon I\right)^{-1}\overline{X}_+^{\mathrm{T}}\gamma, \end{aligned} \tag{8}$$

where $z_k = \left[w_k^{\mathrm{T}}\ b_k\right]^{\mathrm{T}}, (k = +,-)$.

The determination of a new class of samples depends on the distance between the sample points and the two hyper-planes, which can be described as:

$$\text{Class } i = \arg\ \min_{k=+,-} \frac{\left|w_k^{\mathrm{T}} + b_k\right|}{\left\|w_k\right\|} \tag{9}$$

Like traditional support vector machines, TWSVM maps nonlinear interfaces in the original feature space to high dimensional space through the kernel function to obtain better classification results.

Figure 1 and Figure 2 are two dimensional non-cross data and cross data in the SVM and TWSVM classification effect diagram respectively. Among them, positive sample is represented by the symbol "+", while negative sample is represented by the symbol "。". Two dimensional non-cross data is shown in Figure 1. Figure 1 (a) is the classification effect chart of SVM. The classification hyper-plane can separate the two kinds of data and satisfy the maximal margin. Figure 1 (b) is the classification effect chart of TWSVM. Not the same as the traditional SVM, TWSVM eventually gets two non-parallel classification hyper-planes, in which the solid line is represented for positive class classification plane and the dashed line for negative class classification plane. Two dimensional cross data is shown in Figure 2. Figure 2 (a) is the classification effect chart of SVM. It can be seen that there is only one classification plane in linear SVM, which cannot separate the two classes of samples. In Figure 2 (b), TWSVM using two classification planes can efficiently identify two kinds of samples.

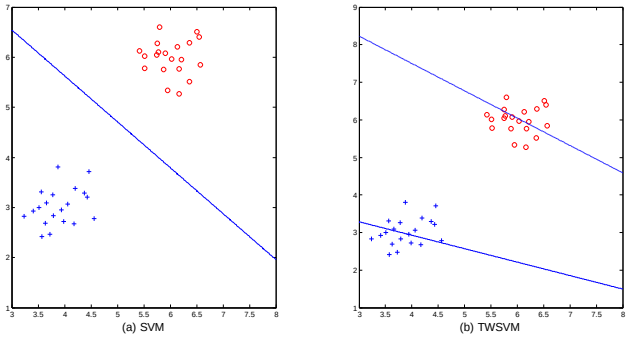


Fig. 1. Classification effect of non-crossing data in SVM and TWSVM. The symbol "+" and the circle "○" represent two kinds of sample points. The solid line in (a) is a support vector machine classification surface to satisfy the maximum interval theory. The dotted lines and solid lines in figure (b) are classification surface of two types of samples in TWSVM

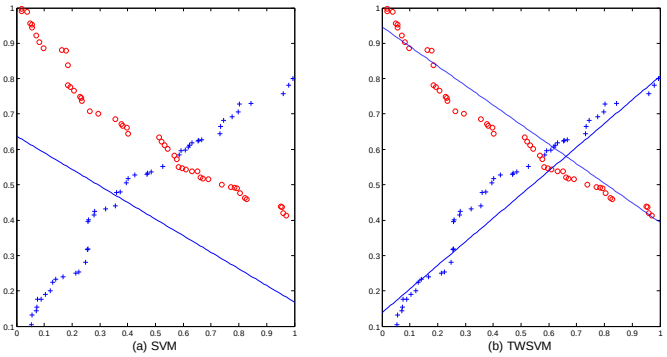


Fig. 2. Classification effect of crossing data in SVM and TWSVM. Figure (a) and (b) are classified as SVM and TWSVM respectively. For cross data, parallel hyper-plane theory of SVM cannot separate the two kinds of data, while non-parallel hyper-plane theory of TWSVM can separate the two kinds of data efficiently

4. The Proposed Approach

4.1. Re-balancing the Data

Over-sampling increases a few samples of the minority, which lead to expand the samples size, increase the training time and easily lead to over-fitting. Under-sampling deletes some samples of the majority. Although the training time is shortened, this method may remove some of the samples which are important for classification in the process of deleting the samples of the majority. In the highly imbalanced dataset, the removal of too many samples leads to serious loss of information, poor sample representation, and a serious departure from the initial data distribution. In this paper, we introduce a hybrid re-sampling technology to balance imbalanced datasets. On the one hand, SMOTE algorithm is used to synthesize new samples for minority class. On the other hand, we use the OSS algorithm to reduce the number of majority samples that have little impact on the classification.

SMOTE is an over-sampling method, the main idea of which is to insert the artificial data in a close distance between the minority samples to increase virtual samples for the minority class. The specific algorithm is as follows: as for every minority class sample of x_i, find k the nearest neighbors, then randomly select one of k neighbor as x_j, finally linear interpolate between x_i and x_j to construct a new minority sample. Figure 3 is an example of SMOTE for single sample. Three nearest neighbors of a given sample point is found. Only an artificial sample made by SMOTE is given in the graph.

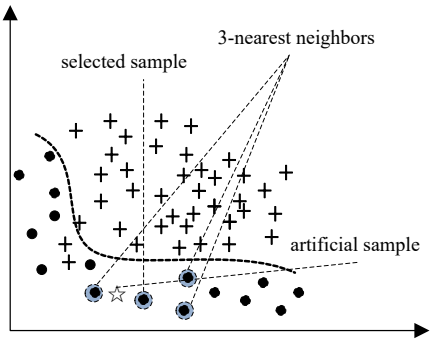


Fig. 3. Example of SMOTE for single sample, in which ● represents minority class, ＋represents majority class, ☆ represents the synthetic sample inserted. Three nearest neighbors of a given sample point is shown and the solid line represents the classification line

OSS algorithm divides majority samples into four groups: noise samples, borderline samples, redundant samples and safe samples. The noise samples are surrounded by the minority class; the borderline samples are close to the boundary; the redundant samples are which can be replaced by other majority class samples and are away from the boundary; the safety samples are which can provide valuable information for classification. Figure 4 shows four groups of samples divided by OSS algorithm specifically. OSS algorithm deletes noise samples and borderline samples to balance the data samples of minority class according to the concept of Tomek Links. Tomek Links algorithm description procedure is as follows. Given a pair of sample points with different sample labels (x_i, x_j) arbitrarily, (x_i, x_j) is called a Tomek Link if no sample x_k exists such that $\mathrm{d}(x_i, x_k) < \mathrm{d}(x_i, x_j)$ or $\mathrm{d}(x_j, x_k) < \mathrm{d}(x_i, x_j)$, where $\mathrm{d}(x_i, x_j)$ is the euclidean distance between x_i and x_j.

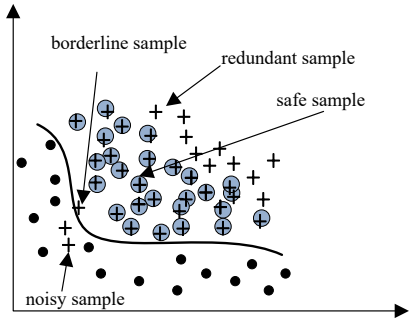


Fig. 4. Four groups of samples divided by OSS algorithm, in which ● represents minority class and ＋represents majority class, the solid line represents the classification line

Figure 5 (a) to (c) shows the effect of SMOTE and OSS. Figure 1 (a) is the distribution of original samples, figure 5 (b) and 5 (c) are distribution after SMOTE and OSS, respectively. From figure 5 we can find that SMOTE method maintains the basic distribution of the original sample, but the virtual sample increased mostly distributed in the original sample with less near the edge. We also can see that OSS method mainly keeps the sample points which are worth to classification.

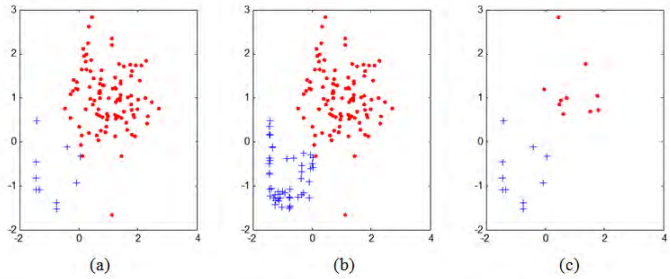


Fig. 5. Distribution graph of different sampling methods. (a) original dataset; (b) dataset distribution after SMOTE; (c) dataset distribution after OSS, in which ● and ＋represent two type of samples respectively

We introduce a hybrid re-sampling technology to balance imbalanced datasets. To be specific, we propose to apply Tomek links to the over-sampled training set as a data cleaning method to decrease over-fitting. Hybrid re-sampling approach can not only get a relatively balanced dataset, but also reduce the overlap between classes, which is beneficial to classification.

4.2. Combining Re-sampling with TWSVM

The scheme of hybrid re-sampling approach is shown as figure 6.

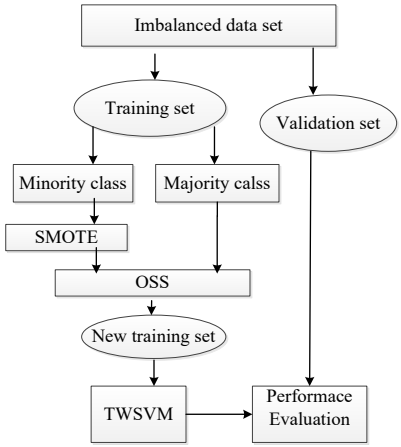


Fig. 6. Scheme of the proposed approach

The proposed approach can be summarized as follows. Firstly, the imbalanced dataset is divided into a training set and a validation set. Among them, the validation set used for cross validation accounts for 20% of the original dataset. Then, SMOTE algorithm is used to increase the minority samples and OSS algorithm is used to decrease the majority samples to get a new relatively balanced training set. Finally, a novel effective classifier TWSVM is used to train the new balanced training set and the validation set which is created initially is used to evaluate the performance of the classifier.

The algorithm for the implementation of our approach is shown in Figure 7.

Algorithm 1.

Input: Original set $\mathbf{S}=\{(\mathbf{x}_i, y_i\}$, $y_i \in \{-1,1\}$ is the label of sample , $\mathbf{x}_i \in \Re^n$, where $i \in [1,n]$

Output: p , is classification performance of TWSVM

1: $\mathbf{T} = 80\% \times \mathbf{S}$, $\mathbf{V} = \mathbf{S} - \mathbf{T}$; /* Randomly selected training set of 80% samples for training set, the rest for validation set*/

2: $\mathbf{M}$= majority($\mathbf{T}$) , $\mathbf{N}$=minority($\mathbf{T}$) /*Get the majority samples and the minority samples from $\mathbf{T}$ */

3:**for each** $\mathbf{x}_i$ in N

4: $\mathbf{x}_j = \text{K-nearset-neigbors}(\mathbf{x}_i)$ /* Find k -th nearest neighbors of $\mathbf{x}_i$ */

5: $\mathbf{x}_{new} = \mathbf{x}_i + r \cdot (\mathbf{x}_j - \mathbf{x}_i)$ /* r is a random number from $[0,1]$ */;

6:end for

7: Noisy set $\mathbf{E} = \varnothing$

8: $\mathbf{T} = \mathbf{T} + \mathbf{x}_{new}$

9: **for each** pair $(\mathbf{x}_i, \mathbf{x}_j)$ in $\mathbf{T}$

10: **if** (class ($\mathbf{x}_i$) !=class($\mathbf{x}_j$)) and

($\notin \mathbf{x}_k$ | $\text{d}(\mathbf{x}_i, \mathbf{x}_k) < \text{d}(\mathbf{x}_i, \mathbf{x}_j)$ **or** $\text{d}(\mathbf{x}_j, \mathbf{x}_k) < \text{d}(\mathbf{x}_i, \mathbf{x}_j)$)

11: $E \leftarrow E \cup \{\mathbf{x}_i, \mathbf{x}_j\}$

12: **end if**

13: $\mathbf{T}_{\mathbf{new}} = \mathbf{T} - \mathbf{E}$

14: end for

15: $model = \text{TWSVM}(\mathbf{T}_{\mathbf{new}})$

16: $p = \text{cross-validation}(model, \mathbf{V})$

17: return

Fig. 7. The algorithm for the implementation of our approach

5. Experiments and Analysis

In our experiments, all the classifiers are implemented in MATLAB 12.0. We employ LIBSVM to carry out SVMs. As for the parameters, we set $c_1 = c_2$ in TWSVM. The nearest neighbor number is 5. We focus on the comparison of our approach with SVM, SVM+OSS, SVM+SMOTE and TWSVM.

5.1. Datasets

In the following, we use eight datasets which have different degree of imbalanced from UCI datasets to verify the effectiveness of the proposed hybrid sampling method with TWSVM. UCI datasets can be obtained from http://archive.ics.uci.edu/ml/. In order to construct imbalanced datasets, we reconstruct the UCI datasets. With multiple classes of datasets, we merge some classes or just get two classes. For each dataset, the size of samples, attribution and imbalanced ratio are listed. The specific description about these datasets is summarized in Table 1. N_n and N_p denote the number of samples in the majority and the minority class respectively. Imbalance ratio is defined as N_n / N_p. From table 1, we can see that the datasets are very different in imbalance ratio.

Table 1. Datasets

Datasets	Samples (N_n / N_p)	Attributions	Imbalance ratio
Pima	768 (500/268)	8	1.87
Germen	1000 (700/300)	13	2.33
Haberman	306 (225/81)	3	2.78
Glass7	214 (185/29)	4	6.38
Satimage4	6435 (5809/626)	36	9.28
Vowel	990 (900/90)	9	10.0
Letter	200000(19266/734)	10	26.25
Yeast	1484(1440/44)	11	32.73

5.2. Evaluation Criteria

In general, it usually takes the classification accuracy as the evaluation criteria among the traditional classification methods. However, it is not reasonable to evaluate the performance of the classifier according to the classification accuracy as for the imbalanced data-sets. Because when the proportion of the minority is very low, even all the minority samples are divided into majority, the total accuracy is still very high. But this kind of classifier is not practical. For the issue of the imbalance datasets, there have already been new evaluation criteria such as $F-measure$ and $G-mean$, which are based on the confusion matrix. Confusion matrix is shown in Table 2.

Table 2. Confusion Matrix

	Predicted positive class	***Predicted negative class***
Actual positive class	TP (true positive)	FN (false negative)
Actual negative class	FP (false positive)	TN (true negative)

In this paper, we use $F-measure$ and $G-mean$ as the evaluation measure, defined as follows:

$$precision = TP \big/ (TP + FN) \tag{10}$$

$$recall = TP \big/ (TP + TN) \tag{11}$$

$$F-measure = \frac{(1+\beta^2) \times recall \times precision}{\beta^2 \times recall + precision} \tag{12}$$

$$G-mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{13}$$

where β is as a parameter and is desirable to 1 in general. The greater the value of $F-measure$, the better classification performance of minority class samples. Because $G-mean$ is based on the accuracy of both classes, it can be used to measure the overall classification performance of the system. In our paper, we utilize *F-measure* and *G-mean* to evaluate the performance of our method by comparing our method with other methods.

5.3. Results and Discussions

In our experiments, we choose Gaussian kernel and use a grid search strategy. The experiment is repeated 10 times for each dataset. Finally, we take the average of 10 experiments for the experimental results.

Table 3 shows the training time classifiers on eight benchmark datasets. From table 3, we can see that OSS algorithm takes the least amount of time, which is consistent with the theoretical analysis. As a kind of under-sampling technique, OSS selects a portion samples from the majority class to balance dataset and decreases the training time. The most time consuming method is the SMOTE algorithm because SMOTE increases the number of minority class samples. For small size datasets such as Glass7 and Haberman, the computing time of SVM and TWSVM is comparable, while for large datasets such as Letter, TWSVM is faster than SVM. Our proposed algorithm is second only to TWSVM in time consuming.

Table 3. The training time of classifiers on benchmark datasets

Datasets	SVM	SVM+OSS	SVM+SMOTE	TWSVM	Our Approach
Pima	3.157	2,154	7.413	1.458	1.947
Germen	28.543	12.422	34.415	7.457	7.498
Haberman	0.457	0.211	0.654	0.138	0.105
Glass7	0.376	0.269	0.557	0.139	0.138
Satimage4	14.675	9.447	18.123	4.116	3.779
Vowel	0.659	0.557	1.214	0.325	0.221
Letter	107.129	97.63	126.258	25.698	30.781
Yeast	0.978	0.615	1.460	0.387	0.526

Experimental results are shown as follows. Figure 8 is performance in $F-measure$ for imbalanced datasets. Figure 9 is performance in $G-mean$ for different datasets. From the experimental results, we can find that the performance of TWSVM is better than SVM in general. For different datasets, the results of SMOTE and OSS are different. On the datasets Haberman and Satimage4, the classification performance of OSS is better than that of SMOTE algorithm. The effect of SMOTE algorithm is better than OSS for other datasets. Compared with TWSVM, SMOTE, or OSS, the hybrid sampling method with TWSVM classification algorithm in this paper is optimal on $F-measure$ and $G-mean$. Specifically, on dataset with low balance rate such as Pima and German, the method proposed in this paper has a different degree of improvement in $F-measure$ and $G-mean$ compared with other algorithms. On the highly imbalanced dataset such as Letter and Yeast, the hybrid re-sampling method with TWSVM also has a good performance. The improvement of $F-measure$ and $G-mean$ shows that this method can not only improve the overall classification performance of the imbalanced data, but also improve the classification performance of the minority class.

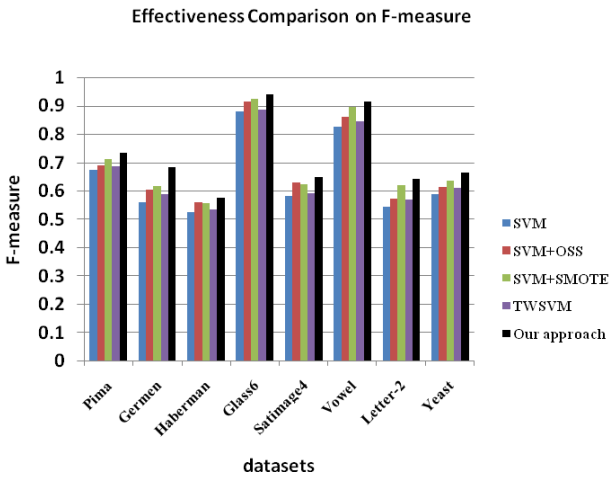


Fig. 8. Effectiveness Comparison on F-measure

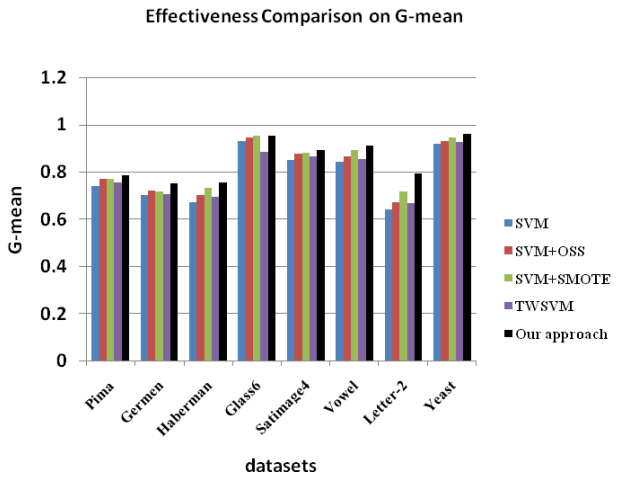


Fig. 9. Effectiveness Comparison on G-mean

6. Conclusion

There are a lot of imbalanced data in practical applications and traditional classification approaches have a low recognition rate for the minority class. An integrated sampling technique, which utilizes SMOTE algorithm and OSS algorithm to balance the training data, combined with the TWSVM classifier is proposed to deal with imbalanced data classification in this paper. As a popular classifier, TWSVM can deal with datasets which SVM is unable to handle, and its computational efficiency is much higher than SVM. Experimental results show that the method of hybrid re-sampling method with TWSVM is feasible and effective. In the experiments, we also find that SMOTE algorithm based on K nearest neighbors is limited to the range of positive samples, which will easily result in over-fitting in practical classification. Therefore, it is the next step to propose a new algorithm with good effect and fast convergence speed. At the same time, we discuss the two classification problem in this paper and multi-class imbalanced data classification is worthy of further study.

Acknowledgments. This work is supported by The 985 Project Funding of Sun Yat-sen University, Australian Research Council Discovery Projects Funding DP150104871, Youth Innovation Talent Project of Guangdong Province (No.2015KQNCX172), Science and Technology Project of Jiangmen City (No.2015[138], No.2016[189]) and Youth Foundation of Wuyi University (No.2015zk11).

References

1. Vapnik, V. N.: The Nature of Statistical Learning Theory. Springer, New York, NY, USA. (1995)
2. Deng, N.Y., Tian, Y. J., Zhang, C. H.: Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions. CRC Press. (2012)
3. Khemchandani, R., Chandra, S.: Twin support vector machines for pattern classification. IEEE Transactions on Pattern analysis and machine Intelligence, Vol. 29, No.5, 905-910. (2007)
4. Shao, Y. H., Zhang, C. H., Wang, X. B., Deng, N. Y.: Improvements on twin support vector machines. IEEE Transactions on Neural Networks, Vol.22, No.6,962-968. (2011)
5. Kumar, M. A., Gopal, M.: Least squares twin support vector machines for pattern classification.Expert Systems with Applications, Vol.36, No.4,7535-7543. (2009)
6. Peng, X.: TPMSVM: A novel twin parametric-margin support vector machine for pattern recognition. Pattern Recognition, Vol.44, No.10, 2678-2692. (2011)
7. Shao, Y. H., Chen, W. J., Deng, N.Y.: Nparallel hyperplane support vector machine for binary classification problems. Information Sciences, Vol. 263, No. 3,22–35.(2014)
8. Petrick, N., Chan, H. P., Sahiner, B., Wei, D.: An adaptive density-weightedcontrast enhancementfilter for mammographic breast mass detection. IEEETransactions on Medical Imaging,Vol. 15, No. 1, 59–67. (1996)
9. Fawcett, T., Provost, F.: Adaptive fraud detection. Data Mining Knowledge Discovery,Vol. 1, No. 3, 291–316. (1997)
10. Pednault, E. P. D., Rosen, B. K., Apte, C.: Handling Imbalanced Data Sets inInsurance Risk Modeling. IBM Research Report RC-21731.(2000)
11. Hulse, J. V., Khoshgoftaar, T. M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. The Twenty-fourth International Conference on Machine Learning, DBLP, 935–942. (2007)
12. Chawla, N. V., Japkowicz, N., Kotcz, A.: Editorial: Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations News letter, Vol. 6, No. 1, 1–6. (2004)
13. Wang, S., Yao, X.: MultiClass Imbalance Problems: Analysis and Potential Solutions. IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man and Cybernetics Society, Vol. 42, No. 4, 1119-1130. (2012）
14. Wang, Q.: A Hybrid Sampling SVM Approach toImbalanced Data Classification. Abstract and Applied Analysis,Vol. 2014, No.5, 22–35. (2014)
15. Chawla, N. V., Bowyer, K.W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique. Journalof Artificial Intelligence Research, Vol. 16, No. 1, 321–357. (2002)
16. Han, H., Wang, W. Y., Mao, B. H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Lecture Notes in Computer Science, Vol. 3644, Springer-Verlag, Berlin Heidelberg New York, 878–887. (2005)
17. He, H., Bai,Y., Garcia, E. A., Li, S.: Adasyn: Adaptive syntheticsampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks, 1322–1328. (2008)
18. Gao, M., Hong, X., Chen, S., Harris, C. J.: PDFOS: PDF estimation based over-sampling for imbalanced two-class problems. Neurocomputing, Vol. 138, No. 11,7535-7543. (2012)
19. Zhang. H., Li, M.: RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. Information Fusion, Vol. 20, No. 1, 99-116. (2014)
20. Abdi, L., Hashemi,S.: To combat multi-class imbalanced problems by means of over-sampling techniques.IEEE Transactions on Knowledge and Data Engineering, Vol.28, No. 1, 238-251. (2016)

21. Das, B., Krishnan, N. C., Cook, D. J.: RACOG and wRACOG: Two Probabilistic Oversampling Techniques. IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 1, 222-234. (2015)
22. He, H. B., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 9, 1263–1284. (2009)
23. Kubat, M., Matwin, S., Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In Proceedings of the 14th International Conference on Machine Learning. Nashville, Tennessee, USA, 179–186. (2000)
24. Yen, S.J., Lee, Y. S.: Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications An International Journal, Vol. 36, No. 3, 5718 – 5727. (2009)
25. Ng, W. W., Hu, J., Yeung, D. S., Roli, F.: Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems. IEEE Transactions on Cybernetics, Vol.45, No. 11, 2402-2412. (2014)
26. Fan, Q., Wang, Z., Gao, D.: One-sided Dynamic Undersampling No-Propagation Neural Networks for imbalance problem. Engineering Applications of Artificial Intelligence. Vol. 53(C), 62-73. (2016)
27. Lin, M., Tang, K., Yao, X.: Dynamic Sampling Approach to Training Neural Networks for Multiclass Imbalance Classification. IEEE Transactions on Neural Networks and Learning Systems, Vol. 24, No. 4, 647-660. (2013)
28. Zhou, Z. H., Liu, X. Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, Vol.18, No. 1, 63-77. (2006)
29. Castro, C. L., Braga, A. P.: Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. IEEE Transactions on Neural Networks and Learning Systems, Vol. 24, No. 6,888-899. (2013)
30. Liu, X. Y., Wu, J., Zhou, Z. H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics, Vol. 39, No.2, 539-550. (2009)
31. Chen, S., He, H., Garcia, E. A.: RAMOBoost: Ranked Minority Oversampling in Boosting. IEEE Transactions on Neural Networks, Vol. 21, No.10, 1624-1642. (2010)
32. Galar, M., Fernández, A., Barrenechea, E., Bustince, H.: A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems Man and Cybernetics Part C Applications and Reviews, Vol. 42, No. 4,463-484. (2012)
33. Shao, Y. H., Chen, W. J., Zhang, J. J., Wang, Z., Deng, N. Y.: Anefficient weighted Lagrangian twin support vector machine forimbalanced data classification. Pattern Recognition, Vol. 47, No.9, 3158 – 3167. (2014)
34. University of California-Irvine, [Online]. Available: http://archive.ics.uci.edu/ml/(current April 2013)
35. Cao, L., Shen, H.: Combining Re-sampling with Twin Support Vector Machine for Imbalanced Data Classification. In Proceedings of 17th International Conference on Parallel and Distributed Computing, Applications and Technologies, Guangzhou, 325-329. (2016)
36. Tang, Y., Zhang. Y. Q., Chawla, N. V., et al.: SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics, Vol. 39, No.1, 281 – 288. (2009)
37. Liu, Y., Yu, X., Huang, J. X., et al.: Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Information Processing and Management, Vol. 47, No.4, 617-631. (2011)
38. Fu, J. H., Lee, S. L.: Certainty-based active learning for sampling imbalanced datasets. Neurocomputing, Vol. 119, No.16, 350-35. (2013)

39. Chawla, N. V., Lazarevic, A., Hall, L. O., et al.: SMOTEBoost: Improving Prediction of the Minority Class in Boosting. Lecture Notes in Computer Science, Vol. 2838, Springer-Verlag, Berlin Heidelberg New York,107-119. (2003)
40. Cateni, S., Colla, V., Vannucci, M.: A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing, Vol. 135, No.8, 32-41. (2014)
41. Sun, Y., Kamel, M. S., Wong, A. K. C., et al.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, Vol. 40, No.12, 3358-3378. (2007)
42. Woniak, M., Grana, M,, Corchado, E.: A survey of multiple classifier systems as hybrid systems. Information Fusion, Vol. 16, No.1, 3-17. (2014)
43. Akbani, R., Kwek, S., Japkowicz, N.: Applying Support Vector Machines to Imbalanced Datasets. Lecture Notes in Computer Science, Vol. 3201, Springer-Verlag, Berlin Heidelberg New York, 39-50.(2004)

This paper is a rewritten and extended version of an earlier conference paper [35] presented at the 17th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2016)

Lu Cao received the B.S. degrees in Electrical Information Engineering from Changjiang University in 2004 and the M.S degree in Sun Yat-sen University in 2006. She joined Wuyi University since 2006. She is currently working towards her Ph.D. in Sun Yat-sen University since 2014. Her research interests include machine learning, pattern recognition and their applications in imbalanced problems.

Hong Shen is Professor (Chair) of Computer Science in University of Adelaide, Australia, and ”1000 People Plan” Professor and Director of Advanced Computing Institute in Sun Yat-Sen University, China. He received Ph.Lic. and Ph.D. degrees from Abo Akademi University, Finland, M.Eng. degree from University of Science and Technology of China, and B.Eng. degree from Beijing University of Science and Technology, all in Computer Science. He was Professor and Chair of the Computer Networks Laboratory in Japan Advanced Institute of Science and Technology (JAIST) during 2001-2006, and Professor (Chair) of Compute Science at Griffith University, Australia, where he taught 9 years since 1992. With main research interests in parallel and distributed computing, algorithms, data mining, privacy preserving computing and high performance networks, he has published more than 300 papers including over 100 papers in international journals such as a variety of IEEE and ACM transactions. Prof. Shen received many honors/awards including China National Endowed Expert of “1000 People Plan” (2010) and Chinese Academy of Sciences “Hundred Talents” (2005). He served on the editorial board of numerous journals and chaired several conference.

Received: December 21, 2016; Accepted: May 15, 2017.

Computer Science and Information Systems 14(3):597–609 https://doi.org/10.2298/CSIS170201018H

Promising Techniques for Anomaly Detection on Network Traffic

Hui Tian[1,2], Jingtian Liu[1] and Meimei Ding[1]

[1] School of Electronics and Information Engineering,
Beijing Jiaotong Univeristy
tianhh@gmail.com, {16120019, 15120019}@bjtu.edu.cn
[2] School of Computer Science,
University of Adelaide

Abstract. In various networks, anomaly may happen due to network breakdown, intrusion detection, and end-to-end traffic changes. To detect these anomalies is important in diagnosis, fault report, capacity plan and so on. However, it's challenging to detect these anomalies with high accuracy rate and time efficiency. Existing works are mainly classified into two streams, anomaly detection on link traffic and on global traffic. In this paper we discuss various anomaly detection methods on both types of traffic and compare their performance.

Keywords: diffusion wavelet, principal component analysis, anomaly detection.

1. Introduction

Traditional studies on network traffic mainly focused on single-link traffic analysis in temporal domain within an ISP network. At present, researchers have made great progress in the research on self-similar stochastic processes, long-range dependence, heavy-tailed distributions, and so on. However, most of these researches focus on partial links [1] or limited number of Internet terminals, regarding network traffic as a time-domain signal. But analysis on signal link or several links only is not enough to capture traffic characteristics of the global network accurately.

Many researchers start to study the global network data in recent years. The global network data can be described by a Traffic Matrix (TM) where each component in the matrix represents an end-to-end traffic flow. There have been many anomaly detection approaches for global traffic or partial end-to-end flows in TMs, statistic-based, traditional wavelet-based, machine learning, data mining, neural network, and so on. We will introduce anomaly detection methods in both scenarios and compare their performance.

There are mainly three types of anomalies studied by existing methods. The first one is anomaly caused by link/node disconnection, which then results in the changes of topology and also the end-to-end users' TMs. The second type of anomaly is caused by DDoS attack [21], which occurs to the network in a distributed way. DDoS attack can instantly sends vast data into the target host by controlling or combine with other hosts, leading the target system to be crowded, finally making the target system paralyzed. So it is necessary to develop the approach to diagnose DDoS attacks efficiently to ensure

network security. The third type is due to the change of end-to-end user demands, which does not bring direct harm to the network. But detection on this is beneficial to network prediction and capacity plan etc. In this paper, anomaly detection targets the first two types of anomalies.

We proposed Diffusion Wavelet (DW)-based and Principal Component Analysis (PCA)-based anomaly detection methods. These two methods are efficient in detecting global traffic anomaly when TMs are available. These two techniques are based on different ideas, but both are effective for sparse matrix analysis. We will compare their performance in experiments and discuss their promising applications. The main contribution of the paper are listed as follows.

- Various anomaly detection methods are introduced and their detection accuracy rate are all analyzed.
- Two promising techniques for global network anomaly detection are given in details.
- The experiments are conducted to compare all methods and their application scenarios and performance are analyzed.

The remainder of the paper is organized as below. In 2nd section, we introduce all related works. Section 3 gives a detailed introduction on techniques used in temporal netflow for partial links. Section 4 introduces several schemes used in anomaly detection for global network traffic. DW-based technique and PCA-based technique are given in details. Section 5, we describe the test network and define a metric for comparing all algorithms' performance. Conclusion is given in the last section.

2. Related Works

Before our work in traffic data analysis, Chandola and Baerjee et. al have discussed anomaly detection techniques in different research areas and application areas in [22]. They defined the anomaly detection problem as the problem of identifying patterns in data that do not conform to a well-defined notion of a normal behavior. They listed the challenges of anomaly detection though it appears to be simple problem, which includes, anomaly pattern's vague boundary, malicious actions' fake normal behavior, availability of validated training sets, different abnormal criterion in different scenarios and so on. Therefore they've done an extensive survey on existing techniques and application domains. These included classification-based, clustering-based, information theory-based, statistics-based techniques. The application areas covered cyber-intrusion detection, fraud detection, medical anomaly detection, industrial damage detection, image processing, textual anomaly detection, and sensor networks.

Another broad review of anomaly detection techniques for numeric as well as symbolic data is presented by Agyemang et al. [23] in 2006. Hodge and Austin [24] in 2004 provided an extensive survey of anomaly detection techniques developed in machine learning and statistical domains. All of these do not cover the emerging techniques introduced in this paper which are based on Principal Component Analysis and Diffusion Wavelets respectively. Existing survey work are focused on particular application data of local view, which are different from what we include in this paper in the global/systematic view.

Existing work on detecting anomaly locally mainly set a prober in a particular position in the network. The anomaly is not hard to be detected based on local data flow analysis by using existing techniques mentioned in above survey papers or more recent papers. We study the performance of a representative method based on statistical data, Generalized Left-to-right Reduce (GLR) [2, 3, 9]. When the end-to-end traffic matrix is known which means a global view of traffic may be available for the system, it's valuable to develop methods to detect anomalies for the global traffic data. In this case, more complicated analysis is involved. Existing work include methods based on Relative Entropy [4], Sketch [5], Non-negative Matrix Factorization (NMF) [6], Principal Component Analysis (PCA) [20], Diffusion Wavelets [7, 8].

We will compare all mentioned techniques and evaluate their performance by detection accuracy rate in this paper.

3. Anomaly Detection for Temporal Traffic Flows

In this section, we briefly give an introduction on anomaly detection for temporal traffic flows. This includes mainly statistics-based anomaly detection for single-link data [2], wavelet transform-based analysis used for several monitored links [3], and Relative entropy based anomaly detection method [4]. We discuss their advantages and disadvantages, and finally introduce how these techniques could be combined with new methods for global data anomaly detection.

G. X. Jia et al. proposed an anomaly detection method on time-series network flow data [2]. They firstly study flow data with 5-minute interval to study network traffic characteristic sequence. Then, the historical sequence of anomaly degree can be figured out by computing anomaly degree on subsequent network traffic characteristic within each time window. Lastly, anomaly is detected by comparing the anomaly degree at current moment with historical data. Network anomaly can thus be detected and alarmed in time. The method is effective for Distributed Denial of Service (DDoS), Worm virus and other intrusion attacks in data flow. It works for single network flow analysis while not for global network traffic. The main benefit is it may guarantee its efficiency in monitoring the specific links, the limitation is that it cannot detect anomaly happened in other location of the networks.

Signal analysis based anomaly detection approach is proposed by P. Barford et al. in [3]. Anomaly can be detected by monitoring local variance of filtered data. Firstly, wavelet transform is applied to data flow. Then, signal in low, middle and high-frequency respectively can be obtained by comprehensive analysis on wavelet coefficients. Low-frequency signal is obtained by comprehensive analysis on coefficients of the 9th and higher layers, able to capture long-term pattern and anomaly of traffic well, whose time scale is usually a few days or weeks. Mid-frequency signal is obtained by comprehensive analysis on the 6th, 7th and 8th layers, able to capture normal change of traffic. High-frequency signal can be obtained by comprehensive analysis on the first 5 layers, able to capture short-term change of traffic. The coefficient is set to zero if its absolute value is less than threshold in the first 5 layers. Lastly, local variability for middle and high-frequency is obtained by calculating their variance within moving window respectively. So a comprehensive variable can be obtained through combining the variance of middle and high-frequency by weighted sum.

Anomaly is detected according to the variable and threshold. The method is effective for temporal data flow of limited links. It is not sensitive to other links' traffic changes.

Relative entropy based anomaly detection method is proposed by D. Y. Zhang et al. in [4]. The change of network traffic can be reflected well by information entropy. First, fractal dimension and lamination are applied to network parameters. Then, the sequence on entropy of network parameters are studied within a moving window. Lastly, anomaly can be detected by comparing relative entropy with a threshold.

There are other techniques for link data anomaly detection [14-16], which are not as timely and efficiently as listed above. All these techniques regard traffic as a one-dimensional signal in temporal domain. But in practice, many traffic volume anomalies at the link traffic level may occur at one or more links. They are often overwhelmed within normal traffic patterns, caused by the high level of traffic volume aggregation on backbone links. Therefore, it is quite hard to discover anomalies at the link level. We thus mainly analyze global traffic and detect anomaly for more complicated data collected in end-to-end terminals. These cannot be attained by above mentioned techniques which are mainly developed for one-dimension temporal data.

We study techniques for global traffic [5, 6] and compare them with two promising techniques proposed in our papers [7, 8, 20]. For anomaly detection on global traffic, time efficiency is a big concern due to the huge amount of data and the time cost in collecting and dealing with these data. Thus the complexity of the algorithms developed for global data anomaly detection is a critical criteria to measure the performance of proposed techniques. Existing techniques and our proposed techniques will be studied in the same view of accuracy rate and complexity in this paper.

4. Anomaly Detection on Global Traffic

We first introduce existing anomaly detection methods for global data, and then describe how diffusion wavelet-based technique and Principal Component Analysis (PCA)-based technique are applied in global traffic data analysis for anomaly detection.

4.1. Existing Anomaly Detection Methods

Sketch based network anomaly detection approach is proposed in [5]. A. Li et al. build compact summaries of the traffic data using the notion of sketches. They record the key network traffic information into summary data structure in every circle online. An IP address traceability network anomaly detection method is proposed in this work. First, network traffic is represented as sketch information in every circle. Forecast value is produced by Exponentially Weighted Moving Average (EWMA) in every circle. Then, error sketch between the observed value and the forecast value can be figured out. Lastly, anomaly is detected according to mean standard deviation of the error sketch. The approach can be well applied to DDoS attack.

Non-negative Matrix Factorization (NMF) based anomaly detection method is given in [6]. X. Wei et al. firstly applied non-negative subspace method to TMs. They then reconstruct TM and compute the reconstruction error. Lastly, anomaly is detected by

employing Shewhart control chart based on reconstruction error. The method is applicable to one or multi-dimensional data.

4.2. DW-Based Anomaly Detection Method

Diffusion Wavelet (DW) based anomaly detection method is explored in [7, 8]. DW is an effective Multi-Resolution Analysis tool suitable for the global TMs. In comparison with traditional wavelets, DW has the following advantages. Firstly, TMs are sparse matrices, a small number of DW coefficients preserves the most energy of the original TM. TMs' property can be represented by this small number of DW coefficients efficiently. Secondly, the error between the reconstructed TM and original TM is in order of 10^{-9} which is extremely small. Thirdly, DW is closely related to network topology, because the diffusion operator is obtained by the Laplacian of network adjacency matrix. The transformed matrices imply the information of network topology. This allows DW to be developed in applications of anomaly localization.

The TM is in high dimension usually. By doing DW transform, the original TM can be resolved into matrices in low dimension, where *V* denotes the approximation matrix and *W* the detail matrix after DW transform. We apply 2-dimension DW to TM analysis. W. Willinger etc. have proved that 15% of the DW coefficients can preserve over 90% of the TM energy on average in [10], therefore the characteristic of the original TM can be represented accurately by a small quantity of coefficients, which are considered as the most significant metrics. Because of this important property of the TM matrix, sparsity, DW-based methods and PCA-based methods can work efficiently.

The ratio of the energy retained in the coefficient matrix to the energy of the original TM is defined as *energy proportion P*:

$$P = \frac{CEnergy}{TMEnergy} = \frac{\sum_{i=1}^{n} |\lambda_i|^2}{\sum_{j=1}^{m} |\lambda_j|^2} \tag{1}$$

TMEnergy represents the energy of the original TM and *CEnergy* corresponds to the energy of the coefficient matrix. λ_j is the eigenvalue of the original TM and λ_i corresponds to the coefficient matrix. To explore energy proportion occupied by various levels, we conduct experiments where 20000 TMs are used. The energy proportion occupied by the 1st~5th level approximate coefficient matrices are denoted by P_{VV1}, P_{VV2}, P_{VV3}, P_{VV4}, and P_{VV5} respectively. P_{WW4} and P_{WW5} correspond to the energy proportions captured respectively by the 4th level detail coefficient matrix C_{WW4} and the 5th level detail coefficient matrix C_{WW5}. It is obtained that $P_{VV1} = P_{VV2} = P_{VV3} = 1$, the average of P_{VV4}, P_{VV5},P_{WW4}, P_{WW5} is 0.7231,0.5475, 0.2299, 0.1243 respectively. Thus, as the level increases, the energy proportion is smaller and smaller. The energy of approximate coefficient matrix is larger than detail coefficient matrix at the same level. It is observed that the energy of *VV1, VV2* and *VV3* is the same as the original TM, which means *VV1, VV2* and *VV3* lose no eigenvalues and can't reduce dimensionality. So it can be concluded that the 4th level approximate coefficient matrix contains the most characteristics of the original TM, reducing the dimension at the same time. Therefore, we study the 4th level approximate coefficient matrix C_{VV4} mainly which does not lose any critical information of the original TM. We also study higher levels

where the approximate matrix is in lower dimension and contain useful information though.

In paper [7], we studied different diffusion operators and their applications in diffusion wavelets. In paper [8] we proposed a method where the anomalies of single-node disconnection and Distributed Denial of Service (DDoS) attack can be detected by using the anomaly degree based on DW coefficients. Based on the analysis results by DW, Hurst index and dynamic thresholds are used to improve the detection performance. Hurst index can be regarded as the most significant parameter to reflect the self-similarity characteristic of network traffic. This procedure is shown in Fig. 1.

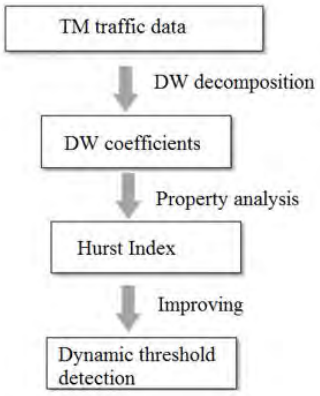


Fig. 1. Procedure in DW-based technique

Dynamic threshold may be further combined with long-range dependence (LRD) [18] with short-range dependence (SRD) [19] in order to improve high detection rate and low false alarm rate. In this case, a higher computation complexity would be involved. Thus, all detection techniques have to be find a balanced traded-off between the accuracy rate and the complexity. The anomaly detection performance of DW-based methods will be compared with all other methods in the following section.

4.3. PCA-Based Anomaly Detection Method

Similar to DW, PCA also works well for a sparse TM in large dimension. PCA is an algorithm for dimension reduction and multivariate analysis. It was first applied in data compression, image processing, neural networks, data mining, and pattern recognition. The widespread use of PCA is mainly due to its three significant characteristics. First, after high-dimensional data is compressed into a set of low-dimensional data, the mean square error of the reconstructed data is inversely proportional to the dimension. Second, the model is stable without adjusting parameters in the process. Third, for given parameters, compression and decompression are easy to conduct.

We proposed a PCA-based method to detect anomalies in [20]. In this paper, we showed that the PCA-based approaches can carry out an effective analysis of OD flows by separating network traffic into a normal subspace and an abnormal subspace. Based on the analysis results, we developed a novel detection method for node disconnection

and DDoS attacks in a backbone network by selecting two significant parameters from OD flows. This approach is able to detect not only single-node anomalies but also multi-node anomalies by parameter improvement, with a high accuracy rate and a low false-alarm rate.

For a TM where each component represents an OD traffic flow, PCA works as a coordinate transformation scheme [17] that maps a given high-dimensional set of samples onto new axes, called principal axes or principal components. The principal components have the following features. The first principal component lies in the direction of maximum variance of the samples. The second principal component corresponds to the direction of maximum variance in the remaining data, except for the variance represented by the first component. The other principal components obtain the maximum variance within the remaining data. All these principal components are orthogonal. Thus, the principal axes are sorted by the amount of data variance that they capture, in descending order.

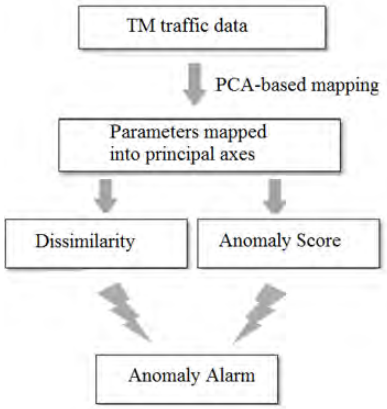


Fig. 2. Procedure in PCA-based technique

As in Fig. 2, we then define two parameters: Dissimilarity, $\overline{d}$ and Anomaly Score, R. R represents the degree to which the projection of the sample to the 1st principal axis deviates from the mean state $R_i = \frac{|u_{1,i}|}{\frac{1}{M}\sum_{j=1}^{M} u_{1,j}}$ where i, j=1,2, ⋯, 12. $u_{1,i}$ denotes the projection of a sample at the i-th instant onto the 1st principal axis. The parameter $\overline{d_i}$ is defined as the mean value of the dissimilarity between the sample at time i and any other sample, given by $\overline{d_i} = \frac{1}{M-1}\sum_{j=1}^{M} d(n_i, n_j)$ where i, j=1,2, ⋯, 12, $i \neq j$, M=12, and M is the number of samples in the set of experiments. Both parameters are regarded as significant parameters for detecting anomalies. By analyzing these two parameters together, the anomaly is recognized and reported.

5. Network Model and Comparison

All above methods work in different scenarios, but their performance can be compared. This section will introduce the testing network and also the metrics to measure their performance.

5.1. Network Model

To compare all above methods, we use the below Abilene network in Fig.3. The Abilene network topology [25] is the backbone network in America, where each node denotes an American state and each edge describes the amount of flow between two nodes. There is an open source for traffic flow recorded in Abilene website for research purpose. Many routing and management algorithms are tested by using data from Abilene network.

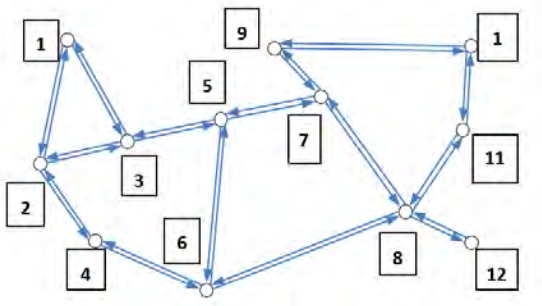


Fig. 3. Network topology of Abilene

TM describes traffic volume from one end in a network to another during regular interval, which is generally denoted by a two-dimensional data structure $T_t(i, j)$ that represents the traffic volume from node i to node j during the interval $[t, t+\Delta t)$.

The datasets used in our experiments are from 2003-2004 open data of the Abilene network, as there is no accessible data source for more recent years. Though the data source is old, DW and PCA-based methods shall show similar performance to experiments on today's data, due to the common sparsity of these datasets. They can also serve as the test data for all above methods so that their performance are comparable on the same testing data. The interval of the obtained TM in Abilene is 5 minutes [7, 8]. We collect one TM sample of size 12*12 in 5 minutes. So we get 12 TM samples in one hour duration and thus 288 samples in one day. One TM sample was presented in a 144*1 vector in PCA-based analysis domain. We collect 12 TM samples and form every 12 TM samples together to be a matrix of 144*12. If the sampling interval is 5 min, we would say the time window is in resolution of 5 min. We can also extend the sampling interval, saying, 10 minutes, 1 hour or 2 hours. 12 continuous samples are selected per group during a longer time window at a coarse resolution. When this interval is too long, large fluctuations caused by the dynamics of network

traffic may affect the detection result. If the window is too short, PCA is applied more frequently, resulting in a massive time overhead.

After having such a TM of size 144*12, we apply PCA to project this high-dimension data to 1st principal axis, 2nd axis, and so on. In most cases 1st principal axis' projection is good enough to conduct an effective detection in our experiment. We then study the similarity of projected data in low dimension as described below.

5.2. Similarity of Data

All above methods are tested on the same datasets, however, the testing results may vary if using the datasets of different days. The testing results, though comparable in performance, are found to be sensitive to the testing data no matter which method is used. The anomaly is recognized based on statistics of all past datasets, so it's important to study the property of the historical datasets.

We use the similarity to describe the property of the datasets and study their Probability Density Function (PDF). The similarity measures how similar the datasets are. If the similarity of the data is greater, the data distribution is more concentrated, depicting as a slender bell-shaped curve. Similarity of original data τ can be described by the ratio of the peak H of PDF to the width W of 98% confidence interval.

$$\tau = \frac{H}{W} \tag{2}$$

PDF is denoted by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

Where x denotes data value, μdenotes mean value of data, and σ denotes standard deviation of data. It is discovered that when x equals to μ, the maximum of probability density is achieved, that is, $H = 1/\sqrt{2\pi}\sigma$. So the value of H is related to standard deviation.

The similarity measured by τ describes the property of original data. The detection accuracy rate is supposed to be higher if the training data are more similar. This has been shown by using PCA-based method as an example in Table 1.

Table 1. Similarity and accuracy rate

L	τ_1	τ_2	τ	Accuracy Rate
1h	100	100	100	93.87%
2h	94.36	86.00	86.00	90.95%
3h	77.57	80.70	77.57	84.26%
4h	63.88	69.35	63.88	66.39%
5h	63.45	60.01	60.01	60.12%

In paper [20], Principal Component Analysis (PCA) based network anomaly detection approach applies two characteristic parameters, R and $\bar{d}$. The similarity property of the datasets is determined by the minimum of the similarity of these two

statistic parameters, which is represented by τ=min{τ_1, τ_2}. Parameters τ_1 and τ_2 are calculated based on R and $\bar{d}$ respectively, through studying on traffic volume within different moving windows. L denotes the width of moving window. The sampling interval of TM is 5 minutes. If L is larger, the data are more likely to fluctuate, which means, the similarity of the data within the window is smaller. Table 1 gives the standardized τ_1 and τ_2, and their corresponding accuracy rate.

The accuracy rate of anomaly detection varies as the original data changes. They are sensitive to the similarity of the original data as we assumed beforehand. The more similar the original data is, the easier the anomaly may be detected. Experiments on PCA-based method clearly show this. This is the same as the results demonstrated in existing methods in [5-8].

From Table 1, it is also seen that PCA based anomaly detection method gives acceptable accuracy result for data with the similarity τ_1 greater than 77.57. When the similarity of original data is too small, the accuracy rate is very low. In this case, we can reduce the window width to improve the accuracy. This, however, affects the time efficiency because the PCA analysis is running at a certain complexity within a small time period and then move forward. The anomaly detection is thus running slowly.

5.3. Comparison Metrics

Now with the same similarity of the data, we compare the detection performance of different methods in Abilene network. The results are shown in Table 2. The application scenarios show which scenarios the method is applicable to. There are four metrics used to compare different methods, computational complexity, detection rate, false alarm rate, and accuracy rate. They are defined based on four types of reports.

True Negative, TN, means the normal sample is reported to be normal.

True Positive, TP, means the abnormal sample is reported to be abnormal.

False Negative, FN, means the abnormal sample is reported to be normal.

False Positive, FP, means the normal sample is reported to be abnormal.

The detection rate counts the ratio that the anomaly is detected among all the abnormal samples, that is, *Detection Rate* = TP/(TP+FN). The false alarm rate gives the ratio that the number of FP samples to the total number of normal samples, *False Alarm Rate* = FP/(TN+FP). The accuracy rate shows the ratio that the number of correctly reported samples to the total number of samples, that is, *Accuracy Rate* = (TP+TN)/(TP+TN+FP+FN).

Table 2 gives the complexity and accuracy rate for all algorithms discussed in this paper. K denotes the dimension of data. D denotes the dimension of normal subspace and T denotes the number of iteration. Although the accuracy rate of Netflow time-series in paper [2] and signal analysis anomaly detection methods in paper [3] is higher, they only work on one-dimension data flow, not applicable to the global traffic whole network. DW-based and PCA-based detection method can be applied to one-dimension data flow and multi-dimension data. Among all methods for global traffic monitoring, PCA-based method shows the best performance in detection rate, false alarm rate and accuracy rate. Of course, it trades off the complexity, which also means it's not so timely compared to Entropy-based and NMF-based methods.

Table 2. Comparison of anomaly detection methods

Method	Complexity	Detection Rate	False Alarm Rate	Accuracy Rate	Application scenarios
Netflow Time-series Based [2]	$o(n^2)$	96.39%	4.42%	95.07%	One-dimension data Anomaly Detection
Signal Analysis Based [3]	$o(n^2)$	98.15%	5.04%	96.94%	1-dimension data anomaly detection
Relative Entropy Based [4]	$o(nK\log n)$	89.57%	4.83%	89.06%	multi-dimension data anomaly detection
Sketch Based [5]	$o(n^2\log K)$	95.07%	10.61%	90.72%	multi-dimension data anomaly detection
NMF Based [6]	$o(nKDT)$	94.36%	7.94%	90.18%	1 and multi-dimension data anomaly detection
DW Based [7, 8]	$o(n^2\log^2 n)$	93.78%	12.55%	88.78%	2-dimension traffic matrix anomaly detection and localization
PCA Based [20]	$o(Kn^2)$	100%	6.02%	94.27%	1 and multi-dimension data anomaly detection

In practice, we may select the anomaly detection method based on above comparison results. It is always wanted to have a method which is in low complexity and false alarm rate and high detection and accuracy rate. However, in practice, we have to find a trade-off among these methods according to requirements on time efficiency, computation complexity and accuracy rate.

6. Conclusion

Anomaly detection in backbone networks attract more and more attention due to the increasing concerns of network security. In this paper, we introduce various methods for anomaly detection on single-link and global traffic. Compared with single-link traffic monitoring, anomaly detection on the whole network is more meaningful, but also more difficult. Among all these methods for global traffic, we prove that DW-based and PCA-based methods [7, 8] are the most promising methods. Based on our experiments

results, they both can deal with global TM of the network powerfully. According to the same metric tested on the same test data in Abilene network, the PCA-based method shows the best performance for anomaly detection on global traffic data. But the results are sensitive to the time window. For a better accuracy, the time window cannot be too wide. It is found the samples usually should be formed within 3 hours to form its basis. This makes PCA-based method not so time efficient. DW-based methods are potential to be developed in anomaly localization. The other important parameter may be used based on DW-based analysis results, such as Hurst parameter, which shall improve its accuracy rate.

Acknowledgement. This work is supported by Research Initiative Grant of Australian Research Council Discovery Projects funding DP150104871, Beijing natural science funding No. 4172045.

References

1. Zhang Y., Ge Z., Diggavi S.: Internet Traffic and Multiresolution Analysis. Markov Processes and Related Topics. A Festschrift for Thomas G. Kurtz. Institute of Mathematical Statistics, 215-234. (2008)
2. Jia G. X., Yang B., Chen Z. X., and Peng L. Z. L.: Detecting network anomalies based on NetFlow time series. Computer Engineering and Applications. (2006)
3. Barford P., Kline J., Plonka D., and Ron A.: A signal analysis of network traffic anomalies. Proceedings of Internet Measurement Workshop, 71-82. (2002)
4. Zhang D. Y.: Network traffic anomaly detection based on relative entropy. Journal of Nanjing University of Posts and Telecommunications. (2012)
5. Li A., Han Y., Zhou B., Han W., and Jia Y.: Detecting hidden anomalies using sketch for high-speed network data stream monitoring. Applied Mathematics and Information Sciences, Vol. 6, No. 3, 759-765. (2012)
6. Wei X. L., Chen M., Zhang G. M., and Huang J. J.: NMF-NAD: detecting network-wide traffic anomaly based on NMF. Journal on Communications, Vol. 33, No. 4, 54-61. (2012)
7. Tian H., Zhong B. Z., and Shen H.: Diffusion wavelet-based analysis on traffic matrices by different diffusion operators. Computers & Electrical Engineering, Vol. 40, No. 6, 1874-1882. (2014)
8. Sun T., Tian H., and Mei X.: Anomaly detection and localization by diffusion wavelet-based analysis on traffic matrix. Computer Science and Information Systems, Vol. 12, No. 4, 1361–1374.(2015)
9. Thottan M., Ji C.: Statistical Detection of Enterprise Network Problems. Journal of Network and Systems Management, Vol. 7, No. 7, 27-45. (1999)
10. Liu B. S., Li Y. J., Hou Y. P.: The identification and correction of outlier based on wavelet transform of traffic flow. International Conference on Wavelet Analysis and Pattern Recognition,1498 - 1503.(2007)
11. Willinger W., Rincón D., Roughan M.: Towards A Meaningful MRA of Traffic Matrices. IMC Proceedings of ACM Sigcomm Conference on Internet Measurement, 331-336. (2008).
12. Wei L., Ghorbani A A.: Network Anomaly Detection Based on Wavelet Analysis. Journal on Advances in Signal Processing, Vol. 1, 1-16. (2009)
13. Coifman R. and Maggioni M.: Diffusion Wavelets. Applied and Computational Harmonic Analysis Vol. 24, No. 3, 329–353. (2008)
14. Hellerstein J., Zhang F., and Shahabuddin P.: A statistical approach to predictive detection. The International Journal of Computer and Telecommunications Networking.(2001)
15. Hamerly G., Elkan C.: Bayesian approaches to failure prediction for disk drives. ICML,1-9. (2001)

16. Shen K., Zhong M., Li C.: I/O System Performance Debugging Using Model-driven Anomaly Characterization. 4th USENIX Conference on File and Storage Technologies, 309-322. (2005)
17. Lakhina A., M. Crovella M., Diot C.: Diagnosing network-wide traffic anomalies. Computer Communication Review, Vol. 34, No. 4, 219-230. (2004)
18. Klivansky S M., Mukherjee A., Song C.: On Long-Range Dependence in NSFNET Traffic. Georgia Institute of Technology. (2000)
19. De Lima A B., Lipas M., De Mello F L.: A Generator of Tele-traffic with Long and Short-Range Dependence. International Symposium on Personal, Indoor and Mobile Radio Communications, 1-6. (2007)
20. Ding M. and Tian H.: PCA-based network traffic anomaly detection. Tsinghua Science and Technology, Vol. 21, No. 5, 500-509. (2016)
21. Ren X Y., Wang R C., Wang H Y.: Design and Realization of Software for Guard against DDoS Based on Self-Similar and Optimization Filter. Journal of China Universities of Posts and Telecommunications, Vol. 13 No. 13, 44-48. (2006)
22. Chandola V., Banerjee A., Kumar V.: Anomaly Detection for Discrete Sequences: A Survey. IEEE Transactions on Knowledge and Data Engineering, Vol. 24 No. 5:823-839. (2012)
23. Agyemang, M., Barker, K., and Alhajj R.: A comprehensive survey of numeric and symbolic outlier mining techniques. Intelligent Data Analysis Vol. 10, No. 6, 521–538. (2006)
24. Hodge, V. and Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review 22, 2, 85-126. (2004)
25. Abilene network: https://uit.stanford.edu/service/network/internet2/abileneAgrawal

Hui Tian, Associate Professor in School of Electronics and Information Engineering, Beijing Jiaotong University. She received B. Eng. and M. Eng. degrees from Xidian University, China and Ph.D. from Japan Advanced Institute of Science and Technology. Her research interests include network performance evaluation, telecommunications and wireless sensor networks.

Jingtian Liu received B.E. degree from University of Jinan, China. She is currently a master student in Beijing Jiaotong University, China. Her research interests are privacy preserving computing.

Meimei Ding received B.E. degree from Changchun University, China. She is currently a master student in Beijing Jiaotong University, China. Her research interests are network performance analysis.

Received: January 2, 2017; Accepted: May 25, 2017.

Computer Science and Information Systems 14(3):611–627 https://doi.org/10.2298/CSIS170202019T

BHyberCube: a MapReduce Aware Heterogeneous Architecture for Data Center

Tao Jiang, Huaxi Gu, Kun Wang, Xiaoshan Yu, Yunfeng Lu

State Key Laboratory of ISN, Xidian University, Xi'an, China
taojiang127@foxmail.com

Abstract. Some applications, like MapReduce, ask for heterogeneous network in data center network. However, the traditional network topologies, like fat tree and BCube, are homogeneous. MapReduce is a distributed data processing application. In this paper, we propose a BHyberCube network (BHC), which is a new heterogeneous network for MapReduce. Heterogeneous nodes and scalability issues are addressed considering the implementation of MapReduce in the existing topologies. Mathematical model is established to demonstrate the procedure of building a BHC. Comparisons of BHC and other topologies show the good properties BHC possesses for MapReduce. We also do simulations of BHC in multi-job injection and different probability of worker servers' communications scenarios respectively. The result and analysis show that the BHC could be a viable interconnection topology in today's data center for MapReduce.

Keywords: Data center, MapReduce, topology.

1. Introduction

In a data center network, up to a few thousands of servers are interconnected via switches to form the network infrastructure. Data center networks (DCN) possess the characteristic of high performance computing (HPC) and mass storage naturally [1]. Based on these properties, data center is used as distributed storage and computing infrastructures for some online applications such as search, social networks, E-learning [2], and web 2.0 technology [3]. In addition, these data centers also support infrastructure services, such as distributed file systems (e.g., GFS [4, 5] and Chubby [6]), structured storage (e.g., BigTable [7], and Megastore [8]), distributed execution engine (e.g., MapReduce, Dryad and percolator) and large computing units' schedulers (e.g., Omega [9]). Traditional resource efficient architecture has become a barrier to meet the diverse application requirements, and it is inevitable that the future network should be application driven [10]. A data center should be equipped with specific infrastructure services to manage and process massive data efficiently [11]. MapReduce is one of the most important distributed execution engines for data processing. MapReduce works by dividing input files into chunks and processing these in a series of parallelizable steps in a good control and execution model. MapReduce is used by companies such as Facebook, IBM, and Google to process or analyze massive data sets [12].

In recent years, the scale of data center is growing at an exponential rate. Some Internet service providers, like Microsoft, are even doubling the number of servers every 14 months, exceeding Moore's Law. Additionally, diverse services emerged in data centers, calls for an improvement of the topological performances of a data center network, including scalability and reliability, etc. But the current DCN interconnects all the servers using a tree hierarchy of edge-switches, core-switches or core-routers generally. It is increasingly difficult to meet the requirements, such as scalability and high network capability. As some solutions, several new DCN architectures have been proposed, such as DCell [13], FiConn [14], and BCube [15]. These architectures have optimized some fundamental topological properties and provide good scalability and reliability. However, considering the distributed data processing mechanisms running on them, they may not have good performance. There are two main reasons. First, many distributed data processing mechanisms, especially MapReduce, require that all servers being partitioned into master servers and worker servers [16]. However, most data center architectures treat all the servers equally [17]. Second, as suggested by the name, mapping and reducing constitute the essential phases for a MapReduce job. Therefore, it requires a strong inner relationship among the servers that execute these operations to exchange the intermediate results. However, these new architectures ignore this relationship in MapReduce job. Obviously, we need dedicated data center architecture to meet users' increasing new service requirements in a complex MapReduce.

In this paper, a new network, called BHyberCube network (BHC) is proposed. BHC is a recursively defined topology to interconnect servers. Each worker server connects several other worker servers in a hypercube unit and one master server. Each master server not only connects several worker servers, but also connects other master servers via a high level switch. The interconnection relationships among master servers and worker servers are determined according to the procedure of MapReduce. A high-level BHC is recursively built from many low-level ones. Due to its heterogeneous architecture, it is well suited to support the data processing procedure of MapReduce. The evaluation and analysis results show that BHC has good topological performance with scalability.

A routing algorithm designed for BHC is also proposed in this paper. This routing algorithm is designed for four scenarios for MapReduce on BHC, routing between a master worker and its worker servers, routing between two worker servers belonging to the same master server, routing between master servers and routing between two worker servers belonging to different smallest recursive units. This routing algorithm is designed to utilize the recursively-defined structure, and accelerate the procedure of MapReduce by loop iterations.

The rest of this paper is organized as follows. Section 2 introduces the related work and our motivation. Section 3 proposes the physical structure and a construction method for BHC and evaluates several topological properties of BHC. Section 4 describes the routing algorithm for MapReduce on the BHC. Section 5 shows the procedure of MapReduce executing in BHC. Section 6 presents simulation results of multi-job injection and different probability of worker servers with dependency relationship. Section 7 concludes this paper.

2. Related Work

In these section, we will introduce the existing datacenters architectures, and the details of MapReduce. The motivation will be also demonstrated.

2.1. Data Center Network Architectures

Existing datacenters generally adopt traditional tree architectures, like Fat tree [18] to interconnect servers [19] [20]. In 0, a generic Fat tree network is presented. This architecture supports a variety of links between the aggregation switches and the core switches, which makes it an architecture with high connectivity and reliability. However, this traditional tree architecture does not scale well.

Some of architectures, like DCell [13] and BCube [15], are recursively constructed, as demonstrated in Figure 1. A high-level structure utilizes a lower-level structure as a unit and connects many such units by means of a given recursive rule [21]. One of this recursive rule's advantages is that more servers can be added into a hierarchical DCN without destroying the existing structure when the level of a network is increasing. Hence, the hierarchical topology is scalable naturally.

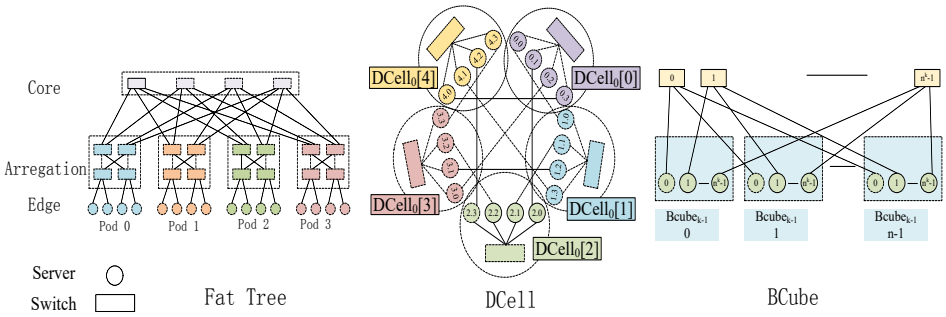


Fig. 1. Fat Tree, DCell and BCube architectures in DCN

2.2. MapReduce

The MapReduce paradigm has emerged as a highly successful programming model for large-scale data-intensive computing applications [22]. A complex MapReduce procedure processes a sequence of jobs, and each job consists of a map phase and a reduce phase [23 24].A unit based on MapReduce is composed of two server types: a master server and several worker servers. A master server controls many worker servers in executing map and reduce tasks. The master server coordinates MapReduce jobs. The worker server is responsible for running map tasks and reduces tasks. The map phase performs a map function where the master server partitions the input datasets into multiple even-sized smaller chunks and distributes them to the worker servers. Each chunk of the input is first processed by a map task, which will generate an enormous amount of intermediate (key, value) pairs on the local disks and report the keys and their

locations to the master server. The master node then partitions the (key, value) into different worker servers based on the keys. The reduce tasks will be activated to first pull the data from the map worker servers, and then apply a reduce function to the list of (key, value) pairs on each key [25]. Reduce tasks merge the intermediate values with the same key by means of predefined reduce programs and then generate the output values.

Considering implementation of MapReduce, the existing architectures addressed above, do not support the distribution data management or processing mechanisms like MapReduce very well. The reasons are as follows:

Homogeneous nodes. Existing DCN architectures do not partition servers into master servers and worker servers [26]. They simply assume that all servers possess the same function and interconnect all the servers in the same way. However, servers are classified into masters and workers based on the different functions in MapReduce [27]. Therefore, servers of different roles should be interconnected in dedicated ways.

Collective communication. In the MapReduce procedure, a master will control several worker servers, and assign different tasks to different worker servers simultaneously [28]. And among these worker servers, they will collect and transmit the intermediate information. Heavy collective operations communications happen in these phases. Hence, the topology for MapReduce should have a good performance on collective communications.

Network diameter. The topological properties should be sufficiently suitable in the DCN with the expanding of their scales. There will be a large number of data transmissions in a complex MapReduce [29]. Hence, it requires a low network diameter to shorten the transmission length between any pair of servers when the network is scaling up [30].

The BHC is motivated from the above analysis. It will treat the servers as a master server or a worker server, according to the function they will perform in the MapReduce. Leveraging the deployment of homogeneous nodes, BHC strongly supports the collective communications in the mapping and reducing phases. Furthermore, BHC employs recursive units, resulting a relatively low network diameter when the network scaling.

3. The BHC Architecture

Heterogeneous nodes, collective communications and network diameter are the main focus on network topologies proposed for DCN to support MapReduce. Servers are classified into masters and workers based on the different functions in MapReduce [27]. Servers in different roles should be interconnected in dedicated ways. If each master server interconnects its worker servers, it will improve collective communication greatly. Units are also implemented because they support collective communication naturally. The recursively defined architecture is implemented to reduce the network diameter when the scale increases.

Based on these observations, BHC is proposed for DCN to support MapReduce. BHC is a recursively-defined architecture with units attached.

3.1. BHC Architecture Specification

BHC uses servers, equipped with multiple network ports, and switches to construct its recursively defined architecture. In BHC, servers and switches are connected via communication links, which are assumed to be bidirectional. A high-level BHC is constructed from low-level BHCs. BHC_k ($k \geqslant 0$) denotes a level-k BHC. The smallest recursive unit of BHC and how to construct a high-level BHC recursively is presented as follows:

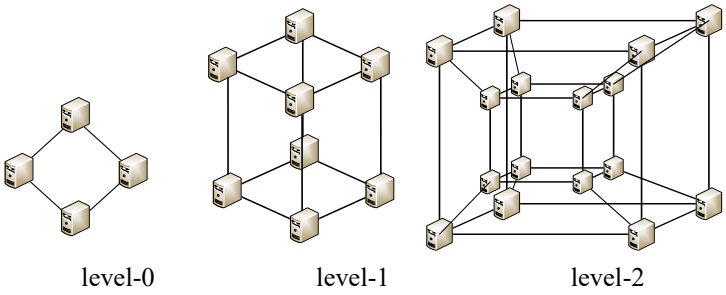


Fig. 2. level-0, 1, 2 worker units

1. Smallest recursive unit and worker units

BHC_0 is the smallest recursive unit. Meanwhile, it is also the building block to construct larger BHCs. It has W worker servers, M master servers also M switches with *P+1* ports, and *P* worker units.

The worker unit is constructed by several worker servers. These worker servers are interconnected by a *level-k* hypercube, for some different applications' requirements. 0 illustrates the *level-0, 1, 2* worker units.

In the BHC_0, each master server connects to a switch. The master servers do not connect each other directly. Neither do the switches. Each switch connects to *P* worker servers in each worker unit. So the number of worker unit in the BHC_0 is *P*.

The construction of a BHC_0 is as follows. A BHC_0 is constructed from P worker units and 2^{i+2} switches. 2^{i+2} switches are numbered from 0 to $2^{i+2}-1$. The *P* worker units are numbered from *0* to *P-1* and the worker servers in each worker unit are numbered from *0* to $2^{i+2}-1$. The *jth* ($j \in [0, P-1]$) port of the *ith* ($i \in [0, 2^{i+2}-1]$) switch is connected to the *ith* ($i \in [0, 2^{i+2}-1]$) worker servers in the *jth* ($j \in [0, P-1]$) worker unit.

There are four advantages for designing the smallest recursive unit in such a way.

Scalability. The worker unit is designed as a hypercube, which makes the worker unit increase exponentially. It means that BHC can scale the worker servers quickly and efficiently, to support different applications' requirement.

Collective communication. The worker servers can transmit intermediate information in the worker unit, which will support good collective communication performance.

Heterogeneous nodes. This recursive unit treats servers as masters and workers naturally, compared with the current recursive units, like in BCube and DCell.

Low percentage of switches. The switch connects with more worker servers, and a master server can control as many worker servers as possible. The evaluation will be presented in the following sections.

2. Bild BHC: the Procedure

As assumed above, the BHC_0 has W worker servers, M master servers also connect M switches with $P+1$ ports, and P worker units. Besides, the worker unit is assumed in the level i ($i \geqslant 0$). More generally, a BHC_k ($k \geqslant 1$) is constructed from P BHC_{k-1} and P *level-k* $P+1$-port switches.

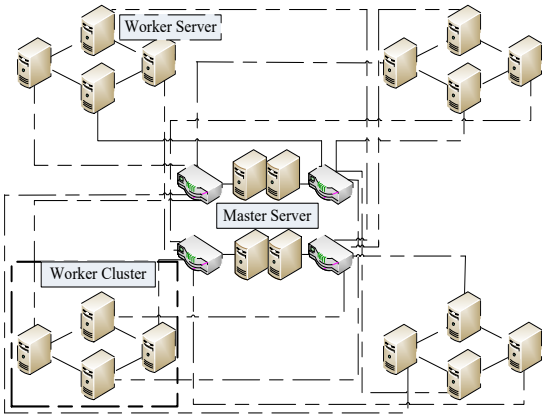


Fig. 3. An example of BHC_1 with $k=1$, $P=4$ and $i=0$

The construction of a BHC_k is as follows. BHC_{k-1} is numbered from *0* to *P-1*, the 2^{i+2} *level-k* switches from *0* to 2^{i+2}*-1* and the 2^{i+2} master servers in each BHC_{k-1} are numbered from *0* to 2^{i+2}*-1*. The *ith* ($i \in [0, 2^{i+2}-1]$) master servers in the *jth* ($j \in [0,P-1]$) BHC_{k-1} is connected to the *jth* ($j \in [0,P-1]$) port of the *ith* ($i \in [0, 2^{i+2}-1]$) *level-k* switch. 0 illustrates an example with $k=1$, $P=4$ and $i=0$, and 0 shows an example of BHC_k.

3.2. Properties of BHC

For a high-level BHC_k, it is constructed in the same way as stated above. If BHC_{k-1} has been built and each BHC_{k-1} has M^{k-1} master servers and W^{k-1} worker servers. Each BHC_{k-1} is treated as a virtual node, and fully connects these virtual nodes to form a BHC_k.

Theorem 1.

The number of master servers in a BHC_k is M_k, and $M_k = P^K 2^{i+2}$;

The number of worker servers in a BHC_k is W_k, and $W_k = P^{K+1} 2^{i+2}$

Based on the recursively defined structure of BHC, M_k depends on the P and M_{k-1}, and W_k also depends on the P and W_{k-1}. Equations 1 and 2 can be derived as follows:

The number of master servers in a BHC_k is M_K:

$$M_k = P \times M_{k-1} = \prod_{i=1}^{k} P \times M_0 = P^K 2^{i+2} \tag{1}$$

The number of worker servers in a BHC_k is W_K:

$$W_k = P \times W_{k-1} = \prod_{i=1}^{k} P \times W_0 = P^{K+1} 2^{i+2} \tag{2}$$

Theorem 1 shows that the number of worker servers and master servers scales based on the number of switch's ports and the worker unit's level. For example, when *K=3, P=6* and i=2, a BHC_3 have as many as 3456 master servers and 20736 worker servers

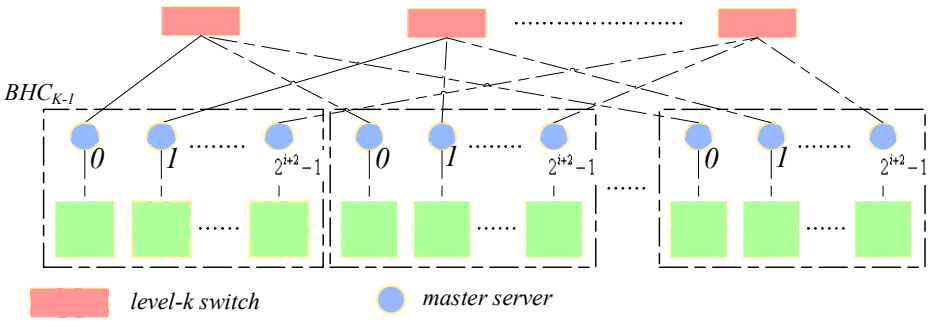


Fig. 4. BHC_k, a BHyberCube network

Bisection width denotes the minimal number of links to be removed to partition a network into parts of equal size. A large bisection width implies high network capacity and a more resilient structure against failures.

Theorem 2. The bisection width of BHC is $P \times 2^{i+1}$.

As the procedure of building a BHC_k addressed above, each of the *P level-k* switches has 2^{i+2} links connected to the *level-k-1* switches. Hence, it is easy to figure out that the bisection width of BHC is $P \times 2^{i+1}$.

Theorem 3. The network diameter of BHC is

$$3 < D_{BHC} < P + 2 \tag{3}$$

The network diameter is the longest path between any two servers. In the uniform traffic model, the maximal number of hops between any two master servers in a BHC_k is *P,* if they are in hi BHC_0*s.* For BHC_0, the maximal number of hops is 2. While in the MapReduce application, the master server distributes job chunks to the workers, and the workers process a map task to generate intermediate and report intermediate to the master. The master partitions the intermediate into workers and workers process a reduce task to generate the output values. Hence, three hops is the minimal distance to complete a job. *Theorem 3* is proven.

$Master_{K,j}$ denotes the any master server in a BHC_K, and *s* denote the sequence of a BHC_k which contains $Master_{K,j}$ in the BHC_K. The value of s is given as follows:

Theorem 4. The sequence of BHC_K $Master_{K,j}$ belongs to in the BHC_K is

$$s = j / P^k 2^{i+2} \tag{4}$$

According to Theorem 1 and Equation 1, the number of master servers is $P^K 2^{i+2}$ in a BHC_K. j is assumed as the sequence of $Master_{K,j}$ in the BHC_K. Therefore, s is the sequence of BHC_K $Master_{K,j}$ belongs to in the BHC_K.

Theorem 5. The number of master servers between $Master_{K,x}$ and $Master_{K,y}$ is

$$|x-y| = P^K 2^{i+2} \tag{5}$$

We assume that two master servers, denoted as $Master_{K,x}$ and $Master_{K,y}$, connect to the same switch at $level_{k+1}$, and belongs to a pair of adjacent BHC_ks in a BHC_{k+1}.

$|x\text{-}y|$ means the absolute value of x minus y in the Equation 5. Based on recursive rules, in this pair of adjacent BHC_ks, $Master_{K,x}$ and $Master_{K,y}$ are the only two master servers connected to the same switch at $level_{k+1}$. For a BHC_{k+1}, the number of switches at $level_{k+1}$ is $P \times 2^{i+2}$, and other master servers between $Master_{K,x}$ and $Master_{K,y}$, are connected to the other $P \times 2^{i+2} - 1$ switches. So the number of master servers between $Master_{K,x}$ and $Master_{K,y}$ is $P \times 2^{i+2} - 1$, namely $|x\text{-}y| = P \times 2^{i+2}$.

4. Routing in a BHC

According to the procedure of MapReduce in the DCN and the roles of the servers in MapReduce, the routing algorithm is designed for four scenarios [31]. The first is the routing algorithm between a master server and its worker servers, used for assigning map and reduce tasks. The second one is the routing algorithm between two worker servers that are controlled by the same master server, used for transmitting intermediate data. The third one is the routing algorithm among master servers, used for assigning jobs. The fourth one is the routing algorithm between two worker servers that belong to different smallest recursive units, used for transmitting the necessary data that are not stored on local disks. Because there are only one or two hops in the second routings, which can be addressed only in the smallest recursive unit, this paper mainly focuses on the third and fourth scenarios.

```
Algorithm 1 :AssigningJobs (int j, int L)
List ServersSought;
for  l=0; l<L; l++
 if Master_{I,y}'s worker severs hold the data for Job_l ;
   assign Job_l to Master_{I,y};
   Master_{I,y}.RoutingPath={ Master_{I,j} };
  Master_{I,y}.RoutingPath=FindRouting1(I- 1,j ,y);
  ServersSought.add(Master_{I,y} );
```

4.1. Master-to-master Routing

The routing algorithm among master servers depends on the job-assigning scheme of MapReduce service [32]. A master server that receives a multijob MapReduce request sends each job to the nearest master server, which controls the worker servers containing the necessary data for the job.

Algorithms 1 and 2 are proposed to implement the master-to-master routing algorithm for assigning MapReduce jobs on BHC. Here $Master_{I,j}$ means the *jth* master server in the BHC_I. $Master_{I,j}$ is supposed to receives a MapReduce service request, which needs to be assigned to L ($L > 1$) master servers. Algorithm 1 demonstrates the algorithm of assigning jobs. In Algorithm 1, Job_l ($0 \leqslant l \leqslant L$) denotes the jobs that need to be assigned to a master server. Algorithm 1 first finds the master server, denoted as $Master_{I,y}$ which controls the worker servers with the data required by Job_l. It then assigns Job_l to $Master_{I,y}$ and finds a routing path from $Master_{I,j}$ to $Master_{I,y}$ by citing Algorithm 2 and adds the routing path to the Path attribute of $Master_{I,y}$. Finally, it adds $Master_{I,y}$ to the object list *ServersSought*.

Algorithm 2 is designed to find a master-to-master routing path for assigning jobs. Algorithm 2 recursively records each node in the routing path from $Master_{I,j}$ to $Master_{I,y}$ from level I to level *0*, For level I, Algorithm 2 takes $Master_{I,j}$ and $Master_{I,y}$ as the source and destination nodes of the routing path, respectively. It determines if $Master_{I,j}$ and $Master_{I,y}$ connect to the same switch at level I through *Theorem 3*. Otherwise, according to *Theorem 4* and *Equation 3* and *4*, Algorithm 2 records the master server, namely $Master_{I,x}$, which not only connects to the same switch at level I with $Master_{I,j}$, but also belongs to the same BHC_{I-1} with $Master_{I,y}$. Above process is performed again for level *I-1*.with taking $Master_{I,x}$ as the new source node, also denoted as $Master_{I,j}$. This process is performed recursively until $Master_{I,y}$ is taken as the new source node or $Master_{I,j}$ and $Master_{I,y}$ belong to the same BHC_0. For the latter event, if the number of hops from $Master_{I,j}$ to $Master_{I,y}$ is larger than one, minimal master servers are further recorded in order in the routing path from $Master_{I,j}$ to $Master_{I,y}$. Otherwise, Algorithm 2 just records $Master_{I,y}$ as the last node and returns the whole routing path.

```
Algorithm 2 : FindRouting1 (int f , int j , int y)
int k = 0; int x = 0;
for  i = f ; i ⩾ 0; i--
   if  i > 0
     if j/P×2^(i+2)  ≠ y/j/P×2^(i+2);
       int h = (y-j) /P×2^(i+2);
       x = j + h ×P×2^(i+2);
       add Master_I,x to Master_I,y.RoutingPath;
       if  x == y
         return Master_I,y.RoutingPath;
      k = i ;
   break;
   if  i = 0
     if  j-y >2;
       for  x = j - 2; x>y; x-=2
         add Master_I,x to Master_I,y.RoutingPath;
       if  y -j >2
         for  x = j +2; x <y; x+=2
         add Master_I,x to Master_I,y.RoutingPath;
     add Master_I,y to Master_I,y.RoutingPath;
     return Master_I,y.RoutingPath;
FindRouting1 (k,x,y);
```

4.2. Worker-to-worker Routing

A worker server may need the data stored at another worker server, when it is executing a map or reduce task. Based on the master-to-master routing algorithm, Algorithm 3 is proposed as the worker-to-worker routing algorithm in BHC. Algorithm 3 adds two worker servers and the routing path between their master servers. Routing path can be obtained from Algorithm 2. $Worker_{j,m1}$ denotes any worker server which are controlled by $Master_{I,j}$, and $Worker_{y,m2}$ denotes any worker server controlled by $Master_{I,y}$. $Worker_{j,m1}$ and $Worker_{y,m2}$ are assumed not to be controlled by the same master server.

```
Algorithm 3: FindRouting2 (int j , int y , int m1, int m2)
  Worker_{j,m1}.RoutingPath = { Worker_{j,m1}, Master_{I,j} };
  Worker_{j,m1}.RoutingPath = FindRouting1(I-1,j,y);
  add Worker_{y,m2} to Worker_{j,m1}.RoutingPath;
  return Worker_{j,m1}.RoutingPath;
```

5. Map and Reduce on BHC

Based on routing algorithm described above, the jobs of a complex MapReduce are assigned to several master servers. These master servers will control a number of worker servers to execute the received jobs. Map and reduce operations are involved in the execution of each job. This procedure is demonstrated in Figure 5.

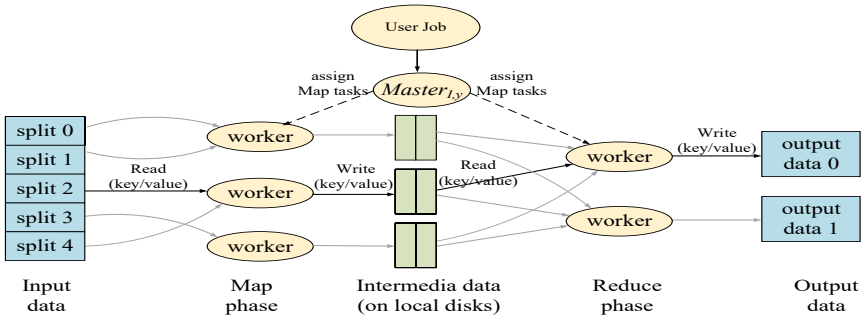


Fig. 5. The procedure of Map and Reduce on BHC

5.1. Map on BHC

Suppose that $Master_{I,y}$ receives a job. The number of map tasks is determined by the number of data chunks that job needs to process. The default mapping approach, which consists of three steps, is one map task for one data chunk. In the first step, $Master_{I,y}$ chooses some idle worker servers, named map worker servers, and assigns a map task to each of them. In the second step, map worker servers divide the corresponding input data into intermediate key/value pairs by means of predefined map programs and store

the intermediate data on local hard disks. In the third step, map worker servers feedback the keys of intermediate data to $Master_{I,y}$ and then send the number of waiting tasks in their local queues, namely their state information, to the corresponding master servers.

5.2. Reduce on BHC

The number of reduce tasks is determined by the types of intermediate data' keys. One reduce task can process one or several types of key/value pairs. But one type of key/value pairs is usually processed by only one reduce task. The default reducing approach consists of four steps. In the first step, $Master_{I,y}$ chooses some idle or not busy worker servers, named reduce worker servers, and assigns a reduce task to each of them. In the second step, according to the types of keys of their received reduce tasks, reduce worker servers fetch the intermediate data from the corresponding map worker servers. In the third step, reduce worker servers merge the same type of key/value pairs by means of predefined reduce programs to generate output values. In the fourth step, reduce worker servers feedback the output values to $Master_{I,y}$. They also send their state information to the corresponding master servers. The output data of some jobs might be the input data of other jobs. When $Master_{I,y}$ has finished its job, it sends the result directly to the master server that receives the next job, namely the next object in the object list FindedServers, which is derived from Algorithm 1. The routing algorithm between $Master_{I,y}$ and that master server can be obtained by means of Algorithm 2. The master server that executes the final job forwards its result to $Master_{I,j}$ through the routing path recorded in its Path attribute.

6. Properties and Simulation

In this section, a comparison of properties of BHC and BCube is demonstrated. Additionally, the performance of two scenarios: different numbers of jobs and different probability of communications in worker servers is shown respectively.

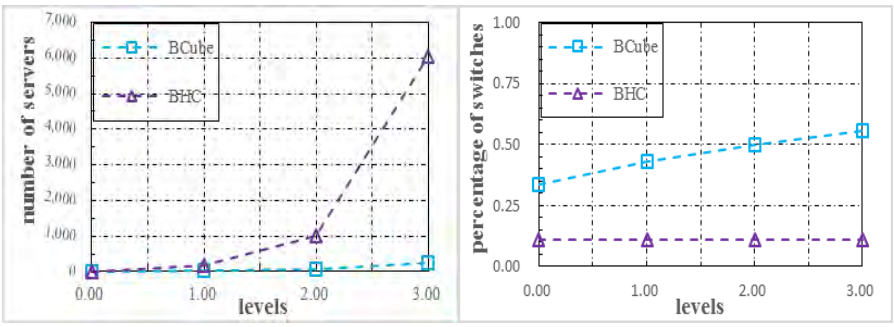


Fig. 6. The network size and percentage of switches of BCube and BHC in case that $N = 4$ servers in $BCube_0$ and $P = 8$ and $i = 0$ in BHC_i.

6.1. Comparison of Properties

Based on the Theorem 1, the network size and percentage of switches can be figured out with the level increasing. The comparison between BCube and BHC is illustrated by Figure 6.

Figure 6 shows the number of servers versus the number of levels in the network. The scalability of BHC is better than the BCube structure, when the level is higher than 3.

Figure 6 also shows the percentage of switches in BCube and BHC. The result of BHC is lower than BCube, which implies that BHC needs much less switches than BCube while the same number of servers can be connected.

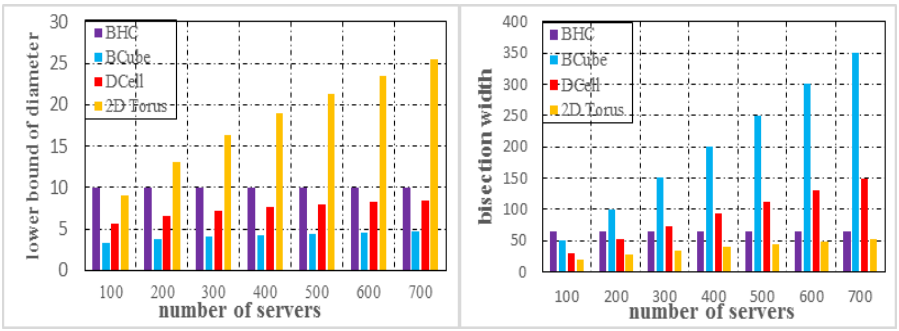


Fig. 7. The comparison of bisection width and diameter in BCube, DCell, 2D Torus and BHC in case that $N = 4$ servers in $BCube_0$ and $DCell_0$ and $P = 8$ and $i = 2$ in BHC_i.

Based on the Theorem 2 and 3, the diameter and bisection width can be figured out with the number of servers increasing. The comparison of bisection width is illustrated by Figure 7.

Figure 7 illustrates the comparison of bisection width in BCube, DCell, 2D Torus and BHC. BHC has the highest bisection width with about 100 servers. However, as network scale growing, the BCube's bisection is the highest and BHC still keeps a fixed value, and is just better than 2D Torus.

The comparison of network diameter illustrates the comparison of diameter in BCube, DCell, 2D Torus and BHC. The diameter of BHC is close to BCube and DCell, but 2D Torus gets worse when the network size grows.

Table 1. Parameters in simulations

Parameter	Value
Traffic pattern	Uniform
Switching mechanism	Wormhole
Packet length(flits)	3
Flit length(bits)	256
Cycle period(ns)	50
Number of Virtual channels	8
Offered load(flits/cycle/node)	0.01~0.4

6.2. Simulations on BHC

In this section, a simulator based on OPNET is built to evaluate the performance of BHC on MapReduce. Every simulated node is configured to use wormhole switching mechanism. The routing algorithm presented in section 4 is implemented in the simulations. We set $P = 4$, $i = 0$ and $K=2$ to build a BHC_2 with 64 master servers and 256 worker servers. Our results quantify two metrics: ETE (End to End) delay and throughput. The ETE delay is the elapsed time (in ns) between the generation of a packet at a source host and its delivery at a destination host. The throughput sum of the data rates (in Gaps) that are delivered to all terminals in a network. The simulation parameters are set as Table 1. Offered load denotes the traffic injection of each node in per cycle.

Performance of different numbers of jobs. The more jobs are injected in the network in the same time, the heavier pressure is exerted on the network. It is essential to test different numbers of jobs injected in the same time on BHC. BHC_0 is the unity to be injected. Hence, BHC_0 is chose as our test unity.

Figure 8 plots the throughput and ETE delay in different numbers of jobs. The single job's saturation point is offered load = 0.2 and the dual jobs' and four jobs' saturation point is about offered load = 0.05, which is about 1/4 of single job's saturation point. This verifies the theoretical value. The results also imply that BHC has a graceful performance even all of the master servers are injected jobs at the same time.

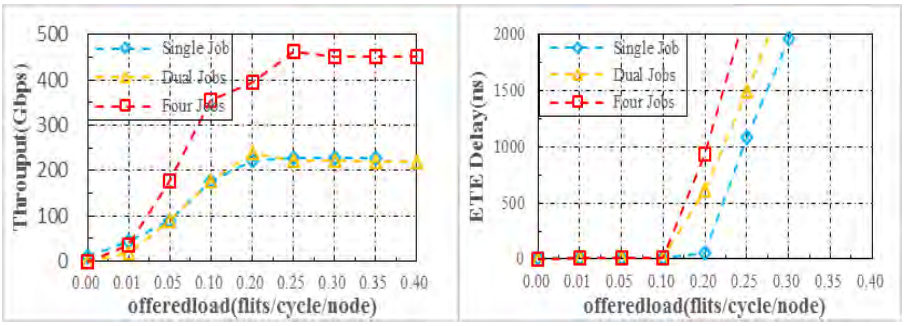


Fig. 8. The comparison of throughput and ETE delay in different numbers of jobs

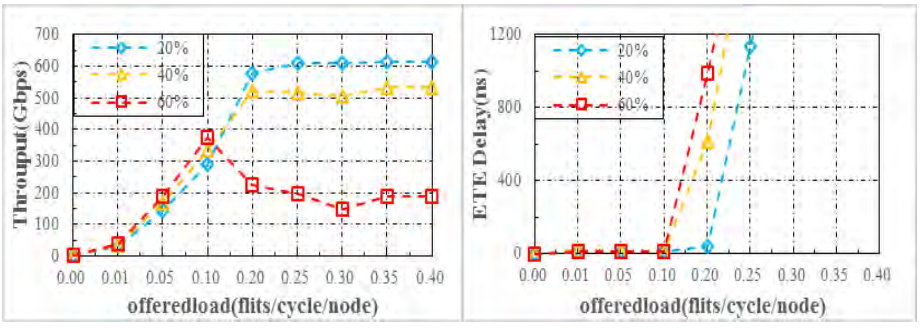


Fig. 9. The comparison of throughput and ETE delay in different P

Performance of different probability of communications in worker servers. As proposed in section IV, when a worker server is executing a map or reduce task, it may need the data stored at another worker server. *P* is supposed as the probability that a worker server needs the intermediate data stored at another worker server. If the *P* is getting higher, the more communications will happen in the worker servers. Hence, the simulation results under different *P* can be one of the metrics of worker-to-worker communication performance.

Figure 9 plots the throughput and ETE delay in different *P* of workers servers' communications. With the P growth, the performance of BHC is deteriorating. For example, when offered load is 0.3, the throughput with *P = 20%* is three times of the throughput with *P = 60%*. When *P* is getting higher, it is more probable that the a worker server, executing a map or reduce task, needs the data stored at the another worker server, and this dependency relationship will reduce the efficiency of handling a job, even turn into congestion to deteriorate the performance. Hence, this dependency relationship should be cut down as much as possible in practical, and we can improve the routing algorithm of BHC in future work.

7. Conclusion

Several new DCN architectures have been proposed to improve the topological properties of data centers, however, they do not match well with the specific requirements of some dedicated applications. This paper presents a MapReduce-supported DCN network, named BHC. Through comprehensive analysis and evaluation, BHC is a scalable topology with excellent topological properties and communication performance. It is proven that BHC is competent for MapReduce under different traffic characteristics. The simulation results show that BHC has a graceful performance in multi-job injection. But when the worker servers have a high probability (60% or higher) of dependency relationship, the performance is deteriorating because the efficiency of handling a job is dropping, even resulting in congestion. Hence, this dependency relationship should be cut down as much as possible in practical.

Acknowledgments. This work was supported by the National Science Foundation of China Grant No.61472300, the Fundamental Research Funds for the Central Universities Grant No. JB150318, the 111 Project Grant No.B08038

References

1. Dede, E., Fadika, Z., Govindaraju, M., Ramakrishnan, L.: Benchmarking MapReduce implementations under different application scenarios. Future Generation Computer Systems, 36, 389-399. (2014)
2. Klašnja-Milićević, A., Vesin, B., Ivanović, M., Budimac, Z.: E-Learning personalization based on hybrid recommendation strategy and learning style identification. Computers & Education, 56(3), 885-899. (2011)
3. Zdravkova, K., Ivanović, M., Putnik, Z.: Experience of integrating web 2.0 technologies. Educational Technology Research and Development, 60(2), 361-381. (2012)

4. McKusick, K., Quinlan, S.: GFS: evolution on fast-forward. Communications of the ACM, 53(3), 42-49. (2010)
5. Ghemawat, S., Gobioff, H., Leung, S. T.: The Google file system. In ACM SIGOPS operating systems review (Vol. 37, No. 5, pp. 29-43). ACM. (2003)
6. Burrows, M.: The Chubby lock service for loosely-coupled distributed systems. In Proceedings of the 7th symposium on Operating systems design and implementation (pp. 335-350). USENIX Association. (2006)
7. Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Gruber, R. E.: Bigtable: A distributed storage system for structured data. ACM Transactions on Computer Systems (TOCS), 26(2), 4. (2008)
8. Baker, J., Bond, C., Corbett, J. C., Furman, J. J., Khorlin, A., Larson, J., Yushprakh, V.: Megastore: Providing scalable, highly available storage for interactive services. In CIDR (Vol. 11, pp. 223-234). (2011, January)
9. Schwarzkopf, M., Konwinski, A., Abd-El-Malek, M., Wilkes, J.: Omega: flexible, scalable schedulers for large compute clusters. In Proceedings of the 8th ACM European Conference on Computer Systems (pp. 351-364). ACM. (2013,)
10. Wang, Y., Lin, D., Li, C., Zhang, J., Liu, P., Hu, C., Zhang, G.: Application Driven Network: providing On-Demand Services for Applications. In Proceedings of the 2016 conference on ACM SIGCOMM 2016 Conference (pp. 617-618). ACM. (2016)
11. Xu, K., Qu, Y., Yang, K.: A tutorial on the internet of things: from a heterogeneous network integration perspective. IEEE Network, 30(2), 102-108. (2016)
12. Fehér, P., Asztalos, M., Vajk, T., Mészáros, T., Lengyel, L. Detecting subgraph isomorphism with MapReduce. The Journal of Supercomputing, 1-42.
13. Guo, C., Wu, H., Tan, K., Shi, L., Zhang, Y., Lu, S.: Dcell: a scalable and fault-tolerant network structure for data centers. In ACM SIGCOMM Computer Communication Review (Vol. 38, No. 4, pp. 75-86). ACM. (2008)
14. Li, D., Guo, C., Wu, H., Tan, K., Zhang, Y., Lu, S.: FiConn: Using backup port for server interconnection in data centers. In Infocom 2009, ieee (pp. 2276-2285). IEEE. (2009)
15. Guo, C., Lu, G., Li, D., Wu, H., Zhang, X., Shi, Y., Lu, S.: BCube: a high performance, server-centric network architecture for modular data centers. ACM SIGCOMM Computer Communication Review, 39(4), 63-74. (2009)
16. Liu, B., Huang, K., Li, J., Zhou, M.: An incremental and distributed inference method for large-scale ontologies based on mapreduce paradigm. IEEE transactions on cybernetics, 45(1), 53-64. (2015)
17. Mohammed, E. A., Far, B. H., Naugler, C.: Applications of the mapreduce programming framework to clinical big data analysis: current landscape and future trends. BioData Mining, 7(1), 22. (2014).
18. Al-Fares, M., Loukissas, A., Vahdat, A.: A scalable, commodity data center network architecture. ACM SIGCOMM Computer Communication Review, 38(4), 63-74. (2008).
19. Ding, M., Tian, H.: PCA-based network Traffic anomaly detection. Tsinghua Science and Technology, 21(5), 500-509. (2016)
20. Kandula, S., Sengupta, S., Greenberg, A., Patel, P., Chaiken, R. The nature of data center traffic: measurements & analysis. ACM SIGCOMM Conference on Internet Measurement Conference (Vol.9, pp.202-208). ACM. (2009).
21. Guo, D., Chen, T., Li, D., Liu, Y.: BCN: Expansible network structures for data centers using hierarchical compound graphs. INFOCOM 2011. IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 10-15 April 2011, Shanghai, China (Vol.21, pp.61-65). DBLP. (2011).
22. Wang, L., Tao, J., Ranjan, R., Marten, H., Streit, A., Chen, J., et al.: G-hadoop: mapreduce across distributed data centers for data-intensive computing. Future Generation Computer Systems, 29(3), 739–750. (2013).

23. Morla, R., Gonçalves, P., Barbosa, J. G.: High-performance network traffic analysis for continuous batch intrusion detection. Journal of Supercomputing, 72(11), 1-22. (2016).
24. Cohen, J.: Graph twiddling in a mapreduce world. Computing in Science & Engineering, 11(4), 29-41. (2009).
25. Fadika, Z., Dede, E., Govindaraju, M., Ramakrishnan, L. Mariane: using mapreduce in hpc environments. Future Generation Computer Systems, 36(3), 379-388. (2014)
26. Slagter, K., Hsu, C. H., Chung, Y. C., Zhang, D.: An improved partitioning mechanism for optimizing massive data analysis using mapreduce. The Journal of Supercomputing, 66(1), 539-555. (2013)
27. Jiang, H., Chen, Y., Qiao, Z., Li, K. C., Ro, W., Gaudiot, J. L.: Accelerating mapreduce framework on multi-gpu systems. Cluster Computing, 17(2), 293-301. (2014)
28. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113. (2008).
29. Kaashoek, F., Morris, R., Mao, Y.: Optimizing mapreduce for multicore architectures. (2010).
30. Yu, Z., Xiang, D., Wang, X.: Balancing virtual channel utilization for deadlock-free routing in torus networks. The Journal of Supercomputing, 71(8), 3094-3115. (2015)
31. Lin, X. Y., Chung, Y. C.: Master–worker model for mapreduce paradigm on the tile64 many-core platform. Future Generation Computer Systems, 36(3), 19–30. (2014).
32. Chen, R., Chen, H., Zang, B.: Tiled-MapReduce: optimizing resource usages of data-parallel applications on multicore with tiling. International Conference on Parallel Architectures and Compilation Techniques (pp.523-534). IEEE Computer Society. (2010).

Tao Jiang received the B.E. degree in Electronics and Communications Engineering from Xidian University in 2015. Now he is doing the M.E. Programme in Telecommunication and information system in the State key lab of ISN, Xidian University. His main research interests are related to optical interconnected networks and data center networks.

Huaxi Gu received B.E., M.E., and Ph.D. in Telecommunication Engineering and Telecommunication and Information Systems from Xidian University, Xidian in 2000, 2003 and 2005 respectively. He is a Full Professor in the State Key Laboratory of ISN, Telecommunication Department, Xidian University, Xidian, China. His current interests include interconnection networks, networks on chip and optical intrachip communication. He has more than 100 publications in refereed journals and conferences. He has been working as a reviewer of IEEE Transaction on Computer, IEEE Transactions on Dependable and Secure Computing, IEEE System Journal, IEEE Communication Letters, Information Sciences, Journal of Supercomputing, Journal of System Architecture, Journal of Parallel and Distributed Computing, Microprocessors and Microsystems etc.

Kun Wang received the B.E. degree and M.E. degree in Computer Science and Technology from Xidian University, Xi'an in 2003 and 2006 respectively. Now she is a lecturer in the Dept. of Computer Science, Xidian University, Xi'an China. Her Current interests include high performance computing and cloud computing, the network vitulization technology.

Xiaoshan Yu received the M.E. degree in Electronics and Communications Engineering from Xidian University in 2013. Now he is doing the Ph.D. Programme in Telecommunication and information system in the State key lab of ISN, Xidian University. His main research interests are related to optical interconnected networks, data center networks.

Yunfeng Lu received the bachelor's degree in Information Engineering from Jilin University in 2016. Now he is doing the M.S. degree in Telecommunication and information system in the State key lab of ISN, Xidian University. His main research interests are related to optical interconnected networks, high performance computing.

Received: February 2, 2017; Accepted: June 15, 2017.

Computer Science and Information Systems 14(3):629–641 https://doi.org/10.2298/CSIS170212020J

Click-Boosted Graph Ranking for Image Retrieval

Jun Wu[1,2], Yu He[1], Xiaohong Qin[1], Na Zhao[2], and Yingpeng Sang[3]

[1] School of Computer and Information Technology, Beijing Jiaotong University
Beijing 10044, China
{wuj, 15120398, 14120420}@bjtu.edu.cn
[2] Logistics and E-commerce College, Zhejiang Wanli University
Ningbo, 315100, China
zhaona@zwu.edu.cn
[3] School of Information Science and Technology, Sun Yat-Sen University
Guangzhou 510275, China
sangyp@mail.sysu.edu.cn

Abstract. Graph ranking is one popular and successful technique for image retrieval, but its effectiveness is often limited by the well-known semantic gap. To bridge this gap, one of the current trends is to leverage the click-through data associated with images to facilitate the graph-based image ranking. However, the sparse and noisy properties of the image click-through data make the exploration of such resource challenging. Towards this end, this paper propose a novel click-boosted graph ranking framework for image retrieval, which consists of two coupled components. Concretely, the first one is a click predictor based on matrix factorization with visual regularization, in order to alleviate the sparseness of the click-through data. The second component is a soft-label graph ranker that conducts the image ranking by using the enriched click-through data noise-tolerantly. Extensive experiments for the tasks of click predicting and image ranking validate the effectiveness of the proposed methods in comparison to several existing approaches.

Keywords: Image Retrieval, Click-Through Data, Graph Ranking, Matrix Factorization.

1. Introduction

Graph ranking [34] has received increasing attention in recent years due to its superiority in various visual ranking tasks, such as natural image search [5], video search [7], shape retrieval [2], cross-media retrieval [30], 3D object retrieval [1], etc. Unlike traditional visual ranking that considers only the pairwise similarity between visual documents, graph-based visual ranking aims to explore the intrinsic manifold structure collectively hidden in visual documents, hoping to refine the similarity measure. Despite these successes, the performance of graph-based visual ranking is still limited by the well-known semantic gap existing between low-level image pixels captured by machines and high-level semantic concepts perceived by humans, especially when the visual targets are dispersed in the feature space.

In order to boost the performance of image retrieval and overcome the semantic gap, a relevance feedback mechanism [35] is incorporated into the graph-based ranking framework [4, 27], which encourages the user to label a few images returned as either positive or

negative in terms of whether they are relevant to user's query or not. The labeled instances is then used to refine the ranking model towards the user's query intends. However, it is not easy to obtain sufficient and explicit user feedback as users are often reluctant to provide enough feedbacks to search engines. It is noted that search engines can record queries issued by users and the corresponding clicked images. Although the clicked images cannot reflect the explicit user preference on relevance of particular query-image pairs, they statistically indicate the implicit relationship between individual images in the ranked list and the given query [33]. Therefore, we can regard the click-through data associated with images as the 'implicit' feedbacks based on an assumption that, in a same query session, most clicked images are relevant to the given query, and the reliability of this assumption has been empirically validated by [18].

We consider a particular ranking scenario where the click-through data is sparse and inaccurate. Such a scenario is pervasively presented in the image retrieval problem for which some methods [8, 33, 11] employ the click-through data as implicit feedbacks for image ranking. Inevitably, the sparsity may lead to an underfitting ranker, and the inaccuracy may further mislead the ranker. Towards this end, this paper presents a Click-Boosted Graph Ranking (CBGR) approach for effective image retrieval based on a preliminary work [12], which incorporates click enriching with noise-tolerant graph ranking within an unified framework. The main contributions are two-fold. For one thing, a Visual Regularized Matrix Factorization (VRMF) method is proposed to enrich the click-through data. For another, a Soft-Label Graph Ranking (SLGR) technique is developed to leverage the enriched click-through data noise-tolerantly. An empirical study on the tasks of click predicting and image ranking shows encouraging results in comparison to several exiting approaches.

The rest of this paper is organized as follows. Section 2 reviews the related works. Section 3 presents the proposed CBGR method. Section 4 reports on the experiments. Finally, section 5 concludes this paper and raises the problem for future works.

2. Related Work

We briefly group related work into three dimensions: graph ranking, collaborative image retrieval and collaborative filtering, and introduce them separately in the following subsections.

2.1. Graph Ranking

Graph ranking has been extensively studied in the multimedia retrieval area. Its main idea is to describe the dataset as a graph and then decide the importance of each vertex based on local or global structure drawn from the graph. One canonical graph-based ranking technique is the Manifold Ranking (MR) algorithm [34], and He et al. [5] first applied MR to image retrieval. Its limitations are addressed by latter research efforts. For example, Wang et al. [24] improved the MR accuracy using a k-regular nearest neighbor graph that minimizes the sum of edge weights and balances the edges in the graph as well. Wu et al. [27] proposed a self-immunizing MR algorithm that uses an elastic kNN graph to exploit unlabeled images safely. Wang et al [25] proposed a multi-manifold ranking method, which jointly exploits multiple visual modalities to encode the image ranking results. Xu

et al. [29] proposed an efficient MR solution based on scalable graph structure to handle large-scale image datasets. In addition, the users' feedbacks are easily exploited by MR method, and previous studies have shown that MR is one of the most successful ranking approaches for the image retrieval with relevance feedback [5, 27, 29].

Note that most existing methods of graph ranking receive the supervision signals provided by the users directly. Differently, in this work the supervision signal, derived from the click-through data, is noisy, and directly using that may degenerates the retrieval performance. Hence this work presents a soft-label graph ranking solution for the noise-tolerance purpose.

2.2. Collaborative Image Retrieval

Collaborative Image Retrieval (CIR) regards the click-through data associated with images as the long-term experience and leverages it to boost the short-term learning with relevance feedback. For example, Yin et al. [31] exploited the long-term experiences to select the optimal online ranker from a set of candidates based on reinforcement learning. Hoi et al. [6] regarded the query log as the 'side information', and then, taking that as constraints, learned a distance metric form a mixture of labeled and unlabeled images. Su et al. [19] suggested discovering the navigation patterns from query logs, and using the patterns to facilitate new searching tasks. Wu et al. [28] proposed a multi-view manifold ranking method, which simultaneously exploits the visual and click features to encode the image ranking results.

In contrast, our proposed approach requires no users' feedbacks once the query has been issued. Alternatively, it automatically derives implicit feedbacks from the click-through data. This is motivated by empirical evidence suggesting that few users are willing to perform any form of feedback to improve their search results.

2.3. Collaborative Filtering

Collaborative filtering (CF) [20] is a family of algorithms popularly-used in recommendation systems. Depending on how the data of user-item rating matrix are processed, two types of methods, neighbor based and latent factor based, can be differentiated.

Neighbor-based methods use similarity measures to select users (or items) that are similar to the active user (or the target item). Then, the prediction is calculated from the ratings of these neighbors. Most of neighbor-based approaches can be further categorized as user-based or item-based depending on whether the process of finding neighbors is focused on similar users [14] or items [16]. Latent factor based methods are an alternative approach that tries to explain the ratings by characterizing both users and items on a few factors inferred from the user-item rating matrix. Matrix Factorization (MF) [9] might be one of the most promising techniques due to its excellent performance, as witnessed by the Netflix contest. During the past years, plenty of research effort has been made to further improve its effectiveness and efficiency, including maximum margin MF [13], Bayesian MF [15], online MF [10] and parallel MF [3], etc. Besides user-item rating matrix, a current trend is to leverage the plentiful side-information around user and item dimensions to enhance MF performance [17].

In the scenario of this work, the user-image clicking matrix is very similar to the user-item rating matrix in recommender system. Inspired by recommender system, we consider

MF for the purpose of click prediction. Different from the traditional recommendation problems, the 'items' in our scenario are the database images, which have plentiful visual content. Considering this, we present a VRMF algorithm that can exploit the images' visual information to improve the prediction accuracy.

3. The Proposed CBGR Approach

Our CBGR approach is developed based on two intuitions. At first, a 'good' visual ranker should be able to exploit the implicit feedback (click-through data) rather than the explicit feedback for the purpose of alleviating the user's labeling burden. Furthermore, the ranker should be able to handle the sparse and noisy properties of the click-through data.

As mentioned, our CBGR approach consists of two key components, i.e., VRMF and SLGR. We start with the description of notations, then elaborate the details of VRMF and SLGR, and lastly present an algorithmic framework of our CBGR approach.

3.1. Preliminaries

Let $\mathcal{X} = \{\mathbf{x}_i, i = 1, \cdots, n\}$ denote the image dataset, where each $\mathbf{x}_i \in \mathbb{R}^d$ is a visual feature vector. To discover the geometrical subacute (manifold), we build a neighborhood graph on $\mathcal{X}$, and define $\mathbf{W} \in \mathbb{R}_+^{n \times n}$ as the corresponding the affinity matrix with element W_{ij} storing the weight of edge between $\mathbf{x_i}$ and $\mathbf{x_j}$. Normally the weight is calculated using a Gaussian kernel

$$W_{ij} = \exp\left(-\frac{d^2(\mathbf{x}_i, \mathbf{x}_j)}{\sigma^2}\right) \tag{1}$$

if $i \in \mathcal{N}(j)$ or $j \in \mathcal{N}(i)$, otherwise $W_{ij} = 0$, where $\mathcal{N}(i)$ denotes a k nearest neighbor set of image i. Typically, k is a small number (e.g. a small fraction of n), $d(\mathbf{a}, \mathbf{b})$ is a distance metric between two vectors $\mathbf{a}$ and $\mathbf{b}$ (suggested by [4], L1 distance is considered), and σ is the bandwidth parameter that can be tuned by local scaling technique, the effectiveness of which has been verified in the clustering [32] and ranking [27] tasks.

The click-through data is represented by a user-image clicking matrix $\mathbf{R} \in \{0, 1\}^{m \times n}$ whose rows correspond to the users and columns correspond to the images. If an image has been clicked in a query session, the corresponding cell is assigned to value 1, i.e. $R_{ij} = 1$, otherwise $R_{ij} = 0$.

3.2. VRMF: Visual Regularized Matrix Factorization

We refer to the problem of click prediction as Matrix Factorization (MF), which is to learn low-rank representations (also referred as latent factors) of users and images from the information of the user-image clicking matrix, and then further employs the latent factors to predict new clicks between users and images. Also, to elevate the accuracy of click prediction, the visual information of images is taken as the regularization term and incorporated into the MF framework, thus called Visual Regularized MF (VRMF). Let $\mathbf{U} \in \mathbb{R}^{f \times m}$ and $\mathbf{V} \in \mathbb{R}^{f \times n}$ be two matrices of latent factors, and our VRMF method is to

Algorithm 1 Gradient Descent Process for Solving $\mathbf{U}_{*i}$ and $\mathbf{V}_{*j}$

1: Initialize $t = 0$, $\eta = 1$, $\mathbf{U}^{(0)}$ and $\mathbf{V}^{(0)}$;
2: **while** $t \leq T$ **do**
3: Update $\mathbf{U}_{*i}^{(t+1)} = \mathbf{U}_{*i}^{(t)} - \eta_t \frac{\partial F}{\partial \mathbf{U}_{*i}^{(t)}}$ and $\mathbf{V}_{*j}^{(t+1)} = \mathbf{V}_{*j}^{(t)} - \eta_t \frac{\partial F}{\partial \mathbf{V}_{*j}^{(t)}}$
4: If $F(\mathbf{U}_{*i}^{(t+1)}, \mathbf{V}_{*j}^{(t+1)}) < F(\mathbf{U}_{*i}^{(t)}, \mathbf{V}_{*j}^{(t)})$, $\eta_{t+1} = 2\eta_t$; otherwise, $\eta_{t+1} = \eta_t/2$;
5: $t = t + 1$;
6: **end while**

recovery the user-image clicking matrix $\hat{\mathbf{R}} = \mathbf{U}^T\mathbf{V}$ by solving the following optimization problem

$$\begin{aligned} \underset{\mathbf{U},\mathbf{V}}{\text{minimize}}\ F(\mathbf{U},\mathbf{V}) = & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - \mathbf{U}_{*i}^T\mathbf{V}_{*j})^2 \\ & + \frac{\alpha}{2}\sum_{j=1}^{n}\sum_{k\in\mathcal{N}(j)} W_{jk}\|\mathbf{V}_{*j} - \mathbf{V}_{*k}\|^2 \\ & + \frac{\beta}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) \end{aligned} \tag{2}$$

where $\mathbf{U}_{*i}$ is a column vector of $\mathbf{U}$, representing the latent factors of user i, likewise, and $\mathbf{V}_{*j}$ represents the latent factors of image j. I_{ij} is an indicator function that is equal to 1 if $R_{ij} = 1$, otherwise 0. $\| \bullet \|_F$ denotes the Frobenius norm of a matrix, and α and β are two free parameters. The first term of above cost function is the fitting constraint that ensures the learned $\hat{\mathbf{R}}$ to be consistent with the observed user-image matrix, the second term is the smoothness constraint that makes the visually similar images having similar latent factors, and the last term is the regularizer that is to alleviate model overfitting.

We adopt a gradient descent process to solve Eq. (2). By differentiating F with respect to $\mathbf{U}_{*\mathbf{i}}$ and $\mathbf{V}_{*\mathbf{j}}$, we have

$$\frac{\partial F}{\partial \mathbf{U}_{*i}} = \sum_{j=1}^{n} I_{ij}(R_{ij} - \mathbf{U}_{*i}^T\mathbf{V}_{*j})\mathbf{V}_{*j} + \beta\mathbf{U}_{*i}, \tag{3}$$

and

$$\frac{\partial F}{\partial \mathbf{V}_{*j}} = \sum_{i=1}^{m} I_{ij}(R_{ij} - \mathbf{U}_{*i}^T\mathbf{V}_{*j})\mathbf{U}_{*i} + \alpha\sum_{k\in\mathcal{N}(j)} W_{jk}(\mathbf{V}_{*j} - \mathbf{V}_{*k}) + \beta\mathbf{V}_{*j}. \tag{4}$$

In the gradient descent process, we dynamically adapt the step-size η in order to accelerate the process while guaranteeing its convergence. Denote by $\mathbf{U}_{*i}^{(t)}$ and $\mathbf{V}_{*j}^{(t)}$ the values of $\mathbf{U}_{*i}$ and $\mathbf{V}_{*j}$ in the t-th turn of the iterative process. If $F(\mathbf{U}_{*i}^{(t+1)}, \mathbf{V}_{*j}^{(t+1)}) < F(\mathbf{U}_{*i}^{(t)}, \mathbf{V}_{*j}^{(t)})$, i.e. the cost function obtained after gradient descent is reduced, then we double the step-size; otherwise, we halve the step-size and do not recompute $\mathbf{U}_{*i}^{(t+1)}$ and $\mathbf{V}_{*j}^{(t+1)}$. The process is illustrated in Algorithm 1.

3.3. GRSL: Graph Ranking with Soft Labels

Our GRSL method is developed based on MR [34], and we slightly modify it for the noise-tolerance purpose. Let $\mathbf{r} : \mathcal{X} \rightarrow \mathbb{R}$ be a ranking function that assigns to each image instance $\mathbf{x}_i$ a ranking score r_i. We can view $\mathbf{r}$ as a vector $\mathbf{r} = [r_1, \cdots, r_n]^T$. We also define a label vector $\mathbf{y} = [y_1, \cdots, y_n]^T$, in which $y_i = 1$ if $\mathbf{x}_i$ is a query, and $y_i = 0$ otherwise. The cost function associated with $\mathbf{r}$ is defined to be

$$\underset{\mathbf{r}}{\text{minimize}}\, Q(\mathbf{r}) = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}\left(\frac{r_i}{\sqrt{D_{ii}}} - \frac{r_j}{\sqrt{D_{jj}}}\right)^2 + \frac{\lambda}{2}\|\mathbf{r} - \mathbf{y}\|^2 \tag{5}$$

where λ is a regularization parameter, and $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} W_{ij}$. The first term in the cost function is a smoothness constraint, which make the nearby examples in visual space having close ranking scores. The second term is a fitting constraint, which makes the ranking result fitting to the label assignment.

By differentiating $Q(\mathbf{r})$ and set it to zero, we can easily get the following closed solution

$$\mathbf{r}^* = (\mathbf{I} - \mu\mathbf{S})^{-1}\mathbf{y} \tag{6}$$

where $\mu = 1/(1+\lambda)$, $\mathbf{I}$ is an identity matrix, and $\mathbf{S}$ is the symmetrical normalization of $\mathbf{D}$, i.e.

$$\mathbf{S} = \mathbf{D}^{1/2}\mathbf{W}\mathbf{D}^{1/2}. \tag{7}$$

We can directly use the above closed form solution to compute the ranking scores of examples. However, in large scale problems, we prefer to use the iteration solution

$$\mathbf{r}(t+1) = \mu\mathbf{S}\mathbf{r}(t) + (1-\mu)\mathbf{y}. \tag{8}$$

As illustrated by Eq. (6) and (8), it is noted that the label vector plays an important role in image ranking, and a dense $\mathbf{y}$ is desirable to derive $\mathbf{r}^*$. In the regular MR [34], only one labeled instance (i.e. the user's query) is concerned, which is hardly to achieve satisfactory ranking result. A few works, such as [4], [5], [27], and [28], etc., take the online relevance feedback mechanism into consideration for the label vector enrichment, but it is unpractical as mentioned before.

Different from previous studies, our idea is to enrich the label vector using the click-through data without any user intervention. The intuition behind our idea is that, when two images are clicked in a same query session, they often share a certain similar meaning and we expect different users to express similar judgments on them. Based on this assumption, the (hidden) correlation between any two images can be inferred by analyzing the judgements (clicks) made by different users on them. Given the user's query q, the correlation of it to each database image is defined by Jacquard coefficient based on the enriched query log matrix $\hat{\mathbf{R}}$

$$Sim(q,i) = \frac{|\mathcal{A}(\hat{\mathbf{R}}_{*q}) \cap \mathcal{A}(\hat{\mathbf{R}}_{*i})|}{|\mathcal{A}(\hat{\mathbf{R}}_{*q}) \cup \mathcal{A}(\hat{\mathbf{R}}_{*i})|} \tag{9}$$

where $\mathcal{A}(\mathbf{a})$ denotes a set composed of the nonzero elements of a binary vector, $\hat{\mathbf{R}}_{*i} \in \{0,1\}^m$ is a column vector of $\hat{\mathbf{R}}$, recording the clicks imposed by different users on image i, and $|\bullet|$ denotes the size of a set. Intuitively, we can directly predict $y_i = 1$ if $Sim(q,i)$ is highly positive. Although this idea can be straightly handled by MR, it may suffer from performance degradation as erroneous click-through data. In particular, in our scheme, more noises may be introduced by MF. To this end, we treat the labels in two different ways for the fault-tolerance purpose. In details, the user's query is treated as the 'hard-labeled' instance, while the images predicted by analyzing the click-through data are treated as 'soft-labeled' instances, i.e. $y_i = Sim(q,i) \in [0,1]$, where the magnitude of the label represents the confidence of the assigned label.

3.4. Algorithmic Framework

So far, we can assemble the proposed VRMF and GRSL methods into the CBGR framework for image retrieval, which ranks the database images with respect to the user's query based on visual features and click-through data. We outline this algorithmic framework in below.

1. Build a neighborhood graph on $\mathcal{X}$, and compute the corresponding affinity matrix $\mathbf{W}$ by Eq. (1) and the normalized one $\mathbf{S}$ by Eq. (7);
2. Compute the enriched user-image matrix $\hat{\mathbf{R}}$ based on $\mathbf{R}$ and $\mathbf{S}$ using Algorithm 1;
3. Compute the soft-label vector $\mathbf{y}$ based on q and $\hat{\mathbf{R}}$ by Eq. (9);
4. Compute the ranking-score vector $\mathbf{r}$ based on $\mathbf{S}$ and $\mathbf{y}$ by Eq. (6) or (8);

Note that the step 1 and step 2 can be implemented offline, and therefore our CBGR approach can be quite efficient in processing online queries. Note that our CBGR approach mainly focuses on processing the in-sample queries, but it can be easily extended to handle the out-of-sample queries. For example, when a completely new query arrives, we can apply a strategy named one-click query expansion [21] to transform the out-of-sample query into a in-sample query with very few user efforts.

4. Experiments

In this section, we first introduce our experimental settings, and then present the experimental results that validate the effectiveness of our approach. The experiments actually contain two parts. In the first part, we will compare our VRMF method with those CF methods that can be used for the task of click prediction. In the second part, we compare our CBGR approach with several existing graph ranking methods for the task of image retrieval.

4.1. Experimental Setup

We employ the '10K Images' dataset[1] which is publicly available on the web to make our experiments reproducible. The images are from 100 semantic categories, with 100 images per category. Three kinds of visual features are extracted to represent the images,

[1] http://www.datatang.com/data/44353. The dataset was firstly used in [26].

Table 1. The P@N measures of our VRMF method compared with several exiting CF approaches.

	N=10	20	30	40	50
user-based	0.659	0.582	0.508	0.462	0.423
item-based	0.681	0.612	0.527	0.448	0.38
regular MF	0.722	0.661	0.596	0.532	0.473
ItemVisual	0.734	0.682	0.61	0.558	0.52
VRMF	**0.75**	**0.702**	**0.652**	**0.593**	**0.534**

Table 2. The F1@N measures of our VRMF method compared with several exiting CF approaches.

	N=10	20	30	40	50
user-based	0.158	0.248	0.294	0.324	0.343
item-based	0.163	0.26	0.305	0.312	0.307
regular MF	0.18	0.294	0.38	0.387	0.384
ItemVisual	0.178	0.298	0.385	0.398	0.395
VRMF	**0.185**	**0.31**	**0.392**	**0.435**	**0.43**

including a 64-dimensional color histogram, an 18-dimensional wavelet-based texture and a 5-dimensional edge direction histogram [26].

A click-through dataset consisting of 1000 query sessions is used in experiments, which is simulated based on the ground truth of image dataset. The average number of clicks in each query session is 20. Also, we randomly add 12% noise into the click-through dataset to approach the real-world search scenario [1].

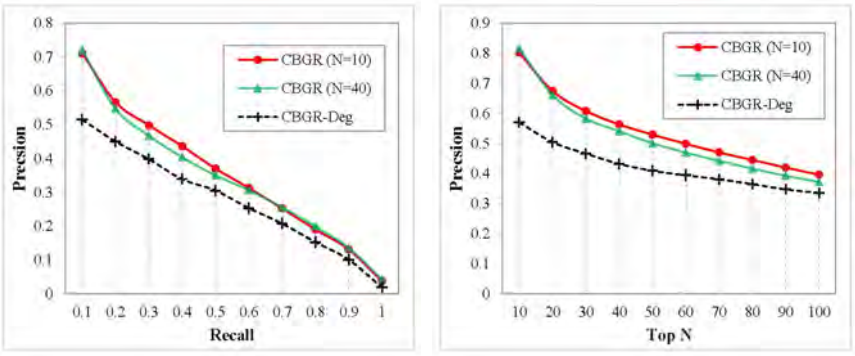
Essentially, our click prediction solution acts for the image recommendation task, while the image ranking problem is equivalent to the image retrieval task. Many measures are commonly used to evaluate both recommendation and retrieval tasks, such as precision and recall. In the top N recommendation scenario, precision and recall are often summarized as the F1 measure. Similarly, in the retrieval scenario, PR (Precision-Recall) graph is widely used to depict the relationship between precision and recall, and it could be further summarized as the MAP (Mean Average Precision) measure. In addition, for many web applications, only the top returned images can attract users' interests, so the precision at top N (P@N) metric is significant to evaluate the image recommendation and retrieval performance.

To evaluate the average performance of image retrieval methods, a query set with 200 images is equally sampled from all semantic categories, i.e. two images are randomly picked from each category.

4.2. Experimental Results for The Task of Click Prediction

In this part, we compare our VRMF method with several existing CF approaches that can be used for the task of click prediction, including user-based CF [14], item-based CF [16],

[1] Empirical study reported by [18] showed that the satisfaction rate of the image click-through data is approximately equal to 88%.



(a) (b)

Fig. 1. The performance of our CBGR approach on different enriching scales in terms of (a) PR graphs and (b) P@N curves.

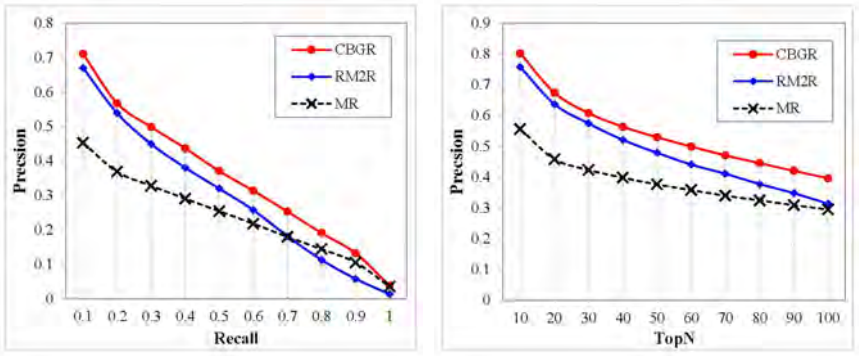
and regular MF method [9]. In addition, the method appeared in our previous conference version [12] is also included in the comparisons. Essentially our previous method is a item-based CF solution with visual side-information, so we term it as ItemVisual.

For the proposed VRMF method, there are two parameters, i.e., α and β (see Eq. (2)). We tune the two parameters through 5-fold cross-validation, and the best settings are $\alpha = 0.06$ and $\beta = 0.04$. For the iteration runs T (see the Algorithm1), we set it to 1000. In our experiments, we found that this value can lead to a well convergence of the optimization process.

Table 1 and Table 2 respectively print the P@N and F1@N measures of our VRMF method compared with several other approaches, where the best performance has been boldfaced. From the experimental results, the following interesting observations are revealed. First, by examining all methods, the two methods using the visual side-information (VRMF and ItemVisual) outperform the three baseline methods (user-based, item-based and regular MF), which verifies the usefulness of the visual side-information. Second, by comparing the three baseline methods, the regular MF performs better than the user-based and item-based CF, which demonstrates the superiority of the latent factor models. At last, the proposed VRMF method achieves the best performance among all comparing approaches. It well demonstrates that combining the latent factor model with the visual side-information is effective and beneficial to the task of click prediction.

4.3. Experimental Results for The Task of Image Retrieval

Furthermore, we evaluate the performance of our CBGR approach on two enriched click-through datasets which are respectively attained when $N = 10$ (with highest precision) and $N = 40$ (with highest F1 measure). To verify whether the click prediction solution is beneficial to image retrieval or not, we compare our CBGR approach with a its degenerated variant termed CBGR-Deg. The CBGR-Deg method is almost same with the CBGR approach, except the former directly use the click-through data without enrichment.



(a) (b)

Fig. 2. The performance of our CBGR approach compared with several existing methods in terms of (a) PR graphs and (b) P@N curves.

For the our CBGR approach, there is only one parameter, i.e., the μ used in Eq. (8). For convenience, we set $\mu = 0.01$, consistent with the previous experiences [34, 4, 29, 27].

Figure 1 print the PR graphs and P@N curves of all comparing methods. By examining the experimental results, we observe that both CBGR ($N = 10$) and CBGR ($N = 40$) approaches outperform the CBGR-Deg method, which verifies the effectiveness of the click prediction solution used by our CBGR approach. Further, by comparing the CBGR ($N = 10$) and CBGR ($N = 40$) methods, we found that their performance curves are very similar to each other. Based on this observation, we prefer to set $N = 10$ for the computational efficiency purpose.

At last, we compare our CBGR approach with a CIT scheme named RM2R [28]. RM2R is a two-view graph ranking model which exploits both visual and click features to encode the ranking results. Moreover, the conventional MR method [5] as a baseline is also included in comparisons to verify whether exploiting the click-through data is beneficial to the task of image retrieval or not. To be fair, all above three graph ranking methods take the local scaling trick [32] to tune the bandwidth parameter used by the Gaussian kernel.

Figure 2 plots the PR graphs and P@N curves of our approach compared with other two methods. We found that the methods using both visual feature and the click-through data perform better than the baseline MR method, which verifies the benefit of exploiting the click-through data for the task of image retrieval. It is impressive that the performance of our CBGR approach is always the best among all comparing methods. It is worth noting that the performance of the RM2R method evaluated in our experiments is not as good as the results reported by [28]. The main reason is that we evaluate it without using relevance feedback as done in [28], and thus its performance drops. This observation reveals that leveraging the click-through data as supervision signal is more effective than as feature set, when only a user's query is available.

5. Conclusions

Existing image search engines usually suffer from imperfect results caused by the well-known semantic-gap. Although many studies on learning with the click-through data have been conducted to address this problem, the improvement in performance is limited as little effort on investigating both sparseness and noisiness of the click-through data. This paper presents a novel CBGR method that aims at noise-resistantly leveraging the click-through data to boost graph-based visual ranking. Concretely, we first propose a VRMF method to enrich the click-through data, and then develop a GRSL solution for fault-tolerant image ranking. Experimental study validates the superiority of the proposed techniques in comparison to some representative approaches.

In the future, we will take more side-information (e.g., the social relationships between users and the surrounding text information of images) into consideration in order to further enhance the effectiveness of the click predicting and visual ranking. In addition, inspired by the data stream mining techniques [22, 23], another extension of this work is to study the incremental solutions to the tasks of click predicting and graph ranking.

Acknowledgment. The authors would like to thank the anonymous reviewers for their constructive suggestions. This work was supported in part by the 'Natural Science Foundation of China' (61370070 and 61671048), the 'Fundamental Research Funds for the Central Universities' (2015JBM029), and the 'Research Foundation for Talents of Beijing Jiaotong University' (2015RC008). The corresponding author of this paper is Na Zhao (email: zhaona@zwu.edu.cn).

References

1. Bai, S., Bai, X.: Sparse contextual activation for efficient visual re-ranking. IEEE Trans. Image Process. 25(3), 1056–1069 (2016)
2. Bai, X., Wang, B., Yao, C., Liu, W., Tu, Z.: Co-transduction for shape retrieval. IEEE Trans. Image Process. 21(5), 2747–2757 (2012)
3. Chin, W., Zhuang, Y., Juan, Y., Lin, C.: A fast parallel stochastic gradient method for matrix factorization in shared memory systems. ACM Trans. Intel. Syst. Tec. 6(1), 2:1–2:24 (2015)
4. He, J., Li, M., Zhang, H., Tong, H., Zhang, C.S.: Manifold-ranking based image retrieval. In: ACM Multimedia. pp. 9–16 (2004)
5. He, J., Li, M., Zhang, H., Tong, H., Zhang, C.: Generalized manifold-ranking-based image retrieval. IEEE Trans. Image Process. 15(10), 3170–3177 (2006)
6. Hoi, S., Liu, W., Chang, S.F.: Semi-supervised distance metric learning for collaborative image retrieval. In: CVPR. pp. 1–7 (2008)
7. Hsu, W.H.: Video search reranking through random walk over document-level context graph. In: ACM Multimedia. pp. 971–980 (2007)
8. Jain, V., Varma, M.: Learning to re-rank: Query-dependent image re-ranking using click data. In: WWW. pp. 277–286 (2011)
9. Koren, Y., Bell, R., Volinsky, C., et al.: Matrix factorization techniques for recommender systems. Computer 42(8), 30–37 (2009)
10. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J Mach. Learn. Res. 11, 19–60 (2010)
11. Pan, Y., Yao, T., Mei, T., Li, H., Ngo, C.W., Rui, Y.: Click-through-based cross-view learning for image search. In: ACM SIGIR. pp. 717–726 (2014)
12. Qin, X., He, Y., Wu, J., Sang, Y.: Leveraging click completion for graph-based image ranking. In: PDCAT. pp. 155–160 (2016)

13. Rennie, J.D.M., Srebro, N.: Fast maximum margin matrix factorization for collaborative prediction. In: ICML. pp. 713–719 (2005)
14. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. In: CSCW. pp. 175–186 (1994)
15. Salakhutdinov, R., Mnih, A.: Bayesian probabilistic matrix factorization using markov chain monte carlo. In: ICML. pp. 880–887 (2008)
16. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: WWW. pp. 285–295 (2001)
17. Shi, Y., Larson, M., Hanjalic, A.: Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. ACM Comput. Surv. 47(1), 3:1–3:45 (2014)
18. Smith, G., Brien, C., Ashman, H.: Evaluating implicit judgments from image search clickthrough data. J. Am. Soc. Inf. Sci. Tec. 63(12), 2451–2462 (2012)
19. Su, J., Huang, W., Yu, P., Tseng, V.: Efficient relevance feedback for content-based image retrieval by mining user navigation patterns. IEEE Trans. Knowl. Data Eng. 23(3), 360–372 (2011)
20. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artificial Intellegence 2009, 421425:1–421425:19 (2009)
21. Tang, X., Liu, K., Cui, J., Wen, F., Wang, X.: Intentsearch: Capturing user intention for one-click internet image search. IEEE Trans Pattern Anal.Mach. Intell. 34(7), 1342–1353 (2012)
22. Wang, H., Wu, J., Pan, S., Zhang, P., Chen, L.: Towards large-scale social networks with online diffusion provenance detection. Comput. Netw. 114, 154 – 166 (2017)
23. Wang, H., Zhang, P., Zhu, X., Tsang, I., Chen, L., Zhang, C., Wu, X.: Incremental subgraph feature selection for graph classification. IEEE Trans. Knowl. Data Eng. 29(1), 128–142 (2017)
24. Wang, M., Li, H., Tao, D., Lu, K., Wu, X.: Multimodal graph-based reranking for web image search. IEEE Trans. Image Process. 21(11), 4649–4661 (2012)
25. Wang, Y., Cheema, M.A., Lin, X., Zhang, Q.: Multi-manifold ranking: Using multiple features for better image retrieval. In: PAKDD. LNAI, vol. 7819, pp. 449–460 (2013)
26. Wu, J., Shen, H., Li, Y.D., Xiao, Z.B., Lu, M.Y., Wang, C.L.: Learning a hybrid similarity measure for image retrieval. Pattern Recogn. 46(11), 2927–2939 (2013)
27. Wu, J., Li, Y., Feng, S., Shen, H.: Self-immunizing manifold ranking for image retrieval. In: PAKDD. LNAI, vol. 7819, pp. 92–100 (2013)
28. Wu, J., Yuan, J., Luo, J.: Robust multi-view manifold ranking for image retrieval. In: PAKDD. LNAI, vol. 9652, pp. 92–103 (2016)
29. Xu, B., Bu, J., Chen, C., Wang, C., Cai, D., He, X.: Emr: A efficient manifold ranking model for content-based image retrieval. IEEE Trans. Knowl. Data Eng. 27(1), 102–114 (2014)
30. Yang, Y., Xu, D., Nie, F., Luo, J., Zhuang, Y.: Ranking with local regression and global alignment for cross media retrieval. In: ACM Multimedia. pp. 175–184 (2009)
31. Yin, P.Y., Bhanu, B., K.-C., C., Dong, A.: Integrating relevance feedback techniques for image retrieval using reinforcement learning. IEEE Trans. Pattern Anal. Mach. Intell. 27(10), 1536–1551 (2005)
32. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: NIPS. pp. 1601–1608 (2004)
33. Zhang, Y., Yang, X., Mei, T.: Image search reranking with query-dependent click-based relevance feedback. IEEE Trans. Image Process. 23(10), 4448–4459 (2014)
34. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Scholkopf, B.: Ranking on data manifolds. In: NIPS. pp. 169–176 (2003)
35. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: A comprehensive review. Multimedia Syst. 8(6), 536–544 (2003)

Jun Wu received the Ph.D. degree in computer science from the Dalian Maritime University, China, in 2010. He was a visiting scholar with the Department of Computer Science,

University of Rochester, from 2015-2016. He is currently an Associate Professor with the School of Computer and Information Technology, Beijing Jiaotong University, China. His current research interests include multimedia retrieval, recommendation and large-scale computing.

Yu He received the B.E. degree in software engineering from the Dalian Jiaotong University, China, in 2015. He is currently pursuing the Masters degree with the School of Computer and Information Technology, Beijing Jiaotong University, China. His research mainly focuses on the machine learning algorithms for recommender systems.

Xiaohong Qin received the Master degree in computer science from the Beijing Jiaotong University, China, in 2017. She is currently a software engineer with the China Unicom Corporation, China. Her research mainly focuses on the image retrieval.

Na Zhao received the Ph.D. degree in transportation planning and management from the Dalian Maritime University, China, in 2008. She is currently an Associate Professor with the School of Logistics and E-Commerce, Zhejiang Wanli University, China. Her current research mainly focuses on the optimization problems in the transportation planning.

Yingpeng Sang received her Ph.D. degree in computer science from the Japan Advance Institute of Science and Technology, in 2007. He was also a Postdoctoral Research Fellow with the University with the School of Computer Science, University of Adelaide, Australia, from 2007 to 2010. He is currently an Associate Professor the School of Data and Computer Science, Sun Yat-sen University, China. His research interests include privacy-preserving problems in databases, data mining, and other networking scenarios.

Received: February 12, 2017; Accepted: July 23, 2017.

Computer Science and Information Systems 14(3):643–660 https://doi.org/10.2298/CSIS170301021Y

A Weighted Mutual Information Biclustering Algorithm for Gene Expression Data

Yidong Li[1], Wenhua Liu[1], Yankun Jia[1], and Hairong Dong[2]

[1] School of Computer and Information Technology, Beijing Jiaotong University
[2] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University
{ydli, 16112077, 15120402, hrdong}@bjtu.edu.cn

Abstract. Microarrays are one of the latest breakthroughs in experimental molecular biology, which have already provided huge amount of high dimensional genetic data. Traditional clustering methods are difficult to deal with this high dimensional data, whose a subset of genes are co-regulated under a subset of conditions. Biclustering algorithms are introduced to discover local characteristics of gene expression data. In this paper, we present a novel biclustering algorithm, which called Weighted Mutual Information Biclustering algorithm (WMIB) to discover this local characteristics of gene expression data. In our algorithm, we use the weighted mutual information as new similarity measure which can be simultaneously detect complex linear and nonlinear relationships between genes, and our algorithm proposes a new objective function to update weights of each bicluster, which can simultaneously select the conditions set of each bicluster using some rules. We have evaluated our algorithm on yeast gene expression data, the experimental results show that our algorithm can generate larger biclusters with lower mean square residues simultaneously.

Keywords: biclustering, mutual information, gene expression data.

1. Introduction

With the rapid development of bioscience and computer science, Bioinformatics became a newly forming discipline combining bioscience and computer science. With the rise of bioinformatics a series of high-throughput detection techniques have been developed rapidly, such as cDNA microarray experiments and the gene chip technology, which have produced huge amounts of high dimensional gene expression data, one example as shown in Figure 1. Those technologies use the same principle, which uses each pairing of complementary characteristics of the four nucleotides, two pairs of single nucleotide chains which are complementary to each other are formed in a double chain, this process is called hybridization.

Gene is the basic unit of genetic information in organisms, gene expression is using the genetic information stored in DNA, through transcription or translation to perform biological functions. By measuring those expression patterns of genes under different conditions, different development stages or different tissues, we can establish the database of gene expression matrix, then we can analyze and summarize gene expression data better. The analysis of gene expression data helps to explain gene expression mechanism and understand the function of genes, find how did genome be influenced by various factors,

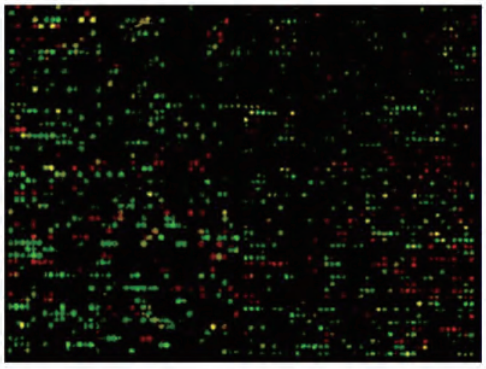


Fig. 1. Gene expression values from cDNA microarray experiments can be represented as heat maps to visualize the result of data analysis

and understand genetic network, then provide information of diseases pathogenesis for further theoretical and applied research at last.

Gene expression data are usually represented by one data matrix, each row is represented as a gene, each column is represented as a condition or sample, as we can see in Figure 2. How to search for potential biological information from high dimensional genetic data becomes an urgent problem ,which need to be solved in data mining technology.

	c1	c2	c3	c4	c5
g1	1	2	3	4	5
g2	2	4	6	8	10
g3	3	6	9	24	25
g4	4	8	12	34	35
g5	5	10	15	4	5
g6	7	12	18	12	15

	c1	c2	c3	c4	c5
g1	1	2	3	4	5
g2	2	4	6	8	10
g3	3	6	9	24	25
g4	4	8	12	34	35
g5	5	10	15	4	5
g6	6	12	18	12	15

Fig. 2. Some example of clustering algorithm. Left: the result of traditional clustering algorithm. Right: result of biclustering algorithm

At present, main methods of analyzing gene expression data are clustering analysis methods. Through clustering analysis, those genes with similar expression patterns are clustered into one category. On this idea, we search for genes with similar patterns, analyze the function of genes, and analyze the transcriptional regulation of genes. The traditional clustering methods can be divided into partitioning method, hierarchical method, density-based method, grid-based method and model-based method. Traditional clustering methods have made some achievements in the analysis of genetic data, but they can only cluster the gene data using global information, the clustering results either contain all rows of data matrix, or all columns of data matrix, but there are usually existing some lo-

cal correlation between genes and conditions for gene expression data. For example, some genes show similar patterns of expression sometimes only under certain subset of conditions, and one gene under different subset of conditions may show different expression patterns. Therefore, traditional clustering algorithms are not very ideal for the analysis of gene expression data in many cases.

The biclustering algorithm as a new method which is introduced to clustering gene expression data from gene dimension and condition dimension simultaneously, which overcomes the limitation of traditional clustering methods. The concept of biclustering algorithm was firstly introduced by Cheng and Church [5] and was applied to the analysis of gene expression data. After that there are emerged a lot of excellent improved biclustering algorithm, those algorithm have achieved considerable results in the biological data mining, such as FLOC algorithm [23], Evolutionary Algorithm [4] and MIB algorithm [8]. The biclustering algorithm is repeatedly clustering from the gene dimension and the condition dimension, and using this local correlation information between genes and conditions to improve the accuracy of clustering results.

Currently most of biclustering algorithms use ordinary Euclidean distance as the similarity measurement between genes, but Euclidean distance can only detect certain linear relationship of gene expression data, and there are existing some complicated nonlinear similarity relationship between biological data. The concept of mutual information comes from information theory, it's commonly used to represent the relationship between information. And when calculated the similarity between genes, different conditions have different effect on the expression pattern of gene information, therefore we set different weight values for different conditions under different biclusters, which used to measure genes' similarity. In this paper our proposed a weighted mutual information biclustering (WMIB) algorithm used the weighted mutual information as the similarity measure of genetic data. Through a series of experiments, we show that our proposed WMIB algorithm has excellent performance, which can not only obtain different types of biclusters, but also ensure that those biclusters have lower mean square residues.

The reminder of this paper is organized as following. Section 2 briefly reviews existing biclustering algorithms in the context of gene expression data. Section 3 defines some theoretical concepts and notations used in our algorithm. Section 4 introduces the framework of our algorithm and details of our algorithm's implementation. Then we further compare with other biclustering algorithm and our experimental results is shown in Section 5. Finally Section 6 contains the conclusion and future work.

2. Related Work

Cheng and Church [5] firstly proposed the concept of biclustering algorithm called CC algorithm, CC algorithm used a greedy iterative searching method to find biclusters, through gradually add or remove rows or columns of genetic data which reduce the mean square residues of biclusters, which get better biclusters after iterations. But CC algorithm could not find overlapping biclusters, Yang et al [23] presented an FLOC algorithm, by calculating the gain function of each action to determine either add or delete one row or column from biclusters. Then some of evolutionary method was proposed, Sefan et al [4] proposed Evolutionary Algorithm (EA) framework which apply some intelligent optimization algorithms to optimize the biclustering result. Pontes presented Evo-Bexpa (Evolutionary

Biclustering based in Expression Patterns) is the first biclustering algorithm in which it is possible to particularize several biclusters features in terms of different objectives. Filipiak [6] proposed HEMBI using an Evolutionary Algorithm to split a data space into a restricted number of regions.

Wang [19]used exhaustive strategies to find biclusters of data, then Liu et al [11] improved algorithm. Tanay [17] proposed a bicluster algorithm called SAMBA that converts biclustering problem into a balanced bipartite graph search problem. Zhu [21] combined simulated annealing technique and particle swam algorithm, presented a simulated annealing particle swarm optimization algorithm. Swarup [14] presented CoBi which used a BiClust tree that needs single pass over the entire dataset to find a set of biologically relevant biclusters. Xu [22] presented an efficient exhaustive algorithm to search contiguous column coherent evolution biclusters in time-series data. Haifa [16] proposed EnumLat algorithm which is the construction of a new tree structure to represent adequately different biclusters discovered during the process of enumeration.

Zhang et al [25][24] proposed a DBF algorithm based on frequent pattern mining. Zhu [26][21] proposed a biclustering algorithm based on hierarchical clustering. Madeira [12] proposed an efficient biclustering algorithm for finding genes with similar patterns in time-series expression data. Rui [15] propose new biclustering algorithms to perform flexible, exhaustive and noise-tolerant biclustering based on sequential patterns (BicSPAM). BicSPAM is the first attempt to deal with order-preserving biclusters that allow for symmetries and that are robust to varying levels of noise. Wang [20] found an UniBic algorithm is to apply the longest common subsequence (LCS) framework to selected pairs of rows in an index matrix derived from an input data matrix to locate a seed for each bicluster to be identified. And some security algorithm [9][10] was proposed for data analysis.

The mean square residue [5] and some valuations criterions that based on the residue are widely used in biclustering algorithms. Teng [18] proposed an average correlation value(ACV) to evaluate the homogeneity of biclusters, which is more reasonable with the co-expression of genes and conditions in biological data. Wassim [2] proposed BiMine algorithm used Average Spearman's rcho (ASR) as evaluation function, later Wassim [3] proposed another evaluation function called Average Correspondence Similarity Index (ACSI) to assess the coherence of given biclusters. Gupta [8] used mutual information to detecting non-linear relationship between genetic data. Aggarwal [1] presented a novel ensemble technique for biclustering solutions using mutual information.

3. Preliminaries

In this section, we will provide notations and preliminaries related to our work.

Gene expression data is usually represented by a data matrix, one row represents one gene and one column represents one condition (or one sample under specific tissues and development stage), each value of matrix represents the expression level of one gene under a specific condition, one row is often referred as a gene expression profile. Analysing the gene expression matrix is used to extract potential biological information. Given the gene expression data, let $G = \{g_1, g_2, g_3, \cdots, g_N\}$ be represented as the set of genes and $C = \{c_1, c_2, c_3, \cdots, c_M\}$ be represented as the set of conditions, where N and M are the number of genes and the number of conditions respectively. Then the expression data can

be represented as a matrix $D_{N\times M}$, where each element value d_{ij} in matrix corresponds to the logarithmic of the relative abundance of the mRNA of one gene g_i under one specific condition c_j.

Definition 1 *Given the gene matrix $\mathbb{G} \times \mathbb{C}$, a bicluster can be defined as a pair (I, J), where I⊂ G be subset of genes G and J ⊂ C be subsets of conditions C.*

A bicluster essentially corresponds to a submatrix in which a subset of genes exhibits consistent patterns under a subset of conditions. For a given gene data expression dataset $\mathbb{G} \times \mathbb{C}$, biclustering algorithm finds a set of submatrixes $(I_1, J_1), \cdots, (I_k, J_k)$ of the matrix $\mathbb{G} \times \mathbb{C}$, $|I_k|$ is the number of specified genes in the k-th bicluster(I,J). A set of biclusters can also be represented as $B = \{B_1, B_2, B_3, \cdots, B_k\}$, where k is the number of biclusters, and B_i is represented as i-th bicluster.

Definition 2 *Given the bicluster (I, J), the volume of a bicluster V_{IJ} is defined as the number of elements d_{ij} in bicluster (I, J) where $i \in I$ and $j \in J$.*

Given the bicluster (I, J), we can have $V_{IJ} = |I| \times |J|$, where $|I|$ and $|J|$ are the number of genes and the number of conditions respectively. Figure 3 shows a gene expression matrix with eight genes and six conditions, for one bicluster, we pick $I = \{g_2, g_3, g_5, g_7\}$ as genes set and $J = \{c_1, c_3, c_5\}$ as conditions set, then the volume of this bicluster is 12.

Definition 3 *Give the bicluster (I, J), d_{ij} is one element value of the bicluster, where $i \in I$ and $j \in J$. The base of one gene d_{iJ} is defined as the average values of gene g_i under certain specified conditions J, it can be calculated by $d_{iJ} = \frac{1}{|J|}\sum_{j\in J} d_{ij}$.*

Similarly, the base of a condition c_{Ij} is defined as the average values of c_j under the specified genes I, it calculated by $d_{Ij} = \frac{1}{|I|}\sum_{i\in I} d_{ij}$. And the base of bicluster d_{ij} can be defined as the average values of each element in bicluster (I, J), calculated by $d_{IJ} = \frac{1}{|I||J|}\sum_{i\in I, j\in J} d_{ij}$.

The base of gene d_{Ij} and the base of condition d_{iJ} may reflect the consistency of information in the corresponding genes or conditions.

Definition 4 *The mean square residue of a bicluster (I,J) can be represented as $r_{IJ} = \frac{1}{V_{IJ}}\sum_{i\in I, j\in J} r_{ij}^2$, where $r_{ij} = d_{ij} - d_{Ij} - d_{iJ} + d_{IJ}$ is the residue of each element d_{ij} in bicluster (I, J).*

The mean square residue of a bicluster can be regard as an important criterion to detect the consistency of the bicluster, the lower the residue, the stronger consistency exhibited by the bicluster. The mean square residue of biclusters are commonly used to evaluate the overall quality of a bicluster.

For example, as show in Figure 3 we have a gene expression matrix with eight genes and six conditions, we pick genes set of $I = \{g_2, g_3, g_5, g_7\}$ and conditions set of $J = \{c_1, c_3, c_5\}$ as one perfect bicluster (I, J). The bases of genes are $d_{2,J} = 40$, $d_{3,J} = 40$, $d_{5,J} = 40$, $d_{7,J} = 40$, the bases of conditions are $d_{I,1} = 50$,$d_{I,2} = 40$, $d_{I,3} = 30$, then the base of bicluster is $d_{I,J} = 40$, so the residue of $d_{1,1}$ obtained by $r_{1,1} = d_{1,1} - d_{I,1} - d_{1,J} + d_{IJ} = 0$, similarly we calculated the residues of other elements in bicluster, finally we can obtain the mean square residue of bicluster (I, J) is 0.

	c1	c2	c3	c4	c5	c6
g1	33	40	45	50	40	70
g2	50	47	40	80	30	33
g3	50	41	40	60	30	44
g4	55	47	80	55	80	70
g5	50	40	40	50	30	55
g6	66	47	45	55	36	44
g7	50	80	40	70	30	46
g8	47	55	45	50	55	44

Fig. 3. The example of a bicluster: all grey color cells represent one bicluster obtained from the dataset, the mean square residue of this bicluster is zero

4. Algorithm Implementation

In this section, we will introduce our proposed WMIB algorithm in detail, which can efficiently and accurately discovered biclusters from gene expression data.

At the beginning of WMIB algorithm, In section 4.1, we wiil introduce the weighted mutual information as new similarity measure between genes, then we will construct a set of seed genes from the dataset as the initial biclusters, which has the least similarity between each seed genes of initial biclusters. In section 4.2, we will use one possibility function to calculate the possibility between each gene from entire data with seed genes of initial biclusters, then we divide genes into corresponding biclusters according those possibility of genes which is greater than the given threshold. Then in section 4.3 we constructed a novel objective function, and by optimizing this objective function we could update the weights of each condition in biclusters, then we remove the conditions set whose have smaller weights in each bicluster. After that we obtained some biclusters according to the updated weights of conditions. Then we can recalculate the new seed genes of each bicluster, and using new seed gene in each bicluster we redivided each gene into biclusters according seed gene in each bicluster and the similarity threshold as show in section 4.4. After completing those steps, in section 4.5 we optimize the obtained biclusters using some novel rules. Finally we will conclude the process of our WMIB algorithm as Figure 4.

4.1. The Construction of Seed Gene Sets

At the beginning of WMIB algorithm, we should construct the seed genes set. At first we initialize the biclusters set B and seed genes set S are empty set, we randomly selected one gene from dataset as the seed gene of the first bicluster, and we add the seed gene into set B and set S, then we calculated the similarity between this seed gene with remaining genes in dataset. Firstly we introduce the measurement used to calculate similarity between genes in our algorithm.

Result of any biclustering algorithm depends on the choice of the similarity measure used. Different similarity measures on the same expression data could produce different results.The mainly similarity measurement used to biclustering gene expression data is Euclidean distance, Mahalanobis distance, and cosine similarity function, but these functions can only measure the linear relationship between genes. However, in gene expression data, there may not only exist a simple linear dependencies between genes, but also exist

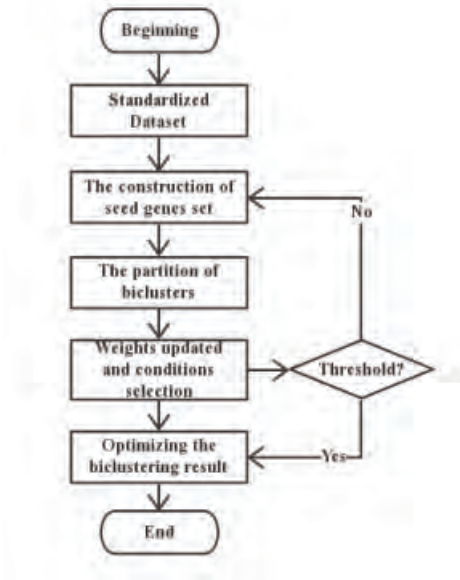


Fig. 4. The process diagram of WMIB algorithm: we briefly present several steps of our algorithm

a complex nonlinear relationship between objects, so these similarity function may not be satisfactory to measure the similarity between genes.

Mutual information is a concept introduced by information theory, which is used to represent the relationship between information. Mutual information had been widely used in many traditional clustering methods, which was proved to be able to detect nonlinear relationships between data, and it is not sensitive to noise data, so in this algorithm we use the mutual information as similarity measure to complex relationship between genes.

The concept of information entropy is a measure of the information contained in the data, the information entropy of a discrete variable X can be defined as follows:

$$H(X) = -\sum p_i log p_i \tag{1}$$

where p_i is the probability of i-th state occurred in X Then the concept of the joint information entropy of two discrete variable X and Y can be defined as:

$$H(X,Y) = -\sum p(x,y) log p(x,y) \tag{2}$$

where $p(x,y)$ is the joint probability of discrete variable X and Y. The definition of mutual information can be defined using the concept of information entropy, the formula of mutual information between two discrete variable X and Y is defined as:

$$M(X,Y) = H(X) + H(Y) - H(X,Y) \tag{3}$$

The calculation of mutual information usually relates to the probability of the random variables' marginal distribution and joint distribution. In most cases these distributions

are unknown, so those requires the estimation of probability density function through the prior knowledge and the statistics. Here we use the Gaussian density function which is commonly used to estimate the probability density distribution of data. For a random variable X, it's probability density estimation as:

$$\hat{p}(X) = \frac{1}{Nh\sqrt{2\pi}} \sum_{i=1}^{N} exp(\frac{-(x - x_i)^2}{2h^2}) \tag{4}$$

The mutual information $M(X, Y)$ is zero if X and Y are independent and it's value is high if they are highly dependent to each other. Supposed that the observations of two random variables X and Y are represented as $\{x_1, x_2, x_3, ..., x_n\}$ and $\{y_1, y_2, y_3, ..., y_n\}$. After brought the probability estimation function of variables into the function of mutual information, the mutual information between X and Y can be represented as:

$$M(X, Y) = \frac{1}{N} \sum_{i=1}^{N} log \frac{\hat{p}(x_i, y_i)}{\hat{p}(x_i) * \hat{p}(y_i)} \tag{5}$$

The weighted measurement is a very common methods in statistics, the weight of a certain index is the relative importance of the index in the overall evaluation.

From the aspect of gene expression data, different conditions shows different influence on expression patterns of gene information in one bicluster, and for different biclusters same conditions may show different effects. In this paper, we apply the weighted thought to the calculation of mutual information to evaluate different effects of conditions.Then we propose a new similarity measurement function, called the weighted mutual information which is represented as:

$$M(X, Y) = \frac{1}{N} \sum_{i=1}^{N} w_i * log \frac{\hat{p}(x_i, y_i)}{\hat{p}(x_i)\hat{p}(y_i)} \tag{6}$$

Most of biclustering algorithms obtain results by choosing the seed genes randomly, but this will lead to the instability of the algorithm, and it can not guarantee that the algorithm will be able to obtain biclusters with consistent volatility. In this paper we use a novel way to initialize the seed genes set.

We start with a random seed gene, then we use the weighted mutual information between the seed and remaining genes to choose the new seed. The next choice of the seed can be a gene whose has the least similarity with the previous seed, and we add this seed into biclusters set B and seed genes set S. Then we obtain the seed genes of each bicluster iteratively.

The next choice of the new seed for next bicluster as follows:

$$\underset{g_j \notin S}{min} \{ \sum_{g_i \in S} M(g_i, g_j, C) \} \tag{7}$$

where g_i and g_j represent i-th and j-th gene in dataset respectively.

Using this initialization method, our algorithm select the gene as the seed of next bicluster whose has the least similarity with previous seeds, which can also avoid the obtained biclusters having high repetition rate. After that we can obtain the initialized biclusters set and the initialized seed genes set. But each bicluster only contains the seed gene, and those seed genes includes all conditions.

4.2. The Fuzzy Partition of Biclusters

After constructed the seed genes set, we partition the set of genes which are most related to the input seed genes set. Firstly we introduce the membership function to calculate the probability of one gene belonging to one bicluster.

From the aspect of gene expression data, different expression patterns may exist in one gene, so one gene may belong to multiple biclusters. General non-fuzzy clustering method is difficult to detect this multiple partition especially when the experimental data is merged by a plurality of experimental data under different conditions.

In order to identify this multiple co-expressed relationship between genes, we use the thought of fuzzy clustering algorithm to complete the partition of biclusters. We first calculated the probability of reminding genes belonging to each bicluster, then the gene is divided into biclusters by the given possibility threshold. The probability function of the gene belonging to the bicluster is defined as:

$$P_{ik} = \sum_{i=1}^{|J_k|} P_{ijk} = \frac{1}{|J_k|} \sum_{i=1}^{|J_k|} M(g_i, g_k, Y_k) \tag{8}$$

where $|J_k|$ represents the number of conditions in bicluster B_k, and $M(g_i, g_k, Y_k)$ is represented as the weighted mutual information between gene g_i and g_k under conditions set Y_k. After we selected genes for each bicluster iteratively, we get the initial fuzzy partition of entire dataset. After such fuzzy partition of genetic data, each gene may simultaneously belong to several different biclusters, or may not belong to any of biclusters. Meanwhile, in order to avoid biclusters obtained by the fuzzy partition with a high overlapping rate, and ensure that those biclusters contain better biological information, we set an overlapping rate threshold. If the size of one biclusters is larger than this threshold then that bicluster will be pruning. After that each gene in the initial fuzzy partition of biclusters still contains all conditions, next we need to complete the selection of conditions.

4.3. Weights Updated and Conditions Selection

When we completed the initial fuzzy partition of entire dataset, we should update weights of conditions for all of biclusters, and complete the selection of condition sets. In our algorithm, we set one weight for each condition, which used to measure the effect of conditions in current bicluster, then through the weights updated to selected conditions of the corresponding bicluster, we remove conditions from biclusters whose weights lower than the weights threshold

Most of the biclustering algorithm use the iterative method to complete selection of condition sets, but the iterative process has high time complexity, and it maybe fill into local minima. Although some biclustering algorithm has proposed some objective function, which is used to obtained some good biclusters, but those objective function only consider to minimize the mean square residue of biclusters, and they are not including the influence of weights for finally biclusters. In this section we propose a new objective function, through optimizing this objective function we can be quickly update the weights of each condition, and we remove conditions from biclusters whose weights lower than the weights threshold to selected conditions for each bicluster.

The mean square residue (MSR) is the most widely criterion to measure the quality of biclustering algorithms, In order to improve accuracy of our biclustering algorithm we combine weights and MSR as the first part of the objective function, by minimizing mean average residue of each bicluster we can have better quality biclusters. The calculation formula of weighted MSR as following:

$$H(I_k, J_k) = \frac{1}{|I_k||J_k|}\sum_{i,j} w_{kj} * (d_{ij} - d_{I_k j} - d_{iJ_k} + d_{I_k J_k})^2 \tag{9}$$

where $d_{iJ_k}, d_{I_k j}, d_{I_k J_k}$ is represented as mean value of gene g_i, condition c_j and bicluster B_k respectively. Then we can obtain the first part formula of new objective function as:

$$R_k = min\sum_{i=1}^{p} w_{kj} \sum_{g_j \in B_k} (d_{ij} - d_{I_k j} - d_{iJ_k} + d_{I_k J_k})^2 \tag{10}$$

The objective function should guarantee that the size of biclusters, if the size of conditions of biclusters too small will lose much important biological information of genetic data. So we use the second part of objective function to control the size of conditions set. We use weights to approximate the size of biclusters, and use some constraint criteria as follows:

$$S_k = \sum_{j=1}^{p} \sqrt{w_{kj}} \quad where \sum_{j=1}^{p} w_{kj} = 1 \tag{11}$$

Since result of two part R_k and S_k are constraint in different ranges, it's inconvenience when optimize the objective function, so we use some constraint criteria to transform objective function, then we obtain the final formula of new objective function as:

$$\Gamma = min \sum_{k=1}^{K} \left\{ \sum_{j=1}^{p} w_{kj} * R_k - \frac{1}{p} \sum_{i=1}^{p} R_k * \frac{\sum_{j=1}^{p} \sqrt{w_{kj}} - 1}{\sqrt{p} - 1} \right\} \tag{12}$$

let w_{kj} be argument of the objective function Γ, we can see that the objective function is a convex function by definition, and we can directly optimize the objective function by the gradient method. Firstly we compute the gradient of the objective function, then we set the gradient to be zero, we can obtain the update formula of weights through transformation:

$$w_{kj} = \frac{1}{4p^2(\sqrt{p} - 1)} \left\{ \frac{1}{\sum\limits_{g_j \in B_k} H_{ik}} \sum_{i=1}^{p} \sum_{g_j \in B_k} R_{ik} \right\}^2 \tag{13}$$

where $H_{ik} = (d_{ij} - d_{I_k j} - d_{iJ_k} + d_{I_k J_k})^2$.

After we update the weights value of each bicluster, we normalize the updated weights using all weights of each bicluster. After that we compared normalized weights of each bicluster with the given weights threshold separately.

Then we select conditions set of each bicluster by setting weights of conditions are zero, whose weights are lower than the given weights threshold. After that we updated all weights of conditions and selected conditions set for each bicluster simultaneously.

$$w_{kj} = \begin{cases} w_{kj} & if\ w_{kj} \geq \gamma \\ 0 & else \end{cases} \tag{14}$$

4.4. The Re-partition of Biclusters

In the last section we have selected conditions set for each bicluster in accordance to the objective function, We need to recalculate the seed genes set of each bicluster, and then cluster data matrix in accordance to the updated set of seed genes and updated weights. Repartition of biclusters can processed by two steps:

Firstly, according to the updated weights and the conditions set, we need to recalculate the seed genes set of each bicluster, the seed gene s_k of k-th bicluster is calculated as :

$$s_k = \frac{1}{|I_k|}\sum_{i \in I_k} d_{ij} \tag{15}$$

Secondly, we use the newly seed genes set to recalculate the probability of genes belonging to each bicluster, then we add the genes into biclusters for whose probability are greater than the given possibility threshold.

After those two steps, we can obtain better biclulsters from gene expression data with more consistent volatility.

4.5. Optimizing Biclustering Results

By repartitioning of the biclusters, we can obtain the new set of biclusters from dataset, but that is not guaranteed that each bicluster has lower mean square residue (MSR). Most of the current biclustering alogorithm use mean square residue as the standard of evaluating biclustering results, but it is not very ideal for the assessment of certain structure of biclusters. In order to get more reasonable structure of biclusters, we use a new kind of weighted mean square residue to optimize biclusters obtained by repartition.

Firstly, we give the calculation formula of weighted mean square residue. $H(I_k, J_k)$ is represented as weighted mean square residue of the bicluster B_k.

$$H(I_k, J_k) = \frac{1}{|I_k||J_k|}\sum_{i,j} w_{kj} * (d_{ij} - wd_{I_k j} - wd_{iJ_k} + wd_{I_k J_k})^2 \tag{16}$$

WR_i and WC_j are represented as weighted mean square residue of the gene g_i and the condition c_j respectively

$$WR_i = \frac{1}{|J_k|}\sum_{j \in J_k} w_{kj} * (d_{ij} - wd_{I_k j} - wd_{iJ_k} + wd_{I_k J_k})^2 \tag{17}$$

$$WC_j = \frac{1}{|I_k|}\sum_{i \in I_k} w_{kj} * (d_{ij} - wd_{I_k j} - wd_{iJ_k} + wd_{I_k J_k})^2 \tag{18}$$

where wd_{iJ_k},$wd_{I_k j}$,$wd_{I_k J_k}$ are represented as weighted mean value of the gene g_i, the condition c_j and the bicluster B_k respectively. and their definition as:

$$label19 wd_{iJ_k} = \frac{1}{|J_k|}\sum_{j \in J_k} w_{kj} d_{ij}, \tag{19}$$

$$label20 wd_{I_k j} = \frac{1}{|I_k|}\sum_{i \in I_k} w_{kj} d_{ij}, \tag{20}$$

$$label21wd_{I_kJ_k} = \frac{1}{|I_k||J_k|}\sum_{i,j} w_{kj}d_{ij} \tag{21}$$

Here we use the weight mean square residue of each bicluster to optimize the biclustering results, which can be divided into two situations:

When one gene is contained in one bicluster, we first assume that this gene is deleted from the bicluster, and we calculate the weighted mean square residue of new bicluster. If the new weighted mean square residue is less than previous weighted mean square residue, then we remove this gene from this bicluster.

When one gene is not contained in one bicluster, we first assume that this gene is added into one bicluster, and we calculate the weighted mean square residue of new bicluster. If the new weighted mean square residue is less than previous weighted mean square residue, then we add this gene into this bicluster.

Cheng and Church [5] have proved that it would not increase the MSR of bicluster if we add one gene into one bicluster, which the MSR of gene is less than the MSR of bicluster. By optimizing the biclustering results obtained by previous section, our algorithm could find out larger biclusters with lower mean square residue, which indicated that we obtain better bicluster with consistent volatility.

4.6. The Processes and Complexity of WMIB Algorithm

In Algorithm 1 we show the main procedure of WMIB algorithm.

Algorithm 1 WMIB Algorithm

Input: Gene Data Matrix D, the number of biclusters α, the possibility threshold β and the weights threshold γ;

Output: Biclusters set B;

1: Initialize the biclusters set B and seed genes set S are empty, and select one gene randomly as the seed of the first bicluster;
2: Calculate weighted mutual information between each gene and seed genes, then select next seed gene according to the formula (7);
3: Calculate the probability P_{ik} of gene g_i belonging to bicluster B_k using the formula (8), then adding this gene g_i into bicluster B_k if its probability P_{ik} greater than the possibility threshold β;
4: Update the weights of comditions for each bicluster using the formula (13), then normalize weights;
5: Set the weights values of conditions to zero if weights less than weights threshold γ, which is used to select conditions set, then re-normalize weights of conditions;
6: Calculate new seed genes for each bicluster using the formula (15), then repartition of data matrix as step $3-4$;
7: Calculate weighted MSR of genes for each bicluster, and optimize the obtained biclusters according to weighted MSR;

In our proposed WMIB algorithm, the main complex process are the construction of seed genes set and the partition of biclusters. When constructed seed genes set, a set of seed gene and the initial biclusters are generated. The time complexity of computing weighted mutual information between genes is $O(M^2)$, where M is the number of

conditions, there are at most $\frac{k(k+1)}{2} \times N$ similarity computions between genes in this process, so the time complexity of construction of seed gene set is $O(k^2 \times N \times M^2)$, where k is the number of biclusters and N is the number of genes. In the second process, a set of biclusters are generated from genetic dataset, there are at most $c \times N$ similarity computions between genes in this process, so the time complexity of this process can represented as $O(k \times N \times M^2)$. Thus, the overall time complexity of our proposed algorithm is $O(k^2 \times N \times M^2)$.

5. Experimental Results

5.1. Dataset and Standardization

In this section, we use the yeast metabolic cycle expression datasets GDS2267 from Gene Expression Omnibus (GEO) database to evluate our proposed WMIB algorithm, the dataset contains 9335 genes and 36 conditions, and it has commonly used to evaluate the performance biclustering algorithm. In this dataset, genetic data is represented as a data matrix, each row is represented as one gene, and each column is represented as one condition. We construct biclusters to find submatrix which have consistency volatility.

In order to reduce the influence of the different attributes of the data or the variance of the data on the biclustering results, and we can compare accurately biclustering results obtained by other main algorithms, we firstly standardized the gene data, following the formula as:

$$g_{ij}^{'} = \frac{g_{ij} - \bar{g}_i}{S_i} \tag{22}$$

where $\bar{g}_i$ is represented as mean value of gene g_i, and S_i is represented as standard deviation of gene g_i.

5.2. Comparison and Visualization

Our WMIB algorithm is implemented with Java programming language and is executed on an AMAX machine. The hardware environment of this experiment as follows: Intel Xeon E5-1620 3.50GHz, 16G memory. The Software environment is Eclipse on Ubuntu operating system.

In order to comprehensively verified the performance our proposed WMIB algorithm, we selected four evaluation criterions as the mean square residue(MSR), average volume, average rows and average columns together measure the performance of biclustering algorithms. The mean square residue of biclustering result is average value of all biclusters' MSR, As the more smaller of mean square residue of biclusters, the consistency of each bicluster is more better. And the average volume is average number of each bicluster's elements, average rows and average columns are the average number of each bicluster's genes and conditions respectively, when the mean square residue of biclustering results are equal, as the average number of genes and average number of conditions become more higher, the performance of biclustering algorithm seems more better.

We compare our algorithm with multiple mainly biclustering algorithm, and the experimental results are shown in Table 1. Note that because of our algorithm has a certain randomness when selected the seed genes, for which we carried out several experiments, the experimental results as shown in Table 1 is the average result selected 30 experiments.

Table 1. Comparison of main biclustering algorithm

Heading level	MSR	Volumes	AvgRow	AvgColumn
DBF algorithm [25]	115.00	1627.00	188.00	11.00
WMIB algorithm	121.842	10509.13	911.46	11.53
IBWMSR algorithm	142.060	8270.06	756.25	10.93
FLOC[23]	187.543	1825.78	195.00	12.80
CC[5]	204.290	1557.98	167.00	12.09
Hierarchical Cluster[26]	220.156	1098.10	171.60	7.90
Multi-objectiveGA[13]	235.000	10302.00	1095.00	9.29
Possibilistic[7]	297.000	22571.00	1736.00	13.00

As we can see from Table 1,Compared with other commonly used biclustering algorithm, our algorithm can produce better quality biclusters from gene dataset, although the mean square residue of our experimental result is relatively larger than DBF algorithm, but the volumes and average number of genes of DBF algorithms are too small,, the volumes of our results are almost seven times larger than DBF algorithm, the results of DBF algorithm may lose abundant genetic information compared with our algorithm. And Compared with other popular biclustering algorithms, the experimental results of our algorithm has the lowest mean square residue than other algorithm, which show that our proposed WMIB algorithm can detect the biclusters with better consistency from gene dataset, and our results have the biggest volumes compared with other biclustering algorithms except Possibilistic biclustering algorithm. Our WMIB algorithm has smaller average volumes compared with Possibilistic algorithms, this is because our algorithm not only can cluster bigger volumes of biclusters, but also can cluster some smaller biclusters from dataset. From above we can proved that the WMIB algorithm has a good performance, it can find biclusters set with highly consistent fluctuation from the high-dimensional genetic data with highly consistent, and it can find larger biclusters meanwhile detecting some small volumes biclusters.

In order to observe the fluctuation trend of the biclusters which obtained by our algorithm directly, we randomly selected 4 biclusters from the result biclusters set and visualized the data of those biclusters.

As we can see from Figure 5, the biclusters obtained by WMIB algorithm has similar fluctuation trend, which can show it's good consistency. Our proposed WMIB algorithm and IBWMSR algorithm exists many similarities, they both use fuzzy cluster to partitioning the dataset, and they both set different weights for different conditions to determine the impact extent for biclusters results. But WMIB algorithm uses weighted mutual information as the similarity metrics between, it can be simultaneously detected complex linear and nonlinear correlation between genes, and IBWMSR algorithm used the weighted Euclidean distance. Compared with the IBWMSR algorithm, WMIB algorithm can find out better co-expression level of biclusters, which have smaller mean square residue, and can guarantee that obtain the larger volume of biclusters.

In order to fully verify that the weighted mutual information can effectively reflect the characteristics of the genetic data as the similarity measure between genes, we compared the mean square residue of two different biclusters set obtained by our algorithm used different similarity measure, Weighted MI represents the biclusters obtained by our

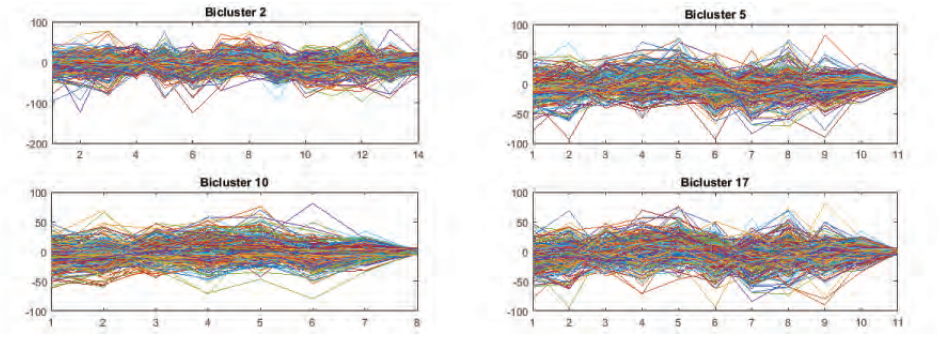


Fig. 5. Fluctuation of Biclusters: we randomly choose four biclusters from experimental results obtained by our biclustering algorithm

algorithm whose used weighted mutual information as the the similarity measure between genes, Weighted ED represents the biclusters obtained by our algorithm which used weighted Euclidean distance as the similarity measure between genes.

Table 2. Comparison of MSR of Different biclusters

Heading level	Weighted MI	Weighted ED
1-th bicluster	109.63	139.08
2-th bicluster	122.03	125.12
3-th bicluster	147.86	145.92
4-th bicluster	112.13	167.26
5-th bicluster	86.76	108.71
6-th bicluster	130.62	140.06
7-th bicluster	101.63	131.61
8-th bicluster	132.43	150.33

As we can see from Table 2, we compared 8 biclusters from two biclusters set, most of mean square residue of Weighted MI biclusters have lower than Weighted ED biclusters, which indicates that using weighted mutual information as similarity measure can effectively reflect the complex linear and nonlinear relationship between genes. It also proved that our algorithm can extract more consistent biclusters from complex genetic data using weighted mutual information as similarity measure, which improve the accuracy and performance of biclustering algorithm.

5.3. Overlapping of Biclusters

Our biclustering algorithm used the fuzzy clustering to partition of gene data, so there may exist a high overlap rate between biclusters. To further investigate the performance of our algorithm, we calculated the overlap rate between biclusters. For two biclusters A and B have N_A and N_B number of elements, respectively, the overlapping rate between

two biclusters is:

$$O_{A,B} = \frac{N_{A \bigcap B}}{(N_A + N_B)/2} * 100$$

where $N_{A \bigcap B}$ is the number of elements belonging to both the bicluster A and B. We

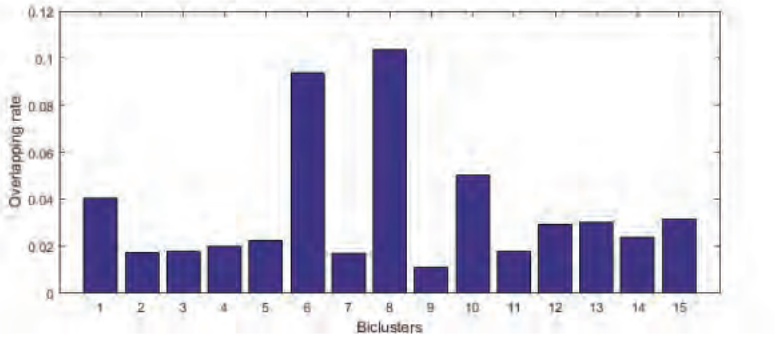


Fig. 6. The histgram of oveppling rate between some biclusters

compared the overlap rate between each bicluster obtained by our algorithm. As we can see from Figure 6, the most highest overlapping rate is still less than 0.2, which shows our algorithm can effectively generate biclusters with reasonable overlapping rate. Those reasonable overlapping rate between biclusters indicate that WMIB algorithm does not generate redundant biclusters from gene dataset. The experimental results show that the WMIB algorithm can successfully cluster better bicluster meanwhile controlling the overlap rate of bicluster in a certain range.

6. Conclusions

How to search for potential biological information from high dimensional gene expression data become an urgent problem to be solved in data mining technology. Biclustering algorithm was introduced to discover biclusters whose subset of genes are co-expressed under subset of conditions. Currently most of biclustering algorithm use Euclidean distance as similarity measure between genes, but it can only detect linear relationship between genes. In this paper, we proposed a new biclustering algorithm called WMIB to find biclusters. In our algorithm we proposed a new weighted mutual information as similarity measure which can be simultaneous detected complex positive, negative correlation and nonlinear relationships between genes. And we constructed a new objective function to optimize biclusters, through weights update and selection of condition sets, which avoid many unnecessary iterations in clustering process and greatly improve efficiency of the biclustering algorithm. Experimental results show that our proposed WMIB algorithm can not only find out biclusters having a low mean square residue, but also generate large capacity biclusters, meanwhile our algorithm can control reasonable overlapping rate between biclusters.

Acknowledgement. This work is supported by National Science Foundation of China Grant No. 61672088, Fundamental Research Funds for the Central Universities No. 2016JBM022 and No. 2015ZBJ007. The corresponding author is Yidong Li.

References

1. Aggarwal, G., Gupta, N.: Bemi bicluster ensemble using mutual information. In: International Conference on Machine Learning and Applications. pp. 321–324 (2013)
2. Ayadi, W., Elloumi, M., Hao, J.K.: A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data. Biodata Mining 2(2), 146–150 (2009)
3. Ayadi, W., Elloumi, M., Hao, J.K.: Bicfinder: a biclustering algorithm for microarray data analysis. Knowledge & Information Systems 30(2), 341–358 (2012)
4. Bleuler, S., Prelic, A., Zitzler, E.: An ea framework for biclustering of gene expression data. In: Evolutionary Computation, 2004. CEC2004. Congress on. pp. 166 – 173 Vol.1 (2004)
5. Cheng, Y., Church, G.M.: Biclustering of expression data. In: International Conference on Intelligent Systems for Molecular Biology ; Ismb International Conference on Intelligent Systems for Molecular Biology. pp. 590–602 (2000)
6. Filipiak, A.M., Kwasnicka, H.: Hierarchical Evolutionary Multi-biclustering. Springer Berlin Heidelberg (2016)
7. Filippone, M., Masulli, F., Rovetta, S., Mitra, S., Banka, H.: Possibilistic approach to biclustering: An application to oligonucleotide microarray data analysis. Lecture Notes in Computer Science 4210, 312–322 (2010)
8. Gupta, N., Aggarwal, S.: Mib: Using mutual information for biclustering gene expression data . Pattern Recognition 43(8), 2692–2697 (2010)
9. Li, Y., Shen, H.: On identity disclosure control for hypergraph-based data publishing. IEEE Transactions on Information Forensics & Security 8(8), 1384–1396 (2013)
10. Li, Y., Shen, H., Lang, C., Dong, H.: Practical anonymity models on protecting private weighted graphs. Neurocomputing 218, 359–370 (2016)
11. Liu, J., Wang, W.: Op-cluster: Clustering by tendency in high dimensional space. In: IEEE International Conference on Data Mining. p. 187 (2003)
12. Madeira, S.C., Oliveira, A.L.: An efficient biclustering algorithm for finding genes with similar patterns in time-series expression data. In: Asia-Pacific Bioinformatics Conference, APBC 2007, 15-17 January 2007, Hong Kong, China. pp. 67–80 (2015)
13. Mitra, S., Banka, H.: Multi-objective evolutionary biclustering of gene expression data. Pattern Recognition 39(12), 2464–2477 (2006)
14. Roy, S., Bhattacharyya, D.K., Kalita, J.K.: Cobi: Pattern based co-regulated biclustering of gene expression data. Pattern Recognition Letters 34(4), 1669–1678 (2013)
15. Rui, H., Madeira, S.C.: Bicspam: flexible biclustering using sequential patterns. Bmc Bioinformatics 15(1), 1–20 (2014)
16. Saber, H.B., Elloumi, M.: An enumerative biclustering algorithm for dna microarray data. In: IEEE International Conference on Data Mining Workshop. pp. 132–138 (2015)
17. Tanay, A., Sharan, R., Kupiec, M., Shamir, R.: Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proceedings of the National Academy of Sciences of the United States of America 101(9), 2981–2986 (2004)
18. Teng, L., Chan, L.: Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. Journal of Signal Processing Systems 50(3), 267–280 (2008)
19. Wang, H.: Clustering by pattern similarity in large data sets. In: ACM SIGMOD International Conference on Management of Data. pp. 394–405 (2002)

20. Wang, Z., Li, G., Robinson, R.W., Huang, X.: Unibic: Sequential row-based biclustering algorithm for analysis of gene expression data. Scientific Reports 6 (2016)
21. Xian, Z., of Computer Science, X.J.D., College, Z., Nanjing, Jiangshu: A gene data biclustering algorithm based on simulated annealing and particle swarm optimization. Computers & Applied Chemistry 30(1), 93–96 (2013)
22. Xu, H., Xue, Y., Lu, Z., Hu, X., Zhao, H., Liao, Z., Li, T.: A new biclustering algorithm for time-series gene expression data analysis. In: Tenth International Conference on Computational Intelligence and Security. pp. 268–272 (2014)
23. Yang, J., Wang, H., Wang, W., Yu, P.: Enhanced biclustering on expression data. In: Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on. pp. 321–327 (2003)
24. Zhang, M., Wen-Hang, G.E.: Overlap bicluster algorithm based on probability. Computer Engineering & Design (2012)
25. Zhang, Z., Teo, A., Ooi, B.C., Tan, K.L.: Mining deterministic biclusters in gene expression data. In: Bioinformatics and Bioengineering, 2004. BIBE 2004. Proceedings. Fourth IEEE Symposium on. pp. 283–290 (2004)
26. Zhu, X., Wei, M.A.: A biclustering algorithm based on hierarchical clustering. Microcomputer Applications (2009)

Yidong Li received the BS degree from Beijing Jiaotong University, the MS and PhD degrees from the University of Adelaide, South Australia. He is currently an associate professor at the School of Computer and Information Technology, Beijing Jiaotong University, China. His research interests include multimedia computing, privacy preserving data publishing, graph/social network analysis, web mining, and distributed computing. He has published more than 60 papers in international journals and conferences, and serves on the program committees of more than 15 international conferences.

Wenhua Liu received the MS degrees from the Shandong University of Science and Technology. She is currently a PhD at the School of Computer and Information Technology, Beijing Jiaotong University, China. She research focuses on Scene classification, Object detection, Data mining. She has published more than 7 papers in international journals and conferences.

Yankun Jia received the BS degree from Qingdao University of Science and Technology. He is currently an master student at the School of Computer and Information Technology, Beijing Jiaotong University, China. His research interests include data mining, and machine learning.

Hairong Dong received the B.S. and M.S. degrees from Zhengzhou University, Zhengzhou, China, in 1996 and 1999, respectively, and the Ph.D. degree from Peking University, Beijing, China, in 2002. She is currently a Professor with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. Her research interests include intelligent transportation systems, automatic train operation. She is a Senior Member of IEEE. She serves as the associate editors of IEEE Transactions on Intelligent Transportation Systems, IEEE Intelligent Transportation Systems Magazine, and ACTA Automatica SINICA.

Received: March 1, 2017; Accepted: August 8, 2017.

Computer Science and Information Systems 14(3):661–684 https://doi.org/10.2298/CSIS170124022Z

An Optimization Scheme for Routing and Scheduling of Concurrent User Requests in Wireless Mesh Networks

Zhanmao Cao[1], Chase Q. Wu[2], and Mark L. Berry[2]

[1] Department of Computer Science, South China Normal University
Guangzhou, Guangdong 510631, China
caozhanmao@m.scnu.edu.cn
[2] Department of Computer Science, New Jersey Institute of Technology
Newark, New Jersey 07102, USA
{chase.wu, mlb32}@njit.edu

Abstract. Multiple-radio multiple-channel (MRMC) wireless mesh networks (WMNs) have been increasingly used to construct the wireless backbone infrastructure for ubiquitous Internet access. These networks often face a challenge to satisfy multiple concurrent user requests for data transfers between different source-destination pairs with various performance requirements. We construct analytical network models and formulate such multi-pair data transfers as a rigorous optimization problem. We propose an optimization scheme for cooperative routing and scheduling together with channel assignment to establish a network path for each request through the selection of appropriate link patterns. The performance superiority of the proposed optimization scheme over existing methods is illustrated by simulation-based experiments in various types of mesh networks.

Keywords: multi-pair paths, compatible paths, multi-radio multi-channel, wireless mesh networks.

1. Introduction

The number of mobile smart terminals is exponentially increasing over years, so is the users' desire for any-where any-time access to the Internet, even in remote rural areas. Recently, Multi-radio multi-channel (MRMC) wireless mesh networks (WMNs) have emerged as a promising solution to provide convenient and ubiquitous broadband access to the Internet. In MRMC WMNs, as router nodes are equipped with multiple interfaces, they may operate in the mode of multiple input multiple output (MIMO). In fact, MRMC represents the main features of the current wireless network infrastructure, where a node interface typically communicates in a dual mode with an omni-directional antenna. Note that a radio may be viewed as the active status of two interfaces on two neighbor nodes over a common channel of wireless media.

MRMC WMNs have brought several important benefits. First of all, they provide significantly more capacity with higher energy efficiency than their predecessors [19]. Secondly, WMNs offer unprecedented flexibility and convenience to expand the covering area by relaying packets hop-by-hop without the support of BS in the mesh mode [6]. Thirdly, the WMN topology remains relatively stable because the nodes are almost static, hence ensuring Quality of Service (QoS) in disparate environments such as a building or

even a smart city. WMNs also enable fast or temporary deployment, which is critical in emergency situations [1].

MRMC WMNs often face a challenge to satisfy multiple concurrent user requests for data transfers between different source-destination pairs (s_i, d_i), $i = 1, \cdots, \rho$, with various performance requirements. Typical examples include a request from an FTP user for transferring data of a certain size z_i from s_i to d_i or any other file transfer request. The multi-pair routing (MPR) problem in WMNs, referred to as WMPR, has a critical impact on the QoS delivered to end users and the utilization of network devices deployed by service providers, especially when resources are limited by a finite number d of antennas and a finite number $|\Omega|$ of orthogonal channels in a given WMN. For illustration purposes, we provide in Fig. 1 an example with four-pair routing, i.e., $(A, D), (B, J), (C, F)$, and (I, H).

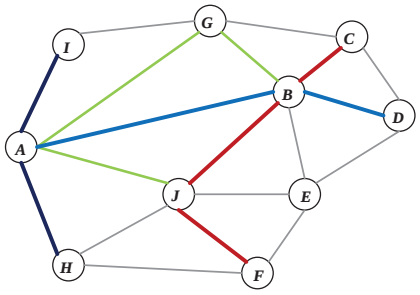


Fig. 1. An example of four compatible paths with joint nodes.

It is important to maximize the utilization of network resources. However, this problem becomes more challenging due to the interference of wireless media. In fact, the approaches to multiple pair shortest path (MPSP) widely adopted in wired networks are not suitable for wireless networks because the wireless radio interference makes this problem a discrete combinatorial one in nature [25]. In existing research, some special constraints are considered, such as edge-disjoint, minimum edge-congestion, or shortest path [4, 27]. A good scheme is needed to solve WMPR for maximum utilization of network (MUN) resources.

WMPR is both practically important and theoretically challenging. Note that each source-destination pair requires a data transfer path, and multi-pair paths may cause interferences to each other. One main goal is to carefully route multiple user requests via different paths and design an efficient scheduling scheme that allows multiple source-destination pairs to simultaneously transmit data packets over their own paths in a cooperative way. Unfortunately, The routing problem of multiple pair paths (MPP) still remains largely unexplored. Even two simplified versions of MPR, multiple pair concurrent paths (MPCP) and multiple pair shortest paths (MPSP), have not been well explored.

A good routing and scheduling scheme should aim to set up as many concurrently active paths as possible, and the links of those paths can be active simultaneously without interferences. This is important to many application scenarios that generate multi-pair data traffic in real time.

Our contributions in this work are two-fold: construct rigorous models to define WMPR, and design efficient algorithms to solve WMPR.

- Network Modeling: We construct analytical network models and formulate WMPR as a rigorous optimization problem to minimize turnaround time under various constraints.
- Cooperative Routing and Scheduling: We design a cooperative routing and scheduling scheme that dispatches multiple user requests along concurrent *compatible paths* without interferences.

The rest of the paper is organized as follows. Section 2 provides a survey of related work. Section 3 constructs network models and formulates the problem. Section 4 designs a cooperative routing and scheduling scheme. Section 5 conducts simulations in triangular meshes and random meshes for performance evaluation.

2. Related Work

As MPR has not been extensively investigated in wireless meshes, we trace several lines of research efforts in wired networks, graph theory, and transportation research.

The MPSP problem has been widely studied in various contexts. Wang *et al.* developed a DLU approach to dense digraph flight scheduling, which is similar to LU decomposition in Carrés algorithm [28]. Their scheme is an algebraic matrix compared with a label-setting method and an LP-based technique.

A weighted digraph problem considers a directed graph with a set of rate demands specified on each source-destination pair $\{s_i, d_i\}$ [16]. Andrews *et al.* designed an almost-tight approximation algorithm for the directed congestion minimization problem. They chose one directed path for every pair (s_i, d_i) to minimize the maximum congestion [5]. However, their work is focused on the theoretical aspect of the problem, without considering the interferences and the limit on channel allocation (CA) in WMNs. Nevertheless, their work at least provided an analysis of the problem's computational complexity and proved the NP-completeness of the maximum utilization problem of MPR. Note that some traffic flows may have joint nodes or even common edges in the network topology.

There exist several research efforts in maximizing the utilization of wireless resources by optimizing throughput or capacity in WMNs. Alicherry *et al.* formulated CA and routing into LP by considering the characteristics of interference, the number of channels, and the number of node radios [2]. Giannoulis *et al.* proposed an iterative method to optimize congestion control by considering CA and traffic distribution [15]. They claimed that the problem is still NP-hard, even in a simpler combinatorial case on CA of MPR in multi-radio networks with a given set of rate demands. After decomposing congestion control into two stages, they formulated the MRMC congest control problem as MRMC-CC, and their optimization method considered CA between multiple pairs only, not for MPCP.

MPP in wired networks was discussed with constraints of minimum edge-congestion or maximum utility of networks [4, 5]. As a necessary step of WMPR in wireless networks, the MPP is even more complex because more constraints have to be considered for

optimization of scheduling, routing, CA, and interference avoidance [25]. Even a simple combinatorial problem involving only one aspect may not have an exact optimal solution. For example, CA to meet a given set of traffic rate demands is NP-hard [24].

We would like to point out that WMPP is different from the traditional disjoint multi-path problem [27], which is focused on finding concurrent paths between one source-destination pair to improve transmission speed and reliability. Different from "multi-path" in [3] or "multipath" in [18], WMPP sets up multiple concurrent compatible paths, each for a different pair of nodes.

Traditional shortest path algorithms are not adequate to solve the problem under study. Even if we do not take into account CA and interference factors, the multiple pair shortest paths (MPSP) problem has already been proved to be NP-hard for edge congestion minimization [4]. Furthermore, these conclusions are only based on a subproblem space or a simplified case of MPP, not yet considering all aspects. For example, the model by Schumacher *et al.* does not take interference into account. In WMNs, MPCP is highly related to interference, routing, link scheduling as well as CA. Consequently, wireless MPCP (WMPCP) is at least as difficult as MPCP in wired networks. On the other hand, MPSP is NP-complete according to Karp [17]. Even only considering CA for real-time data flows, the problem is NP-complete, because CA can be reduced to the 3-partition problem [7]. The problem to perform joint scheduling and routing to achieve maximum utilization of network (MUN) resources is also NP-complete.

As discussed above, WMPR is a challenging problem and has not been thoroughly investigated. In this paper, we tackle this problem considering wireless network resources for minimum turnaround time.

3. Cost Models for WMPR

WMPR aims to achieve the maximum utilization of wireless resources or the minimum turnaround time for user requests. We consider an almost static network topology since routers are always pre-deployed and almost fixed during their operation. To facilitate a rigorous formulation of the problem, we provide below some preliminaries and notations for both the models and the algorithm to be designed.

3.1. Preliminaries

We consider an MRMC WMN structure Γ with a set V of static routers. Each router is equipped with ξ interfaces. There are q orthogonal channels $\{c_1, c_2, \cdots, c_q\}$ that are globally available, each of which has a bandwidth ϖ_i, $i = 1, 2, \cdots, q$. A router node u transmits data to its neighbor node v over a specific channel c_i. The communication link between them is denoted by a directed edge $u\overrightarrow{c_i}v$, or simply $l_{c_i}^{(u,v)}$.

Interference is an inherent nature of wireless networks. At any particular point of time, any node is allowed to send or receive data over a channel only if there is no conflict with other working nodes. Various cases of interference are discussed and categorized in [11]. With initial selected links, let S_C denote the set of possible candidate sender nodes, and let R_C denote the set of possible candidate receiver nodes. The conditions to avoid wireless interference are given in Table 1.

Table 1. Conditions for simultaneous links over one channel.

Three classes of neighbor pairs		Link interference-free conditions	
End nodes from S_C	End nodes from R_C	Necessary Condition	Sufficient Condition
S_i, S_j		$\forall i \neq j, d(S_i, S_j) \geq 2$ (1)	
	R_i, R_j	$\forall i \neq j, d(R_i, R_j) \geq 1$ (2)	(1) $\wedge$ (2) $\wedge$ (3)
S_i	R_j	$\forall i \neq j, d(S_i, R_j) > 1$ (3)	

We consider a set of traffic requests $\Lambda = \{\lambda_i | i = 1, \cdots, \rho\}$, where $\lambda_i = \{(s_i, d_i), z_i\}$, and z_i denotes the size of data to be transferred as in an FTP or any other file transfer request. In some other contexts, an explicit bandwidth may be requested.

A general solution to WMPR first computes a network path to transmit data packets of each traffic request from its source node to its destination node, followed by scheduling and CA. A network path is often defined as a finite hop-by-hop node sequence for packet forwarding. Any two nodes with a direct link in the sequence are considered as neighbors. For a pair (s_i, d_i), its network path is in the form $p_{(s_i,d_i)} = \{s_i, v_{i_1}, \cdots, v_{i_{h^i-1}}, d_i\}$ of length h^i. Given a routing path for λ_i, scheduling is to assign a transmitting time segment to each component link on the computed path, and CA is to assign a channel to each selected link according to the schedule. The scheduling turnaround for path $p_{(s_i,d_i)}$ is illustrated in Fig. 2.

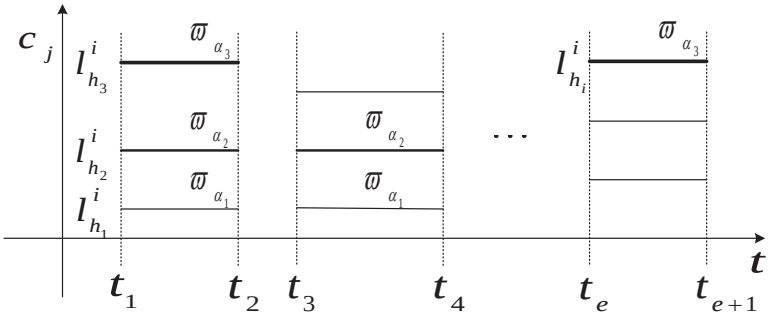


Fig. 2. An example of path scheduling over multiple channels.

Suppose that all user traffic requests are completed after total θ time segments. The s-th time segment has a duration of $t_{s+1} - t_s$. The traffic requests, scheduled links, and link bandwidths have the following relation:

$$\sum_{i=1}^{\rho} z_i \cdot h^i = \sum_{s=1}^{T} \sum_{i=1}^{\rho} \sum_{h=1}^{h^i} l^{i,h} \cdot \varpi_{i_h} \cdot (t_{s+1} - t_s), \tag{1}$$

where $l^{i,h} \cdot \varpi_{i_h}$ denotes the link bandwidth of the h-th hop of path $p_{(s_i,d_i)}$. Note that link $l^{i,h}$ has two statuses: one is active or scheduled, where $l^{i,h} = 1$; and the other is idle or unscheduled, where $l^{i,h} = 0$. An active link contains several critical parameters: 1) a pair of sender and receiver, 2) the channel it operates over, and 3) the interference-free relation.

If $l^{i,h}$ is assigned with a channel c, we use operator $c()$ to denote the CA as $c(l^{i,h})$. If $l^{i,h}$ gets channel c, we denote it as $l_c^{i,h}$. If $l_c^{i,h}$ is scheduled at time t, we set $l_{c,t}^{i,h} = 1$; otherwise, $l_{c,t}^{i,h} = 0$. For simplicity, we use $l_{c,t}^{i,h}$ to represent the assigned channel information c, scheduling time slot t, and the hop h of p_i.

For convenience, we tabulate the notations in Table 2.

Table 2. List of symbols and notations

ρ	The number of pairs
(s_i, d_i)	The i^{th} source-destination pair
Ω	The set of available orthogonal channels
c_i	The i^{th} channel in Ω
ϖ_i	The bandwidth of channel c_i
q	$q = \|\Omega\|$
ξ	The number of node interfaces
$p_{(s_i,d_i)}$	The selected path for (s_i, d_i)
θ	The total time segments needed
α	Parameter to set the number of interfaces
β	Parameter to set the number of channels
$l^{i,h}$	The h-th hop or link of $p_{(s_i,d_i)}$
$l_c^{i,h}$	The link $l^{i,h}$ using channel c
V	The set of nodes in the WMN topology
d_v	The number of interfaces equipped on node $v \in V$
N_v	The set of node v's neighbors
E	A set of neighbor pairs among V
$G = (V, E)$	A connected network graph for a WMN
$\widetilde{RSC}$	A joint scheme for routing, scheduling and CA
λ_i	The traffic request of (s_i, d_i): $\lambda_i = \{(s_i, d_i), z_i\}$
Λ	The set of traffic requests of multiple pairs $\{\lambda_i \| i = 1, \cdots, \rho\}$
f_{h,c_j}^i	The flow rate of the h-th hop of $p_{(s_i,d_i)}$ over channel c_j
T	The time period for updating the WMN structure Γ

3.2. WMPR: Multi-Pair Routing in WMNs

Given an MRMC WMN structure Γ, our goal is to maximize the number of communication paths that can be activated simultaneously to minimize the turnaround time for a given user request set Λ.

An Overview Multiple pair paths (MPP) in WMNs may have node intersections as in real-time video communications [13]. For example, in Fig. 1, node B is the intersection of two paths for two pairs C to F and A to D.

Definition 1 *Γ:*
$\Gamma = \{G, I, \Omega, \xi, \{(s_i, d_i)\}\}$, where:

- *$G = (V, E)$ is a mesh graph, while $|V|$ denotes the number of vertices and $|E|$ denotes the number of edges. Both $|V|$ and $|E|$ are constant integers.*
- *I is a relation for wireless interference awareness between vertices of G.*
- *$\Omega = \{c_1, c_2, ..., c_q\}$ is a set of available orthogonal channels for G. c_i has bandwidth ϖ_i.*
- *ξ is a list of integers standing for the numbers of node interfaces. By default, ξ is a constant.*
- *$\{(s_i, d_i)\}, i = 1$ to ρ is a set of multiple source-destination pairs.*
- *$\Lambda = \{\lambda_i\} = \{(s_i, d_i), z_i\}$ is the list of traffic queues corresponding to $\{(s_i, d_i)\}$ list, where $z_i > 0$.*

The substructure $\{G, I, \Omega, \xi\}$ includes the WMN infrastructure, i.e. the resources of the WMN. Meanwhile, $\{(s_i, d_i)\}$ represents multiple pairs with traffic requests $z_i > 0$, and $\rho = |\{(s_i, d_i)\}|$.

Suppose that in a given Γ, there are $|\Omega|$ channels, each channel c_i has bandwidth ϖ_{c_i}, and each router node $v \in V$ is equipped with d interfaces. We consider a set of requests $\Lambda = \{\lambda_i | i = 1, \cdots, \rho\}$, where $\lambda_i = \{(s_i, d_i), z_i\}$ and $z_i > 0$. For each request $\{\lambda_i\}$, to transmit data, we need to consider the following: find a fixed route/path $p_{(s_i,d_i)}$ for each pair, find a cooperative schedule for the links of multi-pair paths without interference, and assign channels to the scheduled links. We denote a joint scheme of these three operations as $\widetilde{RSC}$.

WMPR aims to achieve optimal joint $\widetilde{RSC}$ on routing, scheduling, and CA in a given mesh network Γ. The objectives are to minimize the turnaround time of user requests Λ, and maximize the utilization of wireless resources in serving the data transfers between multiple source-destination pairs $\{(s_i, d_i)\}$ at time t.

Our discussion is facilitated by Cartesian product of graphs (CPG), in which, each orthogonal channel is an independent virtual layer and an MRMC node is a collection of multiple fully connected identity nodes [11], as illustrated in Fig. 3.

Problem Formulation Nodes in broadband WMNs are generally equipped with multiple interfaces. The number of links that each node can use is limited by the number of interfaces and the number of available channels. We use c_i to denote any available channel, and N_v to denote the set of all neighbors of node v. Again, $l_{c_i}^{(v_1,v)}$ denotes a link from v_1 to v over channel c_i. We have the following constraint on the number of links involving node v (including both incoming and outgoing links of v):

$$\sum_{i \neq j}^{|\Omega|} \sum_{v_1, v_2 \in N_v} l_{c_i}^{(v_1,v)} + l_{c_j}^{(v,v_2)} \leq d_v, \forall v \in V, \tag{2}$$

where Ω denotes the set of channels, and d_v is the number of interfaces equipped on node v. We consider several aspects collectively: network topology, node interface, wireless interference, path selection, and CA. At time t, the channels used by node v form a true subset of Ω. $\forall v_1, v_2 \in N_v$, for links $l_{c_{i_1}}^{(v_1,v)}$ and $l_{c_{i_2}}^{(v,v_2)}$, $c_{i_1} \neq c_{i_2}$. Meanwhile, for links $l_{c_{i_1}}^{(v,v_1)}$ and $l_{c_{i_2}}^{(v,v_2)}$, if $v_1 \neq v_2$, then $c_{i_1} \neq c_{i_2}$. Similarly, for links $l_{c_{i_1}}^{(v_1,v)}$ and $l_{c_{i_2}}^{(v_2,v)}$, if $v_1 \neq v_2$, then $c_{i_1} \neq c_{i_2}$. Intuitively, any two links that share a common node v can be active simultaneously only over different channels. Also, the links of neighbor nodes on the same channel must satisfy the interference-free conditions.

The global maximum utilization of network resources requires scheduling as many links as possible in Γ. It is reasonable to maximize the number of links on multi-pair paths. If a schedule Π that achieves a maximum number of links is not a schedule for maximum utilization of networks (MUN), then there must exist another schedule that achieves MUN. Suppose that all links are of the same capacity. There must be another schedule Π' with more links. The problem should satisfy various constraints, such as multiple traffic requests, node interfaces, and free channels. Let ϖ_j denote the bandwidth of channel c_j. The waiting time for service on path $p_{(s_i,d_i)}$ is recorded as t_i. From a global perspective, to minimize the total waiting time, we should minimize the turnaround time.

The scheduled path of a pair (s_i, d_i) is denoted as $p_{(s_i,d_i)}^{sc}$. The turnaround time of $p_{(s_i,d_i)}$ is denoted as T^i. Let $\psi = |\Omega|$. The optimal WMPR problem is defined as follows:

$$
\begin{aligned}
&\min \max\{T^i | i = 1, \cdots, \rho\} \\
&s.t. \\
&\quad l_c^{i,h} = \begin{cases} 1, \text{ active} \\ 0, \text{ inactive} \end{cases} \\
&\quad \sum_{c_i \neq c_j}^{\Omega} \sum_{v_1, v_2 \in N_v} l_{c_i}^{(v_1,v)} + l_{c_j}^{(v,v_2)} \leq \xi, \forall v \in V; \\
&\quad f_{c_j}^{i,h} \leq l_{c_j}^{i,h} \times \varpi_j, \text{for } j = 1 \; to \; \psi; \\
&\quad \forall i, \forall h, \text{at } t, \; \{l_{c_j,t}^{i,h}\} \propto I; \\
&\quad \xi \neq 0; \\
&\quad \Omega \neq \emptyset.
\end{aligned}
\tag{3}
$$

WMPR does not have a polynomial-time exact optimal solution, even in a simple case with one single objective such as routing, scheduling, or CA. For scheduling, it has been proved to be NP-hard to determine an optimal link schedule in multi-hop radio networks [23], even if CA is not considered. For optimal scheduling, link scheduling can be converted into the edge coloring problem, which has been shown to be NP-complete [26]. For CA, which is a extensively studied problem, has been proved to be NP-hard [20]. Even a constrained version, which is a coloring problem, has been proved to be NP-complete [22]. As mentioned above, a simplified subproblem of WMPR is NP-complete. WMPR for MUN is NP-hard as proved by Schumacher *et al.*, when wireless interference is not taken into account. Without interference, it becomes a subproblem space of Γ, where $I = \emptyset$. Actually, by combining all major aspects of interference, CA, routing and scheduling over MRMC, the general problem is far more complex than the above cases that consider only one aspect.

The network topology puts a limit on the number of path options for a pair, while path selection chooses one that is interference-free with other existing paths. For example,

in the channel layered virtual model in Fig. 3, the two paths share node B on different orthogonal channels. In a channel related planar mesh, the two paths successfully transmit through B's diversified identities via channels.

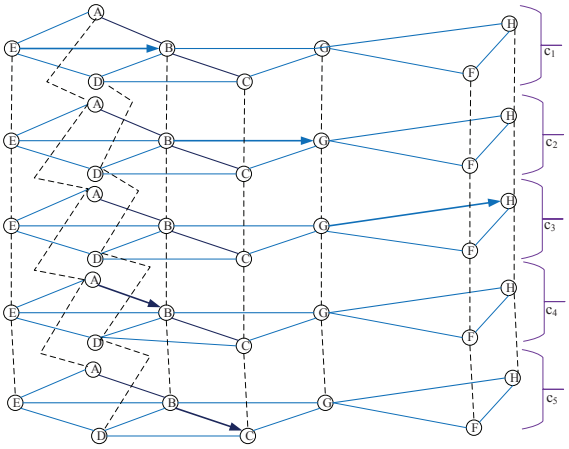


Fig. 3. Two compatible paths in CPG model.

Meanwhile, multiple pairs in a WMN fall in a situation of communication requests in an MRMC WMN defined virtual structure CPG. Note that all links of those active paths, which are simultaneously transmitting data at time t, form a substructure in CPG. We obtain a structure Γ by combining CPG and multiple pairs.

3.3. Compatible Paths

A larger number of concurrently scheduled paths for multiple pairs are able to transmit more data. Hence, it is reasonable to find more links in each channel to combine together for more paths. This idea was proposed for maximum utility of network resources [12]. However, our earlier experiments show that it is prohibitively expensive to find all interference-free link patterns in an arbitrary network topology of size beyond $|V| = 64$.

In Cartesian product of graphs (CPG), a directed path $p_{(s_i,d_i)}$ is called an *active path* at time t, if every component link is active on a distinct interference-free channel [9]. Since the CPG model maps orthogonal channels to corresponding planar meshes, a practical way is to decompose each path into links over different channel layers and choose interference-free links with maximum match to multiple pair paths.

With a minimum number of channels for each path, a good schedule can certainly lead to more active paths. To avoid the situation where several arbitrarily activated paths hog up all resources, the paths should be arranged to match the link patterns in a given mesh. Since link patterns are derived from interference-free links, we may combine links and compatible paths together for optimal performance, by searching for a link pattern from

a certain link of a $p_{(s_i,d_i)}$ over interference-free channels. The link pattern search collects as many links as possible from the selected paths.

At time t, given multiple pairs $(s_i, d_i), i = 1, 2, \cdots, \rho$, and $p_{(s_i,d_i)}$, there must be some paths that can be scheduled concurrently with proper CA, i.e. without interferences and conflicts on the node interface with other existing links. Such paths are called *compatible paths*.

A major characteristic of compatible paths is the concurrent coexistence without conflicts. Each of the compatible paths is an independent packet transmission pipe for (s_i, d_i), $i = 1, 2, \cdots, \rho$. A node equipped with multiple transceivers receives data over one channel, while sending data through another interface over a different channel to guarantee co-existence.

The problem of finding the largest number of compatible paths is at least as hard as the joint path minimum edge congestion MPP, which is NP-hard [4]. With all mixed factors of WMN, such as multiple source-destination pairs, interference along a path, interference between paths, interface limits of those shared nodes, CA, node free interfaces and free channels, as well as the optimal model, we need to consider all properties in a combinatorial way. Path realization is no longer just a path selection problem in an undirected graph, because the interference and CA must be accounted explicitly, compared to the joint path minimum edge congestion MPP.

Free-channel is a true subset of the available channels and it is node related. To a mesh, the available channels of Ω refer to a list of channels that mesh nodes may choose to operate over. In a later stage, to a specific node, the free-channel denotes those channels without a conflict with the already existing assignments so far. For each node, we need to distinguish the currently feasible channels for each node from those initial available channels. The current state includes neighbor node channel settings, channels of the links surrounding the node, impacts from the channels in related paths. If two neighbor nodes have some common free channels, then they can set up a link on one of the free channels at time slot t.

The compatible paths of different pairs can be roughly classified into three classes according to the number of common nodes: i) parallel, if two paths do not share any node; ii) crossing, if two paths share one node; iii) edge-sharing, if two paths share at least one common edge (two or more common nodes). The third case may consume edge resources rapidly and generate a hole of the mesh.

To find compatible paths for multiple source-destination pairs $(s_i, d_i), i = 1, 2, \cdots, \rho$, one needs to consider interference-free constraint, number of node interfaces, as well as orthogonal channels. The link interference-free condition is considered in the sender-receiver distance relation with the node identities or coordinates [9]. Meanwhile, since parallel paths are rare, we pay more attention to the other two cases.

Generally, a common node shared by $\tilde{a}$ paths at time slot t must have more than $2\tilde{a}$ interfaces for operating on $2\tilde{a}$ orthogonal channels. For example, the path $\{A, B, D\}$ for (A, D) shares B with path $\{E, B, G\}$ for (E, G) in Fig. 1. The two paths can not be simultaneously active if they can not satisfy both of these requirements: B has at least 4 interfaces and B has at least 4 free orthogonal channels.

Two Paths with One Overlapped Node In Fig. 1, the shared node B should be equipped with more interfaces than other nodes on the two paths.

The distance matrix M records the shortest path distance of a MRMC mesh. M is essential for further routing scheme, such as the shortest path diversity routing. As an example, the distance matrix M of Fig. 1 is provided in Table 3.

Table 3. The shortest distance matrix of the pairs in Fig. 1.

From To	A	C	D	F
A	0	2	2	2
C	2	0	1	3
D	2	1	0	2
F	2	3	2	0

Suppose that in Fig. 1, a link that makes up one hop of a path is assigned weight 1 (which stands for one-hop transmission). A shortest path of a certain pair can be expressed as a finite ordered node sequence. Paths selected to realize multiple pairs can be verified with the shortest distance matrix. Note that the shortest path is not unique in some cases. For example, in Fig. 1, path C to F can be $p_{(C,F)} = \{C, B, E, F\}$ or $p_{(C,F)} = \{C, D, Q, F\}$.

The overlapped nodes of multi-pair paths may change due to path diversities. In Fig. 1, according to the two paths selected for C to F, the specific intersection node varies. Table 4 provides an example for the cases in Fig. 1. Note that $\{A, E, Q, D\}$ is not a shortest path from A to D.

Table 4. Joint nodes vary with path variations in Fig. 1.

Shared node	Two joint paths involved
B	$\{A, B, D\} \wedge \{C, B, E, F\}$
D	$\{A, B, D\} \wedge \{C, D, Q, F\}$
B	$\{A, B, D\} \wedge \{C, B, Q, F\}$
E	$\{A, E, Q, D\} \wedge \{C, B, E, F\}$

Conflicts Between Paths The number of paths is largely limited by wireless interferences and node interfaces. The paths intersecting with each other compete for the resources of overlapped nodes. If some links of a path cannot be activated because of channel and interface restriction, they may be set to time $t + 1$, and so forth, which is similar to the idea of TDMA.

As it is impossible to compute an optimal WMPP in an exact way, an alternative way is needed to schedule as many links as possible for Λ at slot t. This is reasonable because one objective in Eq. (3) aims to concurrently schedule most links in each channel layer and minimize the maximum turnaround time. We need to know how to combine maximum link patterns and multiple paths. Those nodes shared by several paths are more likely to

run out of free interfaces or channels. For example, no matter what paths are selected for the two pairs of (H, I) and (G, E) in Fig. 1, they always intersect with the path of pair (A, D). Hence, some nodes on $p_{(A,D)}$ are very likely to be overloaded.

Link Patterns At time slot t, a node can be involved in a link if it has common free channels with its neighbors and free transceivers. If a node does not have other idle transceivers, it cannot establish a link to another neighbor, even if it has free channels. Likewise, if a node does not have any interference-free channel because of the current active links and neighbor links, it can not establish another link, even if it has idle transceivers.

WMPP is critical to satisfy specific traffic requests while promising MUN or achieving minimum time. It is essential to explore efficient algorithms for routing and scheduling multi-pair traffic requests.

The total number of compatible paths may be affected by mesh parameters: the topology, available orthogonal channels, mesh size, router type, node interfaces, and interference model. Additionally, it is also limited by the node properties such as power strength and antenna type. For simplicity, we focus our discussion on omnidirectional antennas in the mesh mode, i.e., a sender has only one desired receiver.

A link pattern collects all interference-free links over channel c_i in static wireless meshs [8]. First of all, it makes full use of channel resources. Specifically, given a finite number of orthogonal channels, paths whose lengths are close to the mesh diameter are limited even if some local regions have free resources. To use remaining resources, some shorter paths should be scheduled simultaneously. This necessitates detecting maximal link pattern for channel c_i of planar mesh as in Fig. 1.

With globally pre-computed link patterns, scheduling specific link patterns should be synchronised. After the system collects the information of router nodes, including position, interface number, and radio power, a series of link patterns can be generated using Algorithm 1.

Algorithm 1 facilitates the most links over each channel for $\{(s_i, d_i)\}, i = 1, 2, \cdots, k$. We use $\otimes$ to denote an operation for assigning a specific feasible channel by keeping the least channels used for a path. This operation collects the link patterns together, which are distinguished to each other according to time t and channel c. Obviously, the output of link patterns are dominated by path. The output saves the compatible link patterns for each channel layer. Even the traffic request list can be used to evaluate the heuristic start point for selecting interference links, the fairness should also be considered with maximal link patterns for each channel-layered mesh.

Algorithm 1 collects the interference-free links for desired paths over each channel with a heuristic start link. Different initial links result in different link patterns. Other factors that affect the size of a link pattern include topology and interference model. Furthermore, this algorithm can be modified to compute all possible link patterns without duplication for every heuristic start in a given mesh.

Example 1 *Link $E\overrightarrow{c_1}B$ can coexist with $H\overrightarrow{c_1}F$ or $F\overrightarrow{c_1}H$ in Fig. 3. However, as link $B\overrightarrow{c_2}G$ is located at the center of the pruned mesh, any other links over this channel will be in conflict with it. Hence, link $B\overrightarrow{c_2}G$ exclusively uses channel c_2. Similarly, link $G\overrightarrow{c_3}H$ can coexist with $E\overrightarrow{c_3}D$ or $E\overrightarrow{c_3}A$ or $D\overrightarrow{c_3}E$ or $A\overrightarrow{c_3}E$, at most two links in any*

Algorithm 1 Paths by Link-Patterns
Input: Γ
Output: $\mathcal{P}$

Require: $\{(s_i, d_i)\} \neq \emptyset$
Ensure: $\xi \neq 0$
$\mathcal{P} \Leftarrow \emptyset$;
$Sort(\Lambda)$;
while $i < \rho$ and $t < \mathcal{T}$ **do**
  if $\exists i = i_0$, such that $p_{(s_{i_0}, d_{i_0})}$ satisfies each node on the path has free radios and interference free channels at time slot t **then**
    for $h = 1$ to h_{i_0} **do**
      $L^{i_0,h}$ collects all interference-free links with $l^{i_0,h}$ of $p_{(s_{i_0}, d_{i_0})}$;
      $P_c^{i_0,h} \Leftarrow (\{l^{i_0,h}\} \cup L^{i_0,h}) \otimes c$;
      $P^i \Leftarrow \oplus P_c^{i_0,h}$;
    $i \Leftarrow i + 1$;
  else
    $t \Leftarrow t + 1$;
    Select $(s_{i_1}, d_{i_1}) \in \{(s_i, d_i)\} - \{(s_{i_0}, d_{i_0})\}$;
return $\mathcal{P} = \cup_i P^i$;

combination. Additionally, link $A\overrightarrow{c_4}B$, which shares B with the former decomposed path $p < E, H >$, can coexist with $H\overrightarrow{c_4}G$, or $F\overrightarrow{c_4}G$, or $H\overrightarrow{c_4}F$, or $F\overrightarrow{c_4}H$. Finally, link $B\overrightarrow{c_5}C$ can coexist with $H\overrightarrow{c_5}F$ or $F\overrightarrow{c_5}H$.

The largest number of activated paths is upper bounded by the largest size of those link patterns over all channels. A path is established if every component hop is realized over a certain channel. Given a WMN with the parameters to form a CPG, the number of active paths is obviously no more than the maximum number of all link patterns over all channel layers. Considering the available channels, CA and link cooperation, the number of activated paths may be far less, because it is impossible to have the same maximum number of link patterns for every hop of $p_{(s_i, d_i)}$.

For example, suppose that there are available channels $\{c_1, c_2, c_3\}$ at time t and the maximum link pattern contains 15 links. The maximum number of activated paths cannot surplus 15 even if every link is a certain hop of some path, even assuming that each link of the maximum interference-free link patterns can successfully get its entire path activated with enough over $\{c_1, c_2, c_3\}$.

4. Algorithm for Compatible Paths

To design an efficient algorithm for computing multiple pair paths to simultaneously transmit data packets for realtime services, like video conferences, we need to define several terms clearly.

In MIMO WMNs, a *link* is a transmission connection between a pair of neighbor nodes (sender and receiver) with a traffic request. The sender candidates S_C and receiver candidates R_C are essential to examine interference. The sufficient and necessary conditions for interference-free is discussed in our earlier work [11]. The channels can be

viewed as a checking loop for picking up maximum links while preserving a continuous paths for a specific pair.

We consider triangular meshes and arbitrarily connected graphs. Let T be the time period to update parameters for repeatedly scheduling a certain sequential link pattern, containing slots t_i, $i = 1, 2, ..., T$. For node N_v, we use d_{N_v} to denote its free interface count, and c_{N_v} to denote its available channels. To assign a channel c to a set P of interference-free neighbor pairs, we define a function $c(P)$. If $c(P)$ assigns channel c_k to link group P, it is denoted as P_{c_k}. A path $p_{(s_i,d_i)}$ can transmit traffic λ_i if every one of its link can be active. To assign $l^{i,j}$ a channel, we use a function $c()$. $c(l^{i,j}) = l_c^{i,j}$ assigns a channel to the j-hop link of $p_{(s_i,d_i)}$. Link pattern is generated by recursively expanding partial solutions after interference screening. Algorithm 2 is a resource aware scheme to compute compatible paths based on the optimal model with CPG [10].

Algorithm 2 Compatible Paths
Input: Γ
Output: $\Re$

Ensure: Node clock synchronization in the WMN
1: **for** $v = 1$ to $|V|$ **do**
2: update d_v;
3: update c_v;
4: **for** $i = 1$ to ρ **do**
5: $sort(\Lambda)$ in a decreasing order on key z_i;
6: update $\{(s_i, d_i)\}$ sequence in the order of $sort(\Lambda)$;
7: $sort(\Omega)$ in a decreasing order of $\{\varpi_k\}, k = 1$ to ψ;
8: **for** $t = 0$ to T **do**
9: **while** $i < \rho$ **do**
10: $i = 0$;
11: **for** $j = 0$ to h^i **do**
12: **if** $\exists N_{v_1}, N_{v_2} \in p_{(s_i,d_i)}$ and
(N_{v_1}, N_{v_2}) is the j^{th} hop and
$(d_{N_{v_i}} > 0) \wedge (c_{N_{v_i}} > 0) \wedge (c_{N_{v_1}} \cap c_{N_{v_2}} \neq \emptyset), i \in \{1, 2\}$ **then**
13: $P^{i,j} = P^{i,j} \cup \{(N_{v_1}, N_{v_2})\}$;
14: Sort $\{(P^{i,j})\}$ in a non-decreasing order according to their sizes;
15: $c(P^{i,j}) = P_{c_k}^{i,j}$;
16: $k = k + 1$;
17: **else** {at least one condition is not satisfied}
18: $i = i + 1$;
19: $t = t + 1$;
20: **for all** $t < \mathcal{T}$ **do**
21: activate $c(P_j)$;
22: **return** $\Re = \bigsqcup_{k,t}^{i,j} \{P_{c_k}^{i,j}\}$ for Λ;

We use $c(p_{(s_i,d_i)})$ to denote a directed sequence of links that allow real-time streams. It can be expressed as node sequence combining channel information. For example, $p_{(A,C)}$ can be implemented as links $(A, B)_{c_1}, (B, C)_{c_3}$. Meanwhile, path $p_{(E,H)}$ can be implemented as $(E, B)_{c_5}, (B, G)_{c_7}, (G, H)_{c_2}$ in Fig. 1.

This algorithm attempts to find more active multiple pair paths with more interference-free links over independent orthogonal channels, which naturally leads to a higher level of network resource utilization. The observations of Couto *et al.* [14] provide another perspective to find more compatible links for those selected paths. In fact, they realized that "the shortest path is not sufficient" through two testbed-based experiments, as minimum-hop routing often chooses routes that have significantly less capacity than the best link quality paths.

Path selection does affect the WMN performance. In Fig. 1, suppose that the interface counts of router nodes $\{A, B, C, D, E, F, Q, X, Y, Z\}$ are $\{2, 4, 2, 2, 3, 2, 2, 2, 2, 2\}$, respectively. The multiple pairs are $(A, D), (C, F), (G, J), (G, H), (I, E)$, and the corresponding traffic request queues are $\{4, 3, 2, 2, 2\}$. If we simply select the shortest path, the three paths would be $A\overrightarrow{c_1}B\overrightarrow{c_2}D$, $C\overrightarrow{c_3}B\overrightarrow{c_4}J\overrightarrow{c_5}F$, and $G\overrightarrow{c_5}A\overrightarrow{c_6}J$. However, if we replace path $C\overrightarrow{c_3}B\overrightarrow{c_4}J\overrightarrow{c_5}F$ by $C\overrightarrow{c_3}D\overrightarrow{c_4}E\overrightarrow{c_5}F$, at least the following four paths can become active simultaneously: $A\overrightarrow{c_1}B\overrightarrow{c_2}D$, $C\overrightarrow{c_3}D\overrightarrow{c_4}E\overrightarrow{c_5}F$, $G\overrightarrow{c_5}B\overrightarrow{c_6}J$, and $X\overrightarrow{c_5}A\overrightarrow{c_6}Y$. Note that the number of compatible paths in the second one increases by one, and its link count increases as well, from step 7 to step 9. Therefore, the second scheme is a better choice.

We use V^c to denote the set of nodes that still have free channels in V, and V^r to denote the set of nodes that still have free radio interfaces in V. $d(s_i, r_i) = 1$ means that (s_i, r_i) is a neighbor pair. The remaining free channel set of s_i is denoted as s_i^c.

To find other potential links, the procedure $Game\_Supplement$ is used to exploit chances to maximize the utility of mesh radio and channel resources.

Algorithm 3 $Game\_Supplement$
Input: The updated Γ after Algorithm 2
Output: $\Re'$

Require: $V^c \neq \emptyset$
Ensure: $V^r \neq \emptyset$
1: $\Re' = \emptyset$;
2: **while** $\exists(s_i, r_i) \in V^c \cap V^r \wedge d(s_i, r_i) = 1$ **do**
3: **if** $\lambda_i > 0$ **then**
4: $l_{c_{i_0}}^{(s_i,r_i)}$, where $c_{i_0} \in s_i^c \cap r_i^c$;
5: **if** $l_{c_{i_0}}^{(s_i,r_i)}$ is interference-free to $\Re'$ **then**
6: $\Re' = \Re' \cup \{l_{c_{i_0}}^{(s_i,r_i)}\}$;
7: Exit While;
8: **else** {confliction with other simultaneous links}
9: **while** $s_i^c \cap r_i^c \neq \emptyset$ **do**
10: $c_{i_0} = c_{i_0} + 1$;
11: $(s_i^c \cap r_i^c) = (s_i^c \cap \lambda_i^c - \{c_{i_0}\})$;
12: update V^c;
13: update V^r;
14: **return** $\Re'$;

In fact, the procedure $Game\_Supplement$ is to enlarge the scale of the scheduled links as much as possible. However, $Game\_Supplement$ aims to use those free resources by picking more links. There are chances to get a whole path by taking more hops than the

shortest path of the pair. Suppose the nodes with idle transceiver form V^r after Algorithm 2, and the nodes keeping available channels form V^c.

Assigning a channel c_{i_0} to link λ_i in Algorithm 3 implies confliction avoidance to $\Re$. The paths of final solution, combining Algorithm 2 and 3 form scheduling models for MIMO WMNs. Of course, the Algorithm 3 does not always promise additional path contribution. It actually works when the channel and radio count increases.

Some other major factors on compatible paths are node interface count, available channel count, mesh topology (node relatively position), heuristic initial neighbor pair, and antenna type (omnidirectional antenna, directive antenna, smart antenna, etc.). In real applications, the environments also affect the actual paths [3, 21].

5. Performance

The performance of Algorithm 2 is evaluated both analytically and by simulation. After we estimate the time complexity, we make some simulations on throughput, delay time, as well as statistics of active pairs. While T^r is given, simulations help to understand the performance influenced by different topologies.

5.1. Time Complexity

As mentioned in section 2, MPP in WMN is too hard to solve in exact algorithms. Algorithm 2 attempts to collect as many links as possible for every channel in the mesh to combine wanted paths. Note that $|S_C|$ reduces quickly along with the process of picking out more links without conflicts over a channel.

Let the senders of selected interference free links form a node set $\mathcal{S}$, and receivers form a node set $\mathcal{R}$. All neighbors of $\mathcal{S}$ is denoted as $\mathcal{S}_N$, while All neighbors of $\mathcal{R}$ is denoted as $\mathcal{R}_N$. According to rules as Table 1, for next link to add into the current link pattern, the selection range is given by S_C and R_C.
The next link sender candidates are in (4):

$$S_C = S_C - S - R - \mathcal{S}_N - \mathcal{R}_N. \tag{4}$$

The corresponding receiver candidate set is in (5):

$$R_C = R_C - \mathcal{S} - \mathcal{R} - \mathcal{S}_N. \tag{5}$$

We denote the average node degree of a given mesh as D_{Ave}. For example, in a triangular mesh, $D_{Ave} = 6$, while in a grid, $D_{Ave} = 4$.

An approximate estimate to the size of $|S_C| = |S_C| - 2D_{Ave}$. At the initial step,

$$|S_C| \approx |V| - 2D_{Ave}. \tag{6}$$

Generally, the recursion equation for the size of S_C is as following:

$$|S_C| = |S_C - \mathcal{S} - \mathcal{R} - \mathcal{S}_N - \mathcal{R}_N| \approx |S_C| - 2D_{Ave}. \tag{7}$$

The recursion equation for size of R_C is as follows:

$$|R_C| = |R_C - \mathcal{S} \cup \mathcal{R} - \mathcal{S}_N| \approx |R_C| - D_{Ave} \times |\mathcal{S}|. \tag{8}$$

At the initial step, the selection for both sender and receiver does not need interference screening. i.e, sender set and receiver set are empty. So, R_C in (8) changes to:

$$|R_C| \approx |V| - D_{Ave} \times |\mathcal{S}|. \tag{9}$$

Any link must have one node in S_C and one node in R_C. In other words, any link is an element of $S_C \times R_C$. Then, compatible links for a channel are a subset of relation $S_C \times R_C$.

In Algorithm 1, for one channel, selecting next compatible link goes on until $(S_C = \emptyset) \vee (R_C = \emptyset)$.

As the link patterns are computed via the heuristic start of traffic requests of multiple pairs, sorting the traffic requests in decreasing order costs at worst $\mathrm{O}(|V| \log |V|)$, even if the graph is complete graph $K_{|V|}$.

Another time consumption task is to screen the interference after updating S_C and R_C. Note that generating S_C and R_C only takes $\mathrm{O}(|V|)$. As for adding a link, two nodes from S_C and R_C must match as neighbors, i.e., the shortest distance of them is 1.

Meanwhile, to select as many links as possible, a preferred next sender is one of 2-hop away from the already selected senders $\mathcal{S}_s$. Those 2-hop away in S_C will be checked for next compatible link in priority. This is to avoid space and spectrum taken up by scattered links, because the algorithm aims to get more links for a link pattern in a path-dominated heuristic way.

The interference checking needs to find a neighbor pair of $(sender, receiver)$, where a new sender is 2-hop away to one of $\mathcal{S}_s$. A new receiver must satisfy the conditions in Table 1, while these two nodes are adjacent by checking the adjacent matrix of the mesh graph. The distance can also be verified by searching the distance matrix. This checking takes time at most of $\mathrm{O}(|V|^2)$.

Note that the size of S_C or R_C becomes smaller quickly according to (4) and (5). Let T'_i represent the i-th updated size of S_C. Let the compatible link computing process finish after k times calling of its procedure, where k is given by:

$$\begin{aligned} &T'_k = T'_{k-1} - 2D_{Ave}, \\ &T'_0 = |V|, \\ &T'_k = 0. \end{aligned} \tag{10}$$

After expanding (10), we obtain k by approximation:

$$k \simeq \left\lfloor \frac{|V|}{2D_{Ave}} \right\rfloor. \tag{11}$$

Now, we come to the conclusion for the time complexity T' of Algorithm 2. Because the next link is always determined after checking the interference, and along with the desired paths of some node pairs with traffic request priority, the link patterns from different channels work together to concatenate those paths.

$$T' < \mathrm{O}(|V|^2) + k \cdot |S_C| \cdot |R_C| < \mathrm{O}(|V|^3). \tag{12}$$

Therefore, the time complexity of Algorithm 2 is $\mathrm{O}(|V|^3)$. In broadband backbone WMNs, nodes are equipped with multiple interfaces, which can be viewed as the upper limit of a node on simultaneous links at a time slot. Hence, D_{Ave} is determined by the number of node interfaces, which is different from the node degree in the topology.

5.2. The simulations

We run several sets of simulations to evaluate the performances of Algorithm 2 over two topologies in Figure 4(b) and Figure 4(a): 61-node triangular mesh and 61-node arbitrary mesh. The arbitrary one is generated by selecting random positions for indexed nodes. In the triangular mesh, there are 4-hop circles surrounding the center node. We use the triangular one with additional aims to facilitate the discussion and illustrate the methods. Meanwhile, the randomly generated one is used to evaluate the robustness or verify the consistency for practical generality.

Furthermore, to evaluate the performance of Algorithm 2, we conduct simulations over these two topologies with variations in the number of channels, the number of nodes, and traffic queue sizes.

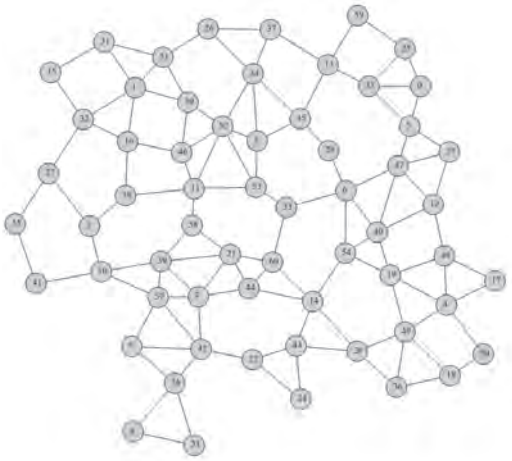
With different multiple pairs, where traffic queues are used to measure the sizes of traffic requests for those node pairs, we conduct simulations to estimate the maximum boundary for the combination cases of the numbers of interfaces and channels. To understand the throughput increase with variations in the available radios $\{4, 8, 12, 16, 20\}$ and the channels $\{8, 16, 32, 64, 128\}$, we run simulations in the situations of MPP for all pairs, MPP for some pairs with path crossing each other, and MPP with less resource competition.

These two network topologies are both virtually deployed in a $100 \times 100\ km^2$ area with available radio number cases $\{4, 8, 12, 16, 20\}$, and channel number cases $\{8, 16, 32, 64, 128\}$. We use T_d to denote the effective transmission distance of certain power strength and I_d to denote the interference distance. We have $T_d < I_d$ and $I_d < 2T_d$. The interference scanning is under the conditions in Table 1. Time duration is set to be $5ms$, packet size is set to be $1MB$, and each link capacity is set to be $10Mb$. A period spans 0.5 second, equally, 100 time slots.

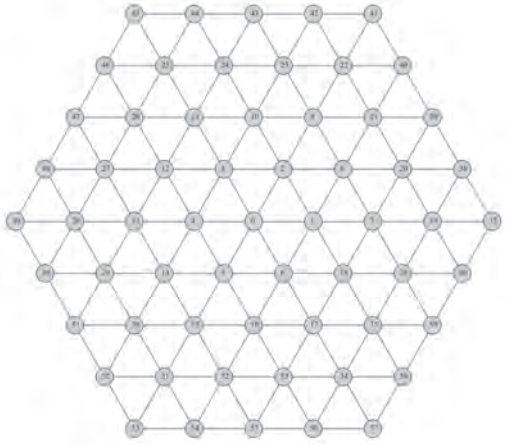
In a combinatorial way, the input instances vary with some parameters such as the number of router node interfaces, the number of available orthogonal channels, the specific multiple pairs and the corresponding traffic queue sizes. We design total 25 combinations of the channel number and the radio number over two topologies to evaluate the performance. The traffic matrix to all pairs is in form $(250)_{61\times61}$.

We first consider a traffic model to simulate the situation of a very busy backbone network, where each node has a traffic request to every others.The traffic matrix to all pairs is $(250)_{61\times61}$.

The throughput comparison over the two topologies is shown in Fig. 5. To be concise, in all figures, we use $(R = Radio_{num}) \wedge (C = Channel_{num})$ to represent a specific combination case, where nodes are equipped with R interfaces, and the mesh operates over C available orthogonal channels. We are able to draw conclusions on the proper relation between radio number and channels for both efficiency and economic purposes: a higher throughput improvement from case $(R = 12) \wedge (C = 128)$ to case $(R = 16) \wedge (C = 128)$ than that from case $(R = 16) \wedge (C = 128)$ to case $(R = 20) \wedge (C = 128)$. This observation is true for both topologies. Then, we conclude that the economically efficient case is $(R = 16) \wedge (C = 128)$. It is clear that the improvements are significant between channel variations for the case $R = 16$. Additionally, the throughput of triangular mesh outperforms that of the random one by $200MB/s$. Meanwhile, we observe that the performance is also stable in the arbitrary mesh, compared with that of the carefully planned triangular mesh.



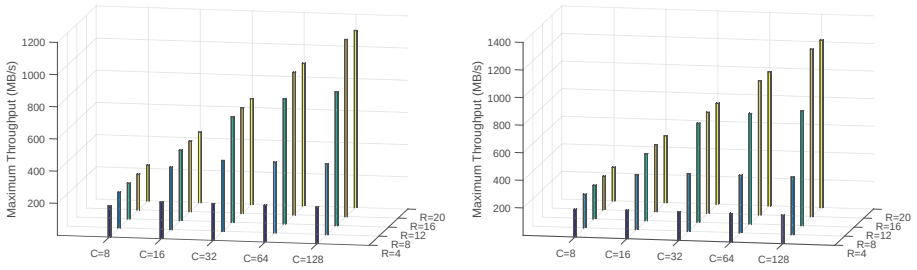
(a) A randomly generated mesh for indexed nodes



(b) A carefully planned triangular mesh

Fig. 4. The arbitrary mesh and the triangular mesh.

To evaluate the general efficiency of the algorithm, the average delays for 25 cases are shown over two topologies in Figure 6. Given a specific traffic request, the combined 25 cases are the elements of $R \times C$, i.e. $\{4, 8, 12, 16, 20\} \times \{8, 16, 32, 64, 128\}$. For example, $(4, 32)$ means that the number of interfaces is 4, and the number of available channels is 32. The average delays in these two topologies are simulated independently. The overall delays are smaller in triangular mesh than that in random mesh, which matches our theoretical expectation. If we only consider the best combined cases, the triangular mesh and the random mesh reveal the common fact: the cases of $R = 20 \wedge R = 16$ with $C = 64 \wedge C = 128$ are efficient, because $R = 16 \wedge C = 64$ or $R = 16 \wedge C = 128$ show less average delays than other cases.

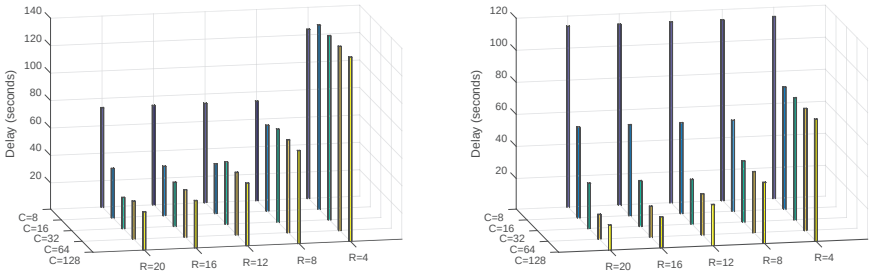


(a) Maximum throughput in the random mesh (b) Maximum throughput in the triangular mesh

Fig. 5. Maximum throughput of Algorithm 2 over two topologies.

The difference of the total used time between triangular mesh and random mesh is significant. As shown in Figure 6, triangular mesh has significant improvements in 25 cases. For example, in case $R = 4 \wedge C = 8$, triangular mesh uses almost only half time of random mesh for the same multiple pairs and traffic queues.

The simulation results also show that the proposed algorithm works efficiently in terms of delay, and it performs better in the triangular mesh than the arbitrary one, as shown in Fig. 6.



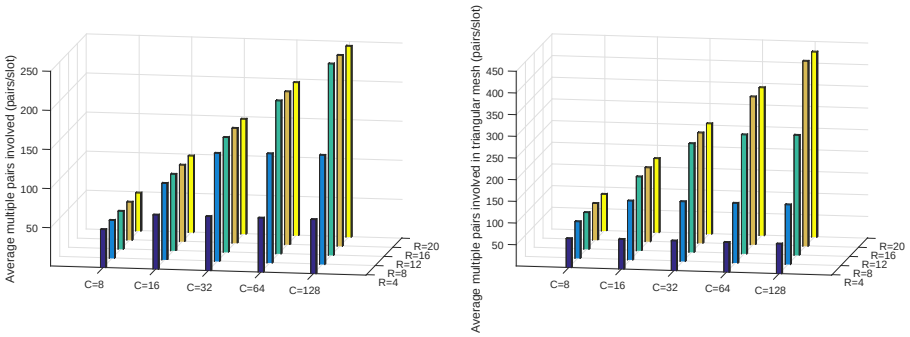
(a) The delay over the random mesh (b) The delay over the triangular mesh

Fig. 6. Average delays of Algorithm 2 over two topologies.

Algorithm 2 is evaluated by 22 random pairs, which are: $(N_{37}, N_0), (N_{41}, N_{43}), (N_{39}, N_{21}), (N_{38}, N_{33}), (N_{57}, N_6), (N_{18}, N_{52}), (N_{55}, N_{10}), (N_{29}, N_{20}), (N_3, N_{36}), (N_{22}, N_9), (N_{25}, N_{44}), (N_{43}, N_{42}), (N_{41}, N_{53}), (N_{39}, N_{38}), (N_{60}, N_{48}), (N_{34}, N_0), (N_{54}, N_0), (N_{45}, N_{57}), (N_{29}, N_0), (N_{16}, N_{20}), (N_{23}, N_{37}), (N_{11}, N_{29})$.

The specific traffic queue sizes of T^r are assigned as: $\alpha = \beta = (70, 60, 50, 40, 30, 20, 10, 70, 60, 50, 40, 30, 20, 10, 80, 50, 60, 70, 40, 30, 20, 30)$. The number of orthogonal channels $|\Omega|$ can be $\{4, 8, 12\}$, and the number d of node interfaces can be one of $\{3, 4, 6\}$. $d = 3$ represents a hexagonal mesh, $d = 4$ represents a grid mesh, and $d = 6$ represents a triangular mesh. Note that in an arbitrary mesh, a node degree is determined by randomly distributed node positions.

The average number of pairs involved We calculate the statistics on the number of pairs involved in the scheduling of each time slot to show the maximum, average and minimum pairs involved. The number of pairs is partially related to the network utility rate, and can be used to evaluate the topology efficiency at the topology planning stage as well. The average number of pairs per time slot is illustrated in Fig. 7, corresponding to the two 61-node topologies, the random and triangular mesh, respectively.



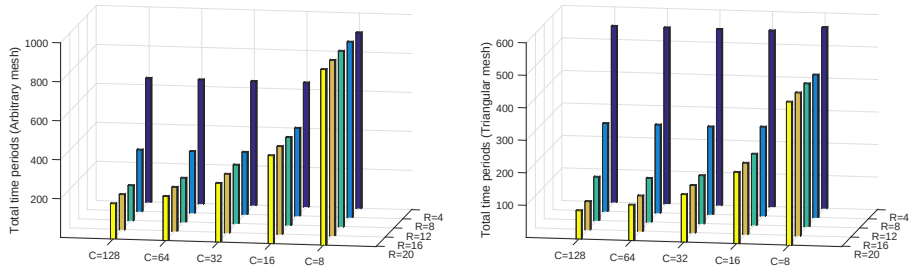
(a) The average number of pairs scheduled over the random mesh

(b) The average number of pairs scheduled over the triangular mesh

Fig. 7. The average number of pairs scheduled by Algorithm 2.

The time used to finish traffic queues The algorithm works on both topologies of triangular mesh and randomly generated mesh. Specifically, Fig. 8 illustrates the total time for traffic $(250)_{61\times61}$ over two topologies, where each node has a request of 250-packet traffic to every other one. We observe that the total time costs are consistent, for both of the triangular mesh and random mesh topologies. Here, one period is set to be 0.5 second, i.e., 100 time slots. Fig. 8(a) and 8(b) plot the total time used for all combined cases in the randomly generated mesh and in the triangular mesh, respectively. Fig. 8 also shows the effects of the number of interfaces and the number of channels. More node interfaces result in less time cost for the specific traffic size. More available channels also result in less time cost. The most efficient case is among $R = 16 \wedge C = 128$, $R = 20 \wedge C = 128$, $R = 16 \wedge C = 64$ and $R = 20 \wedge C = 64$.

We further make a comparison with AODV over the triangular and arbitrary meshes. AODV achieves almost less than half of the throughput achieved by Algorithm 2. A care-



(a) Time used over arbitrary mesh

(b) Time used over triangular mesh

Fig. 8. Time used to transmit the traffic $(250)_{61\times 61}$.

ful investigation into AODV execution processes reveals that AODV does not take into account the remaining resources and the topology updates. For example, even if there is a resource-free shortest path for a given pair, AODV may select another longer path with less resources to forward packets. This random choice certainly degrades its performance in MRMC WMNs. Also, AODV leads to a lower performance in our combinatorial cases as it does not fully consider MRMC situations.

6. Conclusion

WMPP is raised from real applications for data, voice and video transmission in WMNs. With the CPG model and an in-depth analysis, we develop a joint routing and scheduling scheme through channel layered interference-free links, aiming to maximize compatible paths to provide the highest Quality of Service over limited resources. We proposed to decompose multiple paths into channel layered interference-free link patterns to maximize the resource use in MIMO WMNs. Since link patterns mainly contain links of the paths for multiple pairs, maximum compatible paths naturally result in the maximum utilization of network resources for a given problem instance. Extensive simulations over triangular and arbitrary topologies show that the proposed optimization scheme computes maximum link patterns efficiently and exhibits a stable performance, which meets our theoretical expectation. It is our further interest to conduct more extensive simulations for the deployment of BS nodes in triangular and arbitrary meshes.

Acknowledgment. The research is partially funded by China Scholarship Council (CSC) No. 2013-06755013, and is also partially sponsored by U.S. National Science Foundation under Grant No. CNS-1560698 with New Jersey Institute of Technology.

References

1. Akyildiz, I.F., Wang, X., Wang, W.: Wireless mesh networks: a survey. Computer networks 47(4), 445–487 (2005)

2. Alicherry, M., Bhatia, R., Li, L.E.: Joint channel assignment and routing for throughput optimization in multi-radio wireless mesh networks. In: Proceedings of the 11th MobiCom. pp. 58–72. ACM (2005)
3. Andersen, D.G., Snoeren, A.C., Balakrishnan, H.: Best-path vs. multi-path overlay routing. In: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement. pp. 91–100 (2003)
4. Andrews, M., Zhang, L.: Hardness of the undirected congestion minimization problem. SIAM Journal on Computing 37(1), 112–131 (2007)
5. Andrews, M., Zhang, L.: Almost-tight hardness of directed congestion minimization. Journal of the ACM (JACM) 55(6) (2008), article ID 27
6. ASSOCIATION, I.S.: Ieee std. 802.16-2004 for local and metropolitan area networks, part 16: Air interface for fixed and mobile wireless access systems. http://standards.ieee.org/getieee802/download/802.16-2004.pdf (October 2004)
7. Cao, L., Zheng, H.: On the efficiency and complexity of distributed spectrum allocation. In: Cognitive Radio Oriented Wireless Networks and Communications, 2007. CrownCom 2007. 2nd International Conference on. pp. 357–366 (2007)
8. Cao, Z., Peng, L.: Destination-oriented routing and maximum capacity scheduling algorithms in cayley graph model for wireless mesh network. Journal of Convergence Information Technology 5(10), 82–91 (2010)
9. Cao, Z., Tang, J.: Routing methods and scheduling patterns in mimo wmn virtual model. Applied Mechanics and Materials 519, 216–221 (2014)
10. Cao, Z., Wu, C.Q., Berry, M.L.: On routing of multiple concurrent user requests in multi-radio multi-channel wireless mesh networks. In: 17th International Conference on Parallel and Distributed Computing, Applications and Technologies, PDCAT 2016, Guangzhou, China, December 16-18, 2016. pp. 24–29 (2016)
11. Cao, Z., Wu, Q., Zhang, Y., Shiva, S.G., Gu, Y.: On modeling and analysis of mimo wireless mesh networks with triangular overlay topology. Mathematical Problems in Engineering (2015), article ID 185262
12. Cao, Z., Xiao, W., Peng, L.: A mesh$\times$ chain graph model for mimo scheduling in ieee802. 16 wmn. In: Proc. of the 2nd IEEE International Conference on Computer Modeling and Simulation. vol. 2, pp. 547–551 (2010)
13. Cappanera, P., Lenzini, L., Lori, A., Steay, G., Vaglini, G.: Optimal link scheduling for real-time traffic in wireless mesh networks in both per-flow and per-path frameworks. In: World of Wireless Mobile and Multimedia Networks (WoWMoM), 2010 IEEE International Symposium on a. pp. 1–9. IEEE (2010)
14. De Couto, D.S., Aguayo, D., Chambers, B.A., Morris, R.: Performance of multihop wireless networks: Shortest path is not enough. ACM SIGCOMM Computer Communication Review 33(1), 83–88 (2003)
15. Giannoulis, A., Salonidis, T., Knightly, E.: Congestion control and channel assignment in multi-radio wireless mesh networks. In: Sensor, Mesh and Ad Hoc Communications and Networks, 2008. SECON'08. 5th Annual IEEE Communications Society Conference on. pp. 350–358. IEEE (2008)
16. Godsil, C., Royle, G.: Algebraic graph theory, vol. 207. Springer (2001)
17. Karp, R.M.: Reducibility among combinatorial problems. Springer (1972)
18. Laneman, J.N., Tse, D.N., Wornell, G.W.: Cooperative diversity in wireless networks: Efficient protocols and outage behavior. IEEE Transactions on Information Theory 50(12), 3062–3080 (2004)
19. Larsson, E.G., Edfors, O., Tufvesson, F., Marzetta, T.L.: Massive mimo for next generation wireless systems. IEEE Communications Magazine 52(2), 186–195 (2014)
20. McDiarmid, C., Reed, B.: Channel assignment and weighted coloring. Networks 36(2), 114–117 (2000)

21. Nandiraju, N., Nandiraju, D., Agrawal, D.: Multipath routing in wireless mesh networks. In: IEEE international conference on Mobile adhoc and sensor systems (MASS). pp. 741–746. IEEE (2006)
22. Ramachandran, K.N., Belding-Royer, E.M., Almeroth, K.C., Buddhikot, M.M.: Interference-aware channel assignment in multi-radio wireless mesh networks. In: INFOCOM. vol. 6, pp. 1–12 (2006)
23. Ramanathan, S., Lloyd, E.L.: Scheduling algorithms for multihop radio networks. IEEE/ACM Transactions on Networking (TON) 1(2), 166–177 (1993)
24. Raniwala, A., Gopalan, K., Chiueh, T.c.: Centralized channel assignment and routing algorithms for multi-channel wireless mesh networks. ACM SIGMOBILE Mobile Computing and Communications Review 8(2), 50–65 (2004)
25. Schumacher, A., Haanpää, H.: Distributed network utility maximization in wireless networks with a bounded number of paths. In: Proceedings of the 3nd ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks. pp. 96–103. ACM (2008)
26. Sen, A., Huson, M.L.: A new model for scheduling packet radio networks. Wireless Networks 3(1), 71–81 (March 1997)
27. Suurballe, J.W., Tarjan, R.E.: A quick method for finding shortest pairs of disjoint paths. Networks 14(2), 325–336 (1984)
28. Wang, I.L., Johnson, E.L., Sokol, J.S.: A multiple pairs shortest path algorithm. Transportation science 39(4), 465–476 (2005)

Zhanmao Cao received his PhD from School of Computer Science and Technology, South China University of Technology, Guangzhou, China. He is Associate Professor, Dept of Computer Science, South China Normal University. He is interested in developing algorithms for application problems. He proposed algorithms BTA and DC-BTA in multiple sequence alignment of bioinformatics. He is now working on MIMO WMN model, routing and scheduling algorithms.

Chase Qishi Wu received his PhD in Computer Science, Louisiana State University (LSU), Baton Rouge, LA. He is tenured Associate Professor, Department of Computer Science, New Jersey Institute of Technology. He is very active in both research projects and published papers. His research covers a few areas: Big data, data-intensive computing, parallel and distributed computing, high-performance networking, large-scale scientific visualization, wireless sensor networks, cyber security.

Mark L. Berry is a PhD student in Wu's group, Department of Computer Science, New Jersey Institute of Technology.

Received: January 24, 2017; Accepted: July 29, 2017.

Computer Science and Information Systems 14(3):685–702 https://doi.org/10.2298/CSIS170110023C

Construction of Affective Education in Mobile Learning: The Study Based on Learner's Interest and Emotion Recognition

Haijian Chen[1], Yonghui Dai[2,*], Yanjie Feng[2], Bo Jiang[3], Jun Xiao[1], and Ben You[2]

[1] Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, Shanghai 200433, China
{xochj, xiaoj}@shtvu.edu.cn
[2] Management School, Shanghai University of International Business and Economics, Shanghai 201620, China
{daiyonghui, fengyanjie}@suibe.edu.cn, youben022@gmail.com
[3] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China
jiangbo@sui.edu.cn

Abstract. Affective education has been the new educational pattern under modern ubiquitous learning environment. Especially in mobile learning, how to effectively construct affective education to optimize and enhance the teaching effectiveness has attracted many scholars attention. This paper presents the framework of affective education based on learner's interest and emotion recognition. Learner's voice, text and behavior log data are firstly preprocessed, then association rules analysis, SO-PMI (Semantic Orientation-Pointwise Mutual Information) and ANN-DL (Artificial Neural Network with Deep Learning) methods are used to learner's interest mining and emotion recognition. The experimental results show that these methods can effectively recognize the emotion of learners in mobile learning and satisfy the requirements of affective education.

Keywords: Affective education, mobile learning, learner's interest, emotion recognition.

1. Introduction

In recent years, with the development of mobile communication technology and educational technology, profound changes have occurred in the way of learning. Especially, mobile learning model is quickly development. Generally speaking, mobile learning refers to learning facilitated by mobile devices such as mobile phones, tablet PCs or personal media players for distance learning [13][35]. Compared with traditional learning methods, there are two outstanding advantages of mobile learning, one is its learning flexibility, and the other is its abundant learning resources. As mobile learning is not limited to learning time and place, it can allow learner not only make full use of fragmentation time to learn but also easy to share learning content with others, which has brought great convenience to learners. Therefore, this way of learning is loved by more and more contemporary learners. However there are some problems in this way. Learners use mobile devices for learning and they are faced with a lack of emotional machine every day, which is easy to

make learners become indifferent, heartless and emotional imbalance. Therefore, how to improve the situation of the lack of emotion in mobile learning, and carry out the affective education to improve teaching effectiveness is an important research subject.

As we known, the key to the implementation of affective education is learners emotion recognition and interest acquisition. Taking into account the interaction of mobile learning is carried on by the learner's text and voice, we used text affective computing and speech emotion recognition in our study. Because the emotion recognition of mobile learners is closely related to the situational context, emotional state and so on, then it is always difficult to recognize the learners' emotion accurately. Especially, learners voice contains a large number of conversational continuous phrases, which makes the traditional speech emotion recognition method such as six discrete emotion categories cannot get good results. In order to solve the above problems, the three-dimensional PAD emotional model [1] and the neural network method [8] were used to calculate learner's emotion.

This paper is organized as follows. In section 2, related research of affective education, learner's interest and emotion recognition are introduced. In section 3, the framework of learner's emotion recognition and methodology are shown. In section 4, experiment is illustrated. Section 5 is the conclusion of this article.

2. Related works

As a hot research topic in the field of education, affective education has attracted the attention of many scholars. Overall, previous research related to construction of affective education in mobile learning can be summarized into three aspects, namely, affective education theory, interest mining and emotion recognition.

2.1. Affective education

Affective education is used as an educational concept and a part of the educational process, which is concerned with the findings, beliefs, attitudes, and emotions of students with their interpersonal relationships and social skills [29]. Since 1970s, the research of affective education has changed from the initial stage to the development stage, and related works research mainly focused on emotional education theory and affective education model, such as humanistic emotion theory [36], academic achievement emotion theory [41], scaffolding affective education theory [24] and affective education practice model [9] [33]. Among them, humanistic emotion theory emphasizes self-expression, emotion and subjectivity, and it not only pays attention to the development of cognition in teaching process, but also pays more attention to learners emotion, motivation and interest. For example, Connolly used this theory to study on the coaching process, and he thought that communication, self-concept, affect, personal values are the key emphases and strategies for humanistic coaching [4]. Academic achievement emotion theory mainly focuses on the students learning process and the emotion related to achievement [18], for example, learners anger to the homework, or learners disgust to the homework, and the results show that emotion has both positive and negative effects on learners academic achievement [22].

On the research of affective education practice model, Cheng put forward the implementation of affective education in a middle school in Chinas Guangzhou, and three

levels of affective education were described, namely, class-group level, manner-individual level and institutional-whole school level [3]. After that Ghasemaghaei et al introduced a framework for multimodal educational systems and human-computer interaction (HCI) emotion education [10].

2.2. Interest mining

Previous researches on interest mining have focused on log mining and interest modeling. For example, Stamou et al (2009) proposed to get users' preference by analyzing their clicked log (such as query word, browse pages and so on), as well as the semantic similarity from their query words and visited web pages [34]. Xu et al (2009) believed that the behavior of browsing the page contains the attention and interest of the content and it can be used to predict interest, and they pointed out that the user behavior of the relevant interests includes the residence time of browsing the learning page, web link of clicking, and click frequency of a page and so on [38]. Rao et al (2015) extracted user's interests from web log data, including the log of visit time and visit density, and they discussed the technological in data mining and its applications to personalization [28]. Maheswari et al (2015) studied on data preprocessing of web log files and how to predict user's interest, and their research is based on the mining of behavior logs [21].

On the research of learner's interest modeling, Nakatsuji et al (2012) proposed a collaborative filtering method based on time periods and classification for user's interest modeling, their used data were collected from the historical behaviors such as listen to music, users' tweets and visit restaurant, their study showed that their method can get good accuracy in the prediction of interest [25]. Sanchez et al (2013) constructed innovative consumption-modeling system to predict user's interest of TV contents, and their model was established based on Hidden Markov Model and Bayesian inference techniques, experimental results showed that their system was more reliability [30]. Li et al (2014) built users' interest model and offered personalized recommendation according to their reading preferences, their research suggested that the result are associated with long-term and short-term reading preferences [19].

2.3. Emotion recognition

The study of emotion recognition begins with the classification of emotion, and the psychological point of view thinks that the emotion is divided into basic emotion and dimensional space emotion [20]. Among, basic emotion refers to human emotions are divided into fixed categories, for example, the typical classification of six kinds of emotion, that is, anger, disgust, fear, joy, sadness and surprise[31]. Relatively speaking, dimensional space emotion theory holds that human emotion is different position in space. From their point of view, emotion can be divided from one dimension, two dimensions or three dimensions. For example, the widely used PAD three-dimensional emotional model (Pleasure-Displeasure, Arousal-Nonarousal, Dominance-Submissiveness) was divided from three dimensions [6], and it used dimension 'P' to represent someone's evaluation which is positive or negative, dimension 'A' represents the level of neural activation, and dimension 'D' represents the individual's ability to control situations and others people [11]. The three dimensions of continuous emotion classification are shown in Fig.1.

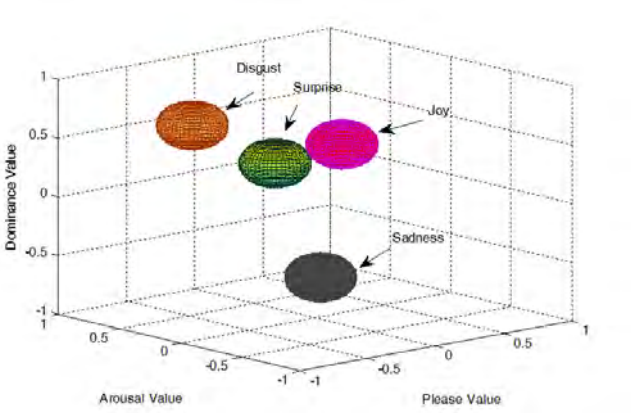


Fig. 1. Three dimensions of continuous emotion classification

Research on emotion recognition has made great progress since affective computing was proposed by Professor Picard in 1997 [27]. Overall, it is focus on text affective computing and speech emotion recognition. The realization of text affective computing is usually based on affective dictionary or machine learning. Because the voice volume, tone, sound speed and so on all contain the individual emotion, then the speech emotion recognition is relatively complex. At present, the main speech recognition methods include ANN (Artificial Neural Network) [39], HMM (Hidden Markov Model), SVM (Support Vector Machine), DBN (Deep Belief Nets) [42], GMM (Gaussian Mixture Model), DTW (Dynamic Time Warping), and mixed method [16]. For example, Schuller (2003) et al selected HMM model as their continuous speech emotion classification model, and their result showed that their method can get 86% recognition rate in recognition of seven discrete emotions [32]. At the same time, Nwe (2003) et al used discrete HMM model of vector quantization to classify six types of basic emotions, their method attained an average accuracy of 78% in the classification of six emotions [26]. Huang et al (2011) proposed a new algorithm that combined GMM and SVM to recognize speech emotion, the result showed that the average recognition rate of their method is 1.7%-3.7% higher than standard GMM method in accuracy [15]. Chavan et al (2012) used SVM to identify the anger, happiness, sadness, surprise and neutral state of the voice, and extracted MFCC as features, and their research obtained the 68% recognition rate [2].

3. Framework and Methodology

Because learning behavior is carried out through mobile devices in mobile learning, which makes the construction of affective education need to take full account of this characteristics. All the time, how to get the learners emotion timely and accurately has been a difficult problem to the implementation of affective education in mobile learning. In order to solve the above problem, the construction of affective education is built on our study, and it is based on learners interest and emotion recognition, which is shown in Fig.2.

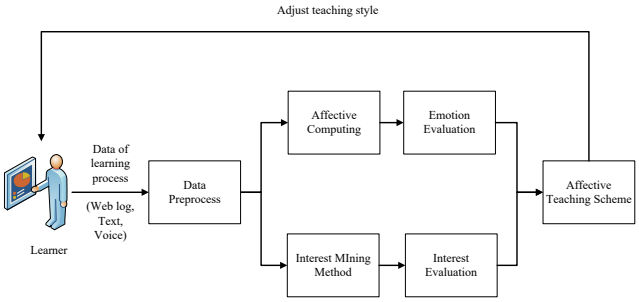


Fig. 2. Framework of affective education

It can be seen from Fig.2 that learner's web log, text and voice data are firstly preprocessed, and then the method of interest mining and affective computing will be used to these data. Once learners interest and emotion are obtained, we can design the affective teaching scheme according to the above result and adjust the teaching style for affective education.

3.1. Interest mining for mobile learning

Generally speaking, if the learner is interested in some learning resource in mobile learning, he or she will usually carries out a series of online activities such as click resource link, add to favorites or post comment and so on. Therefore, if these data are used for mining and analysis, the learners' behavior habits and their interest would be explored. It is well known that data mining method includes classification, clustering, regression analysis, association rules mining and so on. Among, association rules mining can discover the possible association or connection of objects from data, and it is especially good for mobile learners' interest mining. As mobile learners' interests are usually reflected in their learning behaviors, such as the length of their learning time, the times of being clicked of learning resources, the access order of hyperlink of learning resource. So if appropriate data mining methods are used to mine the above data, the interest of mobile learners can be obtained. In order to realize the mining of learners' interest effectively, association rule analysis and ant colony clustering were applied in this study after referring to previous studies.

Association rules analysis The association rule is an implication of the form such as $R\{A\} \rightarrow R\{B\}$. It can be understood that if a transaction contains A, then the transaction is likely to contain B. Among them, A and B are named as the precursor and successor of association rules, AB is called association rule, which is support and confidence. The degree of support in association rules can be calculated as follows.

$$Support(A \rightarrow B) = \frac{R(A) \cap R(B)}{R_{all}} \tag{1}$$

The degree of confidence in association rules can be calculated as follows.

$$Confidence(A \rightarrow B) = \frac{R(A) \cap R(B)}{R(A)} \tag{2}$$

For example, when the value of support threshold is 0.06, a learner's preference for learning content is shown in Fig.3.

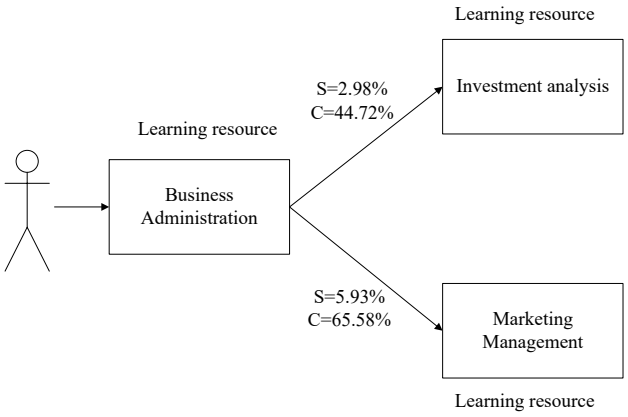


Fig. 3. Relation of learning content preference

It can be seen from Fig.3, the value of confidence of the course of Business administration and marketing management is 65.58%, which means learners are likely to have a preference for marketing management after learning the business administration.

Ant colony clustering algorithm In the course of mobile learning, learners often click on a number of learning resources, then how to effectively count and describe these resources for the purpose of interest mining is a very important task. Ant colony clustering algorithm is a clustering algorithm based on ant cleaning behavior, and the main idea is the process of transporting ants [14]. It can be assumed that the data objects to be clustered are randomly placed on a two-dimensional planar grid, and there are a number of artificial ants that allow them to move randomly in a two-dimensional plane. Each ant determines the probability of handling according to the similarity between the data object and the local environment, if the similarity is higher, the smaller probability of picking up, and the greater probability of dropping, After a certain number of iterations, the same kind of objects are clustered together in the same spatial region, and it realize the self-organizing clustering process. In this paper, an improved ant colony clustering algorithm was used, and it could be described as followings.

```
program AntCluster
  Init number of n Ants and place randomly;
```

```
begin
    repeat
    For all ants do
    Calculate the similarity object of each ant Sn;
    Calculate the probability of Picking up Pp;
    if Pp > threshold
      Pick up object and remember current position;
      Add 1 to number of Load ant;
      Move randomly;
    else
      Dont move;
    end if
    Calculate the probability of Picking up Pd;
    if Pd > threshold
      Put down object and remember current position;
      Add 1 to number of unLoad ant;
      Move randomly;
    end if
    until repeat Maximum times
end.
```

Definition 1 (Similarity) Similarity refers to the comprehensive similarity between an object and other objects in the environment. There are n objects in the data set D, and the similarity of objects is the arithmetic mean of the probability of each attribute of the object, and similarity of S_i is defined as follows.

$$f(S_i) = \frac{1}{n} \sum_{j=1}^{n} p_{ij} \tag{3}$$

Definition 2 (Pick up probability) The probability of picking up for ant is defined as follows.

$$P_p = 1 - \frac{1 - e^{-cf(S_i)}}{1 + e^{-cf(S_i)}} \tag{4}$$

Definition 3 (Dropping probability) The probability of dropping for ant is defined as follows.

$$P_d = 1 - \frac{1 - e^{-cf(S_i)}}{1 + e^{-cf(S_i)}} \tag{5}$$

Among them, the value of P_p and P_d belong to between 0 and 1. And the probability function of pick up and dropping is convex function, c is a constant that is used to adjust convergence speed, once the c value is different, and the function of convergence speed is different.

3.2. Text affective computing

There are usually two ways to implement the text affective computing, one is to rely on the emotion dictionary, and the other is based on machine learning. Considering that the text in mobile learning contains typical domain words, then the former way was used to our study. The process of text affective computing in our study is shown in Fig.4.

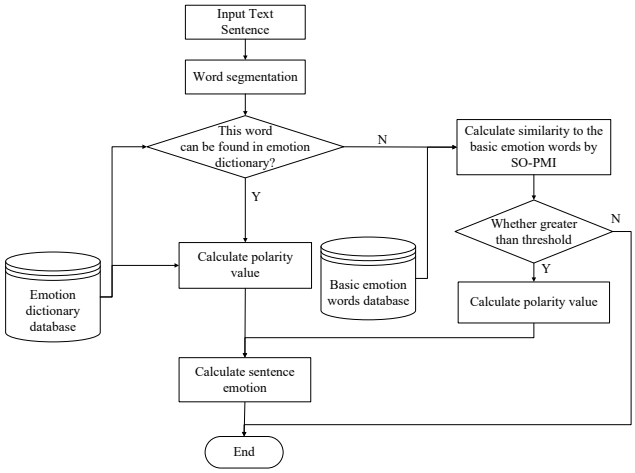


Fig. 4. Process of text affective computing

The process of text affective computing includes word segmentation, POS tagging, matching of emotional dictionary and calculation of emotional similarity and so on. For example, the statement of "Today's course is boring, I'm not interested at all", after the word segmentation and POS tagging, which is shown as follows. "Today /t 's /uj course /n is /v boring /a, I /r 'm /d not /v interested /a at all /d". Because adverbs, verbs and adjectives often contain emotions, then it was selected as candidate affective words. Therefore, 'boring /a', 'not /v', 'interested /a' and 'at all /d' were selected and were calculated according to emotional dictionary. Finally, the calculated results show that the result of this sentence reflects the negative emotions of the learners.

In addition, it can be seen from Fig.4, if the word is not in the emotion dictionary, then SO-PMI (Semantic Orientation-Pointwise Mutual Information) method will be used to calculate its similarity to the basic emotion words. The similarity of two words by SO-PMI is calculated as follows.

$$PMI(Word_1, Word_2) = \log_2 \left(\frac{P(Word_1 \& Word_2)}{P(Word_1)P(Word_2)} \right) \tag{6}$$

Among, $P(Word_1)$ is the probability of $Word_1$ appears independently in the corpus. $P(Word_2)$ refers to the probability of $Word_2$ appears independently in the corpus. And

$P(Word_1\&Word_2)$ refers to the probability of $Word_1$ and $Word_2$ in the corpus at the same time.

3.3. Speech emotion recognition

In the process of mobile learning, some of learners interaction is carried out by their voice. If we can recognize the emotion by their voice, it will be helpful for the implementation of affective education. According to previous research, some feature of speech can represent the characteristics of human, such as volume of voice, short-term zero crossing rate, pitch frequency, formant parameters, and Mel Frequency Cepstrum Coefficient (MFCC) and so on [23]. Based on previous research and our experiments, twenty-four characteristic parameters including rhythm and tone quality were selected as speech emotion recognition, and an acoustic affective computing vector function was built based on the above acoustic characteristic parameters, namely F(n)=[STE,SZR,PV,FF,NVB,VS,MFCC]. Among them, STE means short-time energy (Maximum/Average/Minimum), SZR refers to short time average zero crossing rate (Maximum/Average/Minimum), PV means the value of pitch (Maximum/Average/Minimum), FF means the first formant of voice, NVB means number of voice break, VS means voice speed, and MFCC includes 12 order Mel frequency cepstrum coefficient.

Because the voice of learner in mobile learning often embodies the characteristics of fragmentation and context, traditional speech recognition methods are often difficult to get good effect. Recent studies have demonstrated that three-dimensional PAD emotion model and the ANN-DL method can get effective results [16]. Therefore, once the data of the above speech parameters are collected, the ANN-DL method will be used and the value of PAD will be calculated. For example, if a learner's pad value is [0.47 0.34 0.31] , then we can get the learner's emotional state according to the PAD values for the typical emotion [6]. In order to identify the learner's emotions, some of the characteristic parameters need to be extracted, and some of the speech feature parameters are introduced as follows.

Volume of voice. It refers to the voice of the strength, and it can be regarded as the amplitude of the speech signal. When somebody is in happy or angry state, his volume of amplitude will be higher than that of calm state. On the contrary, if he is in grief and calm state, the volume will decrease the amplitude [37]. Generally, the volume of voice can be calculated by the sum of the voice signal amplitude.

Short-time energy. After the speech signal is divided into frames, the short-time energy of each frame can be calculated. The formula for calculating the short-time energy of frame $f_n(i)$ is as follows.

$$E_n = \sum_{i=0}^{N-1} f_n^2(i) \tag{7}$$

Pitch frequency. It refers to the frequency of the vocal cords. When people begin to speak, the sonant and airflow will pass through the glottis, which make the vocal cords

vibrate. At the same time it will generate excitation pulses and form the pitch frequency, which can be calculated by short-time autocorrelation function $R_n(k)$ or short-time average magnitude difference function $F_n(k)$. Among, $R_n(k)$ can be expressed as follows.

$$R_n(k) = \sum_{i=-\infty}^{+\infty} [x(i)w(n-i)][x(i+k)w(n-i-k)] \tag{8}$$

Where, k is called autocorrelation lag time, n is the N speech segment.
In addition, $F_n(k)$ can be expressed as follows.

$$F_n(k) = \sum_{i=-\infty}^{+\infty} |x_n(i) - x_n(i+k)| \tag{9}$$

Mel Frequency Cepstrum Coefficient. It reflects the sensory judgments of the human ear on the short time amplitude spectrum of voice, and MFCC has been widely used in the field of speech recognition in recent years. The calculation of Mel frequency is expressed as follows.

$$f(Mel) = 2595 * \lg(1 + \frac{f}{700}) \tag{10}$$

The calculation process of MFCC coefficient includes steps as follows.

Step 1. Preprocessing of speech signal. It includes define the sampling length of each frame of the voice sequence (such as N=256), and pretreat each frame of speech signal s(n).

Step 2. Calculation of discrete spectrum power. It gets the spectrum of each frame by the discrete FFT (Fast Fourier Transformation) and calculates the square of the value, and then gets the discrete power spectrum s(n), which is the energy distribution on the spectrum.

Step 3. Power spectrum filtering. This step includes calculate s(n) multiplied M triangular band-pass filter, and get M parameters P_m, where, m =0, 1, ..., M-1.

Step 4. Logarithmic treatment. This step includes calculate the natural logarithm of P_m, and get L_m, where, m =0, 1, ..., M-1.

Step 5. Discrete cosine transforms. This step includes calculate the discrete cosine transform of L_m, and get D_m, where, m =0, 1, ..., M-1, then discard the DC component of D_0, and take D_1, D_2, ..., D_k as the MFCC coefficients.

Deep learning. A lot of fragmented voices bring great difficulty to emotion recognition in mobile learning. The success of deep learning algorithm in various industries has brought inspiration to our research. Overall, the concept of deep learning is based on artificial neural networks, and it is a multilayer perceptron with multiple hidden layers. It is combine low-level features to form a more abstract high-level representation of attribute

categories or features in order to discover the distributed feature representation of data. As a complex machine learning algorithm, deep learning can imitate human beings to solve realistic problems. Especially, it has developed rapidly and achieved great success in speech recognition. For example, it was reported that the accuracy of speech recognition of English in Google machine learning systems has reached 95%, and IFLYTEK Company has got more than 97% rate recognition in Chinese speech. Considering the advantages of deep learning, we used it to our speech recognition study. The structure of ANN-DL (Artificial Neural Network with Deep Learning) is shown in Fig. 5.

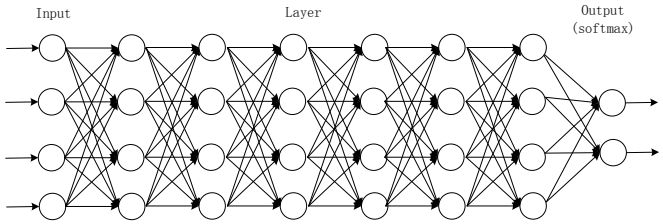


Fig. 5. The network structure of deep learning

Autoencoder is an important part of deep learning technology. As one of unsupervised learning algorithm, it uses the back propagation algorithm and tries to construct an identity function so that the output value is close to the target value. In addition, the gradient descent method is usually used to update the parameters in the Autoencoder algorithm, and the specific weights and bias of the update is shown as follows.

$$U_{ij} = U_{ij} - \alpha \frac{\partial J_{AE+wd}(\theta)}{\partial U_{ij}} \tag{11}$$

$$b_{ij} = b_{ij} - \alpha \frac{\partial J_{AE+wd}(\theta)}{\partial b_{ij}} \tag{12}$$

Where, α means learning rate.

4. Experiment

In this paper, our experimental data comes from a large online learning platform in China (http://www.shlll.net), which has more than 1 million 300 thousand registered learners. Thirty-two learners were randomly selected as subjects, and their text and voice in mobile learning was collected at the same time. In addition, basic speech emotional corpus is composed of CASIA (Chinese Academy of Sciences Institute of Automation) speech corpus [12]and three hundred marked historic sentences voice. Among, CASIA corpus includes happy, sad, angry, surprise, fear, and neutral six different emotional voices with the same semantic texts. On the basis of the establishment of text emotional database, Chinese Affective Dictionary of Information Retrieval Laboratory of Dalian University of Technology was selected as our study, which was created by Prof. Lin et al [40] and it

includes seven categories emotional words, that is, joy, love, anger, sorrow, fear, disgust and surprise.

4.1. Data processing

Generally speaking, learner behavior data needs to be pre-processed firstly in order to get better result, which includes processing of web log, speech signal de-noising and data transformation of interest preference.

Processing of web log. It includes data cleaning, user identification, session identification, formatting output and so on, and it is the key phase of learners interest mining, especially in the process of data cleaning and user identification. If the web log is not handled properly, it can greatly affect the efficiency or accuracy of interest mining. The processing of web log is shown in Fig. 6.

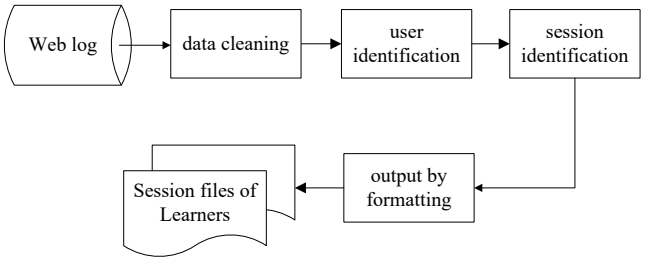


Fig. 6. Processing of web log

Speech signal de-noising. Because environmental noise is inevitably exists, the speech data need a series of pre-processing. Wavelet transform has been widely used to speech pre-processing, so it was used to our study. Firstly, the packet of db5 was selected as wavelet packet, and three level and 'Shannon' of entropy were used to the decomposing of speech signals. After above process, high frequency and low frequency coefficients were separated from initial signals. As the noise signal often exists in high frequency and it need to be discarded. Therefore, once the part of high frequency signals has more than a threshold, it will be discarded and the rest of the signals are recombined into new signals for analysis. Sample of initial signal and de-noising signal are shown in Fig. 7.

Data transformation of interest preference. On the data conversion processing of learning preferences, the threshold value will be set to deal with it. If the value is greater than the threshold value, then the content is the learner's preferences and its value is marked to

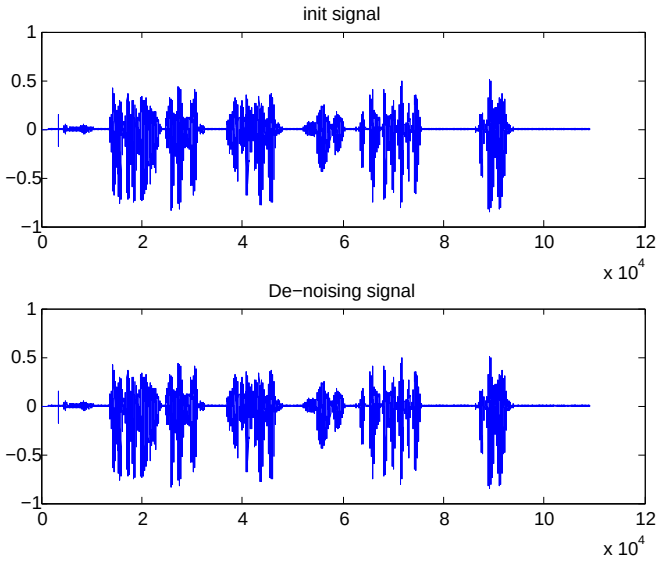


Fig. 7. Contrast of initial signal and de-noising signal

1. Otherwise, it is not the learners preference and its value is marked to 0. The expression is as follows.

$$preference = \begin{cases} 1, \text{is preference} \\ 0, \text{not preference} \end{cases} \tag{13}$$

After the data transformation, all data have become Boolean type and is very convenient to deal with. Taking the content of management course as an example, the preferences of learners are as follows.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} \end{bmatrix} = \begin{bmatrix} 1\ 0\ 1 \\ 1\ 1\ 1 \\ 1\ 1\ 0 \\ \vdots\ \vdots\ \vdots \\ 1\ 1\ 1 \end{bmatrix} \tag{14}$$

In (14), n refers to the number of management course of learners, and the three columns of right of equation represent the learners interest preferences. For example, [1 0 1] indicates that the learner is interested in the first and third courses. However, he or she is not interested in the second course.

4.2. Experimental Results and Analysis

Learner's interest evaluation. According to the association rule analysis and ant colony clustering method, we can get the interest of learners who participate in the experiment.

For example, the course of ”management” includes five knowledge points, namely, planning, organizing, commanding, coordinating and commanding. Some learners exhibit different learning behaviors when they are learning the above knowledge points of the course. We found that a learner spend more time and much number of clicks on learning the knowledge of commanding than any other four knowledge points, especially when he was studying the knowledge of motivational theories and leadership behavior. Then we carried on data mining to this learner's learning behavior data by association rule analysis and ant colony clustering method. And some behavior rules of the learner were discovered such as he likes to browse the story of motivating employees, and he often clicked on the content of controlling chapter when he was studying motivational theories.

If a large number of learners' learning behavior rules are discovered, their interest would be easy to describe. Taking the management course as an example, we collected the learning data from thirty-two learners and described their interests according to five knowledge points of the course. And the interests of two of them are shown in Fig.8. It can be seen that the most interest of No. 1 learner is the knowledge of the commanding, and he has little interested in the knowledge of the planning. However, the most interest of No. 2 learner is the knowledge of the planning. Therefore, if we can get the learners' interest accurately, it will be a great help for the effective implementation of affective education.

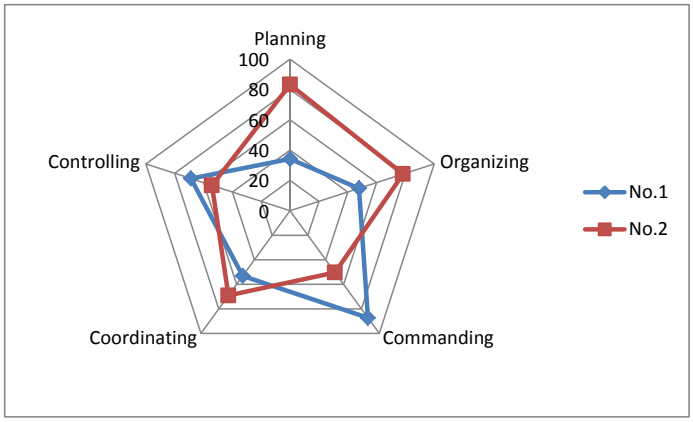


Fig. 8. The learner's interest of knowledge points

Learner's emotion evaluation. According to the result of affective computing, we can get learners' emotion who participates in the experiment. For example, on the recognition of speech emotion, the twenty-four speech feature parameters are extracted and deep neural network method is used for training and speech emotion recognition. By observed, we found learners is easy to show disgust, joy, sadness and surprise emotions in the learning process, then the above four kinds of emotion recognition rate is calculated and it is shown in Fig.9.

The recognition rates for emotion of disgust, joy, sadness and surprise are 86.32%, 91.79%, 83.72% and 90.42%, and the average recognition rate is 88.06%. Overall, this

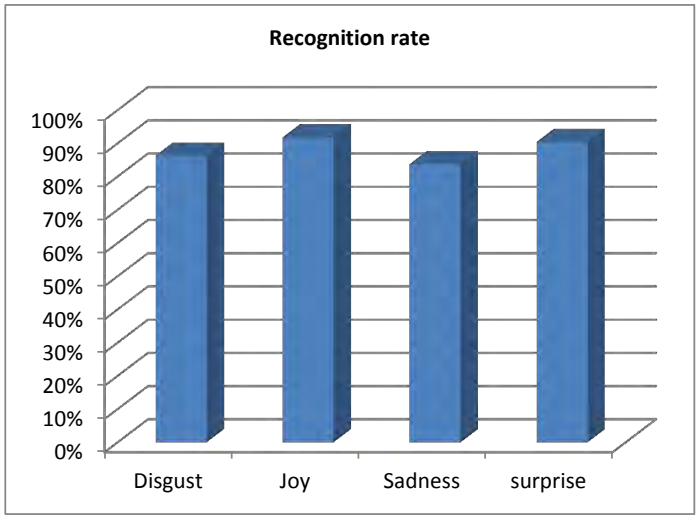


Fig. 9. The recognition rate of emotion states

method can help us to recognize the learner's emotion effectively. In addition, This proposed method has been applied satisfactorily to the affective education of foreign languages and emotional intelligence studies [5][7].

5. Conclusion

From the point of view of learning trend, mobile learning will become more and more popular. At the same time, how to implement affective education is a very meaningful research. Although emotion recognition has made great progress in recent years, however, many challenges still need to be faced and explored, such as the dynamic changes of learners' emotion, or the complex computing of mixed emotion, and so on.

This paper proposes the method of constructing the affective education based on learner's interest and emotion recognition, SO-PMI and ANN-DL methods are applied. Experimental results show that this method is effective and it can reach high recognition rates. From the perspective of future research, face recognition and other biometric verification of the human body can be used to affective education. Especially, combine emotional intelligence [17] with learner's behavior big data to study is the trend of the future research.

Acknowledgments. This work is supported by the Project of Shanghai Higher Education Society (No. GJSL1624), Program of Science and Technology Commission of Shanghai Municipality (No. 13DZ2252200), Project of Shanghai young university teachers training grant program project of Shanghai Municipal Education Commission (No. ZZSUIBE16026) and Project of Qtone Education of Ministry of Education of China (No. 2017YB115). Many thanks to Jinzhao Wang for her assistance to Yonghui Dai who is the corresponding author of this paper.

References

1. Cao, H.T., Li, M.C., Jiao, D., Feng, B.: A method of micro-blog sentiment analysis based on pad model. Advanced Materials Research 659, 186–190 (2013)
2. Chavan, V.M., Gohokar, V.V.: Speech emotion recognition by using svm-classifier. International Journal of Engineering & Advanced Technology (5), 11–15 (2012)
3. Cheng, K.Y.: Implementation of affective education: The case of a middle school in china's guangzhou. Journal of Educational Research 13(2), 47–63 (2010)
4. Connolly, G.J.: Applying humanistic learning theory: The "art" of coaching. Strategies 29(2), 39–41 (2016)
5. Dai, W.H., Duch, W., Abdullah, A.H., Xu, D.R., Chen, Y.S.: Recent advances in learning theory. Computational Intelligence & Neuroscience 2015(4), 1–4 (2015)
6. Dai, W.H., Han, D.M., Dai, Y.H., Xu, D.R.: Emotion recognition and affective computing on vocal social media. Information & Management 52(7), 777–788 (2015)
7. Dai, W.H., Huang, S., Zhou, X., Yu, X., Ivanović, M., Xu, D.R.: Emotional intelligence system for ubiquitous smart foreign language education based on neural mechanism. Jitam 21(3), 65–77 (2014)
8. Dai, Y.H., Han, D.M., Dai, W.H.: Modeling and computing of stock index forecasting based on neural network and markov chain. ScientificWorldJournal. 2014, 1–9 (2014)
9. Fatahi, S., Ghasemaghaee, N.: Design and implementation of an intelligent educational model based on personality and learner's emotion. International Journal of Computer Science & Information Security 7(3), 14–24 (2010)
10. Ghasemaghaei, R., Arya, A., Biddle, R.: The made framework: Multimodal software for affective education. American Journal of Emergency Medicine 8(5), 467–468 (2015)
11. Gong, S.P., Dai, Y.H., Ji, J., Wang, J.Z., Sun, H.: Emotion analysis of telephone complaints from customer based on affective computing. Computational Intelligence & Neuroscience 2015(5), 1–9 (2015)
12. Han, W.J., Li, H.F.: A brief review on emotional speech databases. Intelligent Computer & Applications 3(1), 5–7 (2013)
13. Herrington, A.: Authentic mobile learning in higher education. Aare 8(4), 489–496 (2008)
14. Hu, X.H., Mu, T., Dai, W.H., Hu, H.Z., Dai, G.H.: Analysis of browsing behaviors with ant colony clustering algorithm. Journal of Computers 7(12), 3096–3102 (2012)
15. Huang, Y.M., Zhang, G.B., Dong, F., Peng, D.F.: Speech emotion recognition based on two kind of gmm-ubm multidimensional likelihoods and svm. Application Research of Computers 28(1), 98–101 (2011)
16. Ijjina, E.P., Mohan, C.K.: Hybrid deep neural network model for human action recognition, vol. 46. Elsevier Science Publishers B. V. (2016)
17. Ivanovic, M., Budimac, Z., Radovanovic, M., Kurbalija, V., Dai, W., B?dic?, C., Colhon, M., Ninkovic, S., Mitrovic, D.: Emotional agents - state of the art and applications. Computer Science & Information Systems 12, 47–47 (2015)
18. Kim, C.M., Hodges, C.B.: Effects of an emotion control treatment on academic emotions, motivation and achievement in an online mathematics course. Instructional Science 40(1), 173–192 (2012)
19. Li, L., Zheng, L., Yang, F., Li, T.: Modeling and broadening temporal user interest in personalized news recommendation. Expert Systems with Applications 41(7), 3168–3177 (2014)
20. Luo, Y.J., Wu, J.H.: Emotional and mental control and cognitive strategy research. Journal of Southwestern Normal University (Humanities and Social Sciences Edition 31(2), 26–29 (2005)
21. Maheswari, B.U., Sumathi, P.: Prediction of web users interest on buying behaviour for e-commerce solutions. International Journal of Applied Engineering Research 10(16), 37760–37764 (2015)

22. Mega, C., Ronconi, L., De, B.R.: What makes a good student? how emotions, self-regulated learning, and motivation contribute to academic achievement. Journal of Educational Psychology 106(1), 121–131 (2014)
23. Milton, A., Ghosh, E.S.N., Mohan, S.S., Selvi, S.T.: Analysis towards optimum features and classifiers to recognize the emotion happiness from speech signals. International Journal of Applied Engineering Research 10(12), 32585–32600 (2015)
24. Morcom, V.: Scaffolding social and emotional learning in an elementary classroom community: A sociocultural perspective. International Journal of Educational Research 67, 18–29 (2014)
25. Nakatsuji, M., Fujiwara, Y., Uchiyama, T., Toda, H.: Collaborative filtering by analyzing dynamic user interests modeled by taxonomy. Transactions of the Japanese Society for Artificial Intelligence 28(6), 361–377 (2012)
26. Nwe, T.L., Foo, S.W., Silva, L.C.D.: Speech emotion recognition using hidden markov models. Speech Communication 41(4), 603–623 (2003)
27. Picard, R.W.: Affective Computing. MIT Press, Cambridge, Mass, USA (1997)
28. Rao, K.S., Krishnamurthy, M., Kannan, A.: Extracting the user's interests by using web log data based on web usage mining. Journal of Computational & Theoretical Nanoscience 12(12), 5031–5040 (2015)
29. Romi, S., Katz, Y.J.: Affective education: The nature and characteristics of teachers and students attitudes toward school in israel. Educational Sciences Theory & Practice 25(25), 35–47 (2003)
30. Sanchez, F., Barrilero, M., Alvarez, F., Cisneros, G.: User interest modeling for social tv-recommender systems based on audiovisual consumption. Multimedia Systems 19(6), 493–507 (2013)
31. Schröder, M.: The cognitive structure of emotion. Pure & Applied Geophysics 168(12), 2395–2425 (2011)
32. Schuller, B., Rigoll, G., Lang, M.: Hidden markov model-based speech emotion recognition. In: Proceedings of the 2003 International Conference on Multimedia and Expo, IEEE Computer Society. vol. 2, pp. 401–404 (2003)
33. Siu, K.W.M., Yi, L.W.: Fostering creativity from an emotional perspective: Do teachers recognise and handle students emotions? International Journal of Technology & Design Education 26(1), 105–121 (2016)
34. Stamou, S., Ntoulas, A.: Search personalization through query and page topical analysis. Kluwer Academic Publishers (2009)
35. Valk, J.H., Rashid, A.T., Elder, L.: Using mobile phones to improve educational outcomes: An analysis of evidence from asia. International Review of Research in Open & Distance Learning 11(1), 117–140 (2010)
36. Wang, N.: Highlighting the humanistic spirit in the age of globalization: Humanities education in china. European Review 23(2), 273–285 (2015)
37. Wu, J.D., Ye, S.H.: Driver identification based on voice signal using continuous wavelet transform and artificial neural network techniques. Expert Systems with Applications 36(2), 1061–1069 (2009)
38. Xu, B.: A user interest model based on the analysis of user behaviors. Journal of Intelligence 27(12), 76–78 (2009)
39. Xu, L., Ren, J.S.J., Liu, C., Jia, J.: Deep convolutional neural network for image deconvolution. Advances in Neural Information Processing Systems 2, 1790–1798 (2014)
40. Xu, L.H., Lin, H.F., Pan, Y., Ren, H., Chen, J.M.: Constructing the affective lexicon ontology. Journal of the China society for scientific and technical information 27(2), 180–185 (2008)
41. Zhao, X.: A review of progress and trend of foreign research on emotional education. Comparative Education Review (2013)
42. Zou, C.R., Zhang, X.R., Cheng, Z., Li, Z.: A novel dbn feature fusion model for cross-corpus speech emotion recognition. Journal of Electrical and Computer Engineering 2016(1), 1–11 (2016)

Haijian Chen is an associate professor at the Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, China. He received his Ph.D. in Management Science and Engineering from Shanghai University of Finance and Economics, China in 2015. His current research interests include Educational technology and cloud computing. Contact him at xochj@shtvu.edu.cn.

Yonghui Dai, corresponding author of this work, he is currently a lecturer at the Management School, Shanghai University of International Business and Economics, China. He received his Ph.D. in Management Science and Engineering from Shanghai University of Finance and Economics, China in 2016. His current research interests include Affective Computing, Intelligence Service and Big Data Analysis. His works have appeared in international journals with eighteen papers. Contact him at daiyonghui@suibe.edu.cn.

Yanjie Feng is an associate professor at the Management School, Shanghai University of International Business and Economics, China. She received her Ph.D. in Management Science and Engineering from Shanghai Jiao Tong University, China in 2001. Her research interests include Network information communication and financial risk management. Contact her at fengyanjie@suibe.edu.cn.

Bo Jiang is currently a Ph.D. candidate at the School of Information Management and Engineering, Shanghai University of Finance and Economics, China. His current research interests include Data Mining and Cloud Computing. Contact him at jiangbo@sui.edu.cn.

Jun Xiao is currently a professor at the Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, China. He received his Ph.D. in Educational technology from East China Normal University, China in 2006. His research interests include Educational technology and Management Information System. Contact him at xiaoj@shtvu.edu.cn.

Ben You is currently a Master candidate at the Management School, Shanghai University of International Business and Economics, China. He received his bachelor's degree in Traffic Engineering from Xiamen University of Technology, China in 2016. His current research interests include Data mining and Network behavior analysis. Contact him at youben022@gmail.com.

Received: January 10, 2017; Accepted: August 19, 2017.

Computer Science and Information Systems 14(3):703–718 https://doi.org/10.2298/CSIS170112024H

A Retrieval Algorithm of Encrypted Speech based on Syllable-level Perceptual Hashing

Shaofang He[1,2], Huan Zhao[1,*]

[1] School of Information Science and Engineering, Hunan University
410082 Changsha, China
wxdyzp@sina.com
[2] College of Science, Hunan Agricultural University
410128 Changsha, China
wxdyzp@sina.com
[3] *Corresponding author at: School of Information Science and Engineering, Hunan University
410082 Changsha, China
hzhao@hnu.edu.cn

Abstract. To retrieve voice information in a fast and accurate manner over encrypted speech, this study proposes a retrieval algorithm based on syllable-level perceptual hashing. It implements the function of retrieving speech segment and spoken term over encrypted speech database. Before uploading the speech to the cloud, it needs to embed the digital watermarks (perceptual hashing). In the retrieval process, it does not need search over encrypted speech data directly or decryption, but requires searching the system hash table. Experimental results show that the syllable-level perceptual hashing of the proposed scheme has good discrimination, uniqueness, and perceptual robustness to common speech. In addition, the proposed retrieval algorithm effectively improves the retrieval speed by reducing the matching number of query index. The precision ratio and recall ratio all achieve high under various signal processing.

Keywords: Speech retrieval, Posterior probability, Syllable segmentation, Perceptual hashing.

1. Introduction

Recently, with fast development of multi-media communication, audio and video have been applied more and more widely on the Internet. In particular, the digital audio has virtually become one of the most popular multi-media applications. To satisfy the requirement of large multi-media data management, cloud computing technique presents multi-media services in a new way. Cloud computing is new model of enterprise IT infrastructure which provides on demand high quality application and service from shared pool of computing resources. However, cloud storage servicer is not a trusted third party from a security standpoint. In multi-media applications, many sensitive multi-media data are related to privacy preserving, for example, in the scene of e-health, health related multimedia data is being exponentially generated from healthcare monitoring devices and sensors, coming with it are the challenges on how to efficiently acquire, index, and process such a huge amount of data for effective healthcare and related decision making, while respecting user's data privacy [1]; as well as in the scene of telecommunications, if

the sensitive speech data are stored in the cloud without protecting, it may create issues such as leakage or abuse of personal privacy speech information [2]. Therefore, security protection emerges as an important problem. One of the effective method to protect the security of outsourcing data is data encryption, but it results in the difficulty of encrypted multimedia data retrieval. As we know, encrypted data makes the traditional data utilization service based on plaintext keyword search ineffective.

2. Related works

In last few years, many works have been done for encrypted multimedia database and its retrieval. Qin Liu et al. [3] worked on Secure and privacy preserving keyword searching for cloud storage services which allows the CSP to take part in the decipherment, and returns only files in which user is interested without leaking any information about plaintext. Zhangjie Fu et al.[4] proposed Multi keyword Ranked Search Supporting Synonym Query to overcome the problems of traditional multi keyword scheme and has proposed Two secure schemes to meet up privacy requirements in two threat models as known cipher text model and known background model. The search results achieved when authorized cloud user input the synonym of the predefined keywords, not exact or fuzzy matching keywords. Baojiang Cui et al. [5] worked on Key-Aggregate Searchable Encryption (KASE), in which a data owner only needs to distribute a single key to a user for sharing a large number of documents, and the user only needs to submit a single trapdoor to the cloud for querying the shared documents. Zhangjie Fu et al. [6] proposed flexible and efficient searchable scheme which supports multi keyword and synonym based search. It proposes new text feature weighting function which adds new weighting factor to distinguish keyword on the basis of term frequency keyword and make easy retrieval. Jin Li et al. [7] worked on revocable identity based encryption which offloads all keys generation related operation during key issuing and update, leaving constant no of simple operation so that eligible users can performed locally. All the five schemes mentioned above worked well in encrypted cloud database for retrieval of data files, but as an improvement, Rupali D. Korde et al. [8] suggested new scheme where it was possible for users to upload and download multimedia data.

In multimedia data, privacy-preserving search over encrypted speech data has come into being an important and urgent research field in cloud storage. In the cloud, the rapid increase of speech data size has prompted the need to rapidly and accurately retrieve needed speech data or spoken term from protected speech databases. At present, speech information retrieval over encrypted speech data is in hotspot. Because encrypted speech lose many properties of speech signal, and such loss makes the methods used for plaintext search having highly problematic for encrypted speech retrieval. In traditional retrieval methods, keywords need to be matched exactly, however, the return results will be very less for frequent user access and large number of cloud data. In many existing retrieval methods, a keyword is encrypted as an index and matched with the encrypted data directly. After encryption, keywords lose most of speech features, and the size of encrypted speech data in cloud computing environments is massive, therefore, those algorithms implemented by matching the encrypted keyword and the encrypted data do not possess strong applicability. Ton Kalker first proposed the concept of perceptual hashing in 2001 [9]. Perceptual hashing is described as follows. (1) Bits with little data called perceptual

hash valve can represent multimedia objects with large data; (2) It meets the mapping relationship of multiple objects to one object; (3) For multimedia objects of the same or similar perceptual content, their perceptual hashing sequences are close in mathematical distance [10]. In the field of multimedia information processing and information security, the strong discrimination, uniqueness, and perceptual robustness of perceptual hashing have earned recognition since the concept was presented. The unique characteristics of speech content and mapping speech data to a brief digital digest (called perceptual hashing digest) are the basis of speech perceptual hashing technology. In this technology, a digital representation of multimedia objects is the input data, and the perceptual hashing digest is the output data. For the multimedia information with different contents, its perceptual hashing digest will be significantly differ. In other words, for the multimedia information with the same content regardless of the digital representation, its perceptual hashing digest will remain the same or similarity. The generation of speech perceptual hashing generally involves pretreatment (includes framing and window addition, time-frequency transform), feature extraction, and hash algorithm construction. The method of last two steps make speech perceptual hashing digest different from existing algorithms. Wang et al. put forward a watermark-based perceptual hashing search algorithm over encrypted speech in [10]. In the proposed scheme, the zero-crossing rate is extracted from the digital speech to generate the perceptual hashing as the search digest, which is embedded into the encrypted speech signal; without downloading and decrypting, the search results could be obtained rapidly and accurately by matching and computing the normalized Hamming distance of the perceptual hashing digests between the search target and the extracted one. Based on changes in the characteristics of the time and frequency domain, Hao et al. proposed a speech perceptual hashing algorithm [11]. The scheme also offered good discrimination and robustness and puts forward new ideas for applying perceptual hash technology in large-scale data processing. Recently, after studying existing speech retrieval technologies, Zhao et al. explored a novel perceptual hashing-based retrieval algorithm [12]. In the algorithm, multifractal characteristic of speech data and the technology of piecewise aggregate approximation (PAA) were introduced to generate perceptual hashing sequence. Compared with the methods of [10], [11], the perceptual hashing generated from multifractal characteristics showed better distinctiveness and robustness.

To sum up, the existing methods ([10], [11], [12]) only can search speech segment and need searching the system hashing table completely and matching each index, which makes these methods become inefficient in large-scale data processing. To further improve the retrieval speed, discrimination and perceptual robustness of speech perceptual hashing, the present study proposes a syllable-level perceptual hashing-based retrieval algorithm for encrypted speech. Different from the existing methods, the syllable-level perceptual hashing is introduced in this study for the first time; furthermore, the posterior probability based on acoustic segment models [13] is employed to generate the perceptual hashing digest. Additionally, in the process of retrieving speech segment and spoken term, only the perceptual hashing of equal length and header matching with the target perceptual hashing should be matched in system hash table, and it brings the greatly improvement in retrieval speed.

The remainder of this paper is organized as follows: Section 3 describes the system model of retrieval scenario and a desirable retrieval scheme. Section 4 presents exhaus-

tively the retrieval scheme, which mainly includes the generation of syllable-level perceptual hashing, the retrieval algorithm of speech segment and spoken term. Experimental results and analysis are given in Section 5. Conclusions of the study are drawn in Section 6.

3. System model

As discussed in the introduction, in order to protect speech data privacy, speech need to be encrypted before being transferred to the cloud storage servicer. Speech encryption can be done using state-of-the-art ciphers such as undetermined blind source separation-based dual key speech encryption algorithm [14]. Built upon the established cryptographic speech encryption tools, it is computationally difficult to decrypt speech data. Encryption keeps speech data safe from the server but also makes it difficult for the server to build searchable indexes. A desirable indexing scheme for encrypted speech retrieval, in addition to being efficient and scalable, should retain the similarity between speech pairs. The system model is shown by the left and the retrieval scheme is displayed by the center dash-dotted blocks in Fig.1. The model is mainly composed of generation of encrypted speech with watermarks and retrieval processes. An efficient way of representing speech and potentially enabling fast and scalable search is by the speech perceptual hashing. Before building search index (speech perceptual hashing digest), the posterior probability based on acoustic segment models of speech segment is extracted by employing the method of [13]; meanwhile, syllables are obtained by utilizing the syllable segmentation algorithm [15]. For speech segments, the perceptual hashing sequence of each syllable is generated and embedded into encrypted speech as a digital watermark. The system hash table is formed by the perceptual hashing sequences of all speech segments. In the process of speech retrieval, feature extraction and syllable segmentation of the query speech are conducted, and the perceptual hashing sequences of all the syllables are generated and built the query index (target perceptual hashing). Instead of searching over encrypted speech data directly, the target perceptual hashing digest searches in the system hash table. If the perceptual hash values match successfully, the retrieval result is obtained.

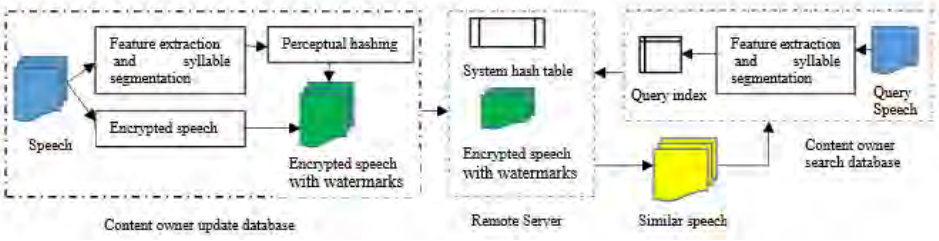


Fig. 1. System model

4. Retrieval scheme

In this section, we consider two retrieval schemes, namely, speech segments retrieval and spoken term retrieval.

4.1. Generation of speech perceptual hashing

For audio retrieval technology based on context, the building of index is one of the key links. It is critical to extract better and shorter digital digest representing audio for enhancing retrieval performance. In the retrieval algorithms of encrypted speech database, perceptual hashing sequence generated from speech features is considered as the index. Generally, the extraction of speech feature for audio signals uses short-time analysis technology. Depending on the extraction method, speech feature involves linear and nonlinear characteristics. There are advantages and disadvantages for linear and nonlinear characteristics of speech signal. Linear characteristics outperform nonlinear features in terms of meaning and computation, but nonlinear features show better robustness for general audio operations, although their extraction is relatively complex [16]. In this work, the posterior probability based on acoustic segment models of speech are chosen for generating speech perceptual hashing. The pending speech data are first divided into ordered syllables by employing a syllable segmentation algorithm. Subsequently, the perceptual hashing value of each syllable is calculated. Finally, the system hash table is constituted by the perceptual hashing sequences of all the syllables.

Supposing a total of t speech segments ($A_1, A_2, \cdots, A_t$) need to generate the encrypted speech with watermarks, and taking the speech segment A for example, the specific generation process of speech perceptual hashing is described as follows:

Step 1. Framing: Pending speech signals A are divided into speech frames with fixed frame lengths. The frame shift is half of the frame length supposing $A = \{a_q, q = 1, 2, ..., n\}$, where q is the frame pointer and n is the total number of frames to be included in A.

Step2. Feature extraction: Through the acoustic segment models, the posterior probability feature vector $P = \{p_1, p_2, ..., p_n\}$ of speech segment A is obtained, where$p_q = \{p_q^1, p_q^2, ..., p_q^D\}, q = 1, 2, ..., n$.

Step 3. Syllable segmentation: Utilizing the syllable segmentation method, the speech data A are divided into ordered syllables$S_i, i = 1, 2, ..., N$, where N is the total number of syllables, supposing each syllable contains up to M frames.

Step 4. Generation of perceptual hashing value: For syllable S_i, whose total frame number is m, $m \leq M$, the posterior probability feature vector is represented by $P_i = \{p_{i1}, p_{i2}, ..., p_{im}\}$, where $p_{iq} = \{p_{iq}^1, p_{iq}^2, ..., p_{iq}^D\}, q = 1, 2, ..., m$. The average value of the D-dimension component is selected to constitute the threshold vector $T = \{T_1, T_2, \cdots, T_m\}$ of perceptual hashing, i.e. $T_q = \frac{1}{D} \sum_{j=1}^{D} p_{iq}^j, q = 1, 2, ..., m$. Comparing the hash threshold vector with D-dimension posterior probability feature vector of each speech frame sequentially, the perceptual hashing sequence H_i of a fixed length being M bits is generated according to formula (1), where $H_i(l, q)$ is the value of the l^{th} row and q^{th} column of perceptual hashing digest. The perceptual hashing sequence of speech signals A is

represented by $H = \{H_1, H_2, ..., H_N\}$.

$$\begin{aligned} &H_i(l,q) = 0, q = 1, 2, ..., M - m \\ &H_i(l,q) = \begin{cases} 1, \ p^l_{i(q-(M-m))} \geq T_{q-(M-m)} \\ 0, \ p^l_{i(q-(M-m))} < T_{q-(M-m)} \end{cases}, \ q = M - m + 1, ..., M \\ &l = 1, 2, ..., D \end{aligned} \tag{1}$$

Step 5. Construction of the system hash table: a system hash table is constructed with the speech perceptual hashing sequences $(H^1, H^2, ..., H^t)$of all the speech segments $(A_1, A_2, , A_t)$.

As discussed in the system model, the perceptual hashing digest needs to be embedded into the encrypted speech as a digital watermark. Because of the modification of encrypted speech resulting in decryption errors (they must be reduced as much as possible), perceptual hashing digests as digital watermarks will be embedded into the least significant bit (LSB) of the encrypted speech data. For syllable S_i, which contains m frames,$m \leq M$, the embedding of perceptual the hashing sequence is described as follows: firstly, in the perceptual hashing sequence H_i with a fixed length M, there are $(M - m)$ zeros before the most significant bit, after removing these zeros, the equivalent perceptual hashing sequence with a length of m bits is obtained; secondly, the sample points of the speech frames are chosen sequentially and converted to binary forms, then, the perceptual hashing value is assigned to the LSB as a digital watermark to produce encrypted speech with watermarks.

4.2. Speech segment retrieval

After the encrypted speech data with digital watermarks is generated and the system hash table is uploaded to the cloud server, the retrieval of speech segment over encrypted speech data can be conducted without decryption as soon as a user sends a retrieval request. Supposing Q is the speech segment to be retrieved, the search process is detailed as follows.

Step 1. The D-dimension posterior probability features of Q are extracted, and syllable segmentation is performed to finally obtain N syllables.

Step 2. For each syllable of Q, a perceptual hashing sequence with a length of M is generated according to the method presented in Section 3.1. The target perceptual hashing sequence $H_Q = \{H_{Q1}, H_{Q2}, ..., H_{QN}\}$that corresponds to the query speech segment is constructed with the perceptual hashing digests of all syllables to be included in Q.

Step 3. The perceptual hashing values with a length of $M \times N$are searched out from the system hash table, supposing one of them is$H_S = \{H_{S1}, H_{S2}, ..., H_{SN}\}$. Before matching H_Q with H_S, the normalized Hamming distance (bit error rate, BER) [12] of perceptual hashing digest between two syllables (such as H_i and H_j) should be first defined, and its formula is displayed as follows:

$$D(H_i, H_j) = \frac{1}{D \times M} \sum_{l=1}^{D} \sum_{q=1}^{M} (H_i(l,q) \oplus H_j(l,q)) \tag{2}$$

Then, the normalized Hamming distance between H_Q and H_S can be calculated according to formula (3).

$$D(H_Q, H_S) = \frac{1}{N}\sum_{i=1}^{N} D(H_{Qi}, H_{Si}) \tag{3}$$

Supposing the similarity threshold is$T', 0 < T' < 0.5$, if$D(H_i, H_j) < T'$, then H_i and H_j are matched successfully; similarly, if$D(H_Q, H_S) < T'$, then H_Q and H_S are matched successfully as well. In the candidate perceptual hashing with a length of$M \times N$, their headers should be matched with H_{Q1}firstly, takeH_S for example, if$D(H_{Q1}, H_{S1}) < T'$, then H_Q and H_S should be matched successively, otherwise, the target perceptual hashing sequence $H_Q = \{H_{Q1}, H_{Q2}, ..., H_{QN}\}$need not to match H_S . It will continue to match the next perceptual hashing with a length of $M \times N$ using the same method. In general, because that the perceptual hashing sequence of each syllable has a fixed length M, the speech segment that corresponds to the perceptual hashing digest with a length of $M \times N$includes N syllables, therefore, the perceptual hashing digests of speech segments without having N syllables do not need matching in the system hash table; furthermore, owning to the speech perceptual hashing sequences matched successfully have the similar header, the candidate perceptual hashing of equal length without similar header do not need matching as well. In this way, it reduces the matching number of retrieval and improves the retrieval efficiency. The illustration is given in Fig.2.

Step 4. The detection results are obtained after the completion of retrieval in the system hash table. The digital watermarks embed in the encrypted speech can be extracted and matched with the perceptual hashing digest of the query speech to verify whether the encrypted speech is damaged or not.

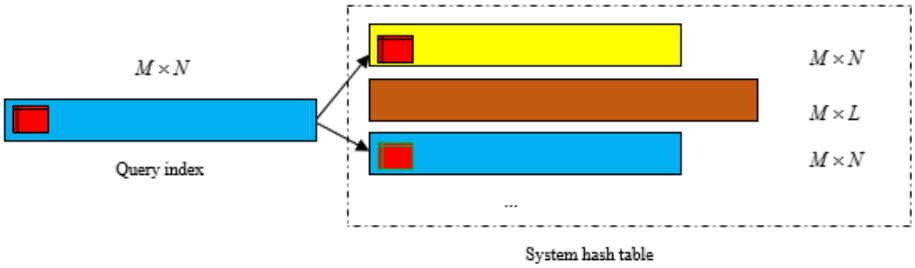


Fig. 2. Illustration of matching selection for speech segment retrieval

4.3. Spoken term retrieval

It is known that the system hash table is constructed with the speech perceptual hashing sequences $(H^1, H^2, ..., H^t)$of all the speech segments $(A_1, A_2, \cdots, A_t)$. The perceptual hashing sequence of speech sentence A_i is$H^i = (H_1^i, H_2^i, ..., H_N^i)$, which is constituted by the perceptual hashing digests of N ordered syllables to be included in A_i. The number of syllables in different speech sentence may be different, and the query spoken term

should be retrieved in the perceptual hashing sequence of each speech sentence. If a user sends a spoken term detection request, supposing K is the query term, the process that K searches in the speech segment A_i is detailed as follows.

Step 1. The D-dimension posterior probability features of K are extracted, and syllable segmentation is performed to obtain L syllables.

Step 2. For each syllable of K, a perceptual hashing sequence with a length of M is generated according to the method presented in Section 3.1, then the target perceptual hashing digest $H_K = \{H_{K1}, H_{K2}, ..., H_{KL}\}$ that corresponds to the query term is constituted by the perceptual hashing sequences of all the syllables in order.

Step 3. The perceptual hashing sequence of speech segment A_i is denoted with$H^i = (H_1^i, H_2^i, ..., H_N^i)$, which generates $N-L+1$ sets $Hv^i = (H_v^i, H_{v+1}^i, ..., H_{v+L-1}^i), v = 1, 2, ..., N-L+1$ to be matched.

Step 4. The target perceptual hashing digest $H_K = \{H_{K1}, H_{K2}, ..., H_{KL}\}$ should be matched to the $N-L+1$ sets in the appropriate order. Before matching, the normalized Hamming distance of perceptual hashing digest between two syllables headers (such as H_{K1} and H_v^i) should be first calculated. Only they are matched successfully, which means their headers are similar, should the target perceptual hashing be matched with candidate perceptual hashing set.

For example, the header of $H_K = \{H_{K1}, H_{K2}, ..., H_{KL}\}$ is first matched to the header of set $H1^i = (H_1^i, H_2^i, ..., H_L^i)$, that is, H_{K1} and H_1^i . Their normalized Hamming distance $D(H_{K1}, H_1^i)$ is calculated according to the formula (2), only it is less than the similarity threshold, should the normalized Hamming distance between the first set and the target perceptual hashing$D(H_K, H1^i)$ be calculated by the formula (3). If $D(H_K, H_1^i) < T'$, then they are matched successfully. The location along with the speech segment Ai should be labeled as one of the retrieval result. Otherwise, the target perceptual hashing H_K is continued to be matched to the next set H_2^i using the same method. The illustration is given in Fig.3.

Step 5. After the completion of retrieval in the system hash table, all the detection results are obtained. In order to verify whether the encrypted speech is damaged, it need to extract the digital watermarks embed in the encrypted speech and match with the perceptual hashing digest of the query speech.

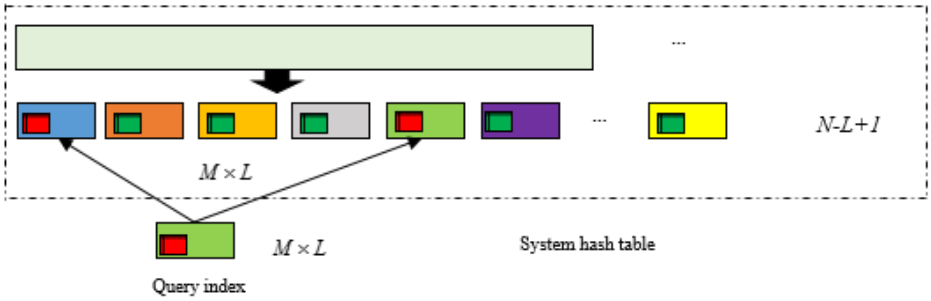


Fig. 3. Illustration of matching selection for spoken term retrieval

5. Experiments and analysis

5.1. Properties of speech perceptual hashing

In the proposed scheme, a binary sequence is used for representing perceptual hashing digest, which have a simple structure and few data. Generally, its mathematical distance is calculated by the normalized Hamming distance (also named BER). Calculating BER can determine whether two perceptual hashing digests represent the same speech. If their BER is less than the preset threshold, then they are deemed to be from the same audio contents. Otherwise, they are deemed to represent different speech contents. The most important properties of perceptual hashing are discrimination and perceptual robustness. In order to clearly describe the properties of perceptual hashing, false accept rate (FAR) is introduced, and it refers to the ratio of speech with different contents that are determined to be the same such that it is accepted by the system [17].

We performed properties of speech segments perceptual hashing experiments on a speech database containing 1,000 different speech segments from 863 Chinese continuous speech database (RASC863). These speech segments with sampling rate of 16 kHz and 16-bit quantization. The sampling signals are added Hamming windows. Each speech is divided into many frames with length of 256 sampling points, and the frame shift is half the length of the frame. Therefore, a frame equals 16 msec. The posterior probability based on acoustic segment models of speech segment is extracted by employing the method of [13]; meanwhile, syllables are obtained by utilizing the syllable segmentation algorithm [15]. The maximum length of syllables is 90 frames, that is $M = 90$, and $D = 64$ in posterior probability feature (there are 64 phonemes in Chinese). Therefore, the dimension of syllable-level perceptual hashing digest is $D \times M$. After that, the perceptual hashing digest of each speech segment is generated using the proposed method (Section 4.1). In order to obtain the statistical characteristics, we conducted a lot of matching calculation. By pairwise matching the generated perceptual hashing value (1,000 999 / 2 = 499,500 matching cases), the statistical results and its histogram of BERs is displayed in Fig.4. Obviously, it can be seen from the figure that the normalized Hamming distance distributes between 0.35 and 0.63, and the result can be approximately 03tted as the Gaussian distribution with the mathematical expectation $\mu = 0.4950$ and standard deviation $\sigma = 0.0352$. Therefore, based on such distribution parameters, the FAR under different thresholds τ (denoted as $R_{FAR}(\tau)$) can be calculated according to formula (4) for the perceptual hashing of speech segments.

$$R_{FAR}(\tau) = P(x < \tau) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{4}$$

Similarly, in the discrimination analysis of syllables perceptual hashing digests, 1,000 syllables are randomly selected from the syllable segmentation results of 1,000 speech segments. By pairwise matching the generated perceptual hashing digests (1,000 999 / 2 = 499,500 matching cases), the statistic results of BERs is obtained. It is found that the BERs distributes between 0.3128 and 0.65, and the results can be approximately fitted as Gaussian distribution with mathematical expectation $\mu = 0.4939$ and standard deviation $\sigma = 0.0463$. The FAR under different thresholds τ for the perceptual hashing of syllable also can be calculated according to formula (4) based on these distribution parameters.

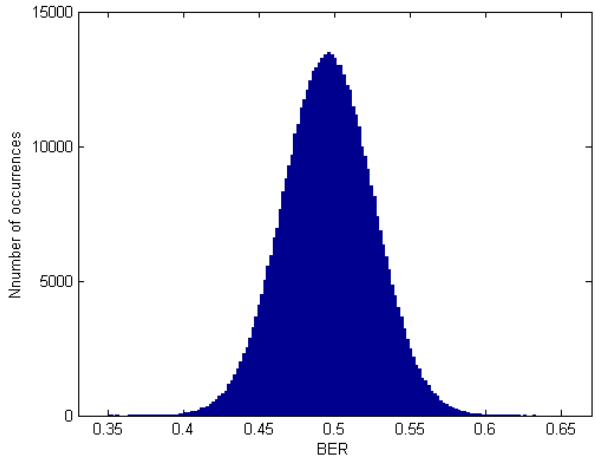


Fig. 4. Statistic histogram of 1,000 speech segments matching results

Given the threshold τ, the lower the value of $R_{FAR}(\tau)$, the better the discrimination of the perceptual hashing scheme. Employing the same speech segments data and comparing the proposed algorithm with those in references [10], [11] and [12], the FARs under different thresholds are calculated and displayed in Table 1. As seen from Table 1, $R_{FAR}(\tau)$ of the adopted scheme is lower than other three algorithms. Therefore, the perceptual hashing of the proposed method have the best properties of uniqueness and discrimination among four schemes. Similarly, according to the distribution parameters of the perceptual hashing of syllable, the FAR under different thresholds for four algorithms can be calculated (Table 2). Obviously, the adopted scheme outperforms other three algorithms in terms of the properties of uniqueness and discrimination. It is remarkable that the properties of uniqueness and discrimination for speech segments are much better than syllables, which result from the length of speech perceptual hashing.

Table 1. Comparison of $R_{FAR}(\tau)$ for the perceptual hashing of 1,000 speech segments

τ	ours	Ref [10]	Ref [11]	Ref [12]
0.02	8.44e-42	5.04e-16	4.18e-19	4.27e-26
0.04	1.60e-38	7.79e-15	1.14e-17	4.25e-24
0.06	2.20e-35	1.07e-13	2.70e-16	3.48e-22
0.08	2.20e-32	1.31e-12	5.59e-15	2.35e-20
0.10	1.59e-29	1.43e-11	1.00e-13	1.30e-18
0.12	8.40e-27	1.39e-10	1.56e-12	5.96e-17
0.14	3.21e-24	1.20e-09	2.12e-11	2.25e-15
0.16	8.90e-22	9.31e-09	2.51e-10	6.97e-14

Table 2. Comparison of $R_{FAR}(\tau)$ for the perceptual hashing of 1,000 syllables

τ	ours	Ref [10]	Ref [11]	Ref [12]
0.02	6.8798e-25	3.3761e-13	4.2121e-18	1.5238e-19
0.06	3.5758e-21	5.0217e-11	4.0171e-15	2.3963e-16
0.10	8.8816e-18	4.7821e-09	2.0809e-12	1.9666e-13
0.14	1.0559e-14	2.9156e-07	5.8549e-10	8.0422e-11

The perceptual robustness of perceptual hashing digest refers that the BER between original speech and the speech under different signal processing (such as noise reduction, compression, resampling, etc.) is less than the preset threshold. By employing four methods above and given the preset threshold 0.005, we used Cool Edit Pro v2.1, Gold Wave v5.68C, and MATLAB R2010b to process the 1,000 syllables, and the average BER between original speech and the speech under different signal processing were listed in Table 3, where the signal processing includes MP3 compression (128kbps), re-quantization (16→8→16bps), decreasing and increasing of amplitude (3dB). From the results listed in Table 3, it can be seen that the average BER of our method is less than the preset threshold under different speech signal processing, which indicates that the proposed perceptual hashing method has good perceptual robustness. Depending on the conclusion that the properties of uniqueness and discrimination for speech segments are much better than syllables, we can infer that the perceptual robustness of speech segments outperforms syllables.

5.2. Performance of speech retrieval

In this paper, we use the recall ratio R and the precision ratio P to evaluate the retrieval performance. In formulas (5) and (6), f_T denotes the number of correct search in the encrypted speech database, f_F denotes error number, and f_L denotes lost number. In the experiments, we generated encrypted speech data with watermarks by employing the undetermined blind source separation-based speech encryption algorithm and the perceptual hashing digests of 1,000 speech segments using the proposed method (they were embedded into the encrypted speech as watermarks). The system hash table was formed by the perceptual hashing digests of 1,000 speech segments. In the process of searching and matching in the system hash table, given the similarity threshold $T', 0 < T' < 0.5$, if the normalized Hamming distance $D(H_Q, H_S) < T'$, the matching succeeds. Obviously, the recall ratio and precision ratio are directly affected by the similarity threshold. In previous experimental results of 1,000 syllables discrimination test, the minimum BER is 0.3128, and in their perceptual robustness test, the maximum BER is 0.0097. Therefore, we chose the similarity threshold T' as 0.25 for avoiding missed detection and achieving a high precision ratio.

$$R = \frac{f_T}{f_T + f_L} \times 100\% \tag{5}$$

$$P = \frac{f_T}{f_T + f_F} \times 100\% \tag{6}$$

Table 3. Comparison of perceptual robustness for 1,000 syllables

Various signal processing	the average BER			
	Ours	Ref [10]	Ref [11]	Ref [12]
MP3	0.0016	0.0093	0.0067	0.0038
Re-quantization	0.0026	0.1895	0.0959	0.0693
Amplitude decrease	0.0042	0.0246	0.0157	0.0139
Amplitude increase	0.0039	0.0498	0.0557	0.0476

100 spoken terms were randomly chosen from 1,000 speech segments as the query speech. Considering that the query fed by user may be corrupted by noise, compression et.al, we processed the query with noise reduction, MP3 compression and re-quantization operation before retrieving. After retrieving in the system hashing table, the recall ratio and precision ratio under different signal processing were shown in Table 4. As can be seen from the table, the number of correct search of the proposed method is high under various signal processing operations, whereas the error number and lost number are small. Obviously, the proposed method have good retrieval performance under various signal processing.

Table 4. Recall and precision ratios under different signal processing

Operation	Noise reduction	MP3	Re-quantization
f_T	98	96	96
f_F	3	4	3
f_L	2	4	2
R	98%	96%	98%
P	97%	96%	97%

Besides, the speech segment retrieval experiments were conducted as well. All the query speech segments were processed by signal processing (shown in Table.2) before retrieving, then their perceptual hashing searched in the system hash table. Take the 600^{th} speech segment for example, it was selected as the retrieval speech and processed by re-quantization (16→8→16bps) operation. The BERs between the perceptual hashing digest of query speech and each perceptual hashing of system hash table were calculated and shown in Fig.5. As can be seen from the figure, apart from the BER between the query speech and the 600^{th} speech segment, all BERs were larger than 0.3. Given the similarity threshold 0.25, only when the BER is less than it the matching succeeds. Additionally, the experimental results of speech segments retrieval were summarized, and it is found that the proposed scheme reached 100% in terms of recall and precision ratios under various signal processing.

Different from references [10], [11] and [12], whose perceptual hashing digest of fixed length was generated for each speech segment, the perceptual hashing sequence of different length will be generated for each speech segment in the proposed algorithm. Supposing the fixed length is 500 frames in references [10], [11] and [12], the retrieval time of query speech segment by employing the adopted algorithm and methods of references

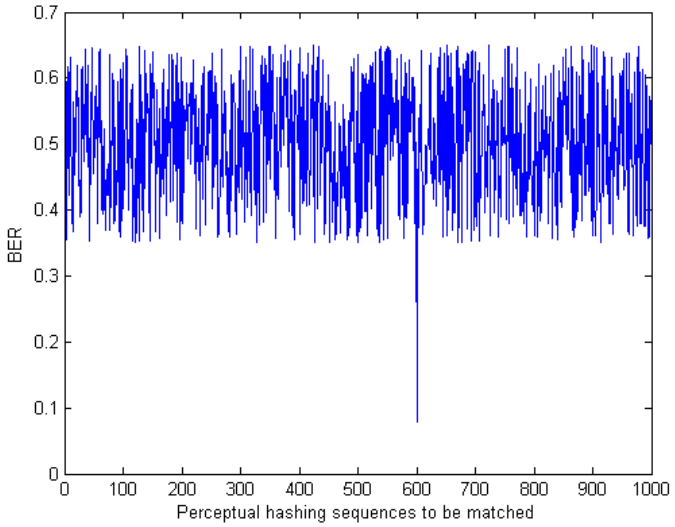


Fig. 5. Matching result of speech segment in system hash table

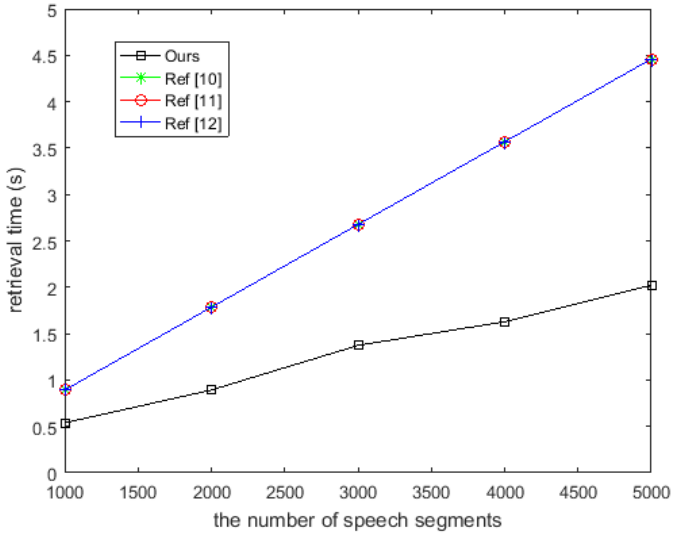


Fig. 6. Comparison of retrieval time for speech segment

[10], [11] and [12] were recorded and displayed in Fig.6. As can be seen from the figure, fixing the number of encrypted speech segments, references [10], [11] and [12] have the same detection time, which due to their same fixed length of perceptual hashing. Compared with them, the proposed method has an advantage in terms of the retrieval speed.

Obviously, as the number of encrypted speech segments increases, the detection time of the proposed algorithm is growing slowly; accordingly, the detection times of references [10], [11] and [12] are growing linearly. This is because that references [10], [11] and [12] need matching all perceptual hashing digests successively in the system hash table; by contrast, the target perceptual hashing only requires matching the perceptual hashing digests of equal length and similar header in proposed method, in other words, if the number of syllables in speech segment is different from that of the query speech, or the number of syllables is the same but without the similar header, then its perceptual hashing sequence does not need to be matched. In this way, the proposed algorithm reduces the matching number of retrieval and improves the retrieval efficiency.

6. Conclusion

Most of the existing retrieval algorithms based on perceptual hashing only can search the speech segments over encrypted speech data, and their retrieval times increase linearly along with the number of encrypted speech segments. If extended them for detecting spoken term, the properties of their perceptual hashing were bad. For the purpose of achieving spoken term retrieval in an encrypted speech database, and further improving the discrimination, uniqueness and perceptual robustness, this study proposes a syllable-level perceptual hashing-based retrieval method. Different from the existing methods, the posterior probability features based on acoustic segment models of syllable are used to generate a perceptual hashing sequence, which is then embedded into encrypted speech as a digital watermark. The perceptual hashing values of syllables obtained from the continuous speech data are constituted the perceptual hashing digest of each speech sentence, and the system hash table is composed of the perceptual hashing sequences of all the speech sentences. Without retrieving the encrypted speech directly or decryption, spoken term retrieval over encrypted speech can implement successfully. In general, the proposed method has three obvious advantages. Firstly, the syllable-level perceptual hashing derived from the posterior probability features based on acoustic segment models show better distinctiveness and robustness than them derived from the time and frequency domain features, which reduces the chance of hash collision & two segments generating the same perceptual hashing values. Moreover, it implements the function of retrieving spoken term over encrypted speech, and effectively improves the retrieval speed by reducing the matching number of query index. Finally, it achieves high recall and precision ratios under various signal processing.

Acknowledgments. This work was supported by the National Natural Science Funds of China (No. 61173106), Key Project Fund of Science and Technology Program of Changsha (No.K1403027-11).

References

1. Yuan X, Wang X, Wang C, et al. Enabling Secure and Fast Indexing for Privacy-Assured Healthcare Monitoring via Compressive Sensing. IEEE Transactions on Multimedia. Vol. 18, 2002 C 2014. (2016)

2. Karan N, Pranav P, Rajesh M. Group Delay Based Methods for Speaker Segregation and its Application in Multimedia Information Retrieval. IEEE Transactions on Multimedia. Vol. 15, 1326 C 1339. (2013)
3. Qin Liu, Guojun Wang , JieWu, Secure and privacy preserving keyword searching for cloud storage services, Journal of Network and Computer Applications, Vol. 35, 927C933. (2013)
4. Zhangjie Fu, Xingming Sun, Zhihua Xia, Lu Zhou, Jiangang Shu, Multi-keyword Ranked Search Supporting Synonym Query over Encrypted Data in Cloud Computing, Proceedings of IEEE. (2013)
5. Baojiang Cui, Zheli Liu and Lingyu Wang, Key-Aggregate Searchable Encryption (KASE) For Group Data Sharing via Cloud Storage, IEEE Transactions On Computers, Vol. 6, No. 1, 1-13. (2014)
6. Zhangjie Fu, Xingming Sun, Nigel Linge, Lu Zhou, Achieving Effective Cloud Search Services: Multikeyword Ranked Search over Encrypted Cloud Data Supporting Synonym Query, IEEE Transactions on Consumer Electronics, Vol. 60, No. 1, 164-172. (2014)
7. Jin Li, Jingwei Li, Xiaofeng Chen, Chunfu Jia, and Wenjing Lou, Identity-Based Encryption with Outsourced Revocation in Cloud Computing, IEEE Transactions On Computers, Vol. 64, No. 2, 425-4371-13. (2015)
8. Rupali D. Korde, Dr. V.M. Thakare. Secure multiple data retrieval over encrypted cloud data. International Journal of Research in Science & Engineering, 330-334. (2016)
9. Kalker T, Haitsma J, Oostveen J C, et al. Issues with digital watermarking and perceptual hashing. Proc SPIE, 189-197. (2001)
10. Wang H, Zhou L, Zhang W, Liu S. Watermarking-based Perceptual Hashing Search over Encrypted Speech. 12th International Workshop on Digital-Forensics and Watermarking (IWDW 2013), Auckland, New Zealand, 1-12. (2013)
11. Hao G Y, Wang H X. Perceptual Speech Hashing Algorithm Based on Time and Frequency Domain Change Characteristics. Symposium on Information, Electronics, and Control Technologies. (2015)
12. Zhao H, He S F. A retrieval algorithm for encrypted speech based on speech perceptual hashing. 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2016), Changsha, China, 1840-1845. (2016)
13. Chan C, Lee L. Model-Based Unsupervised Spoken Term Detection with Spoken Queries. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, No. 7, 1330-1342. (2013)
14. Zhao H, He S F, Chen Z, et al. Dual Key Speech Encryption Algorithm based on Underdetermined BSS. Scientific World Journal, Vol. 1, 57-78. (2014)
15. Andreas S, Neville R, Vikramjit M, et al. Highly accurate phonetic segmentation using boundary correction models and system fusion. IEEE International Conference on Acoustics, Speech & Signal Processing, 5552-5556. (2014)
16. Zhao H, He S F. Analysis of Speech Signals Characteristics based on MF-DFA with Moving Overlapping Windows. Physica A. Vol. 442, 343-349. (2016)
17. Whitman M, Mattord H. Principles of Information Security. Beijing: Tsinghua University Press, 252. (2006)

Shaofang He received her B.Sc. degree in Mathematics and Applied Mathematics and M.S. degree in computational mathematics at Hunan normal University in 2003 and 2006, respectively. Currently, she has received her Ph.D. in Computer Science and Technology of Hunan University. Her current research interests include speech information processing and information security.

Huan Zhao is a professor at the School of Information Science and Engineering, Hunan University. She obtained her B.Sc. degree and M.S. degree in Computer Application Technology at Hunan University in 1989 and 2004, respectively, and completed her Ph.D. in Computer Science and Technology at the same school in 2010. Her current research interests include speech information processing, embedded system design and embedded speech recognition. Prof. Zhao is a Senior Member of China Computer Federation, Governing of Hunan Computer Society, China and China Education Ministry Steering Committee Member of Computer Education on Arts. She has published more than 40 papers and 6 books.

Received: January 12, 2017; Accepted: July 1, 2017.

Computer Science and Information Systems 14(3):719–734 https://doi.org/10.2298/CSIS161220025J

A Novel Link Quality Prediction Algorithm for Wireless Sensor Networks

Chenhao Jia[1], Linlan Liu[1], Xiaole Gu[1], and Manlan Liu[1]

Internet of Things Technology Institute, Nanchang Hangkong University,
330063 Nanchang, China
1127792870@qq.com
liulinlan@nchu.edu.cn
1030697096@qq.com
694531775@qq.com

Abstract. Ahead knowledge of link quality can reduce the energy consumption of wireless sensor networks. In this paper, we propose a cloud reasoning-based link quality prediction algorithm for wireless sensor networks. A large number of link quality samples are collected from different scenarios, and their RSSI, LQI, SNR and PRR parameters are classified by a self-adaptive Gaussian cloud transformation algorithm. Taking the limitation of nodes' resources into consideration, the Apriori algorithm is applied to determine association rules between physical layer and link layer parameters. A cloud reasoning algorithm that considers both short- and long-term time dimensions and current and historical cloud models is then proposed to predict link quality. Compared with the existing window mean exponentially weighted method, the proposed algorithm captures link changes more accurately, facilitating more stable prediction of link quality.

Keywords: wireless sensor networks, link quality prediction, Gaussian cloud transformation.

1. Introduction

Wireless sensor networks (WSNs) is multi-hop, self-organising network formed by a large number of inexpensive micro-sensor nodes which communicate with each other by radio [6]. Their purpose is to collaboratively sense, collect and process information about objects in the network coverage area and pass it on to an observer. As they are usually required to work for long periods, it is important to reduce the nodes' energy consumption. Because of the physical characteristics of the node and the volatile communication environment, the nodes are affected by multipath effects, signal attenuation and signal interference from other wireless communication protocols (Wi-Fi, GSM, Bluetooth). These uncertain spatial and temporal characteristics of data transmission present challenges for the evaluation and prediction of wireless link quality.

Link quality prediction (LQP) plays a fundamental role in WSNs routing protocols, topology control, and energy management, and so on. For instance, an effective mechanism for link quality prediction can help routing protocols choose better link for data transmission, reduce data retransmission requirements and the number of routings, and improve network throughput and the reliability of data transmission. The topology control mechanism in WSNs relies on link quality to eliminate unnecessary links and improve the stability of the network, which is beneficial for prolonging the lifetime of the

whole network. In WSNs energy management applications, LQP can predict changes of the current link, reduce node energy consumption, and improve the efficiency of network communication by selecting the appropriate transmission power.

This paper proposes a cloud reasoning-based link quality prediction algorithm based on multiple parameters, which classifies link quality parameters according to the cloud model. This algorithm overcomes the subjectivity of link quality classification, as different link quality parameters can represent different aspects of link quality. The algorithm, named Apriori, mines association rules between physical parameters and link layer parameter. In order to validate the algorithm, it is compared with the smoothed packet received ratio (SPRR) prediction method using a testbed platform. The results show that the proposed algorithm provides more stable prediction of link quality changes.

This paper makes two main contributions: 1) it uses a cloud model to eliminate subjective factors in the classification of link quality; 2) it proposes a cloud reasoning-based link quality prediction algorithm for wireless sensor networks. In This paper, we select PRR as an indicator of link quality and use cloud model to eliminate subjective factors in the classification of link quality. Based on cloud reasoning process, a novel link quality prediction algorithm is proposed for WSNs. Section 2 analyses the related researches of link quality prediction. Section 3 proposes the key algorithms in predicting the link quality at the next moment. Section 4 describes the experimental scenarios and experimental results to verify the effectiveness of the algorithm. Section 5 makes conclusions.

2. Related Work

Methods of WSNs link quality prediction fall into three categories-based on communication link characteristics, probability estimation and intelligent learning. The physical layer parameters involved in prediction include the received signal strength indication (RSSI), link quality indicator (LQI) and signal-to-noise ratio (SNR). Link layer parameters involved in the prediction include the packet reception ratio (PRR). The prediction methods based on probability estimation predict successful reception probability of the future packets. The prediction methods based on intelligent learning are related to pattern recognition, Bayesian networks, support vector machines (SVM) and so on.

Paper [6] proved the occurrence of SNR patterns resulted by the joint effect of human motion and radio propagation, then it used the cross-correlation to predict (XCoPred) algorithm to predict link quality variation. Paper [18] proposed a link quality prediction model based on supervised learning. It used machine learning to automatically discover correlations between readily-available features and the quality of interest. Paper [12] presented a machine learning based algorithm to link availability prediction in low power WSNs routing. The results of experiments showed that the algorithm has accurate predicting availability of intermediate quality links. Paper [8] proposed 4C based on machine learning, taking the physical parameters and link layer parameters as the input set. It took advantage of a Bayesian classifier, logistic regression and artificial neural network (ANN) to predict link quality, then predicted successful reception probability of the next packet. It had higher precision but lower sensitivity. Paper [9] proposed the temporal adaptive link estimator with no off-line training (TALENT) to predict short-term fluctuation in transition area links. Without prior knowledge and intervention, the algorithm can achieve rapid adaptation to network conditions.

Due to the resource consumption problem caused by asymmetric links, paper [13] applied dual-tree topology to record the receiving and the transmitting link quality parameters of the two nodes. The experiment showed that this algorithm can effectively reduce the hop from the source node to destination node. Paper [11] proposed a built-in learning-based WSNs link quality estimation algorithm that input link quality parameters (RSSI, SINR and PRR), node energy information, the expected number of transmissions (ETX), the expected energy consumption (EEC) and the expected number of retransmissions (ENR) to predict link quality. This method can improve the transmission rate but has poor real-time performance. Paper [3] proposed a link quality prediction algorithm based on fuzzy comprehensive theory and the Bayesian network. While it solved the marginalization problem of link quality, the algorithm required parameters that are independent of each other, and lacked assessment of link quality variability. Paper [10] proposed a generic link quality evaluation framework based on machine learning and was verified by LQI. The experimental results showed that the method is accurate for evaluating stable links, but not fluctuating ones.

Paper [20] applied a neighbourhood-based non-negative matrix factorization algorithm to predict link quality in WSNs. It learned latent features of the nodes from the information of past data transmissions combing with local neighborhood structures. Paper [2] fully considered the reliability, volatility and asymmetry of link and channel quality to construct corresponding link indicators. It applied the theory of fuzzy to obtain a fuzzy link quality estimator (F-LQE). Paper [15] compared reliability, stability and agility performance of ETX, 4-Bit and F-LQE. It noted that F-LQE is more reliable and stable, but has less agility for smart grid environment monitoring. Paper [16] combined a fuzzy theory and SVM to put forward a link quality prediction model that can reduce the effects of noise and outliers. It used a chaotic particle swarm optimisation algorithm to optimise the parameters of the SVM model. The model had better prediction performance than a backward propagation (BP) neural network. Paper [5] applied a Markov chain model to describe packet loss in wireless links, pointing out that link package rates tended to be the same over very short periods of time. Paper [19] proposed a window mean exponentially weighted moving average (WMEWMA) link quality evaluation algorithm based on exponentially weighted moving average (EWMA) filter. This algorithm can predict the next-moment PRR through a historical set of PRRs, but the lack of physical parameters can lead to lower accuracy.

Cloud models are widely applied for evaluation, prediction and algorithm improvement [14] because of their advantages of fuzzy and stochastic processing. Considering the randomness of links and the fuzziness of link quality in WSNs, this paper proposes a novel link quality prediction algorithm. The self-adaption Gaussian transformation algorithm is used to classify link quality parameters. Association rules between physical layer parameters and link layer parameters are mined by the Apriori algorithm [1]. A novel cloud reasoning-based link quality prediction algorithm is proposed by combining historical cloud and current cloud information.

3. Cloud Reasoning-based Link Quality Prediction Algorithm

The cloud reasoning-based link quality prediction algorithm can be divided into three parts: division of link quality parameters, determination of association rules and estab-

lishment of a prediction model. The self-adaptive Gaussian transformation (S_GCT) algorithm is used to classify link quality parameters. The Apriori algorithm is applied to find out rules between the physical-layer parameters (RSSI, SNR, LQI) and the link-layer parameter PRR. A cloud reasoning-based algorithm is proposed to predict link quality. Fig. 1 illustrates a flowchart of the prediction model.

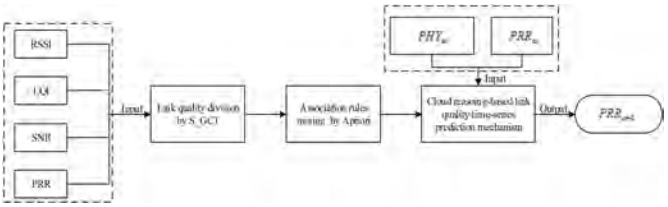


Fig. 1. Cloud reasoning-based link quality prediction algorithm

3.1. Division of Link Quality Parameters

The S_GCT algorithm automatically forms multiple concepts consistent with human cognition, and appropriate granularity based on the statistical distribution of the actual data. This process can reflect the process of human cognition from low to high levels. The self-adaptive Gaussian transformation is iteratively convergent by calling the heuristic Gauss cloud transformation algorithm (H_GCT), and the Gauss cloud transformation strategy is formulated according to the confusion degree (CD). The confusion degree can be used to characterise the degree of dispersion of the Gauss cloud distribution. Larger CDs usually involve greater overlap of adjacent Gauss clouds, hence, concepts are poorly defined. Conversely, smaller CDs involve less overlap, so that concepts easily reach consensus. The threshold of the general CD is 0.5. The CD can be calculated according to Equation (1).

$$CD = \frac{3He}{En} \tag{1}$$

Where *He* represents the hyper-entropy of the cloud model, and *En* represents the entropy of the cloud model.

The S_GCT algorithm specifies the Gauss cloud transformation strategy according to the CD, and the basic area between adjacent concepts will not overlap (CD $\leq$ 0.5) through iteration. The self-adaptive Gaussian transformation algorithm is set out as Algorithm 1.

The data sets of the different parameters can be processed by S_GCT, and the different link quality parameter can be divided into different levels. In this paper, parameters SNR LQI, RSSI and PRR are divided into various levels, such as (SNR1, SNR2, SNR3, SNR4, SNR5), (LQI1, LQI2, LQI3, LQI4, LQI5, LQI6, LQI7), (RSSI1, RSSI2, RSSI3, RSSI4) and (PRR1, PRR2, PRR3, PRR4, PRR5, PRR6, PRR7, PRR8, PRR9, PRR10).

Algorithm 1 S_GCT(p_Eve,β)

Input: p_Eve: sample set for every parameter, β: confusion degree
Output: p_Clu: list of M cloud model parameters

initial M, $M>0$
repeat
    compute Gauss cloud digital features $H\_GCT(p\_Eve, M)$
    add Gauss cloud digital features to the list $p\_Clu.Add(Ex, En, He, CD)$
until p_Eve is empty
repeat
    clear list $p\_Clu.Clear()$
    the Gauss component is reduced by 1 $M = M - 1$
    repeat
        compute Gauss cloud digital features $H\_GCT(p\_Eve, M)$
        add Gauss cloud digital features to the list $p\_Clu.Add(Ex, En, He, CD)$
    until p_Eve is empty
    if $M == 1$ **then** close
    end if
until $CD_k<\beta$

3.2. Association Rules Mining

The link quality prediction algorithm based on cloud reasoning is essentially a kind of regression reasoning prediction. In this paper, the Apriori algorithm is applied to mine the association rules between the physical parameters and link layer parameter.

Firstly, we save the divided data in the database D. Secondly, we find out rules with the Apriori algorithm, like {RSSI1, LQI2, SNR1, PRR1}. The rule is defined as "If RSSI low and LQI low and SNR low then PRR low". More rules will be shown in Table 1.

Table 1. Partial association rules

Rule antecedent			Rule consequent
RSSI1	SNR2	LQI1	PRR1
RSSI1	SNR2	LQI2	PRR1
⋮	⋮	⋮	⋮
RSSI2	SNR4	LQI5	PRR10
RSSI3	SNR5	LQI7	PRR10
RSSI4	SNR6	LQI7	PRR10
RSSI6	SNR7	LQI7	PRR10

3.3. Link Quality Prediction with Time Series

In this paper, a cloud reasoning-based link quality prediction algorithm is used to predict link quality by considering short and long term dimensions. The short term dimension generates the current cloud through the current physical parameters, while the long term dimension generates the historical cloud through the link layer parameter. The current cloud and the historical cloud are integrated into an integrated cloud model which is used to predict link quality [4].

For the current physical layer parameters, we determine the maximum membership degree of association rules by the three condition single-rule cloud generator algorithm (3CSR_CG). Corresponding to the link layer parameters in the maximum membership degree rule, the cloud model is selected as the current cloud. The 3CSR_CG algorithm is shown as Algorithm 2.

Algorithm 2 3CSR_CG(X_1,X_2,X_3,Y,x)

Input: X_1: the first front cloud model parameters, X_2: the second front cloud model parameters, X_3: the third front cloud model parameters, Y: the back cloud model parameters, x: the specific vector

Output: u: membership degree, b: the drop

computer the first random variable $Esx_1 = NORM(Enx_1, pow(Hex_1, 2))$
computer the second random variable $Esx_2 = NORM(Enx_2, pow(Hex_2, 2))$
computer the third random variable $Esx_3 = NORM(Enx_3, pow(Hex_3, 2))$
computer the membership of x $u = membership(x, Esx_1, Esx_2, Esx_3)$
computer the back cloud random variable $Esy = NORM(Eny, pow(Hey, 2))$
if up **then**
    $b = Ey - deviate(u, Esy)$
    close
end if
$b = Ey + deviate(u, Esy)$

We select *N* latest PRR values as drops, and the cloud model is identified as a historical cloud by using the no-degree backward Gaussian cloud algorithm (NB_GCT), shown as Algorithm 3.

Algorithm 3 NB_GCT(p_Eve)

Input: p_Eve: sample set

Output: p_Clu: cloud model parameters

computer sample set mean $Ex = MEAN(p\_Eve)$
computer first order absolute central moment $C = FirstMoment(p\_Eve)$
computer second order absolute central moment $S = SecondMoment(p\_Eve)$
computer entropy of sample set $En = sqrt(pi/2) \times C$
computer hyper entropy of sample set $He = sqrt(S - pow(En, 2))$
compose Gauss cloud digital features $p\_Clu = (Ex, En, He)$

A integrated cloud can be obtained by combining current cloud with historical cloud. Then a set of PRR value can be obtained by using the forward Gaussian cloud algorithm (F_GCT) shown as Algorithm 4.

Algorithm 4 F_GCT(p_Clu,N)

Input: p_Clu: cloud model parameters, N: the number of drop
Output: V: list of drop and membership degree

initial n, $n = 0$
repeat
    computer the random variable of expected value $Es = NORM(En, He)$
    computer the random variable of drop $s = NORM(Ex, Es)$
    computer membership of drop $y = membership(s, p\_Clu)$
    add drop and membership to list $p\_Clu.add(s, y)$
    n increased by 1 $n = n + 1$
until $n \geq N$

The cloud reasoning-based link quality prediction algorithm sets Input_N as the input vector: Input_N=[PRR_N,PHY_N], where PRR_N is the historical PRR value of the latest N moments before the prediction point. Other than PRR_N, which is historical, PHY_N consists of values of RSSI, LQI and SNR at the present time. Input_N is inputed into the prediction algorithm and the drops can be output from Algorithm 2 and Algorithm 4.

4. Experiments and Analysis

The testbed consists of a single-hop network with four TelosB TX (transmitter) nodes and one TelosB RX (receiver) node positioned in outdoor and indoor environments, respectively. These nodes are equipped with a CC2420 wireless transceiver chip designed according to IEEE802.15.4. The RX node is connected to a computer via a USB port. We developed link quality test platform to monitor and analyse the experimental results (Fig. 2). The software tools we used include MATLAB and SPSS. Experimental parameters are shown in Table 2.

4.1. Design of Experimental Scenes and Data Collection

In WSNs, sensing data is easily affected by environmental noise, channel interference and multipath propagation in the process of transmission. In order to guarantee the diversity of the data samples and consider the influence of various interference sources, the experiments are conducted in three scenarios, such as a corridor in a building, a forest on the university campus and a road (Fig. 3). In each scenario, a miniature star WSNs network is deployed to test the link quality.

The corridor scene is used mainly to simulate indoor smart home situation of WSNs. Paper [17] showed that if a Wi-Fi signal is in a WSN area, the WSNs signal will be greatly affected. So indoor scenes are important in assessing WSNs link quality. The campus forest scene is used to simulate applications in field environments where signals are mostly

Table 2. Testbed parameter settings

Parameter	Value
Transmit	0 dBm
Channel	26
Modulation mode	DSSS-O-QPS
Transmission speed	250kbits/s
Number of packets	30
Packet sending interval	0.2 Second
Test cycle	10 Senconds
Number of nodes	4
Location	East,West,South,North

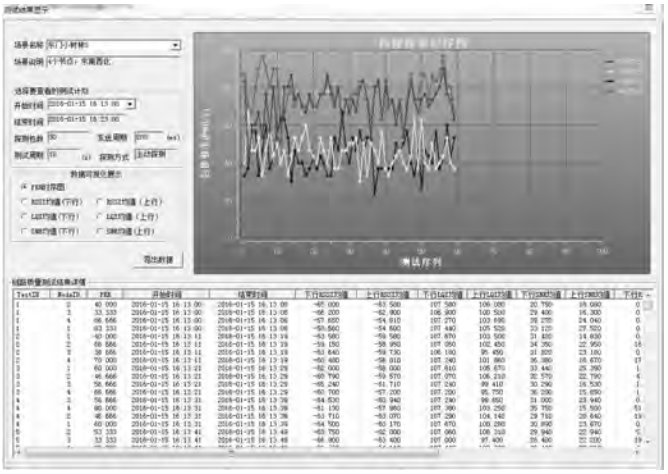
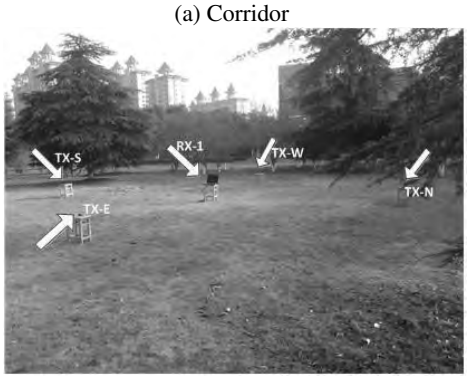


Fig. 2. Link quality testbed platform

(a) Corridor



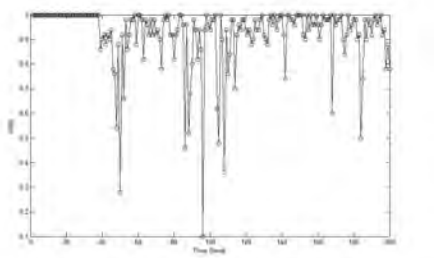
(b) Campus forest

(c) Road

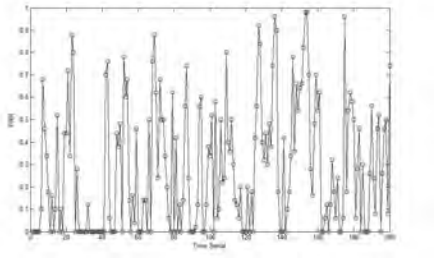
Fig. 3. Experimental scenes

affected by multipath propagation caused by obstacles. The road scene simulates intelligent transportation applications. Here, interference mainly comes from environmental noise, which subjects radio link signals to reflection, refraction and diffraction, etc..

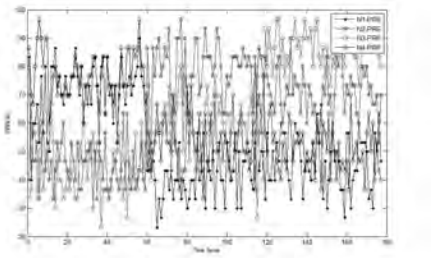
Experimental data are collected from all scenes, and PRR timing diagrams are drawn using MATLAB to facilitate analysis of the WSNs link characteristics (Fig. 4).



(a) Corridor



(b) Road



(c) Campus forest

Fig. 4. The PRR in different scenes

In Fig. 4(a), the time period from samples 30 to 200 is the students' lunchtime. Here, the link quality is largely impacted by mobile phones, other wireless devices and random walking. Fig. 4(b) shows the link PRR to fluctuate violently. The link communication state is very unstable with large burst. Random changes in the number of vehicles on the road, interference and wireless equipment inside vehicles [7], etc., have a direct impact on changes of PRR. The results show that the road scene has relatively large burst, instability and volatility of link quality. In Fig. 4(c), N1-PRR, N2-PRR, N3-PRR, N4-PRR are the

four of the PRR timing diagram. Compared with the road scene, the interference in the campus forest is static, and the sources are mainly multipath propagation caused by trees, stones and other obstacles. Therefore, the volatility of the link quality is relatively large, but the communication link is relatively stable with relatively small burst.

4.2. Experimental Results of Link Quality Parameters Classification

The S_GCT algorithm is used to classify link quality parameters on the basis of different link quality parameter data sets. The corresponding Gauss curves and digital features are shown in Fig. 5 and Table 3 to Table 6. As we can see, different link quality parameter data sets result in different levels.

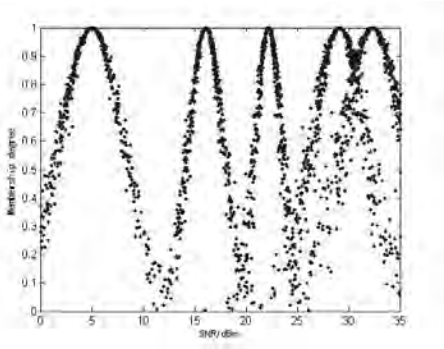
Table 3. Digital features of the cloud classification of SNR

Concept	Expectation	Entropy	Hyper-entropy	CD
Lower	5.0	3.10	0.30	0.29
Low	16.1	1.70	0.20	0.35
Medium	22.2	1.30	0.20	0.46
High	29.1	2.70	0.40	0.44
Higher	32.3	3.0	0.50	0.50

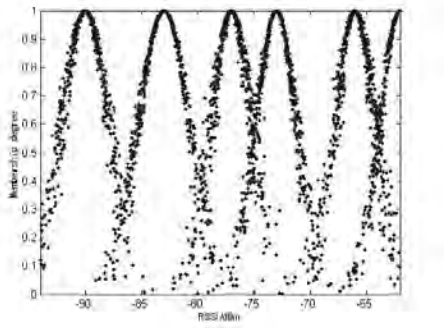
As shown in Table 3, SNR finally generates 5 concepts. The distribution of the first four concepts' expectation is relatively dispersed with less confusion degree. The fifth concept has highest confusion degree and its expectation is close to that of the fourth concept. Meanwhile, it can be inferred that SNR is not particularly good at distinguishing very good links from very good links.

Table 4. Digital features of the cloud classification of RSSI

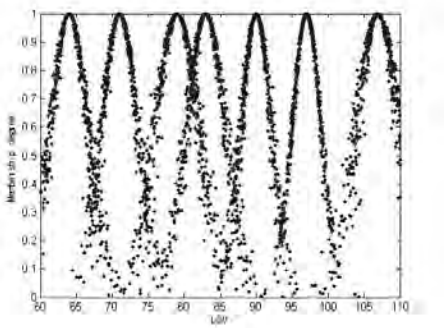
Concept	Expectation	Entropy	Hyper-entropy	CD
Lower	-90	2.10	0.30	0.43
Low	-83	2.30	0.20	0.26
General	-77	1.80	0.30	0.50
Medium	-73	1.80	0.20	0.33
High	-66	1.80	0.30	0.50
Higher	-62	1.80	0.18	0.30



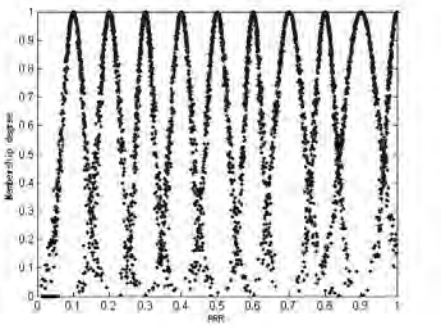
(a) SNR Gauss curve



(b) RSSI Gauss curve



(c) LQI Gauss curve



(d) PRR Gauss curve

Fig. 5. Gauss curves

As shown in Table 4, RSSI finally generates 6 concepts. The distribution of the all concepts' expectation is relatively uniform. When RSSI is low, its confusion degree is relatively small, which shows that RSSI has better ability to distinguish the lower links from the low links. Compared with SNR and LQI, the entropy and hyper-entropy of RSSI is small, which shows that RSSI has better ability to distinguish link quality than SNR and LQI.

Table 5. Digital features of the cloud classification of LQI

Concept	Expectation	Entropy	Hyper-entropy	CD
Very low	64	2.6	0.30	0.35
Lower	71	2.5	0.30	0.36
Low	79	3.0	0.40	0.40
General	83	2.7	0.35	0.39
High	90	2.1	0.25	0.36
Higher	97	2.0	0.15	0.23
Very high	106.9	3.0	0.50	0.50

As shown in Table 5, LQI finally generates 7 concepts. The confusion degree of the all concepts expectation is relatively high. While LQI has a smaller granularity, the expectation of low concept and general concept are close, and LQI is hard to distinguish the lower links from general links.

Table 6. Digital features of cloud classification of PRR

Concept	Expectation	Entropy	Hyper-entropy	CD
Extremely low	0.1	0.030	0.003	0.29
Very low	0.2	0.030	0.040	0.42
Lower	0.3	0.025	0.035	0.42
Low	0.4	0.030	0.003	0.30
General	0.5	0.030	0.003	0.30
Medium	0.6	0.025	0.003	0.36
High	0.7	0.040	0.003	0.26
Higher	0.8	0.030	0.003	0.30
Very high	0.9	0.050	0.003	0.18
Extremely high	1.0	0.030	0.003	0.30

As shown in Table 6, PRR finally generates 10 concepts. The distribution of the all concepts' expectation is uniform very well. The entropy of all concepts is almost same, and the CDs of all concepts are lower, which indicate PRR has good ability to distinguish the link quality.

Compared with SNR, RSSI and LQI, the entropy and the hyper-entropy of PRR are the lowest, which means link layer parameters of WSNs are the best indicator of link quality.

4.3. The Experimental Result

According to the current trend of the link, the cloud reasoning-based link quality prediction algorithm we proposed is practical and effective. In order to verify the effectiveness of the prediction algorithm, we select two 50 groups PRR prediction obtained by SPRR and the proposed prediction algorithm respectively. The SPRR is based on the principle of WMEWMA. Int this experiment, the window size to 5, the factor is set to 0.5, the value of the previous historical time is set to 5. The experimental results are shown in Fig. 6.

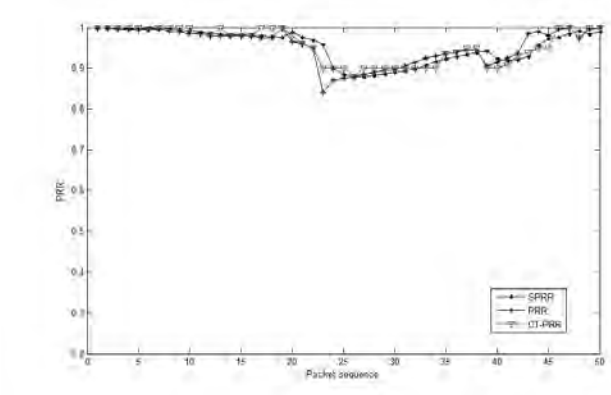


Fig. 6. Comparison of SPRR, PRR and CT-PRR

In Fig. 6, M-PRR represents the experimental measurement values of PRR, and SPRR represents the PRR prediction values based on the SPRR prediction algorithm. CT-PRR represents the PRR prediction values based on cloud reasoning-based link quality prediction algorithm. The experimental results show that the two prediction algorithm have little difference in stability. For the prediction of sensitivity, the CT-PRR values are relatively good, such as predicted values in sample interval 22 to 24 and sample interval 35 to 40. Because the cloud model of the time series takes into account both short- and long-term time dimensions, CT-PRR relative to SPRR can predict the dynamic link more effectively.

5. Conclusions

This paper proposes a method based on intelligent learning for the link quality prediction of WSNs. It introduces link quality parameters classification and prediction methods using

a Gaussian cloud transform algorithm. By establishing association rules between physical layer parameters (RSSI, LQI, SNR) and the link layer parameter PRR, we can use LQI, SNR and RSSI information to predict the future PRR values. Compared with the existing method based on WMEWMA, the experimental results demonstrate that the proposed algorithm more accurately captures link changes, leading to more stable predictions of link quality.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China (Grant No. 61363015, 61762065, 61501218, 61262020), the Jiangxi Natural Science Foundation of China (Grant No. 20171ACB20018, 20171BAB202009, 20171BBH80022), the Key Research Foundation of Education Bureau of Jiangxi Province (Grant No. GJJ150702), and the Innovation Foundation for Postgraduate Student of Jiangxi Province(Grant No. YC2016-S356).

Grateful thanks are due to the participants of the survey for their invaluable help in this study.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: the 20th International Conference on Very Large Data Bases (VLDB), San Francisco, USA, pp. 487–499 (1994)
2. Baccour, N., Kouba, A., Youssef, H., Jama, M.B., Rosrio, D.D., Alves, M., Becker, L.B.: F-lqe: A fuzzy link quality estimator for wireless sensor networks. In: the 7th European Conference on Wireless Sensor Networks (EWSN), Coimbra, Portugal, pp. 240–255 (2010)
3. Guo, Z.Q., Wang, Q., Wan, Y.D., Li, M.H.: A classification prediction mechanism based on comprehensive assessment for wireless link quality. Journal of Computer Research and Development 50(6), pp. 1227–1238 (2013)
4. Jiang, R., Li, D.Y., Chen, H.: Time-series prediction with cloud models in DMDK. Journal of PLA University of Science and Technology 1574(2), pp. 525–530 (2000)
5. Keshavarzian, A., Uysal-Biyikoglu, E., Lal, D., Chintalapudi, K.: From experience with indoor wireless networks: A link quality metric that captures channel memory. IEEE Communications Letters 11(9), pp. 729–731 (2007)
6. Liang, W., Huang, Y., Xu, J.B., Xie, S.Y.: A distributed data secure transmission scheme in wireless sensor network. International Journal of Distributed Sensor Networks 13(4), pp.1–11 (2017)
7. Liang, W., R, Z.Q., Tang, M.D.: A secure-efficient data collection algorithm based on self-adaptive sensing model in mobile internet of vehicles. China Communications 13(2), pp. 121–129 (2016)
8. Liu, T., Cerpa, A.E.: Foresee (4c): Wireless link prediction using link features. In: the 10th International Conference on Information Processing in Sensor Networks (IPSN), Chicago, USA, pp. 294–305 (2011)
9. Liu, T., Cerpa, A.E.: Talent: Temporal adaptive link estimator with no training. In: the 10th ACM Conference on Embedded Network Sensor Systems (SensSys), Toronto, Canada, pp. 253–266 (2012)
10. Marinca, D., Minet, P.: On-line learning and prediction of link quality in wireless sensor networks. In: the 59th Global Communications Conference (GLOBECOM), Austin, USA, pp. 1245–1251 (2015)
11. Mitra, S., Roy, S., Das, A.: Parent selection based on link quality estimation in WSN, In: the 2nd International Conference on Computer and Communication Technologies (ICECCT), Tamil Nadu, India, pp. 1245–1251 (2016)
12. Oh, H.: A link availability predictor for wireless sensor networks. Cs229.stanford.edu (2010). [Online]Available: http://cs229.stanford.edu/proj2009/Haruki.pdf (current July 2017)

13. Pengwon, K., Komolmis, T., Champrasert, P.: Solving asymmetric link problems in WSNs using site link quality estimators and dual-tree topology. In: the 13th International Conference on Electrical Engineering/electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Mai, Thailand, pp. 1–4 (2016)
14. Qiong, Y.E., Li, S.W., Zhang, Y.H., Shu, X.W., Ni, D.P.: Cloud model and application overview. Computer Engineering and Design 32(12), pp. 4198–4201 (2011)
15. Rekik, S., Baccour, N., Jmaiel, M., Drira, K.: Low-power link quality estimation in smart grid environments. In: the 11th International Conference Wireless Communications and Mobile Computing (IWCMC), Dubrovnik, Croatia, pp. 1211–1216 (2015)
16. Shu, J., Tang, J., Liu, L.L., Hu, G., Liu, S.: Fuzzy support vector regression-based link quality prediction model for wireless sensor networks. Journal of Computer Research and Development 52(8), pp. 1842–1851 (2015)
17. Sikora, A., Groza, V.F.: Coexistence of IEEE802.15.4 with other systems in the 2.4 GHz-ISM-Band. In: the 20th Instrumentation and Measurement Technology Conference (IMTC), Ottawa, Canada, pp. 1786–1791 (2006)
18. Wang, Y., Martonosi, M., Peh, L.S.: Predicting link quality using supervised learning in wireless sensor networks. ACM SIGMOBILE Mobile Computing and Communications Review 11(3), pp. 71–83 (2007)
19. Woo, A., Culler, D.: Evaluation of efficient link reliability estimators for low-power wireless networks. In: UCB Technical Report. pp. 1–20 (2003). [Online]Available: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2003/CSD-03-1270.pdf (current July 2017)
20. Zhao, Y., Li, S., Hou, J.: Link quality prediction via a neighborhood-based nonnegative matrix factorization model for wireless sensor networks. International Journal of Distributed Sensor Networks 2015(1), pp. 1–8 (2015)

Chenhao Jia received the Bachelor degree in electronic and information engineering from the Jiangxi University of Science and Technology, Ganzhou, China, in 2015. He is currently a postgraduate, Internet of Things Technology Institute, Nanchang Hangkong University, Nanchang, China. His research is focused on wireless sensor networks.

Linlan Liu received the Bachelor degree in computer science from the National University of Defense Technology, Changsha, China, in 1988. Currently, she is a full Professor, Internet of Things Technology Institute, Nanchang Hangkong University, Nanchang, China. She was a Visiting Scholar at Wilfrid Laurier University, Waterloo, Ontario, Canada. She has authored/coauthored more than 60 papers. Her research interests include wireless sensor networks and embedded system.

Xiaole Gu received the Bachelor degree in electronic engineering from the Nanchang Hangkong University, Nanchang, China, in 2013, He is currently a postgraduate, Internet of Things Technology Institute, Nanchang Hangkong University, Nanchang, China. His research is focused on wireless sensor networks.

Manlan Liu received the Bachelor degree in information engineering from the Linyi University, Linyi, China, in 2015. She is currently a postgraduate, Internet of Things Technology Institute, Nanchang Hangkong University, Nanchang, China. Her research is focused on wireless sensor networks.

Received: December 20, 2016; Accepted: June 1, 2017.

Computer Science and Information Systems 14(3):735–750 https://doi.org/10.2298/CSIS161210026S

Connected Model for Opportunistic Sensor Network Based on Katz Centrality

Jian Shu[1], Lei Xu[1], Shandong Jiang[1], and Lingchong Meng[1]

Internet of Things Technology Institute, Nanchang Hangkong University
Nanchang 330063,China
shujian@nchu.edu.cn
18270717228@163.com
{306315953,282733193}@qq.com

Abstract. Connectivity is an important indicator of network performance. But the opportunistic sensor networks (OSNs) have temporal evolution characteristics, which are hard to modelled with traditional graphs. After analyzing the characteristics of OSNs, this paper constructs OSNs connectivity model based on time graph theory. The overall connectivity degree of the network is defined, and is used to estimate actual network connectivity. We also propose a computing method that uses the adjacency matrix of each snapshot. The simulation results show that network connectivity degree can reflect the overall connectivity of OSNs, which provide a basis for improving the OSNs performance.

Keywords: opportunistic sensor networks, connectivity, network connectivity degree, time graph.

1. Introduction

Opportunistic sensor networks (OSNs) are a type of self-organized network [13] that does not require complete communication paths between source and destination. They utilize contacting opportunities of ferry nodes to achieve data communication so that they can attain the regional information with low cost. They also have high delay, frequent split and intermittent connection characteristics, which are derived from the delay tolerant networks (DTNs).

The trait of the non-fully connection [17] make OSNs dispose the difficult problems when the networks are splitting. Therefore they have a wide application prospect in the hostile environment such as wildlife monitoring, emergency rescue, battlefield information collection, remote networking, vehicle network and so on. After the deployment of nodes, in order to analyze the operating condition of each node and the change of network, we should not only excavate the data from the networks, but also have a further measurement on the connectivity of the OSNs. So it has a great significance for us monitoring network and describing the link of the nodes.

Our research on OSNs connectivity seeks to improve network performance and the success rate of message delivery. We can change the delopyment of nodes and optimize the network design through the overall connectivity model. The key to analyzing connectivity lies on establishing a suitable model and comparing with the message delivery rate. It can also help the network administrator to analyze the condition of the nodes and the environment being monitored.

As shown in Fig. 1, due to the restriction of the regional terrain, the sensing areas are separated into many sub-regions. These sub-regions can only send the messages to the sink node by ferry nodes. The topology and the connectivity change frequently because of the intermittent connection between sub-regions and ferry nodes. Accordingly, it is hard to model OSNs and obtain the connectivity of the whole network accurately using the traditional graph.

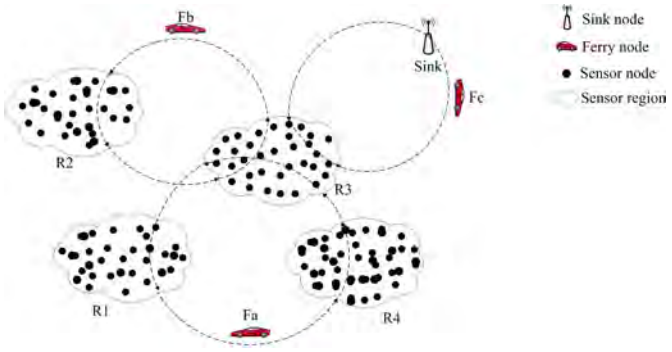


Fig. 1. The scenario of OSNs

In static networks, the links between nodes are stable and the network can be embodied into an undirected static graph by graph theory. So the node degree, the betweenness centrality and the density of the nodes can be obtained through the end-to-end path used to describe the connectivity and the robustness of the network. However, in OSNs, the physical condition and the invalidation of the nodes will lead to a spare distribution of nodes. The frequent mobility of the nodes may lead the end-to-end path becoming invalid for extended periods, which can make it difficult to describe the link relation of the nodes accurately. In this paper we define overall network connectivity with time graph theory by analyzing the opportunistic connection of nodes, and the overall connectivity can be verified through the Opportunistic Network Environment simulator(ONE).

The rest of this paper is structured as follows. In section 2, the previous research that has been performed in this area is analyzed. We propose the definitions of the time graph and the overall network connectivity in the section 3. In section 4, our experimental studies and results are presented. The concluding remarks are summarized in Section 5.

2. Related Works

In the ad hoc network, static graphs are mainly used to study the influence of node degree, the node density and the communication radius on the network connectivity. Based on the percolation-theory Kong et al. [15] found that the dynamic network has the upper and lower bounds on the critical density when it has a phase transition, and the propagation delay between the node has linear correlation with the Euclidean distance.The asymptotic critical transmission radius for k-connectivity in a wireless ad hoc network where nodes are uniformly and independently distributed in a unit area was studied in [26] , which

applied the critical transmission radius to obtain an accurate upper bound for critical k-connectivity neighbor number.

There has been a growing interest in the dense networks. For example, using computer simulations, paper [29]obtained a probability distribution graph of k-vertex-connectivity by casting the nodes randomly in sensor network. Using regression analysis, regression formula of the average connectivity of network and the empirical equation of 3-connected networks were presented. They also discussed the effect on the connectivity by the boundary nodes. These studies have a certain reference for deploying nodes. However, it is difficult to analyze the topology of sparse network like OSNs, which may undergo splitting for an extended period.

The topology of dynamic networks changes frequently with time, making them difficult to model using static methods. Therefore, they are usually modelled using probability statistics. For example, paper [28] utilized the snapshot method and used the network‘s average node degree per unit time as an estimation of the mobile network‘s overall connectivity. A correlation function describing the number of nodes, the node transmission radius and the network connectivity are obtained through curve fitting analysis, which provides the basis for deploying the nodes. Paper [18] studied the connectivity of the mobile ad-hoc networks (MANETs) under the Random Waypoint RWP mobility model in arbitrary convex region. The critical transmitting range is acquired when the network is k-connected, and MANETs have a high requirement for node fault tolerance. Guo L et al. [10] investigated three fundamental characteristics (the node degree distribution, the average node degree and the maximum node degree) of MANETs in presence of radio channel fading. The results are very useful in the study of improving connectivity and the routing protocols. The premise of the MANETs connectivity research is to ensure a complete end-to-end path, which makes the connectivity and the node degree of the network unsuitable for use in OSNs.

Derived from DTNs, OSNs share many similar characteristics. In [11] Harras K A et al. pointed out that the increasing use of wireless devices has created new challenges such as network partitioning and intermittent connectivity. Accordingly, delay tolerant mobile networks (DTMNs) have been proposed to achieve inter-regional communication using a dedicated set of messengers. Additionally, several classes of messenger scheduling algorithms have been developed to improve connectivity and performance. In recent years, with the development of the dynamic network research [1][2][22][24][21][8], characteristics used to analyze static networks have increasingly been applied to dynamic networks. For example, paper [19] referred to the time-varying graph(TVG). [6] Casteigts et al. studied the strongly connected components in time-varying graph and proposed a method to calculate the temporal diameter of the dynamic network. Paper [7] proposed a method to solve the temporal diameter of dynamic network. It proved that the mobile nodes have the “small world“ characteristic in opportunistic network-shorter temporal diameter lead to higher network connectivity. A new temporal distance metric composed of the shortest path and the clustering coefficient was proposed in paper [25]. They demonstrated that, compared with the metrics used with static graphs, this metric can obtain the temporal characteristics more accurately by using a time-varying graph. Paper [27] proposed a new dynamic coding control mechanism to exploit the connection opportunities and optimize network resources. Paper [23] presented a systematic approach to the interdependencies and constructed analogies for the various factors that affect and constrain wireless sensor

networks. However, parameters such as connected component and network diameter do not accurately reflect the overall network connectivity of OSNs.

In order to characterize the connectivity of OSNs precisely, we use the time graph to model the connectivity, and propose the methods to estimate the connectivity based the OSNs' characteristics. A series of time windows are obtained via snapshots, which are then transformed into adjacency matrices. The adjacency matrix set is then used to calculate overall network connectivity.

3. Connectivity Modeling

3.1. Time-Graph Model

In order to obtain the dynamic topology information, the time factor should be considered. Give a dynamic network trace starting at t_{min} and ending at t_{max}. Let $G_t^w(t_{min}, t_{max})$ be a time graph, the snapshot sequences can be expressed as $G_{t_{min}}, G_{t_{min+w}}, \cdots, G_{t_{max}}$, where w is the size of each window in terms of a time unit. Viewed from an abstract level, the topology in OSNs is a series of time windows in a continuous sequence.

Each node is deployed using the wireless sensor network method for monitoring the environment in paper [12]. The density and connectivity in these regions are higher than other simulations, so we can abstract the sensing region into a super node to simplify analysis. The snapshots at t_1, t_2, t_3, t_4 can be viewed from the Fig. 2.

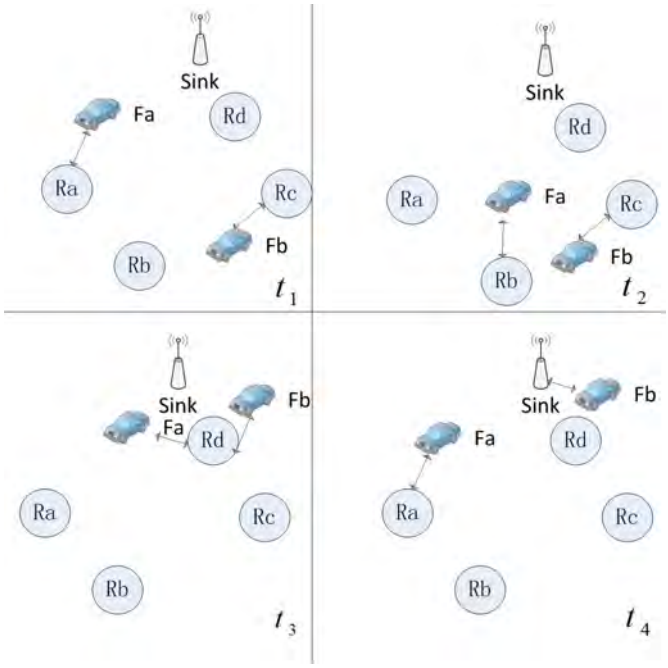


Fig. 2. The snapshots at t_1,t_2,t_3,t_4

Definition 1:Time graph G_t^w

Let G_t^w, $G_{t_{min}}$, $G_{t_{min+w}}$, $\cdots$, $G_{t_{max}}$ denote the ordered graphs under a series of increasing discrete times from t_{min} to t_{max}, and $G(t) = (V(t), E(t))$ denotes the subgraph at the t moment, where $V(t)$ and $E(t)$ are the vertex set and edge set in t moment respectively. For $\forall t \in [t_0, t_\tau], |V(t)| = N$, where N is the number of nodes in network.

N remains approximately invariant during the operation of the network, so we can combine the link information which is acquired from the snapshots and identify the time the link appears. As shown in Fig. 3, although the network is not connected in each snapshot, it can forward the sensing messages from the four sensing regions R_a, R_b,R_c,R_d to the sink node via ferry nodes.

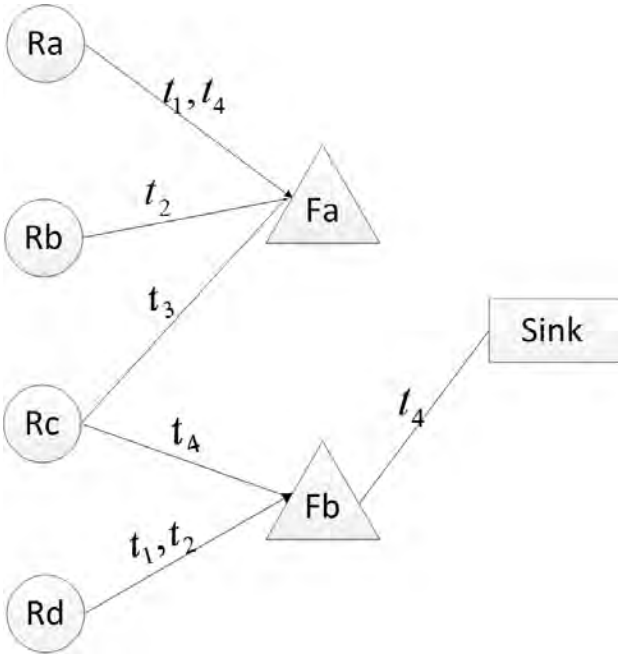


Fig. 3. Time graph model in snapshots corresponding to Fig.2

The transmission paths of the message have typical time-tropism characteristics in the time graph, so we can't use the tradition path to describe it. According to the Fig. 3,the temporal path is defined as follows.

Definition 2:Temporal Path

Let $\{n_1, n_2, n_3, \cdots, n_m\}$ be a set of non-repetitive nodes in $G_{t_{min},t_{max}}$ from t_{min} to t_{max}. If there are non-decreasing time series in $\{t_{min} + w, t_{min} + 2w, \cdots, t_m\}$, for arbitrary two nodes p and q, denoting $n_1 \equiv p$,$n_m \equiv q$, it can be concluded that the temporal path has been existed from p and q, which can be expressed as $R_{pq}(t_{min} + w, t_m)$. In this case, no temporal path exists if $R_{pq}(t_{min} + w, t_m) = \infty$.

The connectivity situation can be reflected by the temporal path as shown in Fig. 3, indicating that the sensing messages can't forward from R_a to R_d.

Definition 3:Temporal Distance

For arbitrary two nodes i and j, temporal distance is the shortest temporal path between i and j from t_0 to t_τ, which is defined as $D_{ij}(t_0, t_\tau)$.

$$D_{ij}(t_0, t_\tau) = Min\{(R_{ij}(t_0, t_\tau))\}. \tag{1}$$

Connectivity efficiency is a metric used to analyze the effectiveness of the static networks [16]. John [25] extended it into dynamic networks and redefined the efficiency of the nodes in a period of time. In this paper, we define the connectivity efficiency based on the characteristics of OSNs, the formula is denoted as follows.

Definition 4:Connectivity Efficiency

$$E_{ij}(t_0, t_\tau) = \begin{cases} 0, & D_{ij}(t_0, t_\tau) = \infty \\ \frac{t_\tau - D_{ij}(t_0, t_\tau)}{t_\tau}, & D_{ij}(t_0, t_\tau) < t_\tau \end{cases}. \tag{2}$$

Where $D_{ij}(t_0, t_\tau)$ represents the temporal distance between i and j. If $E_{ij}(t_0, t_\tau)$ has a bigger value, it means two nodes can form a link in shorter time, and that means a good connectivity between two nodes.

Definition 5:Region Connectivity Efficiency

The mean of the connectivity efficiency of all the nodes in the sensing region is the region connectivity efficiency from t_0 to t_τ , so it can be defined as follows:

$$R_{ij}(t_0, t_\tau) = \frac{1}{N_{R_i}(N_{R_i} - 1)} \sum_{ij} E_{ij}(t_0, t_\tau). \tag{3}$$

Where N_{R_i} is the number of nodes in sensing region R_i .

The region connectivity efficiency can reflect connectivity situation in sensing region.

3.2. The Connectivity of the Whole Network

In OSNs, the sink node is the center node. If a network with good connectivity, each sensing region has good connectivity with the sink node. At this moment, it acts as the core of the network. It is different from sensing nodes and ferry nodes, as it has an obvious centrality. After considering the relationship between the region and the sink node, we use the Katz centrality to reflect the centrality of sink node. Katz Centrality [20][4][14] originates from the social network analysis (SNA) and represents the degree of which one node influenced others. Paper [9] expanded the Katz Centrality to the dynamic network and found that it is extremely suitable for use with sparse networks such as MANETs, Social networks, OSNs and so on. Paper [5] utilized the Katz centrality to assess the capacity of nodes transmitting the messages to the sink node.

Definition 6:Adjacency matrix

Let G(V,E) be an undirected graph, with $V = \{v_1, v_2, \cdots, v_n\}$ the set of nodes and $E = \{e_1, e_2, \cdots, e_m\}$ the edge set . Its adjacency matrix is denoted as follows.

$$A = (a_{ij})_{n \times n}, a_{ij} = \begin{cases} 1, (v_i, v_j) \in E \\ 0, otherwise \end{cases}. \tag{4}$$

In the adjacency matrix A, the m-order power represents the number of paths between nodes with length m, and m also concludes the number of repeated nodes or edges.

$$A \times A \times \cdots \times A = (A^m)_{i,j} = k(m, k \in \mathbf{Z}^+) . \tag{5}$$

Then we should say that there are a total number of k walks with length of m from node i to node j. For any arbitrary node pair i and j, the corresponding number of walks of length $m = 1, 2, 3 \cdots$ is

$$(A^1)_{ij}, (A^2)_{ij}, (A^3)_{ij}, (A^4)_{ij}, \cdots . \tag{6}$$

The total number of walks with length no more than m from node i to all the other nodes in graph G is

$$\sum_{j}^{n} (\sum_{k=1}^{l} A^k)_{i,j} . \tag{7}$$

Generally, those nodes with longer walk length away from the source node may have less compact on the source node. This fact motivated the derivation of Katz Centrality. In this paper, we denote $\alpha(0 < \alpha < 1)$ as the impact factor. Let node j be a l-degree neighbor of node i. The impact factor of node i to node j is assigned with α^l, denoting $S = I + \alpha A + \alpha^2 A^2 + \alpha^3 A^3 + \cdots$. Let $\rho(A)$ represents the spectral radius of A, when $\alpha < 1/\rho(A)$, S is converged to $(I - \alpha A)^{-1}$, so the Katz centrality of i is define as:

$$\sum_{j=1}^{N} [(I - \alpha A)^{-1}]_{ij} . \tag{8}$$

The 'Dynamic Walk' in the temporal network is defined as below.

Definition 7:Dynamic walks

In a non-decreasing time series $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_l$, a dynamic walk of length l consists of a sequence of edges $i \to v_1, v_1 \to v_2, \cdots, v_m \to v_{m+1}, v_l \to j$, that are composed of the dynamic walks only if the r_{th} snapshot satisfies ${A_{im,im+1}}^{[r]} \neq 0$.

However, the formula (7) can't be applied to calculate the repetitive walks when calculating the temporal path. When using meeting opportunities to achieve the message transmission, the sensing messages may appear as repetitive nodes or edges on the transmission direction. So, we can calculate the number of the dynamic walks of length l from t_1 to t_l, multiplying the number with the influence factor α. The dynamic walks with length l can be obtained as follows

$$\alpha^l A(t_1)A(t_2)\cdots A(t_l)(t_1 \leqslant t_2 \leqslant \cdots \leqslant t_l) . \tag{9}$$

Considering all possible dynamic walks, when $\alpha < mint_m \rho(A(tm))$,

$$Q = [I - \alpha A(t_1)]^{-1}[I - \alpha A(t_2)]^{-1} \cdots [I - \alpha A(t_l)]^{-1} . \tag{10}$$

In OSNs, we should consider the reachability of nodes message. In this paper we just consider the message reachability of each sensing region. The calculation is defined as:

$$C = \frac{1}{N_R} \sum_{i \in R} Q_{iS} . \tag{11}$$

Where N_R is the number of the sensing regions in OSNs.

When the network has better performance, the overall network connectivity calculated by formula (11) grows to a bigger value. In order to reflect the connectivity of the current network more accurately, we need to eliminate the influence of the dimension, the variation of the variables and the numerical value, and standardize the overall network connectivity from 0 to 1.Firstly, a logarithmic function transformation method is used to compress the variable scale of the overall network connectivity. The formula is defined as follows, in which C indicates the connectivity of the whole network. C^1 is the compressed value calculated by the logarithmic function transformation method.

$$C^1 = lg(C) \ . \tag{12}$$

After compression of the connectivity, the deviation standardization is adopted to map the connectivity into $[0, 1]$

$$C^* = \frac{C^1 - min}{max - min} \ . \tag{13}$$

Where max and min are the maximum value and the minimum value of the sample data.

(1) Selection of sliding time-window

In order to facilitate network connectivity monitoring, it is very important to understand that connectivity changes over time. Due to the time-evolution characteristics of OSNs, one can obtain the connectivity by computing dynamic walks over different time. So the sliding time-window is utilized to compute the current connectivity.We calculate the success rate of the message delivery in t_2,t_3,t_4,... and find that the temporal paths have relation with the time to live(TTL) of the messages. In this way, if OSNs have better connectivity, there will be more dynamic walks to the sink node, which means more messages can be delivered to the sink nodes. But if the OSNs have bad connectivity, it may exists some dynamic walks which the arrival time is [t_3,t_4] and the departure time is [t_1,t_2] . If we just count the interval [t_1,t_2] and [t_2,t_3] , we can not find dynamic walks in this two interval. And it causes the lower success rate of the message delivery. Therefore we redefined the effective interval in this paper. As it shows in the Fig. 4, the sliding time-window length is 2, which is used to compute the connectivity from t_2 to t_5. The sliding time-window decides the effective time of the dynamic walks in the network.

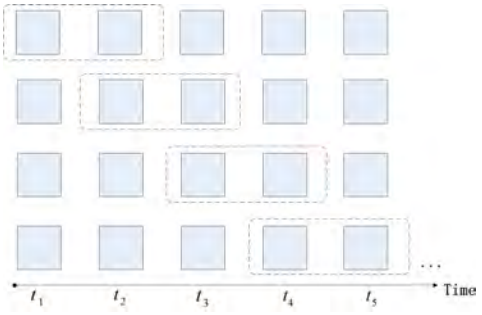


Fig. 4. the sliding time-window with the length 2

(2) Adjacency matrix sequences generation algorithm

Topology information can be acquired from continuous snapshots. However, in OSNs, due to the small number of ferry nodes and sensing regions, there are many non-connected nodes in snapshots and invalid critical-link, such as the link between ferry nodes to sink node, which affect the accuracy of OSNs connectivity.

In this paper, snapshots are generated by the communication records of the ferry nodes. The records are converted into adjacency matrix sequences, which are used as the input parameter for calculating OSNs connectivity, including the ID of the communication nodes, starting and ending time and the like. ONE can provide the format of the record of nodes as shown in Fig. 5.

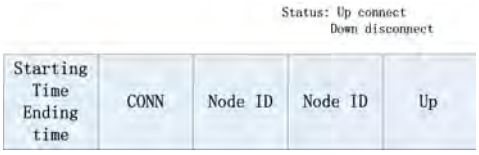


Fig. 5. Node communication record format in ONE

The adjacency matrices are generated from the record of ferry node F_a. In this algorithm, the number of adjacency matrices are determined by the interval time of each snapshot and the statistical period of simulation. The specific algorithm is described as algorithm 1.

Algorithm 1 Adjacency Matrix Seq Build Algorithm

Input: $facr$: F_a's communication record; $[t_0, t_\tau]$: the simulation time; $\triangle\ t$: the interval time;
Output: $amseq$: Adjacency matrix sequences;

1: Traverse facr . Obtain network connection records $facr^{'}$, counting the number $N_{facr'}$ and the dimension of adjacency matrices d;
2: Set $N_{seq} = \frac{(t_{t\tau} - t_0)}{\triangle t}$;
3: Construct a three-dimensional array with N_{seq} , d, d ,initialize k and other variables;
4: Set k=k+1; **if** k$\leq N_{facr'}$, turn to step 6,**then** turn to step 15;
5: **end if**;
6: **if** facr[k].status, turn to step 9,**then** turn to step 4;
7: **end if**;
8: Set the beginning time begin=facr[k].time;
9: **if** k+1$\leq N_{facr'}$, turn to step 11,**then** set end=t and turn to step 13;
10: **end if**;
11: **if** facr[k+1].status=down, turn to step 15,**then** turn to step 6;
12: **end if**;;
13: Calculate low=$\lfloor$begin/$\triangle\ t\rfloor$ and high=$\lfloor$end/$\triangle\ t\rfloor$
14: Set amseq[low][0][0] to amseq[high][d-1][d-1] into 1 with corresponding position and return to step 4;
15: Output the amseq;

As mentioned above, Algorithm 1 abstracts the sensing region into a super node. So we just need to mark the starting time $begin$ (when the ferry nodes contact with the first node), and the ending time end (when the ferry node disconnect with the last node). Due to the better connectivity in the intra-sensing region, it can be considered that the ferry nodes have a connection with regional nodes in a continuous link state from $[begin, end]$. Then set the corresponding nodes of the adjacency matrices into a link state based on the length of time slots. The adjacency matrices can be obtained by traversing the ferry node communication record, which are used to calculate the overall network connectivity.

4. Simulation Experiment

4.1. Simulation Experiment Parameters and Test Indicators

We simulated opportunistic sensor networks using ONE. The experimental scenario is shown as Fig. 6. There are four sensing regions (Ra, Rb, Rc, Rd), two ferry nodes (Fa, Fb) in the scenario. The sensor nodes have a relatively large radius of communication to ensure the interior of the region having a better connectivity. The ferry nodes collects sensing messages from the sensor nodes according to the fixed route line and delivers the messages to the sink node.

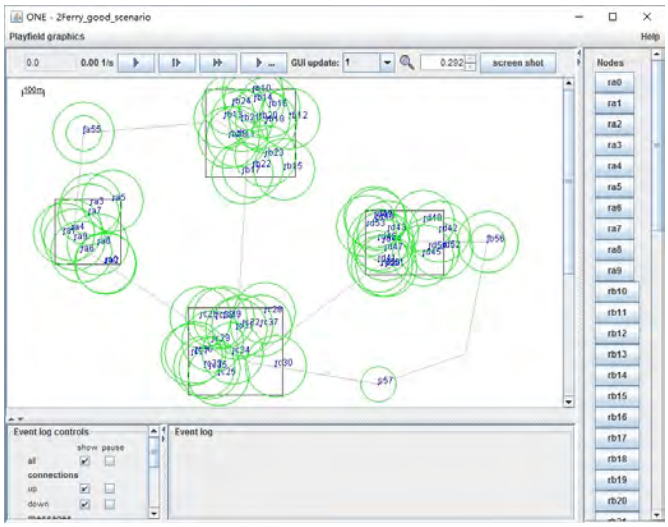


Fig. 6. The scenario of the connectivity of the whole network

The statistic cycle of the simulation is 1200 seconds. During the operation of the network, each sensing area (Ra, Rb, Rc, Rd) will randomly generate a message every 20 seconds. The simulation is carried out for 20 hours. The message delivery rate and the connectivity of the whole network are recorded every 20 minutes. Three connectivity scenarios (good connectivity, general connectivity, poor connectivity) are simulated

by adjusting the movement speed of the ferry node. The detailed parameter settings are shown in Table 2.The region sizes are corresponding to the number of region nodes.

Table 1. Experimental parameter settings

Parameter	Value
Simulation time(h)	20
Simulation period(m)	20
Region size(m*m)	300*295,410*400,430*400,360*295
Number of regional nodes(unit)	10,15,15,15
Ferry node communication radius(m)	80
Message generation intervals (s)	20
Fa movement speed(ms)	1-2,2.5-3,5-6
Fb movement speed(ms)	1-2,3-4,5-6s
TTL(s)	1200

Each sensing area (Ra, Rb, Rc, Rd) randomly generates a message every 20 seconds until the end of the simulation. The messages generated in each region are forwarded to the sink node by the ferry nodes, which taken a certain time. As a result, the messages near the end of a statistical cycle time will fail to deliver to the sink node due to the short time of the message generation. This causes a decrease message delivery success rate. In order to calculate the success rate of message delivery accurately, the following statistical method is used to compute the success rate, as shown in Fig. 7.

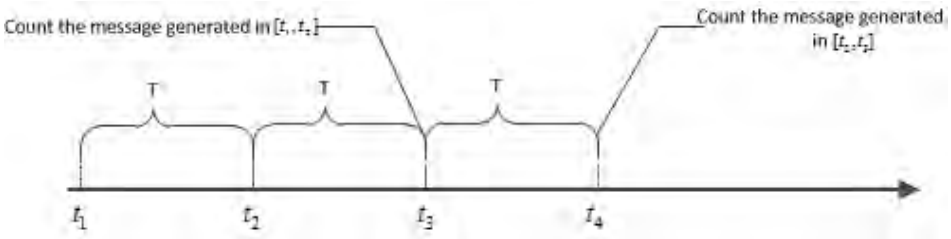


Fig. 7. The statistics of the success of the delivery

The success rate of message delivery at t_3 is generated in the time period $[t_1, t_2]$. Messages are delivered to the sink node in the time period $[t_1, t_3]$. If the sink node receives the sensing messages generated in the time period $[t_2, t_3]$, during the time period $[t_1, t_3]$, it does not count the number of the received message at the current time (t_3), and the packet is counted as the number at the next time (t_4).

In order to verify the overall network connectivity applicability, another experiments was designed in the simulation. We change the static node in regions into mobile nodes with different communication radius (20m, 40m, 60m, 80m) to simulate a network, in which sensing region is extended. The experimental scenario is shown as Fig.6. The detailed parameter settings are shown in Table 2.

Table 2. Experimental parameter settings

Parameter	Value
Simulation time(h)	20
Simulation period(m)	20
Region size(m*m)	600*450,580*390,575*400,610*395
Number of regional nodes(unit)	35,30,20,30
Ferry node communication radius(m)	70
Message generation intervals (s)	20

4.2. Experimental Results

We input the adjacency matrix sequences generated by algorithm 1 into the program and use Matlab to calculate the overall network connectivity. The connectivity of the whole network is shown in Fig. 8.The real success rate of message delivery simulated by ONE is shown in Fig. 9.The message delivery success rate is the important index to represent the real connectivity of the network in paper [3]. So we have compared it with the connectivity calculated in this paper.

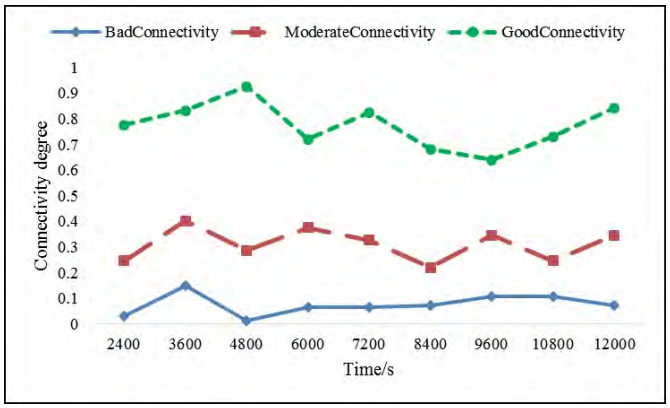


Fig. 8. The simulation effect of the connectivity of the whole network

In Fig.8, the connectivity degree in three connectivity scenarios (good connectivity, general connectivity, poor connectivity) are divided into three different ranges obviously. It can be seen from Figs. 8 and Fig. 9 that the calculated network connectivity is in good agreement with that of the real network. Therefore, the defined model can estimates the network connectivity.

Then we design another experiment to verify the applicability. We extend the sensing region and change the static nodes into mobile nodes in sensing region.By changing the radius of the sensing nodes, the simulation results can be seen from Figure.10, Fig.11 and Fig.12.

Compared Fig.10 and Fig.11, with the decrease of node communication radius in regions, region connectivity efficiency is also decreasing and the message can not spread

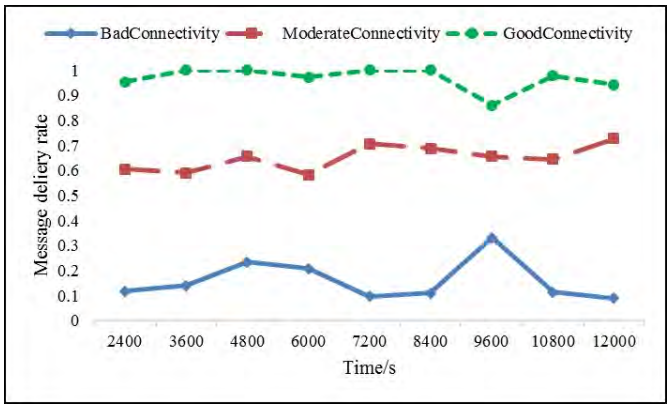


Fig. 9. The success rate of the message delivery

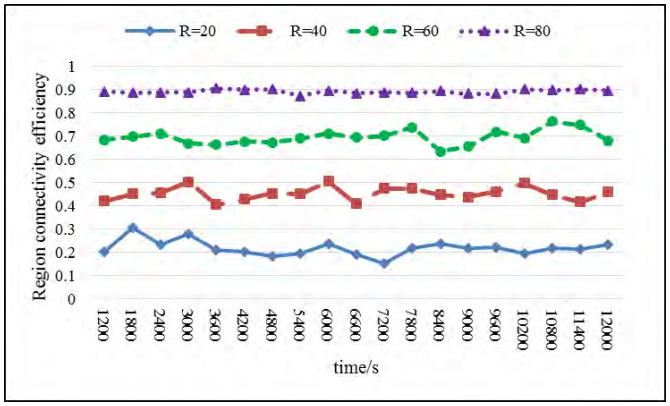


Fig. 10. Simulation results of network region connectivity efficiency

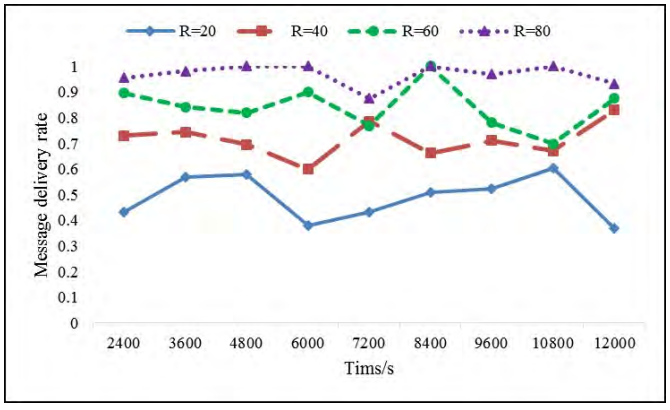


Fig. 11. Simulation results of network message delivery rate

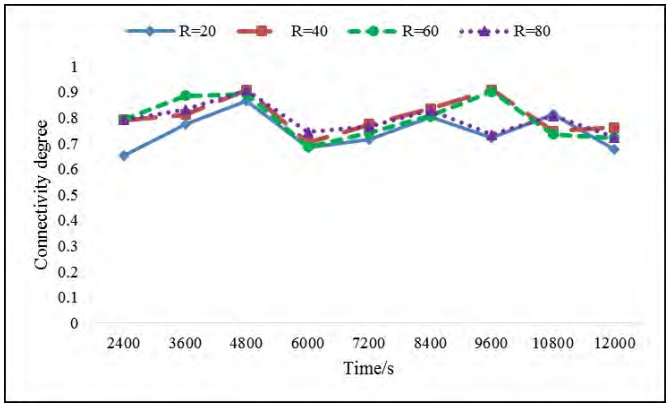


Fig. 12. Simulation results of network connectivity degree

throughout the region when message generates. Although ferry node have gone through the region, it could not obtain all the sensing messages in the region. It causes the decline of the success rate of the message delivery. As it can be seen from the Fig.10, Fig.11 and Fig.12, it can be found that when the connectivity efficiency is better, the calculated overall connectivity is in good agreement with the real network connectivity. The whole network connectivity defined in this paper can reflect the network connectivity in two different scenarios, which shows that the defined connectivity is applicable.

5. Conclusion

Network connectivity is an important metric for measuring network performance. The connectivity of OSNs has a time-evolution characteristic, which makes it difficult to model it with traditional graph models. In this paper, considering the central characteristics of the sink node, the connectivity of OSNs is modelled by time graph, according to the characteristics of OSNs. We define the connectivity of the network based on Katz Centrality in the end. The experimental results show that the proposed network connectivity model can reflect the connectivity of the whole network in different scenarios.

Acknowledgments. This work is supported in part by grants from the National Natural Science Foundation of China (Grant No.61762065,61363015,61501217,61262020),the Jiangxi Natural Science Foundation of China (Grant No.20171ACB20018, 20171BAB202
009, 20171BBH80022), the Key Research Foundation of Education Bureau of Jiangxi Province(Grant No.GJJ150702), and Nanchang Hangkong University Postgraduate Innovation Foundation(Grant No.YC2016069).

Grateful thanks are due to the participants of the survey for their invaluable help in this study.

References

1. Akrida, E.C., Gasieniec, L., Mertzios, G.B., Spirakis, P.G.: Ephemeral networks with random availability of links: diameter and connectivity. In: the 26th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), NY, USA, pp. 267–276 (2014)

2. Barjon, M., Casteigts, A., Chaumette, S., Johnen, C., Neggaz, Y.M.: Testing temporal connectivity in sparse dynamic graphs. (2014). [Online]Avaliable:http://arxiv.org/abs/1404.7634v2 (current July 2017)
3. Boldrini, C., Conti, M., Jacopini, J., Passarella, A.: Hibop: a history based routing protocol for opportunistic networks. In: the 7th International Symposium on World of Wireless, Mobile and Multimedia Networks(WOWMOM), Espoo, Finland, pp.1–12 (2007)
4. Borgatti, S.: Centrality and network flow. Social Networks 27(1), pp.55–71 (2005)
5. Cai, Q.S., Niu, J.W., Qu, G.Z.: Identifying high dissemination capability nodes in opportunistic social networks. In: the 11th Wireless Communications and Networking Conference (WCNC), Shanghai, China, pp. 4445–4450 (2013)
6. Casteigts, A., Flocchini, P., Quattrociocchi, W., Santoro, N.: Time-varying graphs and dynamic networks. International Journal of Parallel, Emergent and Distributed Systems 27(5) pp.346–359 (2012)
7. Chaintreau, A., Mtibaa, A., Massoulie, L., Diot, C.: The diameter of opportunistic mobile networks. Communications Surveys and Tutorials 10(3), pp.74-88 (2008)
8. Gao, L., Yang, J.Y., Qin, G.M.: Methods for pattern mining in dynamic networks and applications. Journal of Software 24(9), pp.2042–2061 (2013).
9. Grindrod, P., Parsons, M.C., Higham, D.J.: Communicability across evolving networks. Physical Review 83(4), pp.46120 (2011)
10. Guo, L., Xu, H., Harfoush, K.: The node degree for wireless ad hoc networks in shadow fading environments. Industrial Electronics and Applications 124(1), pp.815–820 (2011)
11. Harras, K.A., Almeroth, K.C.: Inter-regional messenger scheduling in delay tolerant mobile networks. In: the 6th World of Wireless, Mobile and Multimedia Networks (WoWMoM), NY, USA, pp.93–102 (2006)
12. Huang, Y.T., Chen, Y.C., Huang, J.H., Chen, L.J., Huang, P.: YushanNet: a delay-tolerant wireless sensor network for hiker tracking in Yushan national park. In: the 10th Mobile Data Management: Systems, Services and Middleware(MDM), Taipei, China, pp.379–380 (2009)
13. Jia, J.X., Liu, G.Z., Han D.Z.: DANCER: The routing algorithm in delay tolerant networks based on dynamic and polymorphic combination of dimensions and energy consideration. International Journal of Distributed Sensor Networks 13 (6), pp.1–17 (2017)
14. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), pp.39–43 (1953)
15. Kong, Z., Yeh, E.M.: Connectivity and latency in large-scale wireless networks with unreliable links. In: the 27th IEEE Conference on Computer Communications (INFOCOM), Phoenix, USA, pp.11–15 (2008)
16. Latora, V., Marchiori, M.: Efficient behavior of small-world networks. Physical review letters 87(19), pp.1-4 (2001)
17. Ma, H.D., Yuan, P.Y., Zhao, D.: Research progress on routing problem in mobile opportunistic networks. Journal of software 26(3), pp.600–616 (2015)
18. Min, S., Yan, S., Ye, T., Li, J.D., Zhou, E.H.: On the k-connectivity in mobile ad hoc networks. Acta Electronica Sinica 36(10), pp.1857–1861 (2008)
19. Nicosia, V., Tang, J., Musolesi, M., Russo, G., Mascolo, C., Latora, V.: Components in time-varying graphs. Chaos 22(2), pp.175–187 (2012)
20. Newman, M.: Networks: an introduction. Astronomische Nachrichten 327(8), pp.741-743 (2010)
21. Nicosia, V., Tang, J., Mascolo, C., Musolesi, M., Russo, G., Latora, V.: Graph metrics for temporal networks. Temporal Networks, pp.15-40 (2013). [Online]Avaliable: https://doi.org/10.1007/978-3-642-36461-7-2 (current July 2017)
22. Scellato, S., Leontiadis, I., Mascolo, C., Basu, P., Zafer, M.: Evaluating temporal robustness of mobile networks. IEEE Transactions on Mobile Computing 12(1), pp.105–117 (2013)
23. Snigdh, I., Gupta, N.: Quality of service metrics in wireless sensor networks: a survey. Journal of the Institution of Engineers 97(1), pp.91–96 (2016)

24. Tang, J., Leontiadis, I., Scellato, S., Nicosia, V., Mascolo, C., Musolesi, M.: Applications of temporal graph metrics to real-world networks. Temporal Networks, pp.135–159 (2013). [Online]Avaliable: https://doi.org/10.1007/978-3-642-36461-7-7 (current July 2017)
25. Tang, J., Musolesi, M., Mascolo, C., Latora, V.: Temporal distance metrics for social network analysis. In: the 2nd ACM Workshop on Online Social Networks, NY, USA, pp.31–36 (2009)
26. Wan, P.J., Yi, C.W., Wang, L.: Asymptotic critical transmission radius for k-Connectivity in wireless ad hoc Networks. IEEE Press 56(6), pp.2867–2874 (2010)
27. Wu, D., Wang, Y., Wang, H., Yang, B.: Dynamic coding control in social intermittent connectivity wireless networks. IEEE Transactions on Vehicular Technology 65(9), pp.7634–7646 (2016)
28. Zhao, J.L., Shang, R.Q., Sun, Q.X., Wang, G.X.: Study of the relationship between mobility model of ad hoc network and its connectivity. Journal of Computers 9(4), pp.49–56 (2006)
29. Zhang, Q., Sun, Y.G., Fang, Z.H.: Research on the k-vertex-connectivity reliability in wireless sensor networks. Journal of Transcluction Technology 18(3), pp.439–444 (2005)

Jian Shu is a full Professor at the Nanchang Hangkong University, in China. He received his bachelor degree and master degree in computer science from the Northwestern Polytechnical in 1985 and 1990. His research interests include internet of things, machine learning, and software engineering.

Lei Xu received the Bachelor degree in electronic and information engineering from the Jiangxi University of Science and Technology, Ganzhou, China, in 2015. He is currently a postgraduate, School of Software, Nanchang Hangkong University, Nanchang, China. His research is focused on opportunistic sensor networks.

Shandong Jiang received the Bachelor degree in software engineering from the Nanchang Hangkong Univerisity, Nanchang, China, in 2013. He is currently a postgraduate, School of Software, Nanchang Hangkong University, Nanchang, China. His research is focused on opportunistic sensor networks.

Lingchong Meng received the Bachelor degree in communication engineering from the Qingdao Agricultural University, Qingdao, China, in 2014. He is currently a postgraduate, School of Software, Nanchang Hangkong University, Nanchang, China. His research is focused on opportunistic sensor networks.

Received: December 10, 2016; Accepted: August 10, 2017.

Computer Science and Information Systems 14(3):751–767 https://doi.org/10.2298/CSIS170102027D

An Improved Artificial Bee Colony Algorithm with Elite-Guided Search Equations

Zhenxin Du[1,2], Dezhi Han[1], Guangzhong Liu[1],
Kun Bi[1], and Jianxin Jia[1]

[1] College of Information Engineering,
Shanghai Maritime University,
Shanghai 201306, China
duzhenxinmail@163.com,
dzhan, gzhliu,kunbi@shmtu.edu.cn,
jmakg23@163.com
[2] School of Computer Information Engineering,
Hanshan Normal University,
Chaozhou 521041, China
duzhenxinmail@163.com

Abstract. ABC_elite, a novel artificial bee colony algorithm with elite-guided search equations, has been put forward recently, with relatively good performance compared with other variants of artificial bee colony (ABC) and some non-ABC methods. However, there still exist some drawbacks in ABC_elite. Firstly, the elite solutions employ the same equation as ordinary solutions in the employed bee phase, which may easily result in low success rates for the elite solutions because of relatively large disturbance amplitudes. Secondly, the exploitation ability of ABC_elite is still insufficient, especially in the latter half of the search process. To further improve the performance of ABC_elite, two novel search equations have been proposed in this paper, the first of which is used in the employed bee phase for elite solutions to exploit valuable information of the current best solution, while the second is used in the onlooker bee phase to enhance the exploitation ability of ABC_elite. In addition, in order to better balance exploitation and exploration, a parameter P_o is introduced into the onlooker bee phase to decide which search equation is to be used, the existing search equation of ABC_elite or a new search equation proposed in this paper. By combining the two novel search equations together with the new parameter P_o, an improved ABC_elite (IABC_elite) algorithm is proposed. Based on experiments concerning 22 benchmark functions, IABC_elite has been compared with some other state-of-the-art ABC variants, showing that IABC_elite performs significantly better than ABC_elite on solution quality, robustness, and convergence speed.

Keywords: artificial bee colony, search equations, exploration ability, exploitation ability.

1. Introduction

Many difficult problems can be expressed as optimization problems in real world. Among these problems, however, most of them are often characterized as non-vonvex, discontinuous or non-differentiable. It is difficult to solve such problems with traditional optimization methods. As one of the most popular evolutionary algorithms (EAs), the artificial bee

colony (ABC) algorithm has shown its superior performance in dealing with optimization problems [13], such as the flow shop scheduling problem [22], filter design problem [4], and vehicle routing problem [26].

However, ABC also suffers from slow convergence speed and easily being trapped by local optimum. This is mainly caused by its solution search equations, which is good at exploration but poor at exploitation [1, 5, 11, 21, 28]. In fact, the exploration and the exploitation contradict each other. In order to achieve the excellent performance in solving optimization problems, the main challenge is how to maintain a delicate balance between the exploration and exploitation during the search process [5], and numerous ABC variants have been proposed to improve ABC's performance in this respect. Zhu et al. [28] proposed a gbest-guided ABC (GABC) to exploit the information of the global best individual ($gbest$). In the ABC/best/1 algorithm [10], the information of $gbest$ is also used to enhance the exploitation ability of ABC. Wang et al. [27] proposed a multi-strategy ensemble ABC algorithm, which employs three distinct search equations to form a strategy pool and adaptively choose one of them in different search strategy, thus the balance between exploration and exploitation can be maintained.

Recently, Cui et al. [5] proposed an artificial bee colony algorithm (the ABC_elite) with two novel search equations. One search equation incorporates the beneficial information of elite solutions, which is applied to the employed bee phase, the other one not only exploits the valuable information of the elite solutions, but also employs that of the current best solution used in the onlooker bee phase. Furthermore, the ABC_elite is embedded into depth-first framework to form a new variant of ABC, the DFSABC_elite. Experimental results show that ABC-elite and DFSABC_elite are very effective compared with other recently proposed ABC variants.

However, there still exist some drawbacks in the ABC_elite/DFSABC_elite. Firstly, in the employed bee phase of ABC_elite, the elite solutions employ the same equation as ordinary solutions, easily resulting in the low success rate for the elite solutions because of relatively large disturbance amplitude. In the search equation of ABC_elite, a candidate solution can be treated as the lead individual to explore the search space and produced by adding a scaled disturbance vector to a base vector. But we can draw inspiration from many EAs that the better the fitness value is, the smaller the disturbance amplitude is [3, 17–20, 24]. In a word, the disturbance of ordinary and elite solutions should be treated in a different way. Secondly, in the onlooker bee phase in ABC_elite, the exploitation ability of ABC_elite is still insufficient, especially in the latter half stage of a search process. To balance the exploitation and exploration ability, the search equation in the onlooker bee phase of ABC_elite uses the difference between $gbest$ and a randomly selected ordinary individual X_k as a disturbance vector, which is suitable for the ABC_elite to maintain a good balance between exploration and exploitation in the early stage of a search process, but easily leads to the insufficiency of exploitation ability in the latter half stage of a search process, because the ratio between exploration and exploitation is not constant. Generally speaking, EAs focus on exploration at the early stage and focus on exploitation at the latter half stage, which can also be seen in some other EAs [25].

Based on the above-mentioned considerations, an improved ABC_elite, the IABC_elite has been put forward in the paper. Firstly, inspired by bare-bones particle swarm optimization (PSO) [15], a novel search equation for the elite solutions in the employed bee phase is designed to generate a new candidate solution to exploit the valuable information of

the current best solution. Secondly, a novel search equation is proposed in the onlooker bee phase of ABC_elite to further enhance the exploitation ability of ABC_elite. In addition, in order to obtain a better balance between exploitation and exploration, a parameter P_o is used in the onlooker bee phase to choose a search equation between the original one of ABC_elite or the newly-proposed one. The simplicity of ABC_elite is maintained in the proposed IABC_elite. Moreover, the experiment results concerning 22 benchmark functions have demonstrated its effectiveness in solving complex numerical optimization problems when compared with the ABC_elite, DFSABC_elite and other ABC variants.

The rest of this paper is organized as follows. In Section 2, the original ABC algorithm is presented. In Section 3, the most recently developed ABC variants, the ABC_elite algorithm, is reviewed, which is the basis of the proposed algorithm IABC_elite. In Section 4, the IABC_elite algorithm is proposed based on the two novel solution search equations (i.e., the Eq.(12) and Eq. (13)) and the new introduced search equation selective probability P_o. Section 5 presents and discusses the experimental results. Finally, the conclusion is drawn in Section 6.

2. The original ABC Algorithm

Inspired by the waggle dancing and foraging behaviors of honey bee colonies, the ABC algorithm has been developed. The basic ABC algorithm consists of four sequentially realized phases, i.e. the initialization, the employed bee, the onlooker bee and the scout bee. After the initialization phase, the ABC turns into a loop of the employed bee phase, onlooker bee phase and scout bee phase until the termination condition is satified. The details of each phase are described as follows:

$Initialization\ phase$: At the beginning of the ABC, the initial food sources are generated randomly according to Eq. (1).

$$X_{i,j} = X_j^L + rand_j(X_j^U - X_j^L) \tag{1}$$

where $i = \{1, 2, ..., SN\}$, $j = \{1, 2, ..., D\}$, SN is the number of food sources, and SN is equal to the number of employed bees and onlooker bees. D is the dimensionality (variables) of the search space. X_j^L and X_j^U are the lower and upper bounds of the jth variable respectively. $rand_j$ is a random real number in range of [0,1]. Then, the fitness values of the food sources are calculated by Eq. (2).

$$\begin{array}{l} fit_i = \frac{1}{1+f(X_i)}, f(X_i) \geq 0 \\ fit_i = 1 + |f(X_i)|, f(X_i) < 0 \end{array} \tag{2}$$

where fit_i is the fitness value of the ith food source X_i, and $f(X_i)$ is the objective function value of food source X_i for the optimization problem. In addition, parameter $limit$ should be determined and the parameter $counter$, which records the number of unsuccessful updates, is set to 0 for each food source.

$Employed\ bee\ phase$: Each employed bee will fly to a distinct food source and try to find out a candidate food source in the neighborhood of the corresponding parent food source by using Eq. (3).

$$V_{i,j} = X_{i,j} + \phi_{i,j} \times (X_{i,j} - X_{k,j}) \tag{3}$$

where i, k are picked up from $\{1, 2, ..., SN\}$ randomly, j is randomly selected from $\{1, 2, ..., D\}$, $V_{i,j}$ is the jth dimension of the ith candidate food source (new solution). $X_{i,j}$ is the jth dimension of the ith food source; $X_{k,j}$ is the jth dimension of the kth food source, $\phi_{i,j}$ is a random real number in the range of [-1,1].

After creating a new food source, the fitness value of the candidate food source is calculated by Eq. (2). If the fitness value of candidate food source is better than that of the old one, the candidate food source will replace the old one and is memorized by its employed bee, and the $counter$ of the food source is reset to 0. Otherwise, the $counter$ is increased by 1.

$Onlooker\ bee\ phase$: According to the quality information of the food source shared by the employed bees, each onlooker bee will fly to a food source X_s, which is selected by the roulette wheel, in order to find a candidate food source by using Eq. (3). The selection probability of the ith food source is calculated as Eq. (4). Obviously, the better the fitness value is, the bigger the selection probability is. If a candidate food source V_s obtained by the onlooker bee is better than the food source X_s, X_s will be replaced by the new one, and its $counter$ is reset to 0. Otherwise, its $counter$ is increased by 1.

$$P_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \tag{4}$$

$Scout\ bee\ phase$: The food source with the highest $counter$ value is selected and its $counter$ value is compared with a predefined $limit$ value. If its $counter$ value is bigger than the $limit$ value, the selected food source will be abandoned by its employed bee, and then this employed bee will become a scout bee to seek a new food source randomly according to Eq. (1). After the new food source is obtained, the corresponding $counter$ value is reset to 0, and the scout bee returns to an employed bee. Note that if the jth variable $V_{i,j}$ of the ith candidate food source violates the boundary constraints in the employed bee phase and the onlooker bee phase, it will be reset according to the Eq. (1).

3. The improved ABC variants

As is known to all, the remarkable feature of the ABC depends on its solution search equation that differentiates the algorithm from other EAs. The search equations of ABC play a key role in balancing the exploration and exploitation ability during a search process. However, the search equation of ABC (see Eq. (3)) performs well in exploration but poorly in exploitation [5, 28]. In order to solve this problem, numerous search equations have been proposed to improve ABC's performance.

In the beginning, Zhu et al. [28] proposed a new search equation (GABC), as shown in the Eq. (5) with the information of the global best ($gbest$) to enhance the exploitation ability of the ABC. However, as claimed in [11], the Eq. (5) may cause an oscillation phenomenon and thus may degrade convergence, since the guidance of the last two terms may be in opposite directions. Then Gao et al. [9] proposed a new search equation, as shown in the Eq. (6). Although the information of the current best solution is utilized in the Eq. (6). The candidate solution generated around X_{best} constantly determines its emphasis on exploitation. Therefore, in order to solve these problems in Eq. (5) and (6), they [11] designed a new search equation in the Eq. (7) without any bias to any search direction and under the guidance of the only one term $\phi_{i,j}.(_{r1,j}\text{-}X_{r2,j})$ the oscillation phenomenon can

be effectively avoided. Therefore, the search ability of ABC is improved significantly by Eq. (7). From Eq. (5) to Eq. (7), $\psi_{i,j}$ is a uniform random number in [0,1.5]. $X_{best,j}$ is the jth element of the current best solution. Index k is an integer randomly chosen from $\{1, 2, ..., SN\}$ and different from the base index i. $r1$ and $r2$ are two distinct integers randomly picked up from $\{1, 2, ..., SN\}$, and both of them are different from the base index i.

$$V_{i,j} = X_{i,j} + \phi_{i,j} \times (X_{i,j} - X_{k,j}) + \psi_{i,j}(X_{best,j} - X_{i,j}) \tag{5}$$

$$V_{i,j} = X_{best,j} + \phi_{i,j} \times (X_{i,j} - X_{r1,j}) \tag{6}$$

$$V_{i,j} = X_{r1,j} + \phi_{i,j} \times (X_{r1,j} - X_{r2,j}) \tag{7}$$

Although the Eq. (7) can significantly improve the search ability of ABC, the beneficial information of the population is not fully exploited. Recently, in order to further improve the performance of ABC by utilizing the useful information of some good solutions, Cui et al [5] proposed two novel search equations as follows:

$$V_{i,j} = X_{e,j} + \phi_{i,j} \times (X_{e,j} - X_{k,j}) \tag{8}$$

$$V_{e,j} = \frac{1}{2}(X_{e,j} + X_{best,j}) + \phi_{e,j} \times (X_{best,j} - X_{k,j}) \tag{9}$$

where X_e is randomly chosen from the elite solutions (the top $p.SN$ solutions in current population, $0 < p < 1$)). X_k is randomly chosen from current population. e unequal to k and k unequal to i, X_{best} is the current best solution. $\phi_{i,j}$ and $\phi_{e,j}$ are two random real numbers in [-1,1]. In the ABC_elite, Eq. (8) is used in the employed bee phase, making all solutions learn from elite solutions, and the Eq. (9) is employed in the onlooker bee phase, allowing elite solutions to learn from the current best solution. Moreover, under the guidance from only one term, the Eq. (8) and Eq. (9) can also easily avoid the oscillation phenomenon. In this way, the ABC_elite algorithm can better balance the exploration and exploitation and has shown better performance when compared with other state-of-the-art ABC variants, such as the GABC [28], CABC [11], Best-so-far ABC [2], MABC [10], qABC [14], EABC [12], ABCVSS [23], BABC [8].

4. The proposed Algorithm

From the aforementioned analysis, although ABC_elite has shown excellent performance, it still has some drawbacks. In ABC_elite, all individuals utilize the same search equation in different search stages. To overcome the limitation and enhance the performance of ABC_elite, two novel search equations and a new probability P_o are proposed in this paper. In Section 4.1, inspired from some state-of-the-art PSO variants [15, 18, 19], a novel search equation is proposed based on labor-division strategy in which the elite individuals utilize the new search equation to enhance the exploitation ability. In section 4.2, a more exploitive search equation is proposed. Meanwhile, a probability P_o is introduced to decide which equation is to be selected, the new search equation or the original one. At the end of this section, the complete proposed algorithm is shown.

4.1. The Improvement in Employed Bee Phase

In the Eq. (8), the first term $X_{e,j}$ in the right-hand side is called the base vector, and the second term $\phi_{i,j}.(X_{e,j}\text{-}X_{k,j})$ can be called the disturbance vector. Thus, the candidate solution $V_{i,j}$ in the left hand of the Eq. (8) can be treated as a disturbance to the base vector $X_{e,j}$. However, the disturbance amplitude is obviously too large for elite individuals. The reason is that in the disturbance vector $\phi_{i,j}.(X_{e,j}\text{-}X_{k,j})$, X_e is an elite solution and X_k is a randomly selected ordinary solution. Generally speaking, the fitness of X_e is far better than X_k, thus $\phi_{i,j}.(X_{e,j}\text{-}X_{k,j})$ is moderate for ordinary individuals but relatively large for those elite solutions. Therefore, the success rate of disturbance for elite individuals is very low. The similar conclusion can be found from some other EAs [17–20]. In general, the better the fitness value is, the smaller the disturbance amplitude is [17–20]. In a word, the disturbance amplitude of ordinary and elite solutions should be treated in a different way. PSO [7,16] is another important EA, which is similar to the ABC in evolution mechanism. Kennedy et al. [15] proposed a novel search equation in PSO shown as follows:

$$P_i = \frac{c_1 \times pbest_i + c_2 \times gbest}{c_1 + c_2} \tag{10}$$

Where c_1 and c_2 are two learning coefficients, $pbest$ is the personal best position, $gbest$ is the population best solution found so far.

Based on the Eq. (10), a novel equation is proposed in [19]:

$$X_i = N(\frac{gbest + pbest_i}{2}, |gbest - pbest_i|) \tag{11}$$

where N denotes a Gaussian distribution of mean $(gbest + pbest_i)/2$ and standard deviation $|gbest - pbest_i|$. By using a Gaussian distribution in Eq. (11). The information around $pbest$ and $gbest$ is exploited.

Inspired by Eq. (11), a similar Gaussian search equation of ABC is proposed only for elites in employed bee phase which is shown as follows:

$$V_{i,j} = N(\frac{X_{best,j} + X_{i,j}}{2}, |X_{best,j} - X_{i,j}|) \tag{12}$$

Where $X_{i,j}$ is the jth element of elite X_i; $X_{best,j}$ is the jth element of the global best found so far; j is randomly selected from $\{1, 2, ..., D\}$. By way of the Eq. (12), the elite solutions in employed bee phase search around X_{best}, which can improve the exploitation ability of ABC and the success rate of disturbance for elite solutions.

On the other hand, the ordinary solutions in employed bee phase will still use the same equation as the ABC-elite (i.e., Eq. (8)), which will lay emphasis on exploration. Because ordinary solutions account for the majority of population while elite solutions only account for a small proportion p ($p = 0.1$ in [5]), the employed bee phase will still focus on exploration, which also conform to the design principle of the ABC [13]. Similar to the labor-division strategy in literatures [18] and [19], the ordinary solutions with low fitness can focus on locating the unexplored region, whilst the elite solutions with high fitness can perform local search on the most promising explored regions. In this way, it is beneficial to obtain a better balance between exploration and exploitation for the improved algorithm.

4.2. The Improvement in Employed Bee Phase

In the search Eq. (9) of ABC_elite, $(X_{e,j} + X_{best,j})/2$ in the right-hand side can be called base vector, and the second term $\phi_{e,j}(X_{best,j} - X_{k,j})$ in the right-hand side can be called disturbance vector. The meaning of the Eq. (9) is that the jth element of candidate solution V_e will be produced by imposing the disturbance $\phi_{e,j}(X_{best,j} - X_{k,j})$ on the base vector $(X_{e,j} + X_{best,j})/2$. It is worth noting that only elite solutions in the onlooker bee phase of ABC_elite have a chance of producing candidate solutions, which will enhance the exploitation ability of ABC. In the Eq. (9), three kind of individuals are involved, i.e. the elite individuals X_e, the global best individual X_{best}, and the ordinary individual X_k. Because the fitness value of X_e and X_{best} is generally far better than the ordinary individual X_k, the disturbance vector $\phi_{e,j}(X_{best,j} - X_{k,j})$ is relatively large for base vector $(X_{e,j} + X_{best,j})/2$ .The relatively large disturbance $\phi_{e,j}(X_{best,j} - X_{k,j})$ embodies the exploration ability of ABC_elite, and the excellent $(X_{e,j} + X_{best,j})/2$ embodies the exploitation ability of ABC_elite, thus the balance between exploration and exploitation can be maintained. It can be seen from Fig.1, which is illustrated by literature [5], the candidate solution V_e can be only generated at the red axis, which is closer to the current best solution when $\phi_{e,j}(X_{best,j} - X_{k,j})$ is small, but is far away from the current best solution when X_k is inferior and $\phi_{e,j}(X_{best,j} - X_{k,j}))$ is big.

Therefore, this design can result in the lack of exploitation ability, especially in the mid-late stage of evolution process because the demand of exploitation ability in EAs is not constant from the beginning to the ending. Generally speaking, for an EA, high exploration ability is required in the beginning to find more potential positions, while high exploitation ability is needed for convergence in the end. This conclusion can also be found in some other EAs, one of the most remarkable instance is the wPSO [25], in which linearly diminished weight is used so as to gradually increase the exploitation ability of PSO.

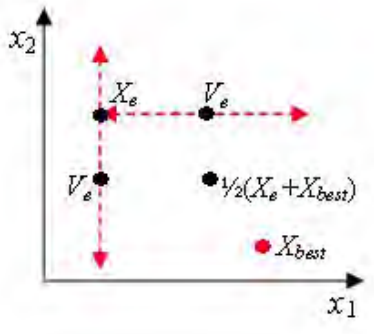


Fig. 1. Evolution process of a solution according to Eq.(9).

Because the randomly selected elite individual $X_{e'}$ has better fitness value than ordinary individual X_k in general and thus $|X_{best,j} - X_{e',j}| < |X_{best,j} - X_{k,j}|$ with a high probability, if X_k is replaced with another randomly selected elite $X_{e'}$ in the Eq. (9), the disturbance of $\phi_{e,j}(X_{best,j} - X_{k,j})$ to the base vector $(X_{e,j} + X_{best,j})/2$ will be diminished, thus the exploitation ability of the Eq. (9) will be strengthened. Based on the above observation, a novel search equation used in the onlooker bee phase is proposed as follows:

$$V_{e,j} = \frac{1}{2}(X_{e,j} + X_{best,j}) + \phi_{e,j}(X_{best,j} - X_{e',j}) \tag{13}$$

Where $X_{e'}$ is a randomly selected elite solution, e' not equal to e; the rest of Eq. (13) is same as that in Eq. (9).

Based on the above analysis, the Eq. (13) has a high exploitation ability than that of the Eq. (9) by imposing a small disturbance $\phi_{e,j}(X_{best,j} - X_{k,j})$ on the base vector $(X_{e,j} + X_{best,j})/2$. However, both the exploration ability and exploitation ability are needed in EAs. If all bees produce new food sources using the Eq. (13), the algorithm can easily get trapped in the local optima when solving complex multi-modal problems. In other words, the Eq. (9) is insufficient in exploitation ability, while Eq. (13) is inadequate in exploration ability. To address this contradiction, we propose a new search mechanism in which the selective probability P_o is introduced to balance the exploration of Eq. (9) and the exploitation of Eq. (13). If the randomly generated number in [0,1] is less than P_o, the Eq. (9) will be executed, otherwise the Eq. (13) will be executed. Because the demand of exploitation ability in EAs is gradually increased, the parameter P_o will be diminished linearly from 1 to 0. (see Lines 20 to 26 in Algorithm 1).

By combining Eq. (8) and (12) used in the employed bee phase, the Eq. (9) and (13) used in the onlooker bee phase and the selective probability P_o used to select the Eq. (9) and (13), an improved ABC_elite, IABC_elite for short, is proposed. The pseudo-code of IABC_elite is given in Algorithm 1.

Compared with the original ABC_elite, the IABC_elite adds no additional computation load, the whole structure of IABC_elite is the same as ABC_elite. The only difference between the two algorithms lies in their search equations. Therefore, the total complexity of the IABC_elite is the same as that of the ABC_elite. Now that the complexity of ABC_elite is $O(D * SN)$) [5], the complexity of IABC_elite is also $O(D * SN)$), which is also the same as original ABC [5].

The major difference between ABC_elite and IABC_elite is that ABC_elite employ only one search equation Eqs. (8) and (9) in the employed bee phase and onlooker bee phase, respectively, while IABC_elite adopts two different search equations in each phase. When the experimental results are analyzed, it is shown that the integration of search equations is a better option than the single search equation used in ABC_elite because each search equation contributes the local search ability or global search ability, thus, the global-local search abilities are better balanced by using different search equations.

5. Experiments and Discussions

To investigate the effectiveness of the proposed algorithm IABC_elite, the IABC_elite algorithm is compared with the original ABC, BABC, ABC_elite, EABC, ABCVSS and

DFSABC_elite. We selected these ABC variants for comparison because the search equation of the basic ABC algorithm is improved in these recently developed methods. DFSABC_elite is a composite algorithm consisting of the ABC_elite and the depth-first framework, showing relatively good performance when compared with other state-of-the-art algorithms.

5.1. Benchmark Functions and Parameter Settings

To analyze and compare the performance and accuracy of the proposed algorithm IABC_elite, a set of 22 benchmark functions with dimension $D = 30$ are used in the experiments. For instance, $f_1 - f_6$ and f_8 are the continuous unimodal functions; f_7 is a discontinuous step function; f_9 is a noisy quartic function. f_{10} is the Rosenbrock function which is unimodal for $D = 2$ and $D = 3$, while it may have multiple optimal solutions when $D > 3$. $f_{11} - f_{22}$ are multi-modal functions, and the number of their local optimal points increases exponentially with the problem dimension. The search range, the global optimal value, the acceptant value of each function and their definitions can be found in the literature [5]. When the objective function value of the best solution obtained by an algorithm in a run is less than the acceptant value, the run is regarded as a successful one.The performance evaluation metrics are the same as those in the literature [5], which are described as follows: (1) The mean and standard deviation of the best objective function value are obtained by each algorithm, which are used to evaluate the quality or accuracy of the solutions obtained by different algorithms. The smaller the value of this metric is, the higher quality/accuracy the solution has; (2) The average FES (AVEN) is required to reach the acceptant value, which is employed to evaluate the convergence speed. The smaller the value of this metric is, the faster the convergence speed is. Note that AVEN will only be calculated for the successful runs. If an algorithm cannot find any solution whose objective function value is smaller than the acceptant value in all runs, AVEN will be denoted by NA; (3) The success rate (SR%) of the 25 independent runs is utilized to evaluate the robustness or reliability of different algorithms. The greater the value of this metric is, the better the robustness/reliability is.

The parameter settings in the two experiments evaluated in the present paper have used the same settings of the ABC_elite [5], and the maximal function evaluation (max_FES) is employed as the termination condition, which is set to 150000. For all the algorithms, SN is set to 50, $D = 30$, $limit = SN.D$; For the ABC-elite and DFSABC_elite, p is set at 0.1. The parameter settings of all the other algorithms are set as suggested in their original papers shown in Table 1. All the algorithms are conducted with 25 independent runs for each test function.

In the two experiments evaluated in this paper, Experiment 1 is used to validate the effectiveness and efficiency of the improved algorithm (IABC_elite). Experiment 2 is used to further evaluate the performance of IABC_elite, when compared to other ABC variants developed recently.

The results of Experiment 1 and Experiment 2 are given in Table 2 and Table 3, respectively. The better results of these two experiments are marked with boldface, and the paired Wilcoxon [6] signed-rank test is used to compare the significance between the two algorithms. The signs-, +,and= denotes that the performance of the corresponding algorithm is worse than, better than and similar to that of the IABC_elite, respectively,

Algorithm 1 The procedure of IABC_elite

Initialization:Generate SN solutions that contain variables according to $Eq.$ (1);
while $Fes < max\_Fes$ **do**
  Select the top $T = p.SN$ solutions as elite solutions from population;
  for $i = 1$ to SN **do**
    //employed bee phase
    if i is an elite solution **then**
      Generate a new candidate solution V_i in the neighborhood of X_i using $Eq.$(12);
    else
      Generate a new candidate solution V_i in neighborhood of X_i using $Eq.$(8);
    end if
    Evaluate the new solution V_i;
    if $f(V_i) < f(X_i)$ **then**
      Replace X_i by V_i;
      counter(i)=0;
    else
      counter(i)= counter(i)+1;
    end if
  end for//end employed bee phase
  for $i = 1$ to SN **do**
    //onlooker bee phase
    Select a solution X_e from elite solutions randomly to search;
    $P_o = 1 - Fes/max\_Fes$;
    if $rand(0, 1) < P_o$ **then**
      Generate a new candidate solution V_e in neighborhood of X_e using $Eq.$(9);
    else
      Select a solution $X_{e'}$ from elite solutions randomly, where e' not equal to e;
      Generate a new candidate solution V_e using $Eq.$(13);
    end if
    Evaluate the new solution V_e;
    if $f(V_e) < f(X_e)$ **then**
      Replace X_e by V_e;
      counter(e)=0;
    else
      counter(e)= counter(e)+1;
    end if
  end for//end onlooker bee phase
  $Fes = Fes + SN$*2;
  Select the solution X_max with max counter value; //Scout bee phase
  if $counter(max) > limit$ **then**
    Replace X_max by a new solution generated according to $Eq.$(1);
    $Fes = Fes + 1, counter(max) = 0$;
  end if//end scout bee phase
end while

Table 1. Parameters setting used in all experiments.

Algorithm	Parameters setting
ABC	$SN = 50, limit = SN \cdot D$
EABC	$SN = 50, limit = SN \cdot D, \mu = 0.3, \delta = 0.3, A = 1$
BABC	$SN = 50, limit = SN \cdot D$
ABCVSS	$SN = 50, limit = SN \cdot D, c = 2$
ABC_elite	$SN = 50, limit = SN \cdot D, p = 0.1$
DFSABC_elite	$SN = 50, limit = SN \cdot D, p = 0.1, r = 1/p$
IABC_elite	$SN = 50, limit = SN \cdot D, p = 0.1$

according to Wilcoxons rank test [6] at a 0.05 significance level. The last row in Table 2 and Table 3 each summarizes the comparison results.

5.2. Benchmark Functions and Parameter Settings

In this experiment, in order to validate the effectiveness and efficiency of IABC_elite, the IABC-elite is compared with the ABC [13], BABC [8], ABC-elite [5] respectively. The results are shown in Table 2.

It can be clearly observed from Table 2 that the IABC_elite outperforms all the other algorithms significantly in most of tested functions in terms of solution accuracy and convergence speed according to mean (std) and AVEN, respectively.

(1) $The\ comparative\ results\ of\ unimodal\ functions$: $f_1 - f_9$ are unimodal functions. For functions $f_1 - f_6$, IABC_elite demonstrates best performance in terms of solution accuracy and convergence speed according to mean(std) and AVEN, respectively. Because functions f_7 and f_8 are easy to solve [5], the solution accuracy of all algorithms of this two functions are similar, but IABC_elite has achieved better results regarding convergence speed. All in all, the results of IABC_elite are better or at least similar to all other compared algorithms in all unimodal functions according to all test metrics.

The advantage of the IABC_elite on unimodal is due to the novel Eq. (12) and Eq. (13), which can further enhance the exploitation ability of ABC_elite.

(2) $The\ comparative\ results\ on\ multimodal\ functions$: In multimodal functions $f_{10} - f_{22}$ of Table 2, IABC_elite also demonstrates good performance. Firstly, in the solution accuracy, the IABC_elite are better than or at least comparable to all other compared algorithms in all multimodal functions except for only 2 functions (f_{10} and f_{18}). Secondly, in the convergence speed AVEN, the IABC_elite performs better than or at least comparable to all its competitors in all multimodal functions only except for the ABC_elite on f_{22}. Thirdly, in the metric SR, the IABC_elite are better than or at least comparable to all other compared algorithms on all multimodal functions. The advantage of IABC_elite on multimodal is due to the introduced parameter P_o, which helps the IABC_elite to maintain a better balance between exploration and exploitation.

The convergence curves of these involved algorithms are shown in Fig.2. Because the length of this paper is limited, only the convergence of 4 functions are given. From Fig.2, it can be seen that the IABC_elite can achieve the fastest convergence speed and best accuracy among the involved 4 algorithms.

Table 2. The comparative results of ABC, BABC, ABC_elite and IABC_elite when D=30.

No.	metric	ABC	BABC	ABC_elite	IABC_elite
f_1	Mean(std)	1.04e-17(1.20e-17)-	1.14e-43(1.77e-43)-	3.33e-50(5.34e-50)-	2.20e-105(7.23e-105)
	SR/AVEN	100/83,702	100/43,530	100/32,166	100/19,617
f_2	Mean(std)	4.38e-10(4.72e-10)-	4.18e-30(3.33e-17)-	2.08e-45(4.36e-45)-	1.31e-102(5.39e-102)
	SR/AVEN	100/136,290	100/83,026	100/45,930	100/25,960
f_3	Mean(std)	1.14e-19(9.89e-20)-	7.40e-15(3.70e-14)-	9.21e-51(8.50e-51)-	2.45e-107(1.17e-106)
	SR/AVEN	100/75,402	100/38,022	100/30,678	100/18,615
f_4	Mean(std)	2.02e-31(5.30e-31)-	4.96e-90(1.54e-89)-	1.69e-95(5.46e-95)-	1.52e-168(1.52e-168)
	SR/AVEN	100/23,578	100/11,222	100/10,662	100/6945
f_5	Mean(std)	7.69e-11(3.04e-11)-	1.61e-24(8.21e-25)-	6.59e-26(3.04e-26)-	9.04e-56(2.07e-55)
	SR/AVEN	100/124,870	100/58,046	100/54,874	100/30,280
f_6	Mean(std)	4.39e+00(1.07e+00)-	1.71e+00(1.15e+00)-	2.66e+00(1.75e+00)-	1.33e-02(1.07e-02)
	SR/AVEN	0/NA	32/122,490	80/104,250	100/68,600
f_7	Mean(std)	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)
	SR/AVEN	100/10,994	100/9426	100/94,740	100/7650
f_8	Mean(std)	7.18e-66(5.21e-73)=	7.18e-66(2.04e-77)=	7.18e-66(1.20e-79)=	7.18e-66(1.19e-81)
	SR/AVEN	100/150	100/150	100/150	100/150
f_9	Mean(std)	6.02e-02(1.09e-2)-	2.70e-02(8.28e-03)-	1.90e-02(4.83e-03)-	1.36e-02(3.70e-03)
	SR/AVEN	100/91,786	100/35,582	100/31,034	100/18,665
f_{10}	Mean(std)	5.45e-02(5.86e-02)+	3.97e-02(4.96e-02)+	1.47e-01(5.18e-01)+	5.6e-01(1.15e+00)
	SR/AVEN	88/11,014	100/83,026	84/78,817	70/65,792
f_{11}	Mean(std)	3.50e-14(1.35e-13)-	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)
	SR/AVEN	100/99,134	100/41,354	100/41,522	100/27,575
f_{12}	Mean(std)	1.70e-12(4.36e-12)-	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)
	SR/AVEN	100/112,080	100/49,050	100/44,206	100/30,175
f_{13}	Mean(std)	2.36e-14(5.62e-14)-	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)
	SR/AVEN	100/94,862	100/42,942	100/39,826	100/30,087
f_{14}	Mean(std)	4.58e-12(1.59e-12)-	2.18e-13(7.80e-13)-	1.16e-12(1.65e-12)-	1.09e-13(3.25e-13)
	SR/AVEN	100/82,946	100/50,418	100/42,794	100/41,826
f_{15}	Mean(std)	4.31e-09(1.85e-09)-	5.65e-15(1.33e-15)=	6.08e-15(7.10e-16)-	5.52e-15(3.21e-16)
	SR/AVEN	100/145,410	100/65,210	100/63,606	100/35,210
f_{16}	Mean(std)	1.03e-18(6.90e-19)-	8.98e-14(4.49e-13)-	1.57e-32(5.59e-48)=	1.57e-32(3.42e-48)
	SR/AVEN	100/77,346	100/40,542	100/30,362	100/17,660
f_{17}	Mean(std)	4.88e-18(5.03e-18)-	1.50e-33(8.28e-33)=	1.50e-33(0.00e+00)=	1.50e-33(0.00e+00)
	SR/AVEN	100/86,542	100/40,810	100/32,470	100/19,055
f_{18}	Mean(std)	2.35e-06(1.66e-06)-	3.33e-17(1.28e-16)+	8.88e-18(4.44e-17)+	3.69e-16(8.23e-16)
	SR/AVEN	0/NA	100/55,262	100/57,226	100/42,280
f_{19}	Mean(std)	4.46e-14(5.39e-14)-	1.35e-31(2.23e-47)=	1.35e-31(2.23e-47)=	1.35e-31(2.23e-47)
	SR/AVEN	100/90,558	100/36,362	100/33,206	100/22,180
f_{20}	Mean(std)	2.06e-02(2.35e-02)-	2.63e-05(1.32e-04)-	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)
	SR/AVEN	0/NA	96/80,696	100/72,506	100/28,025
f_{21}	Mean(std)	-78.332(0.00e+00)=	-78.332(1.23e-14)=	-78.332(8.70e-15)=	-78.332(4.61e-15)
	SR/AVEN	100/26,594	100/10,992	100/11,194	100/9530.0
f_{22}	Mean(std)	-29.999(6.36e-04)-	-30.000(1.92e-06)=	-30.000(0.00e+00)=	-30.000(0.00e+00)
	SR/AVEN	100/25,458	100/14,822	100/15,210	100/19,525
+/=/-		1/3/18	2/10/10	2/11/9	–

Table 3. The comparative results of EABC, ABCVSS, DFSABC_elite and IABC_elite when D=30.

No.	metric	EABC	ABCVSS	DFSABC_elite	IABC_elite
f_1	Mean(std)	5.85e-62(2.90e-61)-	2.40e-35(8.54e-35)-	4.14e-82(8.76e-82)-	2.20e-105(7.23e-105)
	SR/AVEN	100/27,982	100/50,526	100/21,410	100/19,617
f_2	Mean(std)	9.26e-60(1.41e-59)-	2.29e-27(9.79e-27)-	5.37e-78(8.66e-78)-	1.31e-102(5.39e-102)
	SR/AVEN	100/39,006	100/78,802	100/28,674	100/25,960
f_3	Mean(std)	4.50e-65(5.16e-65)-	9.40e-37(2.54e-36)-	2.84e-83(4.66e-83)-	2.45e-107(1.17e-106)
	SR/AVEN	100/25,826	100/46,222	100/19,710	100/18,615
f_4	Mean(std)	9.57e-33(3.42e-32)-	4.31e-44(1.40e-43)-	2.41e-110(1.19e-109)-	1.52e-168(1.52e-168)
	SR/AVEN	100/84,180	100/15,818	100/7122	100/6945
f_5	Mean(std)	9.45e-34(8.43e-34)-	7.03e-19(2.18e-18)-	2.06e-42(2.08e-42)-	9.04e-56(2.07e-55)
	SR/AVEN	100/42,198	100/72,958	100/33,426	100/30,280
f_6	Mean(std)	2.43e+01(5.22e+00)-	2.56e-01(9.19e-02)-	5.08e-07(3.69e-07)+	1.33e-02(1.07e-02)
	SR/AVEN	0/NA	100/111,070	100/32,802	100/68,600
f_7	Mean(std)	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)=	0.00e+00(0.00e+0)=	0.00e+00(0.00e+00)
	SR/AVEN	100/7602.0	100/10,042	100/7534	100/7450
f_8	Mean(std)	7.18e-66/(7.49e-67)=	7.18e-66(9.98e-78)=	7.18e-66(3.23e-81)=	7.18e-66(1.19e-81)
	SR/AVEN	100/150	100/150	100/150	100/150
f_9	Mean(std)	1.65e-02(3.68e-03)-	2.57e-02(5.22e-03)-	1.20e-02(3.80e-03)+	1.36e-02(3.70e-03)
	SR/AVEN	100/23,398	100/40,846	100/16,878	100/18,665
f_{10}	Mean(std)	1.14e+00(2.94e+00)-	3.25e-02(4.58e-02)+	3.45e+00(1.45e+01)-	5.6e-01(1.15e+00)
	SR/AVEN	100/85,233	96/86,483	60/58,683	70/65,792
f_{11}	Mean(std)	3.82e-02(1.91e-01)-	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)
	SR/AVEN	96/34,067	100/51,966	100/27,754	100/27,575
f_{12}	Mean(std)	1.20e-01(3.32e-01)-	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)
	SR/AVEN	88/36,005	100/60,578	100/28,602	100/30,175
f_{13}	Mean(std)	4.29e-08(2.14e-07)-	3.45e-11(1.73e-10)-	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)
	SR/AVEN	96/35,654)	100/69,514	100/31,066	100/30,087
f_{14}	Mean(std)	3.35e-12(8.60e-13)-	1.60e-12(3.45e-13)-	4.37e-13(1.09e-12)-	1.09e-13(3.25e-13)
	SR/AVEN	100/38,454	100/52,906	100/34,430	100/41,826
f_{15}	Mean(std)	2.73e-05(1.36e-04)-	6.50e-15(2.27e-15)=	3.80e-15(1.69e-15)+	5.52e-15(3.21e-15)
	SR/AVEN	96/49,888	100/80,074	100/37,998	100/35,210
f_{16}	Mean(std)	1.57e-32(5.59e-48)=	1.57e-32(5.59e-48)=	1.57e-32(5.59e-48)=	1.57e-32(3.42e-48)
	SR/AVEN	100/24,862	100/46,142	100/18,902	100/17,660
f_{17}	Mean(std)	1.50e-33(0.00e+00)=	1.50e-33(0.00e+00)=	1.50e-33(0.00e+00)=	1.50e-33(0.00e+00)
	SR/AVEN	100/22,540	100/48,154	100/20,970	100/19,055
f_{18}	Mean(std)	6.00e-17(3.41e-16)+	6.26e-18(2.91e-17)+	3.10e-40(1.03e-39)+	3.69e-16(8.23e-16)
	SR/AVEN	100/42,578	100/80,966	100/40,454	100/42,280
f_{19}	Mean(std)	1.35e-31(2.23e-47)=	1.35e-31(2.23e-47)=	1.35e-31(2.23e-47)=	1.35e-31(2.23e-47)
	SR/AVEN	100/26,762	100/48,330	100/24,890	100/22,180
f_{20}	Mean(std)	6.03e-03(1.30e-02)-	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)=	0.00e+00(0.00e+00)
	SR/AVEN	64/58,950	100/93,050	100/55,910	100/28,025
f_{21}	Mean(std)	-78.332(2.90e-15)=	-78.332(1.05e-14)=	-78.332(5.02e-15)=	-78.332(4.61e-15)
	SR/AVEN	100/8538.0	100/13,038	100/6502.0	100/9530.0
f_{22}	Mean(std)	-30.000(1.51e-06)=	-30.000(3.82e-12)=	-30.000(0.00e+00)=	-30.000(0.00e+00)
	SR/AVEN	100/12,602	100/18,726	100/5270.0	100/19,525
+/=/-		1/7/14	2/11/9	4/11/7	–

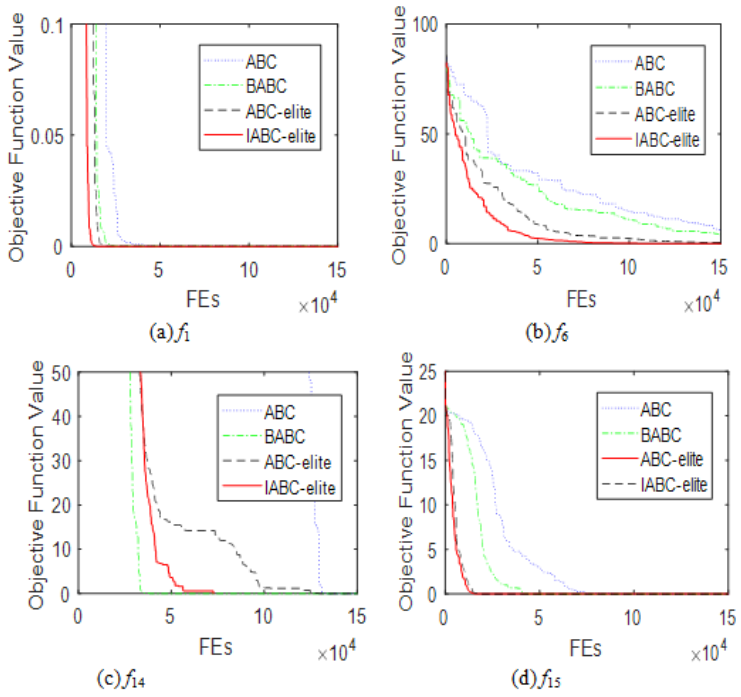


Fig. 2. The convergence curves of ABC, BABC, ABC_elite and IABC_elite on 4 representative test

5.3. Experiment 2: comparison of the IABC_elite and other ABC variants

In this section, in order to further evaluate the performance of IABC_elite, the IABC_elite is compared with 3 recently developed representative ABC variants, i.e.,the EABC [12], ABCVSS [23], DFSABC_elite [5] on all 22 test functions with $30D$. The parameter settings are shown in Table 1, and the termination condition max_FES is the same as experiment 1 ($max\_FES = 150000$). All the compared ABC variants have proposed an improved search equation. It's worth noting that the DFSABC_elite is a composite algorithm consisting of the ABC_elite and depth-first strategy (DFS). The comparative results are shown in Table 3.

(1) $The\ comparative\ results\ on\ unimodal\ functions$:

$f_1 - f_9$ are unimodal functions. For functions $f_1 - f_5$,According to Table 3, the IABC_elite performs significantly better than all compared algorithms regarding solution accuracy (mean(std)) and convergence speed (AVEN), and all algorithms obtain the same results in the success rate (SR). For functions $f_7 - f_8$, although all the algorithms get the similar performance regarding solution accuracy and success rate because $f_7 - f_8$ are easy to solve [5], the convergence speed of the IABC_elite is faster than or at least comparable to all the competitors. For functions f_6 and f_9, the IABC_elite is only second to the DFSABC_elite regarding solution accuracy and convergence speed, while IABC_elite exhibits best success rate, beating all its competitors. In a word, the IABC_elite shows the best overall performance in unimodal functions.

(2) $The\ comparative\ results\ on\ multimodal\ functions$:

$f_{10}-f_{22}$ are multimodal functions. f_{10} is Rosenbrock function and its global optimum is inside a long, narrow, parabolic shaped flat valley, the variables are strongly dependent, and the gradients do not generally point towards the optimum. If the population is guided by the global best solution or some other good solutions, the search will fall into some unpromising areas. Therefore, DFSABC_elite is beaten by all the competitors, even original ABC is also far better than DFSABC_elite in function f_{10}.This phenomenon reflects the defect of DFS strategy used in DFSABC_elite. Because the DFS strategy always search a direction greedily, it tends to result in lacking of randomness of EA and make it trapped into local optima. And the same conclusion can be drawn from literature [5] (see Table 3 of literature [5]). For function f_{10}, the IABC_elite is better than the DFSABC_elite and EABC, but still worse than ABCVSS slightly, regarding solution accuracy.

The last row of the Table 3 summarizes the comparison results. It can be seen that the IABC_elite exhibits significantly advantage when compared with other algorithms. In the comparison with the DFSABC_elite, IABC_elite wins over it in 7 functions, ties in 11 functions while losed on 4 functions regarding solution accuracy. Although the DFSABC_elite has combined with the DFS strategy, IABC_elite still outperform it. Similarly, the IABC_elite performs better than the EABC and ABCVSS on most of the test functions regarding solution accuracy.

Overall, the IABC_elite still performs better than all other algorithms on most of multimodal functions.

6. Conclusions

In order to increase the exploitation ability of the ABC_elite and seek a better balance between the abilities of exploration and exploitation, an improved ABC_elite (the IABC_elite) algorithm is put forward in this paper, combining two novel search equation and a new parameter with ABC_elite. The first search equation is used in employed bee phase, thus the elite solutions and ordinary solutions adopt different search equation. The second search equation is used in the onlooker bee phase to further enhance the exploitation of the ABC_elite. The new parameter P_o is introduced to maintain the balance between the ability of exploration and that of exploitation. The experiment results have shown that the IABC_elite can significantly improve the performance of ABC_elite. When further compared to other state-of-the-art ABC variants, IABC_elite also exhibits the best overall performance.

Acknowledgments. This work has been supported by the National Natural Science Foundation of China (No. 61373028 and No. 61672338).

References

1. B. Akay and D. Karaboga. A modified artificial bee colony algorithm for real-parameter optimization. *Information Sciences*, 192(12):120–142, 2012.
2. A. Banharnsakun, T. Achalakul, and B. Sirinaovakul. The best-so-far selection in artificial bee colony algorithm. *Applied Soft Computing*, 11(2):2888–2901, 2011.
3. J. C. Bansal and H. Sharma. Arya, k. v., et al.: Self-adaptive artificial bee colony. *Optimization*, 63(10):1513–1532, 2014.

4. D. Bose, S. Biswas, and A. V. Vasilakos. Optimal filter design using an improved artificial bee colony algorithm. *Information Sciences*, 281(20):443–461, 2014.
5. L. Cui, G. Li, and Q. Lin. A novel artificial bee colony algorithm with depth-first search framework and elite-guided search equation. *Information Sciences*, 367(22):1012–1044, 2016.
6. J. Derrac and S. Garca. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. volume 1 of *1*, pages 3–18. 2011.
7. R. Eberhart and J. Kennedy. A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS'95.*, pages 39–43. Proceedings of the Sixth International Symposium on.IEEE, 1995.
8. W. Gao, F. T. S. Chan, and L. Huang. Bare bones artificial bee colony algorithm with parameter adaptation and fitness-based neighborhood. *Information Sciences*, 316(18):180–200, 2015.
9. W. Gao and S. Liu. Improved artificial bee colony algorithm for global optimization. *Information Processing Letters*, 111(17):871–882, 2011.
10. W. Gao and S. Liu. A modified artificial bee colony algorithm. *Computers and Operations Research*, 39(3):687–697, 2012.
11. W. Gao, S. Liu, and L. Huang. A novel artificial bee colony algorithm based on modified search equation and orthogonal learning. *IEEE Transactions on Cybernetics*, 43(3):1011–1024, 2013.
12. W. Gao, S. Liu, and L. Huang. Enhancing artificial bee colony algorithm using more information-based search equations. *Information Sciences*, 270(12):112–133, 2014.
13. D. Karaboga. An idea based on honey bee swarm for numerical optimization, 2005. Ereiyes University.
14. D. Karaboga. A quick artificial bee colony (qabc) algorithm and its performance on optimization problems. *Applied Soft Computing*, 23(10):227–238, 2014.
15. J. Kennedy. Bare bones particle swarms. *SIS'03. Proceedings of the 2003 IEEE. IEEE*, (1):80–87, 2003.
16. J. Kennedy. Particle swarm optimization. *Encyclopedia of machine learning. Springer US*, pages 760–766, 2011.
17. G. Li, P. Niu, and X. Xiao. Development and investigation of efficient artificial bee colony algorithm for numerical function optimization. *Applied soft computing*, 12(1):320–332, 2012.
18. W. H. Lim and N. A. M. Isa. Two-layer particle swarm optimization with intelligent division of labor. *Engineering Applications of Artificial Intelligence*, 26(10):2327–2348, 2013.
19. W. H. Lim and N. A. M. Isa. An adaptive two-layer particle swarm optimization with elitist learning strategy. *Information Sciences*, 273(14):49–72, 2014.
20. W. H. Lim and N. A. M. Isa. Adaptive division of labor particle swarm optimization. *Expert Systems with Applications*, 42(14):5887–5903, 2015.
21. J. Luo, Q. Wang, and X. Xiao. A modified artificial bee colony algorithm based on converge-onlookers approach for global optimization. *Applied Mathematics and Computation*, 219(20):10253–10262, 2013.
22. Q. K. Pan, L. Wang, and J. Q. Li. A novel discrete artificial bee colony algorithm forthe hybrid flowshop scheduling problem with makespan minimisation. 45(6):42–56, 2014.
23. Kiran M S, Hakli H, and Gunduz M. Artificial bee colony algorithm with variable search strategy for continuous optimization. *Information Sciences*, 300(8):140–157, 2015.
24. A. Salman, M. G. H. Omran, and M. Clerc. Improving the performance of comprehensive learning particle swarm optimizer. *Journal of Intelligent and Fuzzy Systems*, 30(2):735–746, 2016.
25. Y. Shi and R. Eberhart. A modified particle swarm optimizer. In *Evolutionary Computation Proceedings*, pages 69–73. IEEE World Congress on Computational Intelligence.The 1998 IEEE International Conference on. IEEE, 1998.
26. W. Y. Szeto, Y. Wu, and S. C. Ho. An artificial bee colony algorithm for the capacitated vehicle routing problem. *European Journal of Operational Research*, 215(1):126–135, 2011.

27. H. Wang, Z. Wu, and S. Rahnamayan. Multi-strategy ensemble artificial bee colony algorithm. *Information Sciences*, 279(18):587–603, 2014.
28. G. Zhu and S. Kwong. Gbest-guided artificial bee colony algorithm for numerical function optimization. *Applied Mathematics and Computation*, 217(7):3166–3173, 2010.

Zhenxin Du received the B.S. degree in computer science from Zhejiang Sci-tech University in 2011. He is currently a Ph.D. candidate at Shanghai Maritime University. His main research interests include computation intelligence, data mining and cloud computing.

Dezhi Han (corresponding author) received the Ph.D. degree from Huazhong University of Science and Technology. He is currently a professor of computer science and engineering at Shanghai Maritime University. His research interests include cloud computing, mobile networking and cloud security.

Guangzhong Liu received the Ph.D. degree from China University of Mining and Technology. He is currently a professor of computer science and engineering at Shanghai Maritime University. His specific research interests include underwater acoustic communication technology, mobile networking and network security.

Kun Bi received the Ph.D. degree from University of Science and Technology of China. He is currently a lecture of computer science and engineering at Shanghai Maritime University. His research interests include network security, big data and cloud security.

Jianxin Jia received the M.S. degree from Shanghai Maritime University of computer science and engineering. He is currently pursuing the Ph.D. degree at Shanghai Maritime University. His main research interests include mobile networking, underwater acoustic communication technology and cloud security.

Received: January 2, 2017; Accepted: September 1, 2017.

Computer Science and Information Systems 14(3):769–788 https://doi.org/10.2298/CSIS161217028H

A DDoS Attack Detection System Based on Spark Framework

Dezhi Han, Kun Bi, Han Liu, and Jianxin Jia

College of Information Engineering,
Shanghai Maritime University,
Shanghai 201306, China
{dzhan, kunbi}@shmtu.edu.cn
{jmakg23,onlyoneman}@163.com

Abstract. There are many problems in traditional Distributed Denial of Service (DDoS) attack detection such as low accuracy, low detection speed and so on, which is not suitable for the real time detecting and processing of DDoS attacks in big data environment. This paper proposed a novel DDoS attack detection system based on Spark framework including 3 main algorithms. Based on information entropy, the first one can effectively warn all kinds of DDoS attacks in advance according to the information entropy change of data stream source IP address and destination IP address; With the help of designed dynamic sampling K-Means algorithm, this new detection system improves the attack detection accuracy effectively; Through running dynamic sampling K-Means parallelization algorithm, which can quickly and effectively detect a variety of DDoS attacks in big data environment. The experiment results show that this system can not only early warn DDoS attacks effectively, but also can detect all kinds of DDoS attacks in real time, with low false rate.

Keywords: Distributed Denial of Service (DDoS), Early Warn, Attack Detection, Spark framework, K-Means Algorithm.

1. Introduction

With the high-speed development of Internet, majority users has upgrade the bandwidth especially in some large cities, bandwidth of home users has reached 20M or even higher. Besides, with the popularization of 3G networks and gradual application of 4G networks, mobile internet has entered a booming stage. The rapid growth of private network bandwidth and continuously increasing internet users have posed enormous challenges for network security because the impact will be beyond measure once these high bandwidth network users are controlled by hackers and involved in DDoS (distributed denial of service).

It is indicated in the DDoS attack trend report [2] of Incapsula, a globally renowned CDN service provider, published in 2014 that DDoS attacks increased by 240% in 2014 and the traffic exceeded 100G. In addition, it is pointed out in an recently released analysis report [6] by the company that there are at present about tens of thousands or even millions of dedicated SOHO (small office home office) routers that have become part of BotNet and used by hackers to carry out large-scale DDoS attacks in the present. As found by the survey of losses due to DDoS attacks conducted in 2014 Incapsula, 49% of the DDoS attacks would last for 6 to 24 hours and average economic loss per hour is 40,000 dollars

[5]. During the latter half of 2015, Aliyun security team monitored a total of over 100,000 DDoS attacks, an increase of 32% as compared with that in the first half of 2015. Among them, attacks with a traffic exceeding 300Gbps amounted to 66 times, a rise of 127% than that in the first half of 2015 [7].

Network security incidents occurred frequently in the past two years. On January 21, 2014, DNSPod of Tecent got hijacked, resulting in DNS problems for a large number of domestic users. From December 20 to 21, 2014, a game company with its services deployed at Aliyun suffered DDoS attacks and the peak traffic of 453.8G/s made it the worlds biggest victim of DDoS attacks. On March 26, 2015, GitHub, a famous code hosting site, started to suffer a large scale of DDoS attack, it caused interruption of services in certain areas, and the attack was lasted for over 80 hours. On May 11, 2015, NetEase suffered a new DDoS attack, named, LFA (Link Flooding Attack) [5] which resulted in service interruption of 9 hours and loss of RMB 15 million yuan. Consequently, it is important both in theoretical significance and great economic value to research efficiently and promptly detect, warning, and manage large-traffic DDoS attacks.

In a big data background, highly efficient DDoS attack detection involves computation and processing of massive data, while traditional method of single machine takes much time and cannot meet actual demand. The new distributed stream-oriented computing framework (Spark Streaming) adopts the memory-based parallel computing method, which compared with the traditional computing method based on single-machine file system, significantly enhance the processing data quantity and processing data speed in unit time. Application of Spark Streaming to the real-time analysis system of big data flow network can accelerate the speed and accuracy of detection of DDoS attacks in a big data background.

In this paper, a novel DDoS attack system is proposed to detect DDoS attacks in a big data environment based on Spark framework, which includes 3 main algorithms. Based on information entropy, the first one can effectively warn all kinds of DDoS attacks in advance according to the information entropy change of data stream source IP address and destination IP address; With the help of designed dynamic sampling K-Means algorithm, this new detection system improves the attack detection accuracy effectively; Through dynamic sampling K-Means parallelization algorithm, which can quickly and effectively detect a variety of DDoS attacks in big data environment.The experimental results show that good warning results are obtained and the detection accuracy and speed are obviously superior than traditional DDoS attack detection methods.

The rest of this paper is organized as follows: Section 1 presents the working principle of Spark Streaming; Section 2 describes the DDoS attack warning algorithm design in detail. Section 3 presents the detailed design of improving K-Means parallel algorithm based on dynamics of Spark Streaming; Section 4 introduced the structure and major modules of the DDoS attack detection system. Section 5 presents the simulations and results of proposed DDoS attack detection system; Finally, we conclude this paper in Section 6.

2. Spark Streaming Working Principle

Spark [3], proposed in APMLab in University of California Berkeley, formally opened the source in 2010, became an Apache project in 2013 and a top level project of Apache in

2014. Spark offers solution for the problem of slow computation speed due to storage of intermediate results into the disc during calculation of Hadoop [4]. The ecological system of Spark includes batch processing, stream processing, machine learning, diagram calculating, data analyzing, etc. Compared to Hadoop ecosystem, it is a more comprehensive and suitable distributed computing framework used for big data application scenarios.

RDD [11] (Resilient Distribute Data sets) is not only the core of Spark but also the key for Spark to realizing failure recovery and data dependency. With the simple logic of Lineage, RDD can perfectly solve the dependency between data and data, guarantee good fault tolerance. The RDD can also store intermediate results into the memory which significantly improves the computation speed by reducing disc read and write to the minimum. Especially in iterative computation, the speed is increased by one order of magnitude.

Different from MapReduce in Hadoop, MapReduce of Spark is well packaged into RDD. The operation can be conducted with RDD into two types: transformation and action. The Data in RDD do not exist in their original forms but incorporated in RDD in the forms of their locations; then new RDD can be obtained through different transformation of the data in RDD and we can get the final result action when we perform to start the real calculation.

Spark Streaming [10] is a framework in Spark ecological system used for real-time calculation and its core is also based on RDD. Therefore, it can realize seamless connection with Spark to fuse historical data and real-time data perfectly. The features of Spark Streaming are as follows:

(1) Spark Streaming can realize complex processing logic with short simple codes. Its principle is to divide streaming data into small time intervals (e.g. several seconds), namely, to make the data discrete and transform them into data sets (RDD), then process the RDD in batches and conduct calculation on the RDD, thereby finishing the complex streaming data processing.

(2) Good fault tolerance: Spark Streaming has inherited the fault tolerance feature of RDD. If certain partitions of RDD is lost, computation can be restored based on the lineage information.

(3) Good universality: thanks to the design of RDD, Spark Streaming can realize seamless integration with other modules data of the Spark platform and combine real-time processing and batch processing.

(4) Spark Streaming has external data sources of various types which can be classified into the following two major categories: external file system data (such as HDFS data) and network system streaming data (such as streaming data collected by Kafka, ZeroMQ and Flume). The above features of Spark Streaming make it quite suitable for real-time data analysis against the background of big data.

3. DDoS Attack Early-Warning

It is of great significance to study the DDoS attack early-warning algorithm and early-warning, for they can process the early-warning of DDoS attack, especially in big data environment before DDoS attack do harm to the system, and they will save time for system by eliminating damages to the system caused by large-scale DDoS attack. In this paper,

DDoS attack is early-warned based on abnormal changes of source IP and destination IP information entropy of network data stream.

3.1. Traffic Information Entropy Feature

Entropy is an indicator of diversity and uniformity of the microscopic state which reflects the probability distribution of the system in the microscopic state. It can be seen from the perspective of communication that random interference in a system is unavoidable. Therefore, statistical methods can be adopted to describe characteristics of the communication system. To be specific, take the information source as a collection of random events whose probability of occurrence is similar to uncertainty in the microscopic state in thermodynamics; Calculating probability of occurrence in each information source in the information system to simulate the uncertainty of the system in thermodynamics, thus forming information entropy [12]. Information entropy has similar meaning to entropy in thermodynamics and it is an uncertainty indicator of the information system, which may indicate the amount of information in an information system.

Based on the network traffic information, entropy is defined as shown in Equation (1).

$$H(X) = E[-\log p_i] = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

In above Equation (1), X represents an information source symbol which has n values: $X_1 ... X_i ... X_n$, each value corresponding probabilities are: $P_1 ... P_i ... P_n$, since each source symbol appears independent of each other, so there comes to the equation:

$$\sum_{i=1}^{n} p_i = 1 \tag{2}$$

When DDoS attacks are launched, hundreds of bottled machines will send large streams of data packets to the target and the attacker, in order to hide its position, will randomly produce fake source IP addresses for the attacking packets or adopt more advanced reply flood DDoS attacks. In this case, the amount of requests for source IP addresses monitored by the server will drastically increases and the distribution will be more dispersed. Moreover, there will be a large amount of request flow flocking into certain service ports at the server side, and at the same time, the requests distribution for destination IP addresses which monitored by the server and the destination ports will become concentrated increasingly. When it occurs to the DDoS attacking, the information entropy of destination IP and source IP of the data flow that arrived the attacked server, which can reflect the uncertainty of system by calculating information entropy of destination IP and source IP, that also can be used for the DDoS attack warning in large-scale network traffic.

Fig. 1 and Fig. 2 are shown as the experimental and test conditions of the public server for the authors school network center. In the beginning of the first 100 seconds test time, the public servers to be tested will be attacked by traffic DDoS 30GB, which are issued by multiple clients in the laboratory. From the detecting results of the gateway to connect the public server, DDoS attack flow occurred in 100th seconds and it is detected by the system that the information entropy based on the destination IP and source IP occurs significant changes. The information entropy based on destination IP decreases rapidly, while the information entropy based on source IP increases rapidly. The result may certify that when the information entropy can better reflect the DDoS attack, the server receives

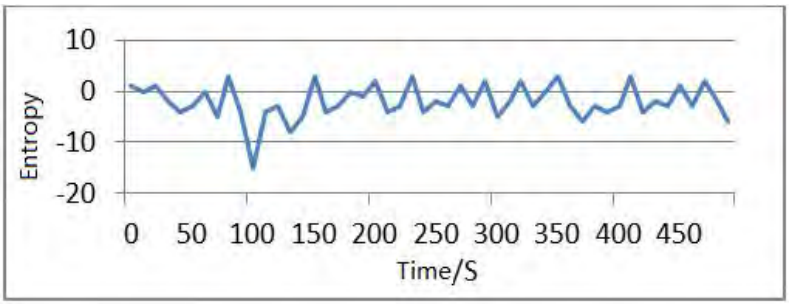


Fig. 1. Information entropy change based on the destination IP.

the uncertainty of the request change range, can be used for the early-warning of DDoS attacks.

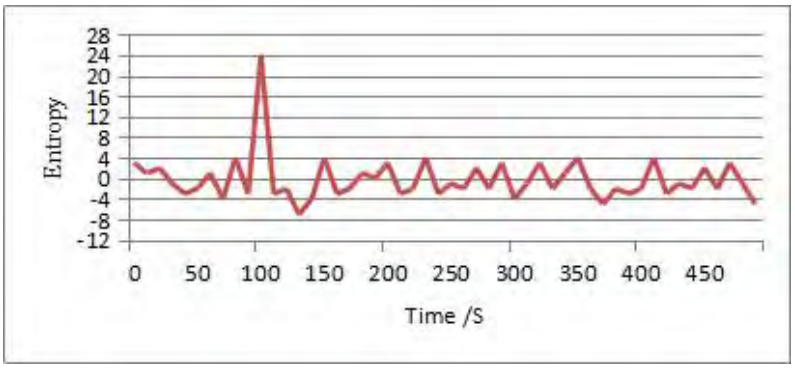


Fig. 2. Information entropy change based on the source IP.

3.2. Design of Network Traffic Model and Early-Warning Algorithm

When DDoS attack happens, the entropy of destination IP and source IP will change largely. Based on the characteristics, the network traffic model was defined. We can analysis the destination IP and sources IP feature on a given time windows based the model. So a DDoS attack early-warning algorithm that based on the information entropy is designed.

First, define a network traffic model, as shown in Fig.3. The traffic model mentioned in this paper includes two kinds of traffic entities, namely Normal (normal request flow) and DDoS (attack flow) under normal circumstances,The detection system collects all the traffic data at a certain timeΔt,and calculated the information entropy of the flow of Δt. Calculate the mean value of the formal flow information entropy of the first $n-1\Delta t$. Calculate the maximum information entropy and mean the difference between the values as an early-warning threshold. When DDoS attacks occur, In theΔt time, the information entropy will change greatly, when the difference of information entropy and the mean value

exceeds the early-warning threshold value, the systems may encounter DDoS attacks and send out the alarm.

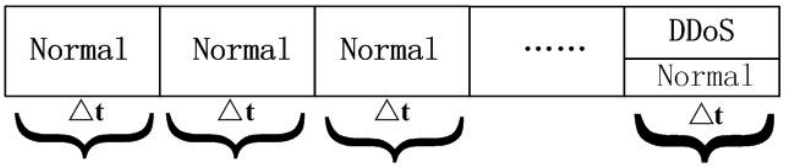


Fig. 3. The flow model.

DDoS attack early-warning algorithm is as follows:

Step 1: StatisticΔt time of all requests, n kinds of different purposes IP (source IP) recorded as X, the number of times per X appears as N.

Step 2: Calculate the probability P of the X emergence.

$$p_i = \frac{N_i}{\sum_{j=1}^{n} N_j} \tag{3}$$

Step 3: Calculate Δt time information entropy $H(X)$.

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{4}$$

Step 4: Calculate the mean value of information entropy of the first $(n-1)\Delta t$.

$$A = \frac{1}{n-1} \sum_{i=1}^{n-1} H(X_i) \tag{5}$$

Step 5: Calculated threshold V, k is the amplification factor, different network environment K value is different, the value is greater than or equal to 1.5.

$$V = (\mathrm{Max}[H(X)] - A) \times k \tag{6}$$

Step 6: Calculate the difference value in Δt, between information entropy and mean.

$$S = H(X) - A \tag{7}$$

Step 7: if $S >= V$, this means issue a DDoS attack alert, and the detection system will start calling DDoS detection module; if $S < V$, this means the entropy change in the normal range, and the network traffic is normal.

There are two key parameters in the network traffic model and early-warning algorithm:

(1) Δt settings, according to the characteristics of DDoS attacks, Δt can be set between 1–10 seconds. The smaller time requests the greater calculating amount when the attacks are detecting, and the detection and treatment effect of DDoS attacks are better;

(2) The calculation of early-warning threshold V, when calculating the V, the network flow and the peak period of network traffic should be fully considered. The key is to set the amplification factor K which can set the value between 1.5 and 2.2. According to experience, the value will automatically set to 2.

4. DDoS Attack Detection

Under the big data environment, traditional single-machine processing methods are not competent to solving the high-speed DDoS attacks because it will cost a great deal of time. The unique RDD internal access mechanism in Spark platform and support provided from Spark Streaming modules for real-time processing will effectively solve the attacking problems caused by DDoS attacks with huge and real-time data flow. Therefore, this paper utilizes K-Means clustering algorithm belonging to category of machine learning and data mining. Besides, improvements to K-Means proposed in this paper will make it suitable for dynamic sampling and parallelization environment, and make it be able to merge sufficiently with Spark Streaming modules of Spark platform. Thus it can adapt to detecting various high-speed DDoS attacks under the big data circumstance.

4.1. Data Preprocessing and Feature Extraction

Faced with a large amount of requested data, DDoS detecting system cannot perform machine learning determination. The data flow texts are produced from diverse networking protocols. However, the detecting algorithm, based on machine learning, requires entering feature vectors including fields with special meanings. Since the proper values must express features of relative requests efficiently and accurately, it is required to carry out pretreatment to request flows. In dimensions of time, space and protocol type, quantification of data flow can make machine recognize and process data. Because data flow of DDoS attack presents strong dependency, certain features describing total flow can be obtained by analyzing existing relationship between current link and before links. On the basis of the features, the thesis will adopt K-Means clustering algorithm to build detecting model of DDoS attack and design related algorithms. According to the features of data flow, the feature extract can be carried out from two parts. The first part is statistics analysis of links during past period t which have the same destination host as current link; the second part is statistics analysis of links during past period t which have same services as current link.

The traffic statistics based on the time are just statistics in the $T1$ time period of the connection, of which relationship refers to the relationship between the other connections in this period and the current connection. In the actual DDoS attack, attackers sometimes use slow attack methods to scan IP and ports.When slow attack scanning frequency is greater than t, the method of time-based traffic statistics cannot get contact between requests.

In this paper, we use a time window to statistics that, in the time window N a current connection with the previous N connection information and set connection information as a feature. According to the characteristics of the specific set of 10 characteristic variables, these characteristic value variables include as follows:

(1) $x1$ represents the number of the current connection with N connection with the same target host, and the value ranges from 0 to 255.

(2) $x2$ represents the number of the same services for the current connection and previous N connections with the same target host, and the value ranges from 0 to 255.

(3) $x3$ represents the ratio of the same service to the current connection and before the N connection has the same target host, and the value ranges from 0 to 1.

(4) $x4$ represents the ratio of the current connection to the previous N connection with the same target host different services, and the value ranges from 0 to 1.

(5) $x5$ represents the ratio of the current connection to the same source port of the previous N connection with the same target host, and the value ranges from 0 to 1.

(6) $x6$ represents the ratio of the same service to the same service as the previous N connection, which is the same as the host, and the value ranges from 0 to 1.

(7) $x7$ represents the ratio of SYN error in links with same as the destination host the same service between current links and the former N links, and the value ranges from 0 to 1.

(8) $x8$ represents the ratio of SYN error in links with same destination host between current links and the previous N links, and the value ranges from 0 to 1.

(9) $x9$ represents the ratio of REJ error in links with same destination host between current links and the previous N links, and the value ranges from 0 to 1.

(10) $x10$ represents the ratio of REJ error in links with same as the destination host the same service between current links and the previous N links, and the value ranges from 0 to 1. By pretreating and extracting of the characteristic value of the normal network data, it can be trained to detect K-Means clustering model and design K-Means clustering algorithms of DDoS attack detection.

4.2. K-Means Clustering

The detecting objective of DDoS attack is to distinguish normal access request flow from abnormal attack flow; in nature, it is a kind of cluster. K-Means is a classic type of objective function clustering algorithm of LAN prototype, which belongs to category of unsupervised learning. In 1967, it was firstly put forward by James MacQueen and then it was popularized in various machine learning fields. The core idea of the algorithm is as follows: firstly, to select k objects at random and every initial object shows the center or average value of a cluster. After successive traversal, distances from the surplus objects to centers of all clusters will be calculated. Then by the comparison of the distances, they will be distributed to center with the smallest distance and calculations of all centers will be performed again. Next repeat the process until the convergence of clustering criterion function. The algorithm flow chart is shown in Fig. 4. The detailed description of the algorithm is as follows:

Input: K, D (Initial sample data)

Output: K clustering centers

Step1: Data set D as the initial sample, the n-dimensional of each point: $d_j = \{x_1, x_2, x_3, ..., x_n\}$. Each one dimension represents a feature vector. Random selection of K objects as initial cluster centers from data set D, the cluster center set is denoted as K.

Step2: Calculate the distance from each point in the D to the K cluster center, according to the minimum, assign the point to the corresponding category, cluster centers corresponding data is denoted .Using Equation (8) to calculate the Euclidean distance.

$$D(k,d) = \sqrt{\sum\nolimits_{i=1}^{n} (x_{ki} - x_{di})^2}, k \in K, d \in D \tag{8}$$

Step3: Cluster center of updated cluster.

$$k = \frac{1}{n} \sum\nolimits_{i=1}^{n} c_i, c_i \in C, k_i \in K, n = \text{Size}(C_k) \tag{9}$$

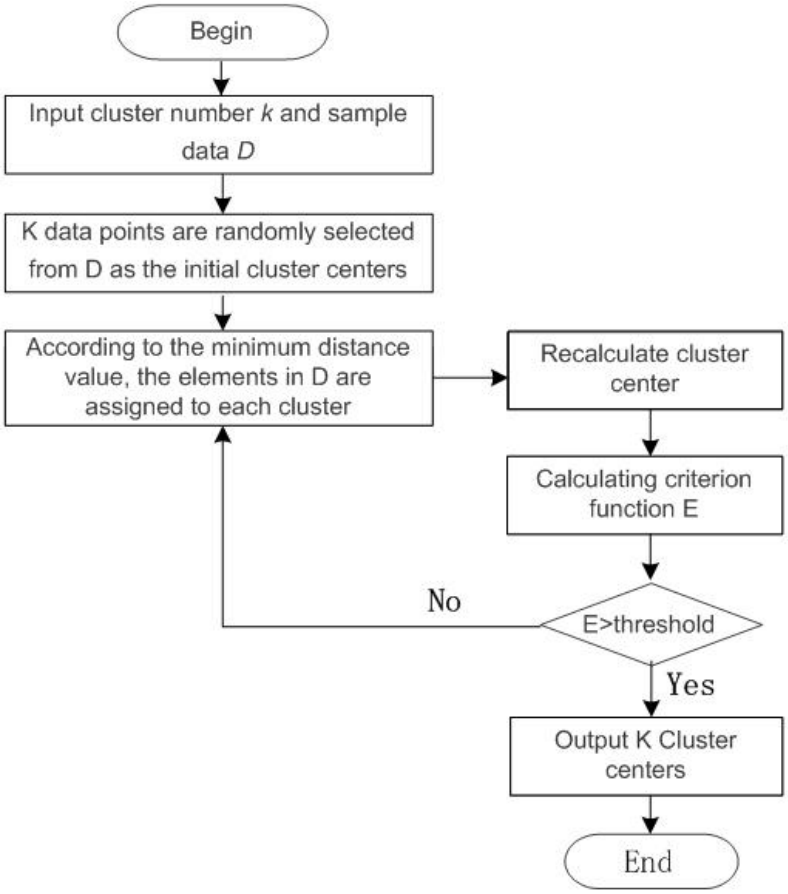


Fig. 4. K-Means flow

Step 4: Calculation criterion function.

$$E = \sum_{i=1}^{k} \sum_{c_j \in C_i} \left(c_j - k_i\right)^2, c_i \in C, k_i \in K \tag{10}$$

Step5: Meet the threshold criterion function exit, otherwise return to step 2.

4.3. Dynamic Sampling K-Means Algorithm

In common clustering algorithm, k points will be selected randomly as the center and the value of k as well as the selection of initial center will have direct influence on consequence of final cluster. If k is selected inappropriately, K-Means algorithm would converge on locally optimal solution, with the result that the correct result would not be obtained. In the DDoS attack detection system, K-Means requires to process a great deal of data mixing with attack flow, which leads to great difficulty to the selection of initial center. In order to solve the problem, dynamic sampling K-Means cluster will be employed to improve the algorithm to meet the demands of DDoS attack detecting system. The algorithm is shown Fig.5.

The main way to improve K-Means algorithm is to select only one point in advance as clustering center to build scale function. The function represents quadratic sum of distance between data point and its clustering center. Then, the clustering results will be converged by continuous iteration of minimum function value. The main theory for the improvement of K-Means algorithm is as follows: firstly, select a point from data set as initial clustering center and add it into dynamic sampling set C, which can be calculated by scale function, then perform circulation N times; secondly, select m points during each circulation and calculate sampling probability $P(X)$ .The meaning of the probability shows that clustering center is easy to be another center when it is more far away from original center because it is relatively disperse. In other words, the selected points should be far away from current clustering center. After iteration, the function value should be calculated again and it is required to update sampling probability for the next time. Afterwards, the overlaps between central point set C of sampling cluster and original sampling set C will act as new sampling set. After the N circulation, a new sampling set C will be produced which has several data. The scale of current data set is far smaller than that of original X and the data are relatively centralized due to the reason that they are filtered. Finally, the common K-Means algorithm of C will be performed and the process will be extremely fast because C is obtained after processing in advance. Meanwhile, the algorithm is improved in time complexity. It adopts the method of iteration replacing convergence threshold and reduces times of iteration, which is important to inspect DDoS attack under the environment of big data by machine learning method.

The specific algorithm is defined as follows:

Definition 1: The scale function $V(X)$ is defined as the formula (11). Where the $D^2(X, C)$ represents the square sum of the distance from the point in the X to the cluster center.

$$V(X) = \sqrt{\sum_{i=1}^{n} D^2(X, C)} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{d} \left(x_j - x_c\right)^2} \tag{11}$$

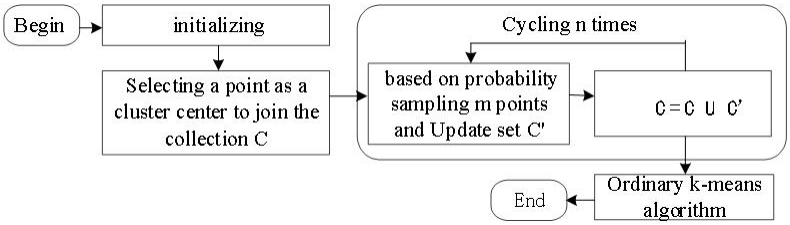


Fig. 5. Dynamic sampling improved K-Means algorithm flow.

Definition 2: The dynamic sampling probability function $P(X)$ is defined as the formula (12).

$$P(X) = \frac{D^2_{\min}(X, C)}{V(X)} \tag{12}$$

Definition 3: initial limited scale function value V, initial sample number $m < k$.

Specific algorithm is as follows:

Input: data set, K

Output: K clustering centers

Step 1: Randomly select one point from the set X to join in the set C.

Step 2: According to formula (11) to calculate the initial limited scale function value of the C, denoted as V.

Step 3: Cycle $\log V = N$, calculate the dynamic sampling probability $P(X)$ according to equation (12), recorded as P. Take out m points from the set X in accordance with the probability of P to join the collection C', calculate $C \bigcup C'$ and denoted by C, end of the cycle.

Step 4: Calculate the clustering center of the set C by using a common K-Means algorithm.

4.4. An Improved K-Means Algorithm for Dynamic Sampling Based on Spark

Ordinary DDoS attack detection algorithm cannot run directly on the Spark platform, according to the principle of Spark, the design of dynamic sampling and improved K-Means algorithm. The specific process is as follows:

(1) Algorithm begins, Master node program obtain the initial data set from the data input source, which is a predefined interface that can obtain data through a variety of ways, such as InputStream, HDFS, local files, etc., this design is convenient for the test of the algorithm. After obtaining the data, the system will convert the data to RDD1, and call the cache method to load the RDD1 to memory, the RDD will act as the data to be processed.

(2) Carry on the segmentation of data, to prepare for the parallelization. The system takes the block as a unit (64MB) to divide the RDD1 into several sub blocks. Then the master node calls the map method, and the large data blocks are allocated to multiple Worker nodes. When worker node receives the data blocks and executes the map instruction of Master, processing the data block. After this step, the String text of the original data set will be converted to DenseVector objects, which are the data that the program can

use directly; the distribution of the data is calculated on each Worker node. When the map method is finished, the RDD1 generates a new RDD2.

(3) Randomly select the initial cluster centers. The program calls the takeSample method, selects one of the RDD2 as the clustering center vector, and creates the RDD3 object.

(4) Begin to enter the cycle process, the program according to the 4.3 section of the implementation of the specific algorithm step 3 to carry out an iterative calculation. In each cycle, according to the definition 1 and definition 2 to recalculate the current sampling probability function P, and then call the takeSample method according to the probability P select the new RDD vector as the center point. After a cycle, the sampling total vectors are $1 + m$, and generate RDD4. Then the system calls the union method, the RDD3 and RDD4 merged into RDD5.

(5) After $\log V$ times will end the cycle, at this time, the number of vectors in the RDD5 is not more than $1 + m * \log V$. This amount is far less than the amount of initial data.

(6) The system will output RDD5 as a result.

The RDD conversion process of the entire sampling phase is shown in Fig.6. In Fig.6, the rounded rectangle frame represents the RDD; the straight rectangular box in the RDD represents the data fragmentation in the RDD, which is spread on a different Worker node; the direction of the arrow indicates the process of the RDD conversion.

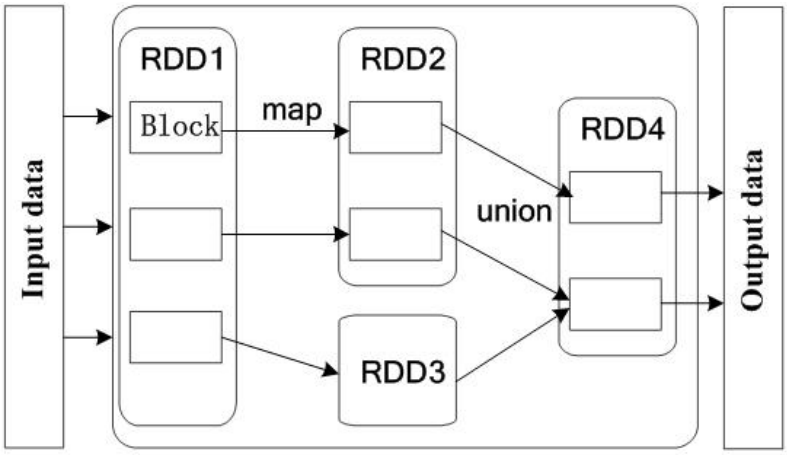


Fig. 6. RDD conversion process of sampling phase.

5. DDoS Attack Detection System

The software structure of DDoS attack detection system based on Spark is shown in Fig.7. The whole system is divided into four modules. These four modules are running on the nodes of Spark cluster and work together to complete the DDoS attack detection.

The detailed design of DDoS attack detection system is shown in Fig. 8. The whole system is running on a distributed cluster. It can not only make full use of Spark technology in management of distributed computing, but also improves the reliability and the processing speed of the system.

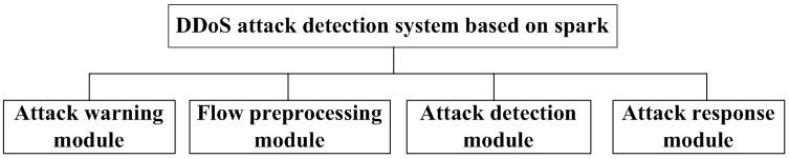


Fig. 7. Structure of DDoS attack detection system based on Spark.

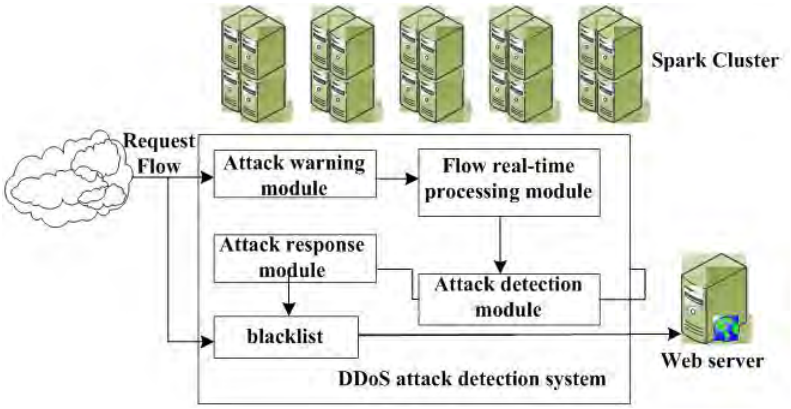


Fig. 8. Frame of DDoS attack detection system.

The system is comprised by attack warning, flow preprocessing, inspection and attack response modules. Attack warning module adopts the early-warning algorithm based on flow information entropy; flow real-time processing module mainly processes warning data flow section by section upon real-time processing framework called Spark Streaming, picking up relative characteristics and outputting the characteristic data to modules of attack detection in order to detect DDoS attacks; attack detecting module adopts DDoS detection algorithm similar to Spark, mainly receiving data from flow processing module, recognizing DDoS attacks according to clustering results and outputting results to module of attack response; the attack response module adds original IP address of DDoS flow detected by attack detecting module into blacklist, then the detecting system will filter attack flow in the list.

6. Experimental Results and Analysis

6.1. DDoS Attack Warning Algorithm Test

The test based on DDoS alarming algorithm of information entropy in the thesis, will be implemented on an e-commerce website. The website uses 3 nodes and each machine employs 8 core CPU with16GB internal storage and 1TB hard disk drive. In terms of software configuration, Spark 1.5.2 version is used to process big data and Java 1.8 version to compile program. The website has a quite large amount of information about access behaviors of users and daily records of access behavior reach over 16 million. The warning test period lasts a week from 8:00 am to 20:00 pm. DDoS attack uses software Autocrat [1] to perform SYN, LAND, FakePing and Furious Ping attack respectively. The test result is that the average warning rate reaches 98.5% while the average error warning rate is only 1.6%. For network server being in peak period, the error warning is mainly caused by various interferences.

6.2. DDoS Attack Detection Algorithm Test

In order to test the improved K-Means algorithm for dynamic sampling, a set of contrast tests under the situation of single-machine operation is designed in the thesis. Firstly, Java language programming is used to verify K-Means algorithm. Meanwhile, the algorithm is adopted to perform clustering analysis of test data and work out the required time for calculation and accuracy of clustering. Secondly, Java language programming is employed to verify the improved K-Means algorithm for dynamic sampling. Besides, the same test data is used to carry out clustering calculation to count the required time and accuracy. The test data is selected from training set with intensive kddcup-99 [8] data. Whats more, the data is also filtered and eight classical properties from the original properties are selected as properties of test data. Totally, 5 groups of data is selected and their data is respectively 10000 for group 1, 50000 for group 2, 100000 for group 3, 200000 for group 4 and 500000 group 5. Data for each group is different in figure but similar in distribution, which is used to perform a contrast test. The results of the test are shown in Fig. 9 and Fig.10.

From Fig.9 and Fig.10, we can see that when the amount of data is less, the dynamic sampling of the improved K-Means algorithm and the common K-Means algorithm is very close to the time. With the increase of the training set size, the advantage of the improved K-Means algorithm of dynamic sampling becomes increasingly distinct. In the case of 500 thousand data sets, the improved algorithm is obviously superior to the ordinary algorithm in time complexity. In terms of accurate rates, the improved K-Means algorithm of dynamic sampling is relatively close to ordinary K-Means algorithm. And the accurate rates in different test sets fluctuate but the fluctuation maintains in a relatively stable range.

In order to test the detection speed and accuracy of the proposed detection algorithm on the Spark cluster, the following experiments are designed: Using the KDD99 data set of the training works (5 million data) as the experimental samples, respectively, 5 groups of data are selected. These data are respectively 1 ten thousand for group 1, 50 ten thousand for group 2, 1 million for group 3, 2 million for group 4 and 5 million for group 5. And

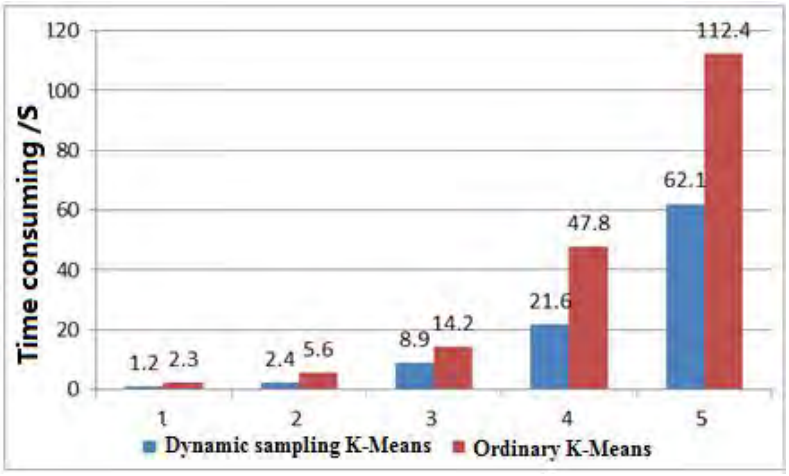


Fig. 9. Comparison of the time-consumption of two algorithms in five experiments.

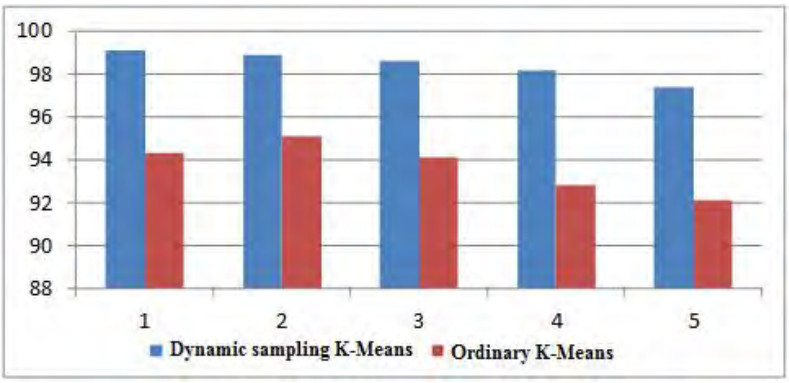


Fig. 10. Comparison of the accuracy rates of two algorithms in five experiments.

Spark cluster adopts 1.5.2 Spark version for the configuration of the software, while Java 1.8 version is used for the preparation of the Spark program.

Three experimental groups are designed in the experiment with the first experimental group using a single algorithm, serial processing of data samples; second experimental group using ordinary K-Means algorithm, parallel processing of data samples and the third experimental group using the improved K-Means algorithm.

Analysis is made on the time consumption, the average time of each round of iteration, and the correct rate in the three experimental groups. The final results are shown in Fig. 11 and Fig. 12.

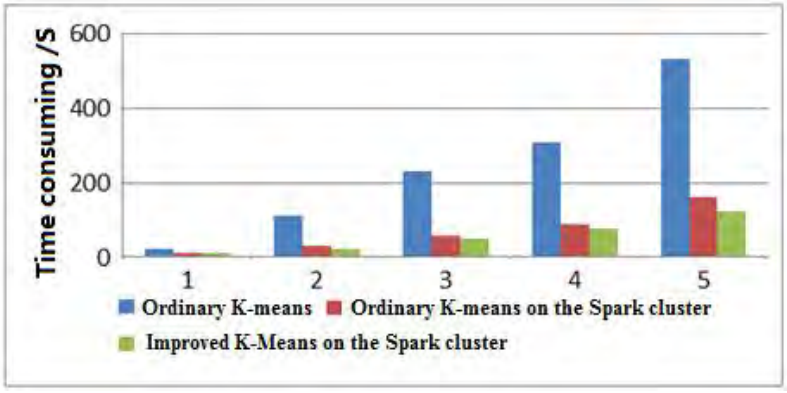


Fig. 11. Comparison of time-consumption for three experimental groups in five experiments.

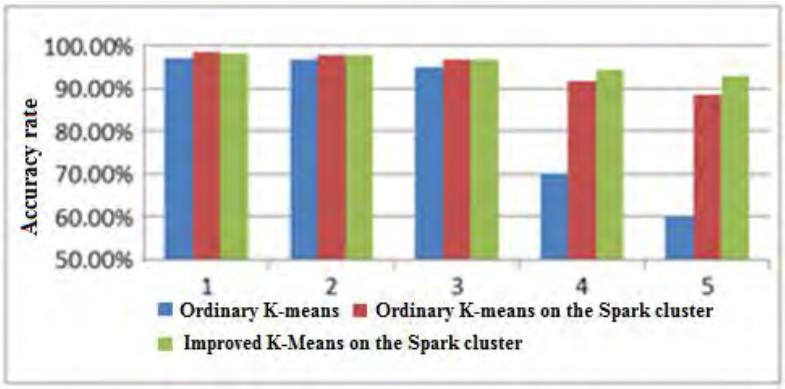


Fig. 12. Comparison of accurate rates for three experimental groups in five experiments.

Through the comparison of the above results, in the case of a small amount of data, it is found that the time difference between the three algorithms is not very large and the ac-

curacy of single machine operation is relatively high, and the advantage of Spark parallel computation is not obvious. When the amount of data is greater than 1million, the time to run a single machine increases dramatically while the accuracy of the data decreases rapidly. At the same time, the advantage of Spark parallel computing is very significant. Compared to the ordinary K-Means algorithm implemented on the Spark cluster, the improved K-Means algorithm has better accuracy and efficiency. This experiment can better reflect the advantage of parallel computing based Spark dynamic sampling platform to achieve an improved K-Means algorithm.

In order to test the system's ability to deal with DDoS attacks, this article through the open source software simulates the large data traffic DDoS attack [1], and starts the detection system to detect and address it. The Experimental design is to launch the attacks on the Web Service that has set up the DDoS attack detection system and the web service that did not build the DDoS attack detection system respectively. The actual impact of DDoS attacks on the server is determined by calculating the Web Service real-time throughput and CPU utilization rate. The final experimental statistics are shown in Fig. 13 and Fig. 14.

As is shown in Fig. 13 and Fig. 14, the throughput of the server increases rapidly after the DDoS attack within 100 seconds, after which, the throughput of the server in experimental group 1 without DDoS detection system falls sharply with CPU occupancy rate close to 100% whereas that of the server in experimental group 2 with DDoS detection system remains at normal level. Thus, it is proved that Web Service without detection system cannot continue to provide the normal service while the one with the detection system still can operate normally when confronted with DDoS attacks.

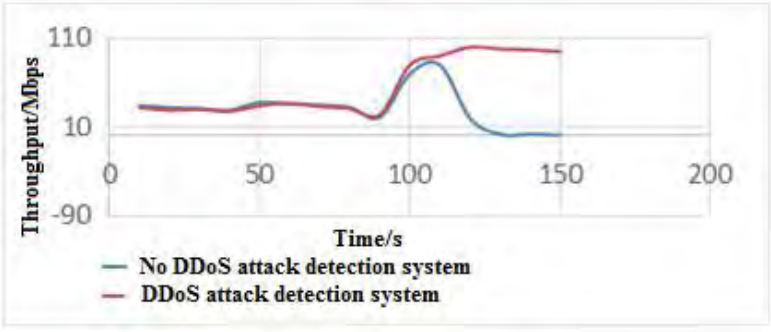


Fig. 13. Throughput Comparison.

6.3. Comparison with the classical DDoS detection method

In order to effectively analyze the performance of the proposed method, the simulation experiment is also used in the training of KDD 99 data sets (5 million data) as the experimental samples. 5 groups of data are selected and their data is respectively 1 ten thousand for group 1, 50 ten thousand for group 2, 1 million for group 3, 2 million for group 4 and 5 million for group 5. Three classical DDoS detection methods are selected after the

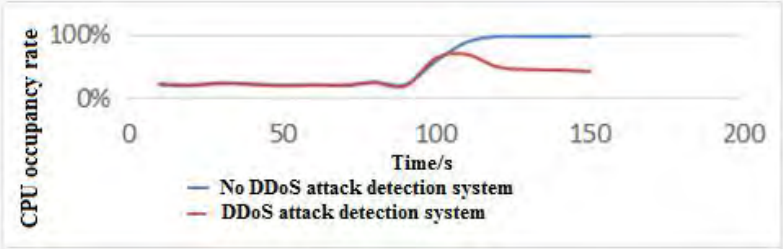


Fig. 14. Occupancy rate comparison.

experiment comparison [9], namely, DDoS attack detection method based on Hurst parameter(indicated by "DC1"), DDoS attack detection method based on nonlinear network flow analysis (DC2) and Wavelet analysis method based on adaptive detection of DDoS attacks ((indicated by "DC3"). By comparing these three classical algorithms with the dynamic improved K-Means method based on Spark in this paper (by "DC4") concerning the average response time, the average recognition rate, the average false rate, the results of the four methods are demonstrated in table 1. According to table 1, the method proposed in this paper is superior to the classical DC1, DC2 and DC3 methods in terms of average response time, average recognition rate, and average false rate.

Table 1. The performance comparison of DDoS detection algorithms.

Comparison Algorithm	DC1	DC2	DC3	DC4
Average response time	6.61	2.21	1.83	0.62
Average recognition rate	87.23 Apache15	93.12	91.64	98.3
Average false rate	3.52	2.13	2.25	1.5

7. Conclusions

In the big data environment, DDoS attacks are becoming one of the biggest threats to network security. Based on the existing researches, this paper designs a DDoS detection system based on Spark, to ensure accuracy in detection. In the meanwhile, the time for detecting DDoS attacks is reduced and the detection efficiency is improved significantly with the advantage of Spark technology.

In the future research work, the following aspects need to be improved:

(1) Spark Framework version iteration is very fast and each version will have new content and more powerful features. In the future research work, we should use the new features of the Spark framework flexibly to improve the efficiency of the system.

(2) For distributed systems, parameter setting is essential. In the future research work, we should do in-depth research in parameter tuning of the Spark framework to improve DDoS attack detection efficiency in big data condition.

(3) The limitation of this research is that it does not study much on tracking attackers in the DDoS detection. In order to prevent DDoS attacks more effectively, the method of investigating the legal liability of the attacker through internet forensics will be studied.

Acknowledgments. This work has been supported by the National Natural Science Foundation of China (No. 61373028 and No. 61672338).

References

1. Ddos attack using common tools (2013), [Online]. Available: http://www.bingdun.com/news/bingdun/8576.htm
2. Incapsula.report:2014 ddos trends-botnet activity is up by 240% (2014), [Online]. Available: https://w-ww.incapsula.com/blog/ddos-threat-landscape-report-2014.html
3. Apache software foundation. apache spark-lightning-fast cluster computing (2015), [Online]. Available: http://spark.apache.org/
4. Apache software foundation. welcome to apache hadoop (2015), [Online]. Available: http://hadoop.apache.org/
5. Incapsula.ddos impact survey reveals the actual cost of ddos attacks (2015), [Online]. Available: htt-ps://www.incapsula.com/blog/ddos-impact-cost-of-ddos-attack.html
6. Incapsula.lax security opens the door for mass-scale abuse of soho routers (2015), [Online]. Available: https://www.incapsula.com/blog/ddos-botnet-soho-router.html
7. Cloud shield internet ddos state and trend report in the second half of 2015 (2016), [Online]. Available: http://wenku.baidu.com/view/747d352f0c22590103029d6f
8. Hettich, S., Bay, S.D.: Kdd cup 1999 data (1999), [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup-99/kddcup99.html
9. Peng, Y., Yan, L.: Approach of ddos attacks prediction and detection with heqpso-svm algorithm based on date center network. Journal of Chinese Computer Systems 36(1), 150–163 (2015)
10. Zaharia, M., Das, T., Li, H., Shenker, S.: Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters. In: HotCloud. pp. 141–146. ACM (2012)
11. Zaharia, M.: Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. in-memory cluster computing. In: USENIX Symposium on Networked Systems Design and Implementation (NSDI), pp. 141–146 (2011)
12. Zhao, H.: DDOS anomaly detection technology research based on the information entropy clustering. Ph.D. thesis, Central South University, Changsha, China (2010)

Dezhi Han (corresponding author) received the Ph.D. degree from Huazhong University of Science and Technology. He is currently a professor of computer science and engineering at Shanghai Maritime University. His research interests include cloud computing, mobile networking and cloud security.

Kun Bi received the Ph.D. degree from University of Science and Technology of China. He is currently a lecture of computer science and engineering at Shanghai Maritime University. His research interests include network security, big data and cloud security.

Han Liu received the M.S. degree from Shanghai Maritime University of computer science and engineering. He is currently a Ph.D. candidate at Shanghai Maritime University. His research interests include cloud computing and cloud security.

Jianxin Jia received the M.S. degree from Shanghai Maritime University of computer science and engineering. He is currently pursuing the Ph.D. degree at Shanghai Maritime University. His main research interests include mobile networking, underwater acoustic communication technology and cloud security.

Received: December 17, 2016; Accepted: August 20, 2017.

Computer Science and Information Systems 14(3):789–804 https://doi.org/10.2298/CSIS170107029Z

A kernel based true online Sarsa(λ) for continuous space control problems

Fei Zhu[1,2], Haijun Zhu[1], Yuchen Fu[3], Donghuo Chen[1], and Xiaoke Zhou[4]

[1] School of Computer Science and Technology, Soochow University
Shizi Street No.1 158 box, 215006, Suzhou, Jiangsu, China
zhufei@suda.edu.cn, 1017942265@qq.com, dhchen@suda.edu.cn
[2] Provincial Key Laboratory for Computer Information Processing Technology, Soochow University
Shizi Street No.1 158 box, 215006, Suzhou, Jiangsu, China
[3] School of Computer Science and Engineering, Changshu Institute of Technology
yuchenfu@suda.edu.cn
[4] University of Basque Country, Spanish
xzhou001@ikasle.ehu.eus

Abstract. Reinforcement learning is an efficient learning method for the control problem by interacting with the environment to get an optimal policy. However, it also faces challenges such as low convergence accuracy and slow convergence. Moreover, conventional reinforcement learning algorithms could hardly solve continuous control problems. The kernel-based method can accelerate convergence speed and improve convergence accuracy; and the policy gradient method is a good way to deal with continuous space problems. We proposed a Sarsa(λ) version of true online time difference algorithm, named True Online Sarsa(λ)(TOSarsa(λ)), on the basis of the clustering-based sample specification method and selective kernel-based value function. The TOSarsa(λ) algorithm has a consistent result with both the forward view and the backward view which ensures to get an optimal policy in less time. Afterwards we also combined TOSarsa(λ) with heuristic dynamic programming. The experiments showed our proposed algorithm worked well in dealing with continuous control problem.

Keywords: reinforcement learning, kernel method, true online, policy gradient, Sarsa(λ).

1. Introduction

Reinforcement learning (RL) is an extremely important class of machine learning algorithm [15]. The agent of reinforcement learning keeps continuous interaction with the unknown environment, and receives feedback, usually called reward, from the environment to improve the behavior of agents so as to form an optimal policy [8]. Reinforcement learning maps the state of the environment to the action of the agent: the agent selects an action, the state changes, and the environment gives an immediate reward as an excitation signal. The goal of intensive learning is to get a maximum long-term cumulative reward from the environment, called return. As a kind highly versatile machine learning framework, reinforcement learning has been extensively studied and applied in many domains, especially in control tasks [14][1][19][6].

In many practical applications, the tasks that have to be solved are often with continuous space problems, where both the state space and the action space are continuous. Most common methods of solving continuous space problems include value function methods [13] and policy search methods [2]. The policy gradient method [16] is a typical policy search algorithm which updates policy parameters in the direction of maximal long-term cumulative reward or the average reward and gets optimal policy distribution. The policy gradient method has two parts: policy evaluation and policy improvement. Reinforcement learning has many fundamental algorithms for policy evaluation is concerned, such as value iteration, policy iteration, Monte Carlo and the time difference method (TD) [9] where the time difference method is an efficient strategy evaluation algorithm. Both the value function in the policy evaluation and the policy function in the policy improvement require function approximation [3]. The policy evaluation and policy improvement of the policy gradient method can be further summarized as the value function approximation and the policy function approximation. In reinforcement learning algorithms, the approximation of the function can be divided into parametric function approximation where the approximator and the number of parameters need to be predefined, and nonparametric function approximation where the approximator and the number of parameters are determined by samples. So nonparametric function approximation has high flexibility, and has better generalization performance. Gaussian function approximation and kernel-based method are nonparametric function approximation methods.

Although conventional reinforcement learning algorithms can deal with online learning problems, most of them have low convergence accuracy and slow convergence speed. The kernel based method is nonparametric function approximation method, and its approximation value function or strategy can alleviate the above problem of reinforcement learning. The policy gradient is an efficient way to deal with continuous space problems. In this paper, we propose an online algorithm that is based on kernel-based policy gradient method to solve continuous space problem. In the Section 2, we introduce the related work, including Markov decision process, reinforcement learning, and policy gradient; in the Section 3, we state how forward view matches backward view; in the Section 4, we introduce a true online time difference algorithm, named TOSarsa(λ); in the Section 5, we combine TOSarsa(λ) with heuristic dynamic programming.

2. Related Work

2.1. Markov Decision Process

Markov Decision Process (MDP) [5] is one of the most influential concepts in reinforcement learning. Markovian property refers that the development of a random process has nothing to do with the history of observation and is only determined by the current state. The state transition probability with a Markovian stochastic process [12] is called the Markov process. By Markov process, a decision is made in accordance with the current state and the action set, affecting the next state of the system, and the successive decision will be determined with the new state.

Normally a Markov Decision Process model can be represented by a tuple $M = < S, A, f, r >$, where:

S is the state space, and $s_t \in S$ denotes the state of the agent at time t;

A is the action space, and $a_t \in A$ denotes the action taken by the agent at time t;

f: $S \times A \longrightarrow$ [0,1] is the state transfer function, and f is usually formalized as the probability of the agent taking action $a_t \in A$ and transferring from the current state $s_t \in S$ to the next state $s_{t+1} \in S$;

ρ: $S \times A \longrightarrow$ R is the reword function which is received when the agent takes action $a_t \in A$ at the state $s_t \in S$ and the state transfers to the next state $s_{t+1} \in S$.

A Markov decision process is often used to model the reinforcement learning problem.

2.2. Reinforcement Learning

Reinforcement learning is based on the idea that the system learns directly from the interaction during the process of approaching the goals. The reinforcement learning framework has five fundamental elements: agent, environment, state, reward, and action, showed as Fig. 1. In the reinforcement learning, an agent, which is also known as a controller, keeps interaction with the environment, generates a state $s_t \in S$, and chooses an action $a_t \in A$ in accordance with a predetermined policy π such that at $a_t = \pi(s_t)$. Consequently, the agent receive an immediate reward $r_{t+1}=\rho(s_t, a_t)$ and gets to a new state s_{t+1}. By continuous trails and optimizing, the agent gets the maximal sum of the rewards as well as an optimal action sequence.

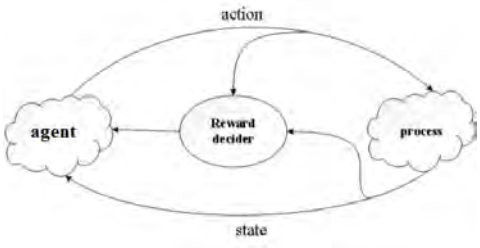


Fig. 1. Framework of reinforcement learning. The agent selects an action; the environment responds to the action, generates new scenes to the agent, and then returns a reward.

The goal of reinforcement learning is to maximize a long-term reward R which is calculated by:

$$\begin{aligned} R &= \mathrm{E}^{\pi}\left\{r_1 + \gamma r_2 + \cdots + \gamma^{T-1} r_T + \cdots\right\} \\ &= \mathrm{E}^{\pi}\left\{\sum_{t=1}^{\infty} \gamma^{t-1} r_t\right\} \end{aligned} \tag{1}$$

where E^{π} is expectation of accumulation of the long term reward, and $\gamma \in$ (0,1] is a discount factor increasing uncertainty on future rewards showing how far sighted the controller is in considering the rewards.

Reinforcement learning algorithms use state value function $V(s)$ to represent the expected rewards of state s under policy π. The value function $V(s)$ is defined as [15]:

$$
\begin{aligned}
V(s) &= \mathrm{E}^{\pi}\left\{\sum_{i=1}^{\infty}\gamma^{i-1}r_{t+i}|s_t = s\right\} \\
&= \mathrm{E}^{\pi}\left\{r_{t+1} + \gamma\sum_{i=1}^{\infty}\gamma^{i-1}r_{t+i+1}|s_t = s\right\} \\
&= \sum_{t=1}^{\infty}\gamma_t r(s_t)
\end{aligned} \tag{2}
$$

Reinforcement learning algorithms also use state state-action function $Q(s,a)$ which represents the accumulated long-term reward from a starting state. State-action function $Q(s,a)$ is defined as[15]:

$$
\begin{aligned}
Q(s,a) &= \mathrm{E}^{\pi}\left\{\sum_{i=1}^{\infty}\gamma^{i-1}r_{t+i}|s_t = s, a_t = a\right\} \\
&= \mathrm{E}^{\pi}\left\{r_{t+1} + \gamma\sum_{i=1}^{\infty}\gamma^{i-1}r_{t+i+1}|s_t = s, a_t = a\right\} \\
&= \sum_{t=1}^{\infty}\gamma_t r(s_t, a_t)
\end{aligned} \tag{3}
$$

Despite that the state value function $V(s)$ and the state action value function $Q(s, a)$ represent long-term returns, they still can be expressed in a form that is relevant to the MDP model and the successive state or state action pair, called one step dynamic. In this way, it is not necessary to wait for the end of the episode to calculate the value of the corresponding value function, but to update the new value function in each step, so that the algorithm has the ability real-time online learning. In addition, the state value function and the state action value function also can be expressed as:

$$
V(s) = \int_{a\in A}\pi(a|s)\,Q(s,a)da \tag{4}
$$

$$
Q(s,a) = \int_{s'\in S} f(s'|s,a)\left[R(s,a,s') + \gamma V(s')\right]ds \tag{5}
$$

As we can see that in the case where the environment model is completely known, the state value function and the state action value function can be transferred to each other seamlessly.

2.3. Policy Gradient

The reinforcement learning method can be categorized as the value function method and the policy gradient method. The typical value function methods include value iteration,

the policy iteration, the Q learning [10], Sarsa algorithm [11] and LSPI algorithm [7]. The policy iteration algorithm computes the optimal policy by repeating policy evaluating and policy improving. The value function method is a generalized iterative algorithm, focusing on the solution of the state action of the value function, and then the strategy is calculated by the value function, commonly by greedy strategy. Unlike the value function method, the policy gradient method represents the strategy directly through a set of policy parameters, rather than indirectly through the value function. The policy gradient method maximizes the cumulative reward function or the average reward by the gradient method to find out the optimal policy parameters, and each update is along the fastest rising direction of the reward function. The updates of policy parameters can be denoted as:

$$\psi = \psi + \alpha \frac{\partial Q(s, a_\psi)}{\partial u_\psi} \frac{\partial u_\psi}{\partial \psi} \tag{6}$$

$$\psi = \psi + \alpha \frac{\partial R}{\partial \psi} \tag{7}$$

The gradient becomes zero when the reward function reaches the local optimal point. The core of the policy gradient method update is the solution of the gradient.

The updates of policy parameters in the policy gradient method can be categorized as deterministic policy and non-deterministic policy. A deterministic policy is a greedy strategy that can deal with continuous action space problems. Because reinforcement learning requires action exploration, deterministic policy cannot be applied individually to reinforcement learning, often with some other method such as ε-greedy method. The non-deterministic policy gradient can solve both discrete and continuous space problems, just being provided with strategy distribution in advance. The Gibbs distribution is often used for discrete space problems as:

$$\pi\left(a|s\right) = \frac{e^{\kappa(s,a)^{\mathrm{T}}\psi}}{\sum\limits_{a' \in A} e^{\kappa(u,a')^{\mathrm{T}}\psi}} \tag{8}$$

While continuous space problem often takes advantage of Gaussian distribution, as:

$$\pi(a|s) = \frac{1}{\sqrt{2\pi\sigma^2(s)}} \exp\left(-\frac{\left(a - \mu(s)\right)^2}{2\sigma^2(s)}\right) \tag{9}$$

$$\mu(s) = \kappa_\mu^\top(s)\psi_\mu \tag{10}$$

$$\sigma(s) = \kappa_\sigma^\top(s)\kappa_\sigma \tag{11}$$

where κ(*s*,*a*) is the kernel of the state action pair (*s*, *a*), μ (s) is the mean value of the Gaussian distribution, $\sigma(s)$ is the standard deviation of the Gaussian distribution, $\psi = \left(\psi_\mu^\top, \psi_\sigma^\top\right)^\top$ is the parameter vector, and $\kappa(s) = \left(\kappa_\mu^\top(s), \kappa_\sigma^\top(s)\right)^\top$ is the kernel vector.

However, policy gradient algorithms are often suffered from the disadvantage brought by large gradient variance, which will affect the algorithm learning speed and convergence

performance. Therefore, in practice, the natural gradient method is used to replace the gradient method, so as to reduce the variance of the gradient, speed up the convergence rate of the algorithm and improve the convergence performance of the algorithm.

3. Forward View and Backward View

As the most important part of the reinforcement learning method, the time difference (TD) method is an effective method to solve the long-term forecasting problem. However, the traditional TD methods have problems in matching forward view and backward view. In this section, we will state how to make the forward view equivalent to backward view, which is a very important foundation of the proposed algorithms.

3.1. Time Difference (TD)

TD method is one of the core algorithms of reinforcement learning. TD method, which is able to learn directly from the raw experience from an unknown environment and update the value function at any time without determining dynamic model of environment in advance. Temporal difference combines the advantages of Monte Carlo method and dynamic programming. It updates the model by estimation based on part of learning rather than final results of the learning. Temporal difference works very well in dealing with real time prediction problems and control problems. Temporal difference learning updates by [15]:

$$V\left(s_{t+1}\right) \leftarrow V\left(s_t\right) + \alpha\left[R_t - V\left(s_t\right)\right] \tag{12}$$

$$Q\left(s_{t+1}, a\right) \leftarrow Q\left(s_t, a\right) + \alpha\left[R_t - Q\left(s_t, a\right)\right] \tag{13}$$

where R_t is return of step t, α is a step size parameter. Temporal difference learning updates V or Q in step $t+1$ using the observed reward r_{t+1} and estimated $V(s_{t+1})$ $Q(s_{t+1}, a_{t+1})$or .

One simple form of time difference algorithm, TD(0), updates the value function using the estimated deviation of a state s at the two time points, before and after. As TD (0) algorithm updates the value function every step, rather than after all steps, the entire update process does not require environment information as many other algorithms do. This advantage of TD(0) algorithm makes it suitable for the online learning task under the unknown environment. In addition, as the value function updating of TD(0) doesn't need to wait until the end of the episode, TD(0) can actually be used for non-episodic tasks, which sharply widens its application range compared to the Monte Carlo algorithm. The TD (0) algorithm is a method of evaluating the strategy. The Q learning algorithm and Sarsa algorithm are the two forms of TD(0).

3.2. TD(λ)

Inspired by the Monte Carlo algorithm, researchers introduced the idea of n-step updating and applied it to time difference. The update of the current value function that is based

on the next state value function and the immediate reward is called a one-step update. Likewise, it is referred as *n*-step update if the update is based on the next *n* steps. The *n*-step update can be defined as:

$$R_{t,\omega}^{(n)} = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n \omega^\top \kappa\left(s_{t+n}\right) \tag{14}$$

As the *V* function value of the current state *s* has a variety of estimates, in the process of algorithm implementation, we often uses weighted average of the different *n* steps, which is called λ-return:

$$R_t^\lambda = (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} R_{t,\omega}^{(n)} + \lambda^{T-t-1} R_{t,\omega}^{(T-t)} \tag{15}$$

where λ is regarded as recession factor, and *T* is the maximum number of steps. It is called λ-return algorithm when using λ-return to update the current state value function:

$$\omega_{t+1} = \omega_t + \alpha \left[R_t^\lambda - \omega^\top \kappa\left(s_t\right)\right] \kappa(s_t) \tag{16}$$

In the reinforcement learning, the above stated view is called forward view. The λ-return algorithm cannot update value function until the end of the episode. Therefore, λ-return algorithm uses the backward view to update the value function, which employs current TD error to update the value function of all states.

TD(λ) introduced the concept of the eligibility trace in backward view. The eligibility trace is essentially a record of the state or state of action recently visited. The cumulative eligibility trace can be defined as:

$$\mathbf{e}_t = \lambda\gamma \mathbf{e}_{t-1} + \kappa(s_t) \tag{17}$$

In the backward view, the TD error δ_t is updated according to the eligibility trace for all state values, as:

$$\omega_{t+1} = \omega_t + \alpha \delta_t \mathbf{e_t} \tag{18}$$

The conventional forward calculates λ return R_t^λ until the end of episode, while the online forward view method is able to calculate λ return at time *t*. This is called a truncated return, as:

$$R_t^{\lambda|t'} = (1-\lambda) \sum_{n=1}^{t'-t-1} \lambda^{n-1} R_{t,\omega_{t+n-1}}^{(n)} + \lambda^{t'-t-1} R_{t,\omega_{t'-1}}^{(t'-t)} \tag{19}$$

4. TOSarsa(λ) Algorithm

In the previous section, we have introduced how to achieve equivalence between forward view and backward view as well as its benefit of doing so. In this section, we will introduce a true online time difference algorithm which uses a clustering-based sample sparsification method [20] and selective kernel-based value function [4] as value function representation.

4.1. TOSarsa(λ) Algorithm Description

The true online time difference algorithm, named True Online State-action-reward-state-action(λ) (TOSarsa(λ)) is based on the effective Sarsa(λ) algorithm and uses Equation (17) as basic form of update equation to calculate eligibility trace, Equation (18) to calculate TD error and Equation (19) to calculate return.

Algorithm 1 True Online State-action-reward-state-action(λ) (TOSarsa(λ))

Input: policy, threshold
Output: optimal policy

Initialize kernel function $\kappa(\cdot,\cdot)$
Initialize sample set S
Set up data dictionary $\boldsymbol{D}$
repeat
  Initialize starting state s_0
  Initialize eligibility trace $\mathbf{e} \leftarrow 0$
  $V(s) \leftarrow \omega^\top \kappa(s)$
  repeat
    $V(s_{t+1}) \leftarrow \omega^\top \kappa(s_{t+1})$
    a $\leftarrow \pi(a|s)$
    Observe r, s
    $\delta_t \leftarrow r_{t+1} + \gamma \omega_t^\top \kappa(s_{t+1}) - \omega_{t-1}^\top \kappa(s_t)$
    $\mathbf{e}_t \leftarrow \gamma\lambda \mathbf{e}_{t-1} + \alpha_t \kappa(s_t) - \alpha_t \gamma\lambda \left[\mathbf{e}_{t-1}^\top \kappa(s_t)\right] \kappa(s_t)$
    $\omega_{t+1} \leftarrow \omega_t + \delta_t \mathbf{e}_t + \alpha_t \left[\omega_{t-1}^\top \kappa(s_t) - \omega_t^\top \kappa(s_t)\right] \kappa(s_t)$
    $\xi \leftarrow \min_{s_i \in D} \left(\kappa(s,s) + \kappa(s_i,s_i) - 2\kappa(s,s_i)\right)$
    Update $\boldsymbol{D}$
    if ξ is greater than a predefined threshold **then**
      $V(s_t) \leftarrow \omega^\top \kappa(s_{t+1})$
      Get ω and **e**
    else
      $V(s_t) \leftarrow V(s_{t+1})$
    end if
    $s_t \leftarrow s_{t+1}$
  until all step of the current episode end
until all episodes end
return optimal policy

4.2. Mountain Car Problem

Mountain car problem [18] is a classic problem in strengthening learning, as shown in Fig. 2. The task of the car is to get to the top of the mountain, the right side of the "star" mark position, as soon as possible. However, as the car is short of power, it is unable to drive to the top of the mountain directly. It has to accelerate back and forth many times to reach a higher position, and then accelerated to reach the end.

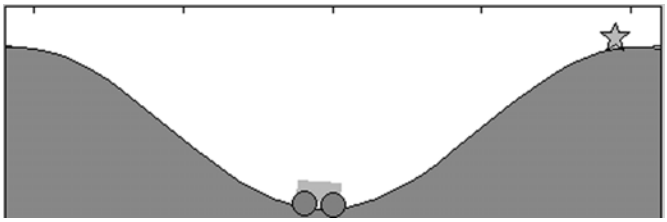


Fig. 2. Diagram of mountain car problem. The task of the car is to get to the top of the mountain, the right side of the "star" mark position, as soon as possible.

We use MDP to model mountain car problem. In the mountain car problem, the state contains two dimensions, the position denoted by p and the speed denoted by *v*. Then state of the car can be represented by a vector $\mathbf{x} = \begin{bmatrix} p \\ v \end{bmatrix}$. The acceleration of the car is in the range of -1 to 1, that is, the action $a \in$ [-1,1]. The curve of the road surface can be expressed by the function

$$h = sin(3p) \tag{20}$$

The state transition function can be expressed as

$$v_{t+1} = \textbf{\textit{bound}}[v_t + 0.001u_t - 0.0025\cos(3p_t)] \tag{21}$$

$$p_{t+1} = \textbf{\textit{bound}}[p_t + 1] \tag{22}$$

where ***bound*** is a function used to limited the value, ***bound***$(v_t) \in$ [-0.07,0.07], ***bound***$(p_t) \in$ [-1.5, 1.5]. The coefficient of gravity acceleration direction is -0.0025.

Sarsa is an effective TD algorithm for control problems. We implemented the Sarsa version of the TOSarsa(λ) algorithm and compared with Sarsa and Sarsa(λ). Fig. 3 shows the control effect of the three algorithms on the initial state's value function.

As it can be seen from Fig 3, in the both initial stage and final stage after convergence, the algorithm TOSarsa(λ) was better than the other two algorithms, Sarsa and Sarsa(λ). The three algorithms are value function methods, and their control policy is directly related to the evaluation of the value function. From the approximation point of view, TOSarsa(λ) got to convergence earlier than the other two. In general, the three algorithms were all effective in dealing with the mountain car problem and TOSarsa(λ)

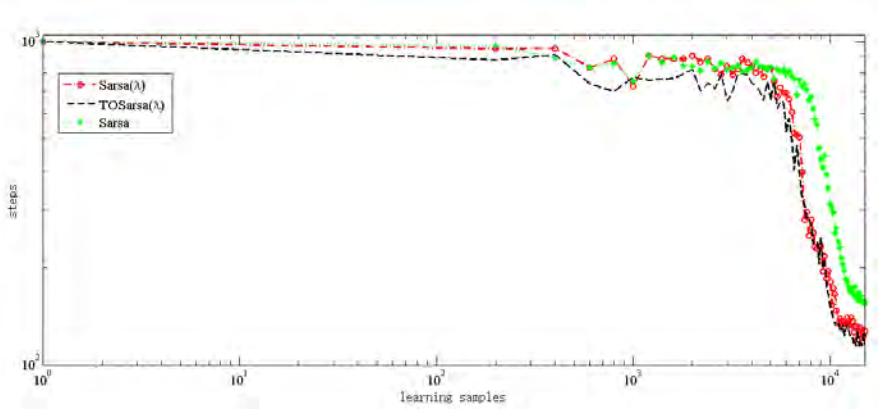


Fig. 3. The approximation effects of the algorithms on the initial state's value function on the initial state's value function.

which had a better strategy to evaluate performance was the best of three. However, all of the three algorithms had fluctuations at the beginning stage because at the initial stage, the data dictionary for the algorithms has not yet been completely established, and the algorithm kept exploration. We used TOSarsa(λ), Sarsa and Sarsa(λ) to solve the mountain car problem for 50 times. The results are shown in Fig. 4, where we can see that TOSarsa(λ) is the fastest in the three algorithms in the process of approximation. Fig. 4 shows the number of episodes required by the three algorithms, TOSarsa(λ), Sarsa and Sarsa(λ), to reach the target in different scenarios. TOSarsa(λ) was superior to the other two in the convergence rate and the convergence result. Moreover, the convergence result of TOSarsa(λ) was more stable.

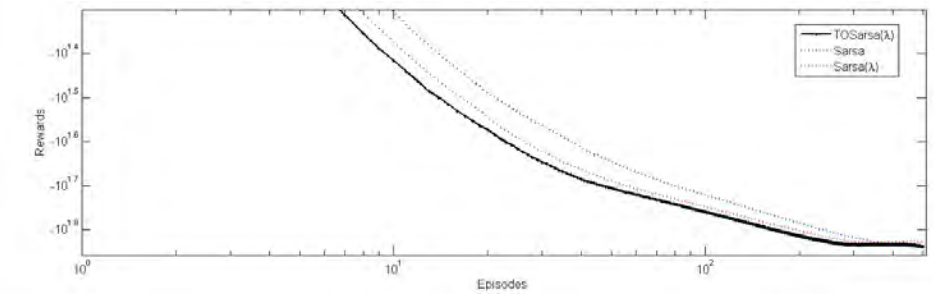


Fig. 4. The number of average steps of three algorithms. The abscissa represents the number of episodes and the ordinate shows the average number of steps.

5. TOSarsa(λ) With Heuristic Dynamic Programming

The dual heuristic dynamic programming (DHDP) algorithm is a method of dealing with continuous action space by neural network. It applies the actor-critics framework, evaluates the strategy in the critics section, and calculates deterministic strategies in the actors section. In this section, we will try to combine TOSarsa(λ) with heuristic dynamic programming.

5.1. TOSHDP Algorithm Description

The TOSarsa(λ) is used to evaluate the derivative of the value function to the state; update policy is updated by using the gradient descent method. The value of the function is:

$$\lambda(s_t) = \frac{\partial V(s_t)}{\partial s_t} = \omega^\top \kappa(s_t) \tag{23}$$

It satisfies the Bellman equation. We take TOSarsa (λ) method to get value of $\lambda(s_t)$, TD error, as:

$$\delta_t = \frac{\partial r_{t+1}}{\partial s_t} + \gamma \left(\frac{\partial s_{t+1}}{\partial s_t} + \frac{\partial s_{t+1}}{\partial a_t} \frac{\partial a_t}{\partial s_t} \right) \omega_t \kappa(s_{t+1}) - \omega_{t-1} \kappa(s_t) \tag{24}$$

As it can be seen from the above equation, the Equation(24) needs to solve $\frac{\partial s_{t+1}}{\partial s_t}$ and $\frac{\partial s_{t+1}}{\partial a_t}$, which requires a complete information of environment or model. The dual heuristic dynamic programming algorithm uses more environment knowledge and has a pretty good performance. In addition, the dual heuristic dynamic programming algorithm algorithm calculates the value of $\frac{\partial a_t}{\partial s_t}$, which is the actor part of the policy function of the derivative. The policy parameters updating as follows:

$$\begin{aligned} \omega_{t+1} &= \omega_t - \beta \Delta \omega_t \\ &= \omega_t - \beta \frac{\partial V(s_{t+1})}{\partial a_t} \frac{\partial a_t}{\partial \omega_t} \\ &= \omega_t - \beta \lambda(s_{t+1}) \frac{\partial s_{t+1}}{\partial u_t} \frac{\partial a_t}{\partial \omega_t} \\ &= \omega_t - \beta \lambda(s_{t+1}) \frac{\partial s_{t+1}}{\partial a_t} \kappa(s_t) \end{aligned} \tag{25}$$

where β is learning step for policy parameters. The following is the algorithm of TOSarsa(λ) with heuristic dynamic programming, where the 8^{th} step of the algorithm is the combination of optimal policy function and ε-greedy.

5.2. Cart Pole Balancing Problem

In this section, we verify the algorithm by cart pole balancing problem [17], which is a very classic continuous problem. There is a car on the horizontal track with a mass of

Algorithm 2 TOSarsa(λ) with heuristic dynamic programming (TOSHDP)))

Input: policy, threshold
Output: optimal policy

Initialize sample set S
Set up data dictionary $\boldsymbol{D}$
repeat
  Initialize starting state s_0
  Initialize eligibility trace $\mathbf{e} \leftarrow 0$
  $\lambda(s_t) \leftarrow \omega^\top \kappa(s_t)$
  repeat
    a $\leftarrow \pi(a|s)$
    Observe r, s
    $\lambda(s_{t+1}) \leftarrow \omega^\top \kappa(s_{t+1})$
    $\delta_t \leftarrow r_{t+1} + \gamma \left(\frac{\partial s_{t+1}}{\partial s_t} + \frac{\partial s_{t+1}}{\partial a_t} \frac{\partial a_t}{\partial s_t} \right) \lambda(s_{t+1}) - \lambda(s_t)$
    $\mathbf{e}_t \leftarrow \gamma\lambda \mathbf{e}_{t-1} + \alpha_t \kappa(s_t) - \alpha_t \gamma\lambda \left[\mathbf{e}_{t-1}^\top \kappa(s_t) \right] \kappa(s_t)$
    $\omega_{t+1} \leftarrow \omega_t - \beta\lambda(s_{t+1}) \frac{\partial s_{t+1}}{\partial a_t} \kappa(s_t)$
    $\xi \leftarrow \min_{s_i \in D} \left(\kappa(s,s) + \kappa(s_i,s_i) - 2\kappa(s,s_i) \right)$
    Update **D**
    if ξ is greater than a predefined threshold **then**
      $V(s_t) \leftarrow \omega^\top \kappa(s_{t+1})$
      Get ω and **e**
    else
      $V(s_t) \leftarrow V(s_{t+1})$
    end if
    $s_t \leftarrow s_{t+1}$
  until all steps of the current episode end
until all episode end
return optimal policy

m=1kg, the length l = 1m. The pole and the car are hinged together. The pole and the vertical direction are at an angle. In order to make the angle of the pole and the vertical direction in $[-36^\circ, 36^\circ]$, where the angle is negative if the pole is on the left side of the vertical line, and the angle is positive if the pole is on the right side of the vertical line. After each time interval Δt = 0.1s, a horizontal force F is applied to the cart, where F is within [-50N, 50N] (negative means the force is to the left, and positive is to the right), and there is a random noise disturbance between [-10N, 10N] when F is applied. All frictional forces were not considered. The task of the agent is to learn a policy so that the angle between the pole and the vertical direction is kept as much as possible in the specified range.

We use MDP to model the cart pole balancing problem. The state of the environment is represented by two variables α and β, where α is the angle formed by the pole and the vertical line, and β is the angular acceleration of the rod. The state space is:

$$S = \{(\alpha, \beta) | \alpha \in [-36^\circ, 36^\circ], \beta \in [-36^\circ, 36^\circ]\} \tag{26}$$

the action space is:

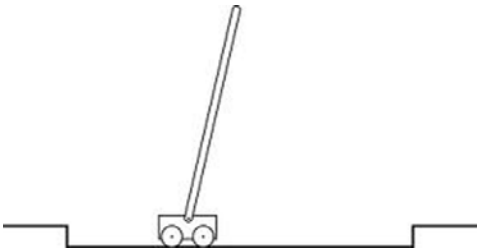


Fig. 5. Cart pole balancing problem diagram.

$$A = \{a|a \in [-50N, 50N]\} \tag{27}$$

Agent exerts force *F* on the cart, and the angular acceleration of the pole is:

$$\xi = \frac{g \sin\alpha + \cos\alpha \left(\frac{-f - ml\beta^2 \sin\theta}{m+M} \right)}{l \left(\frac{4}{3} - \frac{m\cos^2\alpha}{m+M} \right)} \tag{28}$$

where g is the constant of gravitational acceleration, with value 9.81 m/s^2; and f is the value of force *F*. After Δt , the states are $\alpha = \beta + \xi\Delta t$, $\beta = \alpha + \beta\Delta t$, and the reward function is

$$\rho(x,u) = \begin{cases} 1, \ |f(x,u)| < 36^\circ \\ -1, \ |f(x,u)| \geq 36^\circ. \end{cases} \tag{29}$$

The episode ends when the angle between the pole and the vertical line exceeds the given range. If the pole has not fallen and kept standing after 3000 time steps, it is regarded as a successful trial.

We compare TOSHDP with conventional DHP algorithm, where DHP uses two three-layer neural networks for value functions and policy approximation, all of their learning steps are 0.1. The results are shown in Fig. 6.

We can see from Fig. 6 that be seen from the convergence rate of the TOSHDP algorithm is higher than that of the conventional DHP algorithm in the same step size. The TOKDHP algorithm begins to converge at about 200 episodes, while the traditional DHP algorithm requires about 270 episodes to converge. There are mainly three factors caused this. First, kernel method is a more lightweight approximation algorithm than the neural network as the kernel method deals with the nonlinear problem directly by mapping and linear technique, while the neural network deals with the nonlinear problem through the multi-layer nonlinear transformation. Secondly, when the learning step is large, the neural network is easy to fall into the local optimal solution. Thirdly, our approach is more efficient in policy evaluation that was verified in the earlier experiment, resulting in an accelerated effect on the learning of the policy.

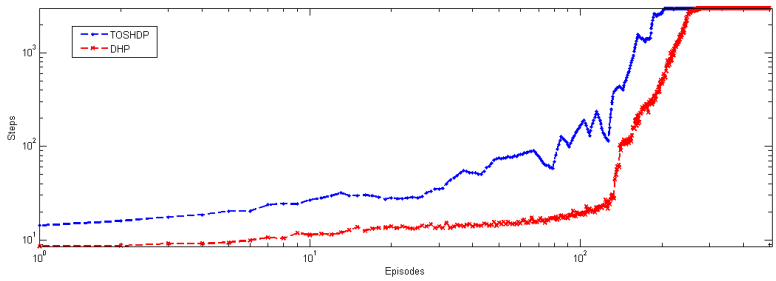


Fig. 6. Results of TOSHDP vs. conventional DHP algorithm where the value of the learning steps were set as 0.1.

6. Conclusion

We propose a true online kernel time difference algorithm, TOSarsa(λ), which employs a clustering-based sample sparsification method and selective kernel-based value function as value function representation. The experiment on mountain car problem showed our algorithm was effective in deal with the typical continuous problems and could speed up strategy search as well.

We combined the proposed TOSarsa(λ) algorithm with the dual heuristic dynamic programming algorithm to improve policy learning speed of policy search algorithms by replacing approximating using neural network method with approximating using kernel method. The experiment on cart pole balancing problem verified that our proposed algorithm really worked. It is a good alternative to deal with continuous action space problems. However, there is still some work to study further, such as how to extend the model to deal with the continuous space problems of unknown environment.

Acknowledgments. This paper is supported by National Natural Science Foundation of China (61303108, 61373094,61772355, 61702055,61602332), Jiangsu Province Natural Science Research University major projects (17KJA520004), Suzhou Industrial application of basic research program part (SYG201422), Provincial Key Laboratory for Computer Information Processing Technology of Soochow University (KJS1524), China Scholarship Council project (201606920013).

References

1. Al-Rawi A, Ng A, Y.A.: Application of reinforcement learning to routing in distributed wireless networks: a review. Artificial Intelligence Review 43(3), 381–416 (2015)
2. Bagnell A, Ng Y, S.J.: Policy search by dynamic programming. In: Advances in Neural Information Processing Systems. pp. 831–838 (2004)
3. Busoniu L, Babuska R, e.a.: Reinforcement learning and dynamic programming using function approximators. CRC Press (2010)
4. Chen X, Gao Y, W.R.: Online selective kernel-based temporal difference learning. IEEE transactions on neural networks and learning systems 24(12), 1944–1956 (2013)

5. E., B.: A markov decision process. Journal of Mathematical Fluid Mechanics 6(1), 65–73 (1957)
6. El I, Feng M, e.a.: Reinforcement learning strategies for decision making in knowledge-based adaptive radiation therapy: application in liver cancer. International Journal of Radiation Oncology Biology Physics 96(2), 38–45 (2016)
7. Ghorbani F, Derhami V, A.M.: Fuzzy least square policy iteration and its mathematical analysis. International Journal of Fuzzy Systems 19(13), 1–14 (2016)
8. H., V.H.: Reinforcement learning, chap. Reinforcement learning in continuous state and action spaces, pp. 207–251. Springer Berlin Heidelberg (2012)
9. K., D.: Reinforcement learning in continuous time and space. Neural Computation 12(1), 210–219 (2000)
10. Kiumarsi B, Lewis L, e.a.: Reinforcement q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. Automatica 50(4), 1167–1175 (2014)
11. Kober J, Bagnell A, P.J.: Reinforcement learning in robotics: a survey. The International Journal of Robotics Research 32(11), 1238–1274 (2013)
12. L, P.: Markov decision processes: discrete stochastic dynamic programming. John Wiley and Sons (2014)
13. M., H.: Value-function approximations for partially observable markov decision processes. Journal of Artificial Intelligence Research 13(1), 33–94 (2011)
14. Scholkopf B, Platt J, H.T.: An application of reinforcement learning to aerobatic helicopter flight. In: Advances in Neural Information Processing Systems. vol. 19, pp. 1–8. Proceedings of the Twentieth Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada (2007)
15. Sutton R, B.G.: Reinforcement learning : an introduction. IEEE Transactions on Neural Networks 16(1), 285–286 (2005)
16. Sutton R, Mcallester D, e.a.: Policy gradient methods for reinforcement learning with function approximation. In: Advances in Neural Information Processing Systems. vol. 12, pp. 1057–1063 (2000)
17. T., P.: Solving the pole balancing problem by means of assembler encoding. Journal of Intelligent and Fuzzy Systems 26(2), 857–868 (2014)
18. Whiteson S, Tanner B, W.A.: The reinforcement learning competitions. AI Magazine 31(2), 81–94 (2010)
19. Yau A, Goh G, e.a.: Application of reinforcement learning to wireless sensor networks: models and algorithms. Computing 97(11), 1045–1075 (2015)
20. Zhu H, Zhu F, e.a.: A kernel-based sarsa(λ) algorithm with clustering-based sample sparsification. In: International Conference on Neural Information Processing. pp. 211–220. Springer International Publishing (2016)

Fei Zhu is a member of China Computer Federation. He is a PhD and an associate professor. His main research interests include machine learning, reinforcement learning, and bioinformatics.

Haijun Zhu is a postgraduate student in the Soochow University. His main research interest is reinforcement learning. He programmed the algorithms and implemented the experiments.

Yuchen Fu (corresponding author) is a member of China Computer Federation. He is a PhD and professor. His research interest covers reinforcement learning, intelligence information processing, and deep Web. He is the corresponding author of this paper.

Donghuo Chen is a member of China Computer Federation. He is a PhD. His research interest includes reinforcement learning, model checking.

Xiaoke Zhou is now an assistant professor of University of Basque Country UPV/EHU, Faculty of Science and Technology, Campus Bizkaia, Spain. He majors in computer science and technology. His main interests include machine learning, artificial intelligence and bioinformatics.

Received: January 7, 2017; Accepted: May 15, 2017.

Computer Science and Information Systems 14(3):805–821 https://doi.org/10.2298/CSIS170413030D

Social evaluation of innovative drugs: A method based on big data analytics

Genghui Dai[1], Xinshuang Fu[2], Weihui Dai[3], and Shengqi Lu[4]

[1] School of Marine Sciences, Sun Yat-Sen University, Guangzhou 200433, China
daigengh@mail2.sysu.edu.cn
[2] School of Management, Shanghai University, Shanghai 200444, China
gracief@126.com
[3] School of Management, Fudan University, Shanghai 200433, China
whdai@fudan.edu.cn
[4] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China
shengqilu@fudan.edu.cn

Abstract. The evaluation of drugs is a professional and time consuming process which involves a series of clinical trials and evidence-based verifications. However, an innovative drug may still suffer from unpredictable risks after coming into market due to the complex circumstances in practical utilization. Owing to the popularization of information networks and social media, big data analytics exhibits a new perspective of social evaluation as the supplementary means on this issue. This paper designed a Hadoop platform for data collection and processing, and explored the social evaluation of innovative drugs based on big data analytics. Through the analysis of mined data and affective computing on online comments, a new Chinese drug extracted from marine organisms can be evaluated comprehensively by the proposed method. Furthermore, the potential utilization of fullerene materials may be considered for improving its curative effects. Research work of this paper provides a big data analytics method for social evaluation of innovative drugs as well as their promising improvements.

Keywords: Big data analytics, social media, innovative drug, social evaluation, marine biology, fullerene materials.

1. Introduction

The full evaluation of drugs includes pre-marketing evaluation and post-marketing evaluation. In pre-marketing evaluation, an innovative drug has to pass IND(Investigational New Drug) and NDA(New Drug Application) procedures before it can be approved of coming into market. It is a professional and time consuming process, for example, the median time of a standard review is 384 days on IND, and 846 days on NDA by China Food and Drug Administration [30]. Although the above evaluation is based on a series of clinical trials and evidence-based verifications, but there are probably some unpredictable risks which may cause serious adverse reactions for an innovative drug in practical utilization [5][25] Therefore, post-marketing evaluation is the necessary and vital part in a full evaluation of innovative drugs.

As we known, current post-marketing evaluation of innovative drugs is mainly based on the statistical analysis of investigated samples or depends on special reporting channels such as the reporting system of health care institutions [25]. Nevertheless, a lot of disadvantages have been found in the existing method, such as limited samples, poor timeliness, inefficiency and the influence of uncertainty factors [2]. Actually, the practical curative effects and adverse reactions of drugs are mostly related to patients' individual conditions, living habits, and environmental factors. Especially for Chinese drugs, a reliable evaluation usually requires the comprehensive reviews from complex circumstances because there are important differences in different cases. It is difficult to be implemented through the current evaluation system.

Owing to the development of information technology and popularization of mobile applications, more and more people share their experiences of daily life by social media, such as shopping, tourism, medical treatment, and so on. In the meantime, the network of social media has appeared as a new platform and provides the valuable repository for scientific research and social study. As one of the most popular topics on social networks, health care and medical treatment attracts extensive concerns, and thereupon expedites the flourishing of various medical and health forums. The valuable information about innovative drugs in practical utilization can be mined from multifarious online comments and posts on the above forums through big data analysis. Therefore, big data analytics exhibits a new perspective of social evaluation of innovative drugs, which can be applied as the supplementary means to post-marketing evaluation. This paper aims to propose a Hadoop platform for data collection and processing from social media, and explore the social evaluation of innovative drugs based on big data analytics. It is organized as follows: Section 2 introduces the related works; Section 3 designs a Hadoop platform and studies the big data analytics from social media; Section 4 proposes the social evaluation method of innovative drugs; and Section 5 is the discussion and conclusion of this paper.

2. Related works

In recent years, big data analytics has been successfully applied in various fields such as financial markets, social management, production and manufacturing, as well as precision medicine, and shows superiorities over traditional methods in many aspects. In modern medicine and pharmacy, the classification of drugs is becoming more complicated than before, beyond the limitations to diseases or symptoms. As well, the ingredients of drugs are no longer invariants [25]. Those circumstances bring new difficulties and risks on the evaluation of innovative drugs.

Generally speaking, the evaluation of an innovative drug is based on a series of clinical trials and the comprehensive reviews on its effects [2][24]. For example, the test of drug allergy is carried out on extracts of natural drugs [11], and pharmacodynamic test is used for the evaluation of genetic engineering products [10]. However, the innovative drugs may still suffer from unpredictable risks after coming into market, and should be evaluated comprehensively through a professional and time consuming process. In recent years, the outbreak of new epidemic diseases such as influenza A (H1N1) has made the evaluation of innovative drugs faced with great challenges. In order to cope with this problem, many solutions have been proposed, one of which is the big data analytics. Up to now, many achievements have been made with the help of big data analytics [18][27]. It also

provides a new research methodology in medical and health fields, such as the analysis of diabetes cases, the study of regional characteristics of infectious diseases, the mining of disease causing factors, and so on. Zhu et al. summarized the research status and progress on the data mining of DNA sequence, and pointed out its significance in biological application [35]. Yue et al. applied data mining technology to study the classification of DNA sequences, and proposed a new judgment method to explore their classifications [33]. Li designed a health risk model for the assessment of Chinese people from the analysis of big data [16]. Karaolis et al. developed a data mining system to study the pathogenic factors of heart disease using association analysis algorithm [12]. Chang et al. adopted artificial neural networks to predict the outcome in the diagnosis of Parkinson's disease [3]. Dreiseit et al. proposed an improved method which combined artificial neural networks with regression analysis and decision tree to estimate the mortality in diseases [7].

In regard to big data analytics of health and medical information from social media, Zhou et al. used machine learning techniques to realize the automatic retrieval of online text information, and established a social medical terminology dictionary [34]. Ye et al. built a corpus of Chinese medicine, and studied the social evaluation of Chinese medicines in United States from the news reports and social media. Their research showed the social trend of increasing interest and attentions to Chinese medicines by American society and people [31]. Sampathkumar et al. applied Hidden Markov Model to analyze the adverse drug reactions based on the information of online healthcare forums, and provided an effective method for early warning of pharmacovigilance [21]. Existing research findings have indicated that the social evaluation of innovative drugs based on big data analytics can timely reveal the underlying influences and undiscovered effects of the above drugs from patients' feeling and their comments, which are hard to be reflected in the regular post-marketing evaluation.

3. Hadoop platform for data collection and processing

Through an analysis of the related works, we found that affective computing on text information is the useful big data analytics for the study of online comments [1][8][32] .In order to establish the big data environment for social evaluation of innovative drugs, we designed a Hadoop platform [14] to complete the data collection and processing, which can efficiently implement subject extraction and sentiment analysis from online comments. Its framework is designed as in Fig.1.

It includes three layers namely information collecting layer, data storage layer and business analysis layer. Firstly, the related text information are collected by web crawlers and sent to the text server group for preprocessing in information collecting layer. Secondly, the above data will be stored in MapReduce and HDFS in the data storage layer through the interface of HDFS [9]. Finally, text subject extraction, sentiment analysis, and other data analysis will be carried out in the business analysis layer, and all of data changes are executed by calling the data interface system such as HDFS and Hive.

3.1. Run mechanism for big data collection and processing

As big data analytics for social evaluation involves the collection and processing of enormous unstructured data from social media, it is necessary to design an efficient run mechanism carefully for dealing with the data. Hadoop platform has good capacity of distributed

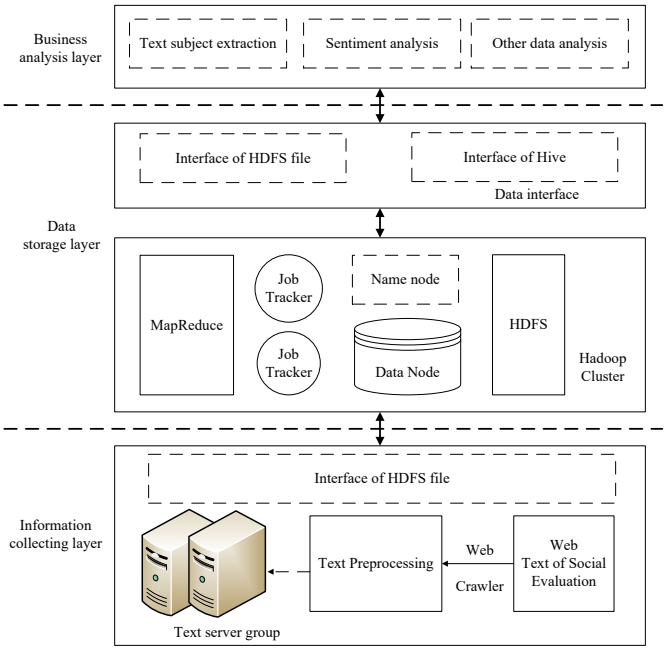


Fig. 1. Hadoop platform for social evaluation of innovative drugs

storage and parallel co-processing. However, its performance depends on the design of an effective run mechanism [22]. The platform includes three main components: master node, client node, and slave cluster, all of which coordinate with each other through the run mechanism to accomplish tasks. In our solution, we designed the run mechanism for big data collection and processing as in Fig. 2.

It can be seen from Fig.2, the Master Node is responsible for job management and resource scheduling, and slave cluster includes a lot of map tasks or reduce tasks for dividing sentiment words, subject extraction and so on. The above run mechanism can be described as follows.

Running mechanism. The running mechanism for data processing includes the following steps.

Step 1. Job submission, Firstly, the client node of the Mapreduce start a JobClient, and send a job with request ID to the JobTracker in Master Node by the JobClient, such as the job of dividing sentiment words.

Step 2. Job initialization, JobTracker puts the job into an internal queue, and hand over the scheduler job for scheduling, and then complete its initialization.

Step 3. Assignment of tasks, JobClient creates the corresponding number of Map tasks and Reduce tasks according to the number of input data, and assigns the Map task and Reduce task to the TaskTracker node in the Slave node.

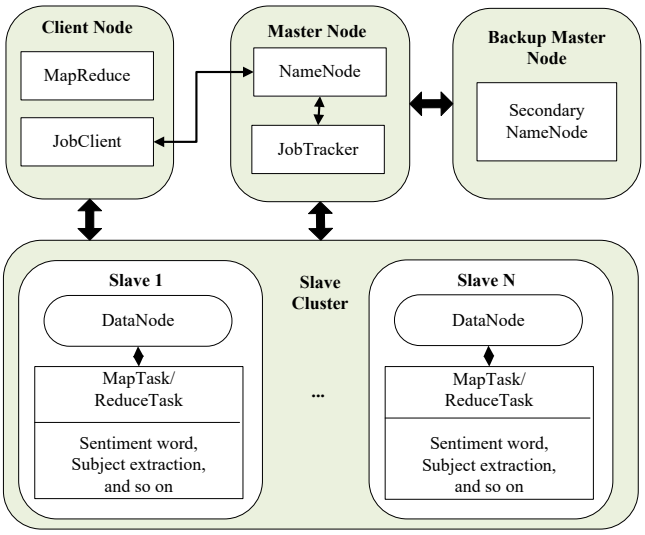


Fig. 2. The run mechanism for big data collection and processing

Step 4. Perform tasks, The TaskTracker node reads the input data stored on the HDFS, at the same time, the TaskRunner task will be created by MapTask and RedcueTask respectively, and the above two tasks will run until the end of task.

In the Hadoop platform, HDFS is responsible for the distributed storage of files in Hadoop cluster, which contains three major parts namely NameNode, DataNode, and Client.

NameNode. It acts as the management role in HDFS and is used to provide a name query service. It is responsible for managing the namespace of the file system, backup and the configuration of the cluster. In addition, the Metadata information stored in NameNode will be loaded into the memory after the NameNode starts.

DataNode. It is the basic unit of file storage, mainly used to save the information block, and will report to NameNode block when the DataNode thread is started, at the same time, it send a heartbeat in every fixed seconds to keep in touch with NameNode. Once NameNode hasnt received heartbeat within a fixed minutes, it means that the DataNode has been lost, and its block should be copied to the other DataNode.

Client. It is a client application to get files in distributed file system, which includes write file, read file and copy file block. The process of read file as follows. Client sends a request to the NameNode to read the file, and the NameNode return the address information of the DataNode that hold the data block, and then the Client calls the read() function to read data from the DataNode. When the Client data read is completed, it will call the close function FSDataInputStream(). In the process of data reading, if Client and DataNode

communication are errors, then Client tries to connect to the next data node. At the same time, the failure of the DataNode will be recorded.

MapReduce. It is responsible for the decomposition of tasks and the summary of the results. The tasks are distributed and completed by each individual node, and all the above nodes belong to a master node, and the final results come from each node through an integration of their intermediate results. The running mechanism of MapReduce is shown as in Fig. 3.

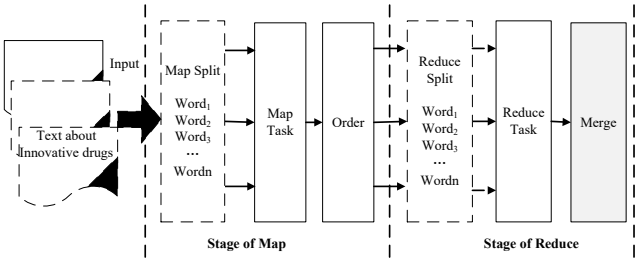


Fig. 3. The processing of the MapReduce

It can be seen from Fig.3, the mechanism of MapReduce include the map and reduce of tasks. In the process of map, data will be split into <key, value> according to the definition of Map function, and will be merged after completing the reduce tasks. It's worthy of mentioning that the Map process and Reduce process can run in parallel.

Task execution. The algorithm of task execution can be described as follows.

```
program Execution of task (Output)
  Init analysis task and hive database connection pool tp;
  begin
      (1) Get the connection from the hive database
          connection tool;
      (2) Connect to Hive, read the task of HQL, and
          send HQL query request to Hive;
      (3) Hive compiles and executes HQL, returns the
          execution result;
      (4) Write the result to local file and upload to
          HDFS path which is assigned by the analysis
          task;
      (5) Read result of the analysis task configuration,
          create new table in the Hive according to the
          configuration;
      (6) Upload the file in step 5 to new table which is
```

```
        create in step 6;
end.
```

3.2. Data analytics for social evaluation

The information about practical utilization of innovative drugs are scattered on microblogs or healthcare forums such as http://www.dxy.com, tieba.baidu.com, 91160.com, and so on in China. The above information are all unstructured texts, for example, the questions and answers, comments on the treatment of a disease or the curative effects of a drug, which are from the patients, family members, and doctors, and usually contain valuable information to be used for social evaluation.

The data analytics for social evaluation of innovative drugs includes text classification and affective computation. The purpose of text classification is to separate and keep the subjective text information for affective computation. It is realized by the subject extraction with a LDA model and the classification based on SVM (Support vector machine) and Bayes classifier. The purpose of affective computation is to calculate the trend and intensity of the above subjective text information for social evaluation. It is realized based on an emotional dictionary, and will be discussed in Section 4 of this paper. The outlined process of data analytics is shown as in Fig. 4.

It can be seen from Fig.4 that the data will be collected from various related websites by crawlers and preprocessed by filter and subject extraction. The specialized subjects will be extracted by LDA algorithm, and then classified by SVM and Bayes classifiers. If it is a subjective text, the affective intensity will be calculated for social evaluation. Otherwise, if it is an objective text, this text will not be processed.

LDA model. In order to extract the related subjects more efficiently, we used LDA model to fulfill this task. LDA model is also called the three layers Bayesian probability model, which includes the layers if words, subjects, and document structures. We hereby divide the above layers into: words, probable subject, and document sets. The matrix model of LDA can be shown in Fig.5 [6].

In Fig. 5, SE refers to the all of social evaluations on innovative drugs, and ϕ refers to the probability distribution of each subject on all terms. Θ expresses the subject distribution of each social evaluation. d_m is the m social evaluation, and w_n is the word of n term, and z_k is the k implicit subject.

In order to obtain the appropriate parameters of LDA model, the preprocessing data are used for training by the following steps:

Step 1. Initialization, randomly assign a subject number z to each of word w from prepared data. Generally, set α is 50/N_{theme}, where, N_{theme} is the number of subject, and β is 0.01.

Step 2. According to the Gibbs Sampling algorithm, collect the subject z from the set of word w, and update this set.

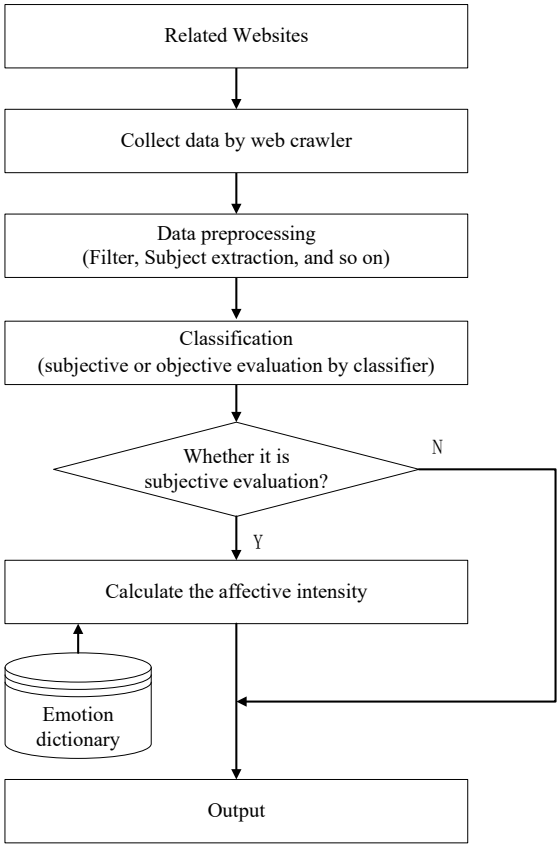


Fig. 4. Process of data analytics for social evaluation

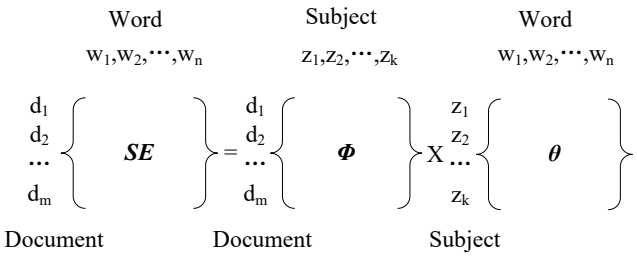


Fig. 5. The matrix model of LDA

Step 3. Repeat step 2 until Gibbs Sampling converges, that is to say, both subject distribution of each comment and word items of each subject are all convergence. After that, the probability distribution function is calculated as follows [19].

$$p(Z_i = k|\overrightarrow{Z}\neg i, \overrightarrow{w}) \infty \frac{n^k_{m,\neg i} + \alpha_k}{\sum_{k=1}^{K}(n^k_{m,\neg i} + \alpha_k)} \bullet \frac{n^t_{k,\neg i} + \beta_t}{\sum_{k=1}^{K}(n^t_{k,\neg i} + \beta_t)} \tag{1}$$

In 2, the probability distribution of subject-topic vector can be described as follows.

$$\theta = \frac{n^k_{m,\neg i} + \alpha_k}{\sum_{k=1}^{K}(n^k_{m,\neg i} + \alpha_k)} \tag{2}$$

As well, the probability distribution of subject-word can be described as follows.

$$\varphi = \frac{n^t_{k,\neg i} + \beta_t}{\sum_{k=1}^{K}(n^t_{k,\neg i} + \beta_t)} \tag{3}$$

Step 4. Calculate the co-occurrence frequency matrix of document-subject-word, and construct the LDA model.

Classified by SVM. Support vector machine (SVM) is a statistical machine learning classification method based on VC dimension theory and structural risk minimization principle. It has been widely used in affective computing on texts and vocal recognition for its superior performance on classification [4][23]. The classification method of subjective or objective comments by SVM is shown as in Fig. 6.

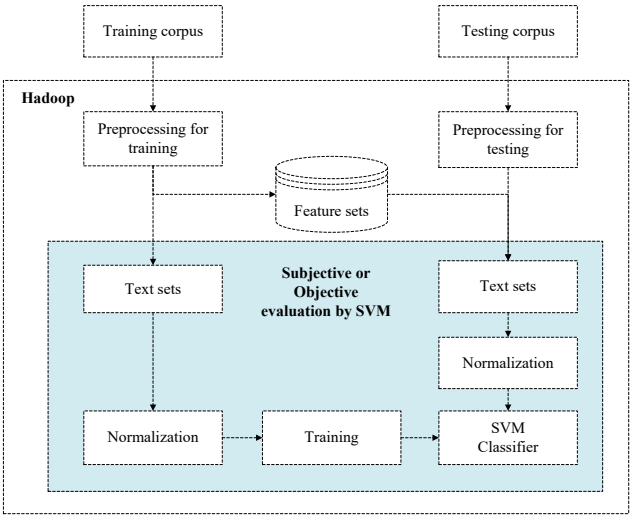


Fig. 6. The classification method by SVM

The classification algorithm can be described as follows.

Set $\{x_i, y_i\}_{i=1}^{n}$ as the set of data sample, where the input data x_i belongs R_d , and the output data $y_i \in$ (-1,1), then the linear discriminant function in d space is $f(x) = \omega \cdot x + b$, and the classification hyperplane equation is $\omega \cdot x + b = 0$. So the method of SVM in a high dimensional space can be described as:

$$y_i[\omega \cdot x + b] = 1 - e_i, i = 1, ..., n \tag{4}$$

Here, ω is the weight, and input x_i is the high dimensional space, b is the error constant, Therefore , the computation of the optimal classification can be converted into dual problem as long as the Lagrange optimization method is used. And the optimal classification function can be expressed as follows

$$f(x) = sgn(\sum_{i=1}^{N} \alpha_i^* y_i K(x_i, X) + b^*) \tag{5}$$

Where, b^* is the threshold of classification, $K(x_i, X)$ is kernel function and it was used the four forms as follows.

RBF kernel function:

$$K(x, y) = exp\frac{-||x - y||^2}{\sigma^2} \tag{6}$$

Linear kernel function:

$$K(x, y) = x^T \cdot y \tag{7}$$

Polynomial kernel function:

$$K(x, y) = [(x \cdot y) + 1]^m \tag{8}$$

Sigmod kernel function:

$$K(x, y) = tanh(u(x \cdot y) + c) \tag{9}$$

Here, the RBF kernel function was used in the SVM in our study.

4. Evaluation method based on big data analytics

As pointed out in this paper, the goal of social evaluation is to provide supplementary information for the comprehensive review on innovative drugs, and makes up the defects of a regular post-marketing evaluation. Therefore, the main role of big data analytics is reflected in the two aspects: new findings of the drug in practical utilization, and feeling and experiences of the drug in practical utilization. It has caused the researchers' attentions that the patients' emotional expressions about a drug possibly indicate its underlying influences and undiscovered effects, as well as the market value. We use affective computing technology to calculate the trend and intensity of emotions from the subjective texts. The above computation is realized based on the emotional dictionary developed by Prof. Lin et al [28], which includes 27,466 emotional words and divided into seven basic categories. At the same time, the collection rules of data should be built in order to get better results.

4.1. Collection rules and word frequency calculation

Collection rules. The valid data can be used for social evaluation should include the complete items: title, content, date of publication, and replying posts. Besides, the data promulgator must be identified, such as patient, family member, or doctor. Table 1 lists the samples of collection data.

Table 1. Sample of data collected

No	Title	Content	Date of publication	Type of promulgator	Count of replying posts
1	cerebral infarction	Butylphthalide is good for the disease · · ·	2017-02-25 13:37:28	Patient	17
2	Haishengsu	Will it affect patient's condition? · · ·	2017-02-24 15:08:42	Family member	22
3	Scopola mine Butylbromide Injection	It is used in the acute gastrointestinal tract · · ·	2017-02-21 15:08:42	Doctor	16
4	Domperidone Tablets	Lead to elevated serum prolactin levels · · ·	2017-02-21 16:08:42	Doctor	15
· · ·	· · ·	· · · · · ·	· · ·	· · ·	

Part-of-Speech. The segmentation methods for Chinese words commonly include forward maximum matching method [29], bidirectional maximum matching method [26] and reverse maximum matching method [20]. We adopted the NLPIR segmentation system [15] for word segmentation and extended it with the POS tagging. Therefore, each word is assigned by a Part-of-Speech as the samples shown in Table 2.

Table 2. Samples assigned by Part-of-Speech

No	Title	Annotation format
1	Nouns	/n
2	Verbs	/v
3	Adjectives	/a
4	Adverb	/d
5	Numerals	/m
6	Punctuation mark	/w
· · ·	· · ·	· · ·

In the processing of word segmentation, if a word is not included in the dictionary, it can't be identified, and should be added to the dictionary by manual. For example, 'Butylphthalide is good for the disease', in which the word of 'Butylphthalide' can't be found in the dictionary, and needs to be added to the dictionary. After processing of segmentation, the online comments still contain a lot of useless words, such as pronouns,

prepositions, determiners, auxiliary, conjunctions, interjections and onomatopoeic words. The above words can't help to extract subjects, but maybe reduce the calculation efficiency of LDA model, and need to be filtered out.

Word frequency calculation. Word frequency calculation is ready for affective computation and evaluation, and fulfilled by the parallel computing on Hadoop platform. Firstly, the type of input and output are built to class Mapper(), and their expressions are as follows: input type is <Object, Text>, and the output type is <Text, IntWritable>. If a task comes up, parallel computing is performed by calling the processes of Map() and Reduce() to complete the word frequency calculation. For example, Fig. 7 shows the word frequency calculation about the comments on fullerene materials.

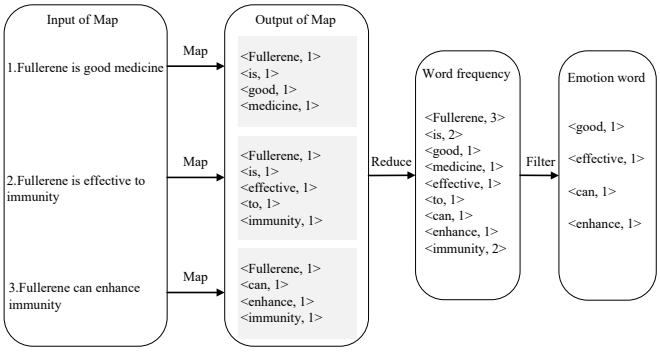


Fig. 7. Word frequency based on parallel computing of MapReduce

It can be seen from Fig.7, there are three tasks about the comments on fullerene materials in the Input of Map, and one task in Reduce. The above three text sections are independently assigned to three map tasks for processing firstly, and then the expression will be transformed into <'word', count value> by the specific function in intermediate process of Output in Map. Among which, count value refers to the total number of a certain word in the text section. Thereafter, the values will be used as input to the Reduce task, and the Reduce task will complete the computation of the total number of occurrences of each word. Finally, the three text sections will be merged into, and output the expression <'word', count value>, such as <immunity, 2> in the case about fullerene materials.

4.2. Subject extraction from online social media

After text preprocessing, the vocabulary dictionary needs to be built as input files for LDA model. In our research, the vocabulary dictionary has more than ten thousand words. Generally speaking, the parameters of the model must be initialized. Here, we set the initial value of the subject number to 5 according to the existing experiences [17]. As

well, we set α equals ten, β is 0.01, and the number of iterations for Gibbs sampling is 2000.

After 2000 iterations of Gibbs sampling, we can get the optimal extraction of the feature words on the five topics. Furthermore, five keywords are extracted respectively on each topic, and the distributions of the above keywords on each topic are shown as in Table 3.

Table 3. Distributions of the keywords on each topic

	Distribution				
1	Comfortable (0.2783)	Good (0.1067)	Body (0.0965)	Anticancer (0.075)	Depression (0.0023)
2	Blood (0.3135)	Complications (0.2149)	Lead to (0.1063)	decline (0.0075)	Form (0.0031)
3	Innovative drugs (0.0843)	Control (0.083)	Blood pressure (0.0645)	Appetite drugs (0.0473)	Dose (0.0163)
4	Eat (0.0873)	Food (0.0584)	Diet (0.1078)	shape (0.0873)	Marine organism (0.0464)
5	Symptom (0.2084)	Study (0.1172)	Technology (0.1070)	Treatment (0.0775)	Development (0.0562)

It can be seen from Table 3 that promulgators on social media pay more attentions to the above five types of subjects. The first subject is 'Effect description', which includes words such as comfortable, anticancer, and so on. At the same time, the other subjects have also been extracted. In addition, the number in the bracket of each word indicates the contribution of the word to this subject.

4.3. Emotional intensity analysis

To facilitate the evaluation, we divide the intensity of emotion into five levels, and assign to the value of 1, 3, 5, 7, and 9 respectively according to previous research [28]. As well, the emotional tendency is also assigned a polarity value. The positive tendency is expressed as 1, and the negative tendency is expressed as -1. The neutral tendency is expressed by 0. After quantized by the above values, the emotions becomes easy to be identified.

Based on big data analytics, we studied the social evaluations of a marine biological medicine and the fullerene materials, which have been reported as with significant curative effects and promising potentials for the treatment of tumors. With the rapid development of ocean resources, marine biological medicine has caused great interest by the developers of innovative drugs due to its natural and special bioactivity. 'Haishengsu', an innovative Chinese drugs extracted from marine organisms, was developed in recent years, and the clinical trials reported its significant anti-tumor effects. This innovative drugs was approved of coming into market in 2013. The emotional tendency and intensity about fullerene materials and 'Haishengsu' are shown as in Table 4.

Table 4. Emotional tendency and intensity about fullerene materials and 'Haishengsu'

First Keywords	Second Keywords	Third Keywords	Emotional intensity	Emotional tendency
Fullerene	Immunity	Lose weight	5	0
		Feel sleepy	3	-1
		Increased resistance	7	1
		Shortness of breath	5	-1
		Vulnerable to the cold	7	-1
Haishengsu	anti-tumor	Significant effect	9	1
		Affect physiological balance	3	-3
		Restrain the disease	5	1
...	...	...	...	...

It can be seen from Table 4, the value of the emotional tendency include three values (-1, 0, 1). From the value of 1, for example, we can deduce that fullerene has the positive function to increase resistance. The value of 0 means it is not associated with weight loss. As well, 'Haishengsu' has significant anti-tumor effects, and can restrain Heptocellular. However, its effects on physiological balance obtained a weak negative evaluation. The above studies show that social evaluations based on big data analytics may offer supplementary information about the innovative drugs in their practical utilization. It is very helpful for taking a comprehensive review on the innovative drugs, as well as for the improvement of the above drugs. Furthermore, the correlative analysis of evaluations indicates that curative effects of 'Haishengsu' are expected to be promisingly improved if combined with the utilization of fullerene. Therefore, big data analytics exhibits a new perspective of not only the new method for social evaluation of innovative drugs, but also the valuable information for promising development and application of the above drugs.

5. Discussion and Conclusion

Innovative drugs play the important role on promoting the progress of medicine and medical treatments. However, the traditional evaluation method of innovative drugs is a time consuming process, and has a lot of defects such as limited samples, poor timeliness, inefficiency, and the influence of uncertainty factors, especially in the face of sudden outbreak of diseases [13].

This paper designed a Hadoop platform and explored the social evaluation method of innovative drugs based on big data analytics. It aimed to provide the supplementary information for a comprehensive review on innovative drugs, as well as to make up the defects of a regular post-marketing evaluation. The main role of big data analytics is reflected in the following two aspects: new findings of the drug in practical utilization, and feeling and experiences of the drug in practical utilization. Research work of this paper provides a big data analytics method for the evaluation of innovative drugs, and as well, the valuable information for improving their promising development and application.

From the perspective of future research, more data sources such as geography and weather information, historical information about the process of treatments, and the accurate analysis methods such as logical reasoning and meta analysis, may be considered in

big data analytics for improving the precision of evaluations and providing more valuable details. In addition, how to use artificial intelligence to enhance the intelligent analysis ability is of great significance in the future researches.

Acknowledgments. This research was supported in part by Qtone Education of Ministry of Education of China (No. 2017YB115) and Shanghai Pujiang Program (No.16PJC007). Many thanks to Dr. Hongzhi Hu for her assistance to Prof. Weihui Dai and Xinshuang Fu who are the joint corresponding authors of this paper.

References

1. Ahmad, K.: Affective computing and sentiment analysis. emotion, metaphor and terminology. IEEE Intelligent Systems 31(2), 102–107 (2016)
2. Buyck, J.M., Tulkens, P.M., Bambeke, F.V.: Pharmacodynamic evaluation of the intracellular activity of antibiotics towards pseudomonas aeruginosa pao1 in a model of thp-1 human monocytes. Antimicrobial Agents & Chemotherapy 57(5), 2310–2318 (2013)
3. Chang, C.W., Gao, G.D., Chen, H., Li, W.X.: Study on the diagnosis of parkinson's disease with artificial neural network. Chinese Journal of Clinical Rehabilitation 7(28), 3818–3819 (2003)
4. Dai, W., Han, D., Dai, Y., Xu, D.: Emotion recognition and affective computing on vocal social media. Information & Management 52(7), 777–788 (2015)
5. Davis, C., Abraham, J.: The socio-political roots of pharmaceutical uncertainty in the evaluation of 'innovative' diabetes drugs in the european union and the us. Social Science & Medicine 72(9), 1574–1581 (2011)
6. Di, L., Du, Y.P.: Application of lda model in microblog user recommendation. Computer Engineering 40(5), 1–6 (2014)
7. Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., Binder, M.: A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of Biomedical Informatics 34(1), 28–36 (2001)
8. Fleuren, W.W., Alkema, W.: Application of text mining in the biomedical domain. Methods 74, 97–106 (2015)
9. Ghazi, M.R., Gangodkar, D.: Hadoop, mapreduce and hdfs: A developers perspective. Procedia Computer Science 48, 45–50 (2015)
10. Ghobrial, O., Derendorf, H., Hillman, J.D.: Pharmacokinetic and pharmacodynamic evaluation of the lantibiotic mu1140. Journal of Pharmaceutical Sciences 99(5), 2521–2528 (2010)
11. Gómez, E., Torres, M.J., Mayorga, C., Blanca, M.: Immunologic evaluation of drug allergy. Allergy Asthma & Immunology Research 4(5), 251–263 (2012)
12. Karaolis, M., Moutiris, J.A., Papaconstantinou, L., Pattichis, C.S.: Association rule analysis for the assessment of the risk of coronary heart events. In: IEEE International Conference on Engineering in Medicine and Biology Society. pp. 6238–6241 (2009)
13. Koukol, O., Kelnarová, I., Cerný, K.: Recent observations of sooty bark disease of sycamore maple in prague (czech republic) and the phylogenetic placement of cryptostroma corticale. Forest Pathology 45(1), 21–27 (2015)
14. Lee, T., Lee, H., Rhee, K.H., Shin, U.: The efficient implementation of distributed indexing with hadoop for digital investigations on big data. Computer Science & Information Systems 11(3), 1037–1054 (2014)
15. Li, X., Zhang, C.: Research on enhancing the effectiveness of the chinese text automatic categorization based on ictclas segmentation method. In: 2013 IEEE 4th International Conference on Software Engineering and Service Science. pp. 267–270 (2013)
16. Li, Y.M.: Research on Chinese health risks model and risk appraisal. Fourth Military Medical University, Xi'an, China. (2011)

17. Magnusson, M., Jonsson, L., Villani, M., Broman, D.: Parallelizing lda using partially collapsed gibbs sampling. Statistics 24(2), 301–327 (2015)
18. Mochón, M.C.: Social network analysis and big data tools applied to the systemic risk supervision. Ijimai 3(6), 34–37 (2016)
19. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M.: Fast collapsed gibbs sampling for latent dirichlet allocation. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, Usa, August. pp. 569–577 (2008)
20. Qu, H.Y., Zhao, W.: A revised bmm and rmm algorithm of chinese automatic words segmentation. Advanced Materials Research 267, 199–204 (2011)
21. Sampathkumar, H., Chen, X.W., Luo, B.: Mining adverse drug reactions from online healthcare forums using hidden markov model. BMC Medical Informatics and Decision Making 14(1), 1–18 (2014)
22. Seo, Y.D., Ahn, J.H.: Hadoop-based integrated monitoring platform for risk prediction using big data. Applied Mechanics & Materials 826, 113–117 (2016)
23. Silva, C., Ribeiro, B.: On text-based mining with active learning and background knowledge using svm. Soft Computing 11(6), 519–530 (2007)
24. Stegenga, H., Chambers, M., Jonsson, P., Thwaites, R., Garner, S.: A framework to guide the use of real-world evidence to support evaluation of relative effectiveness of new medicines. Value in Health 19(7), A488–A488 (2016)
25. Wang, M.: Progress and attention of new drug evaluation research. Journal of Chifeng University 24(4), 19–21 (2008)
26. Wei, X.U., Zhang, M.J., Xiong, Z.H.: Feature point matching based on bidirectional maximal correlation and parallactic restriction. Computer Engineering & Applications 44(28), 155–157 (2008)
27. Woodard, J.: Big data and ag-analytics: An open source, open data platform for agricultural & environmental finance, insurance, and risk. Agricultural Finance Review 76(1), 15–26 (2016)
28. Xu, L.H., Lin, H.F., Pan, Y., Ren, H., Chen, J.M.: Constructing the affective lexicon ontology. Journal of the China Society for Scientific & Technical Information 27(2), 5–7 (2008)
29. Yang, C.C., Luk, J.W.K., Yung, S.K., Yen, J.: Combination and boundary detection approaches on chinese indexing. Journal of the Association for Information Science & Technology 51(4), 340–351 (2000)
30. Yao, X., Ding, J., Liu, Y., Li, P.: The new drug conditional approval process in china: Challenges and opportunities. Clinical Therapeutics 39(5), 1040–1051 (2017)
31. Ye, Q., Wu, Q.: Study on report of american media of traditional chinese medicine situation and key words. AMIA Symposium proceedings 7(8), 626–629 (2014)
32. Yu, H.: Analysis on the subject and emotion of the medical forum of cerebrovascular disease. Beijing Jiaotong University, Beijing, China. (2016)
33. Yue, X., Jing, Y.: Research based on the algorithm of dna sequences data mining. Journal of Biomathematics 24(2), 363–368 (2009)
34. Zhou, L., Srinivasan, P.: Concept space comparisons: Explorations with five health domains. AMIA Symposium proceedings 2005, 874–878 (2005)
35. Zhu, Y.Y., Yun, X.: (dna) sequence data mining technique. Journal of Software 18(11), 2766–2781 (2007)

Genghui Dai is currently a Ph.D. candidate at the School of Marine Sciences, Sun Yat-Sen University, China. He received his master degree in Zoology from East China Normal University, China in 2005. His research interests include marine biology, microbial mechanism, and innovative drugs. Contact him at daigengh@mail2.sysu.edu.cn.

Xinshuang Fu, is currently a lecturer at the School of Management, Shanghai University, China. She received her Ph.D. in Management Science and Engineering from Shanghai University, China in 2013. Her research interests include knowledge management and information science. Contact her at gracief@126.com.

Weihui Dai is currently a professor at Department of Information Management and Information Systems, School of Management, Fudan University, China. He received his Ph.D. in Biomedical Engineering from Zhejiang University, China in 1996. He serves as a committee member of Shanghai Chapter, China Computer Society, and deputy director of Research and Translational Expert Council, Professional Committee of Endovascology, Chinese Medical Doctor Association. His recent research interests include complex system modeling and simulation, social media and intelligent information processing, social neuroscience and emotional intelligence, etc. Dr. Dai became a member of IEEE in 2003, a senior member of China Computer Society in 2004, and a senior member of Chinese Society of Technology Economics in 2004. His works have appeared in international journals with more than 130 papers. Contact him at whdai@fudan.edu.cn.

Shengqi Lu is currently a Ph.D. candidate at the School of Information Management and Engineering, Shanghai University of Finance and Economics, China. He received his master degree in Software Engineering from Fudan University, China in 2006. His current research interests include Artificial Intelligence and Business Intelligence. Contact him at shengqilu@fudan.edu.cn.

Received: April 13, 2017; Accepted: August 25, 2017.

Computer Science and Information Systems 14(3):823–837 https://doi.org/10.2298/CSIS161229031W

Sentiment information Extraction of comparative sentences based on CRF model

Wei Wang, Guodong Xin*, Bailing Wang*, Junheng Huang, and Yang Liu

School of Computer Science and Technology, Harbin Institute of Technology, 150001, Harbin, China
{wwhit, gdxin,wbl, hithjh}@hit.edu.cn, lyylwhhit@126.com

Abstract. Comparative information mining is an important research topic in the sentiment analysis community. A comparative sentence expresses at least one similarity or difference relation between two objects. For example, the comparative sentence "The space of car A is bigger than that of car B and car C" expresses two comparative relations <car A, car B, space, bigger> and <car A, car C, space, bigger>. This paper introduces conditional random fields model to extract Chinese comparative information and focuses on the task of element extraction from comparative sentences. We use the conditional random fields model to combine diverse lexical, syntactic and semantic features derived from the texts. Experiments show that the proposed method is competitive and domain-independent, with promising results.

Keywords: information extraction, comparative sentence, comparative element.

1. Introduction

Whenever people need to make a decision, they commonly want to know about others' views, attitudes and sentiments. A comparative sentence provides an important insight into how an entity or event is compared to other entities or events, which could effectively help people make decisions. For example, *" X旅馆比Y旅馆更干净，尽管房间的价格相同(Hotel X is cleaner than Hotel Y, although its price is the same as Y.)"*. Such opinion about comparison, directly comes from customers, could provide greater help for those who have potential consumer demands, but also help business executives to automatically track the attitudes and emotions of customers in the on-line forums, determine whether the customers are satisfied with their products and services, and capture the information of competitors. Therefore, the development of effective methods to automatically analyze opinions, especially comparative opinions, has become an urgent need [8,22,15,21,12,25,5].

The processing object of comparative sentiment analysis(SA) is comparative sentences in evaluative texts, the task is to extract and analyze the opinion elements in the comparative sentences, including judging the tendency of each comparative relation and extracting the various elements related to the tendency. These elements include compared entities, compared aspects, comparative words and opinion words. For example, the comparative sentence *"X手机比Y 手机有更好的用户体验.（Phone X has better user experience than phone Y.） "*. We extract 'phone X' as a subject entity(SE), 'phone Y' as object entity (OE), 'user experience' as a compared aspect (CA), and 'better' as an opinion phrase (OP) related to 'phone X', 'than' as a comparative keyword (CK).

The primary task of comparative sentiment analysis is to locate and extract the comparative elements in sentences, and then to determine the emotional tendency of the author for different objects according to the extracted contents. Information extraction in comparative sentences is different from that in regular opinion sentences. It extracts the objects with comparative relation and their shared aspects, rather than extracts a single entity or aspect that is directly evaluated by the author. The comparative relations between entities are usually reflected by the comparative words, so this paper introduces the comparative word candidate features and heuristic position features to improve the system's ability to identify compared entities. In addition, the comparative element extraction has the following problems:

Problem 1: How to fully identify phrase-level elements, for example, a product name may be consisted of a brand name and a model name. If we only extract a part of them, it will cause the lack of information. Therefore, we introduce shallow syntactic features to enhance the ability of system to identify phrase-level elements.

Problem 2: How to distinguish between different types of elements, for example, the POS tags for SEs, OEs and CAs are usually nouns or noun phrases. Therefore, it is difficult to distinguish these three types of elements. But Relative positional relations between them and comparative words have some directive functions.

To sum up, in order to construct a general comparative element extraction system, we introduce some linguistic features and heuristic features, such as shallow syntactic features, comparative word candidate features, and heuristic position features. In the case of no increase of domain knowledge, the performance of comparative element extraction is improved effectively, which shows the effectiveness of the proposed method.

The remainder of the paper is organized as follows, section 2 presents related work. Section 3 describes the method for comparative element extraction from comparative sentences. After that, the experiment results and the future directions are given in section 4.

2. Related work

There are many unsupervised methods [6,14,8,22,21,3,24] for aspect term extraction in review texts. Hu and Liu[6] first study the problem, they extract aspect terms through the association rule mining. And then they employ opinion words to mine infrequent aspect terms. Many of subsequent studies use the relationships between opinion words and aspect words to extract the aspect terms and opinions. In Qiu's[14] work, dependency relations are used as key clues, and the dual propagation approach is proposed to extract aspect words and opinion words by propagating information between opinion words and aspect words. The method is a semi-supervised bootstrapping process because the use of opinion word seeds. The purpose of comparative element extraction is to obtain various components associated with the comparative statement. Jindal and Liu first define the comparative element extraction problem [8] , where they deem a comparative sentence that describes a comparative relation that is consisted of five fundamental elements: comparative keyword, two compared entities, compared aspect and comparative type. They present a new method based on sequence rule mining called as "label sequence rule(LSR)" . The LSR method can extract the elements in a single comparative relation. Yang and Ko[22] propose an alternative approach that marks comparative element candidates based on part of speech (POS) type, then constructs POS sequence patterns for each

candidate and treats them as features of machine learning algorithm. These two works are based on context POS information to obtain comparative elements with a certain type of POS. In addition, some researchers use the dictionary including the domain data to mine comparative elements. Xu et al. [21] compile a product dictionary and an attribute dictionary in mobile phone domain by collecting some common product names and attribute names manually in corresponding domain. Feldman et al [3] build a brand dictionary for running shoes and cars respectively and recognize product model by developing a set of regular expressions for the model-names. The approaches based on dictionary are domain-related, which have many limitations.

There are some researchers adopting bootstrapping technique to extract compared entities. Li et al [12] develop a weakly-supervised bootstrapping method for automatic compared entities mining from online comparative questions. In their study, the algorithm starts bootstrapping process with a single extraction pattern. Using it, a set of initial seed comparator pairs are extracted from a question collection. Next, new extraction patterns are generated from comparator pairs and the comparative questions containing comparator pairs. The algorithm iterates until no more new patterns are found from the question collection. In Ding et al study [1], the bootstrapping process starts with a few seed entities. From them, the algorithm iteratively find more entities in a document set. At each iteration, sequence patterns are mined to find more entities based on already found entities. Obviously, their works are weakly-supervised and do not need to label a large number of corpora.

Supervised methods [20,7,23,18,9] often treat aspect and opinion word extraction as a sequence labeling problem, and the Conditional Random Fields (CRF) is one of the most main-stream methods used for sequence labeling tasks. Xing et al [20] select keywords, noun phrases and their position information as the features of CRF model to build up an element extraction model for identifying technical indices in standard technical comparative sentences. Huang et al. [7] first identify compared entities in sentences using CRF model. Then they distinguish compared subjects from compared objects based on the relative position between the entity and keyword. Yin et al.[23] extract aspect terms based on the features of distributed representations of words and dependency paths. They regard multi-hop dependency paths as a sequence of syntactic relations. In learning the embedding features, they not only use the word, but also consider the richer context information, such as neighbor words, and the dependency context information. Wang et al. [18] propose a new model by integrating recursive neural networks and conditional random fields into a unified framework for aspect and opinion term extraction. The recursive neural networks can learns high-level features by utilizing the double propagation of aspect-opinion pairs in the dependency tree. The learned features are input into the CRF model to capture the context of each word for aspect and opinion term extraction.

Representation learning has been successfully applied to natural language processing, such as information extraction, sentiment analysis [19,16,2] and so on. It represents text in different granularities with a low-dimensional dense vector, which includes context semantic information. Wang et al. [19] perform aspect level sentiment classification using an Attention-based LSTM (Long Short-Term Memory) networks. Attention can focus on different parts of a sentence when different aspects are used as input. Tang et al. [16] design a deep memory network with multiple computational layers for aspect level sentiment classification. Each layer of the deep memory network is an attention model with an

external memory to calculate the importance of each context word of a given aspect. Dong et al.[2] employ adaptive recursive neural network(AdaRNN) to perform target-dependent Twitter sentiment classification. AdaRNN uses more than one composition functions and adaptively choose them based on the context and linguistic tags.

In the context of comparative element extraction, there are some scholars converting element extraction task into semantic role labeling task. Wang et al [17] define three types of comparative patterns (eg. <entity> <keyword> <entity> <sentiment word>) to describe the relation among comparative elements. They employ the generalization comparative patterns to label comparative elements. Li [11] constructs semantic role parsing trees by utilizing semantic role labeling package and Stanford parser. They calculate the similarity between two sub-trees to label comparative elements. However, the above works can just obtain elements in a single comparative relation.

The work to determine entities preferred by reviewer has also been explored. Ganapathibhotla and Liu [4] primarily deal with context-sensitive sentiments by exploiting external information available on the Web. In this study, we use the Conditional Random Fields (CRF) learning algorithm to identify comparative elements. Lafferty et al [10] first introduce CRF for segmenting and labeling sequence data.

This paper uses the supervised method to extract comparative elements. Compared with the existing studies, our method makes full use of the various lexical, syntactic and heuristic information of the comparative sentence. And multiple key elements of comparative sentences are extracted at the same time.

3. Methods

3.1. Comparative sentence key concepts

Comparative information plays an important role in dealing with some practical problems, such as decision-making, opinion summarization, etc. Here, we give some basic definitions of comparative information mining(CIM) at sentence level.

Definition (comparative information mining): CIM is a problem of finding the comparison information between entities in text documents, which can be decomposed into the following main subtasks: I. Identify comparative sentences. II. Extract comparative elements and relations.

Definition (comparative sentence): A comparative sentence is a sentence that expresses one or more comparative relations between objects, which means that there may be more than one comparative relation in a sentence.

A comparative sentence can be explicit, e.g., *" X*电视比*Y*电视画面清晰*.(TV X has a clearer picture than TV Y.)"* or implicit, e.g., *"X*手机有摄像功能，而*Y* 手机没有*. (Phone X has a camera function, but phone Y does not have.)"*.

Definition (comparative relation): A comparative relation describes a relation of similarity or difference between two objects on an aspect.

A comparative relation can be formally expressed as a 5-tuple: *(SE, OE, CA, OP, CK)*, which refers to subject entity, object entity, compared aspect, opinion phrase, and comparative keyword. Some elements in a comparative relation can be omitted. For example, in a superlative sentence, object entity is usually being omitted.

Definition (comparative keyword): A comparative keyword is an indicator of comparative relation, for example *'*比*(than) '*, *'*相似*(similar) '*, *'*不同*(different) '*, *'*最*(most)*

’ etc. There are not the specialized morphemes in Chinese, such as the -er/-est suffix, as the comparative characteristics.

Definition (compared entity): A compared entity is an object that is being compared with another object in a sentence, which can be a subject entity or an object entity. A compared entity can be almost anything, e.g., a people, a place, a product, an event, etc.

Definition (compared aspect): A compared aspect is an aspect on which two objects are being compared. An aspect can be explicit or implicit in a sentence, for example, *“*钻石的价格高于珍珠*.(The price of diamonds is higher than that of pearls.)”, and “*钻石比珍珠更昂贵*.(Diamonds are more expensive than pearls.)”*.

There are two main comparative types: gradable and non-gradable. Gradable comparison describes an order relationship of entities with regard to an aspect. For example, sentences comprising phrases such as ‘比…性能更好(better performance than)’, ‘低于(lower than)’, ‘相比…有所提高(improved…compared with)’ are typically classified to gradable comparison. We further divide gradable comparison into two sub-types, greater or less than comparison, and superlative comparison. The latter generally contains phrases such as ‘the most expensive’, ‘the best quality’ etc, for example, the sentence “在所有手机品牌中, iphone是最受欢迎的.(In all mobile phone brands, iphone is the most popular.) ” is a gradable superlative comparison where we extract ‘iphone’ as a subject entity, ‘popular’ as an opinion phrase, and ‘most’ as a comparative keyword.

Non-gradable comparison describes similarity or difference between entities, and does not express the order of entities. We further divide it into three sub-types, similarity comparison, difference comparison and implicit comparison. Non-gradable similarity comparison expresses the similarity of entities by using phrases such as ‘ 和…一样(the same…as, as …as)’, ‘和…相似(similar to, similarity between)’ in a sentence. For example, in a camera review, the sentence “The photo quality of camera X is as good as camera Y.” indicates that the similarity in picture quality between camera X and camera Y. Non-gradable difference comparison states the difference of entities on a certain aspect, and does not grade them. Phrases such as ‘different from’, ‘distinguish from’, ‘difference between’ can be the indicator of such type sentence. For example, in the sentence “The screen size of monitor X is different from that of monitor Y”, the user expresses the difference between monitor X and monitor Y in screen size, without ordering them based on the size of screen. Non-gradable implicit comparison implicitly states the difference of entities on one or more aspects, for example, the sentence “Entity X has aspect A1, but entity Y does not have.”.

3.2. Extraction of comparative elements

Comparative elements In this section, we describe how comparative elements are extracted from comparative sentences. The basic strategy is an integrated lexical, syntactic and semantic features and condition random fields learning approach to extract comparative elements.

There are four types of comparative elements to be extracted in our study: subject entity(*SE*), object entity(*OE*), compared aspect(*CA*), and opinion phrase(*OP*).

Example 1. *“*手机*X*的摄像头比手机*Y*的更好更实用*. (Phone X has a better and more practical camera than phone Y.)”*

Example 2. *“*在所有汽车中，*Z*性能最优越*. (The performance of Z is the most superior in all cars.)”*

In Example 1 sentence, *'手机X(phone X)'* is a SE, *'手机Y(phone Y)'* is an *OE*, *'摄像头(camera)'* is a *CA*, *'更好更实用(better and more practical)'* is a *OP*. In Example 2 sentence, *'Z'* is a *SE*, *'性能(performance)'* is a *CA*, *'优越(superior)'* is a *OP*.

There are two important problems need to be solved in the task of comparative element extraction: i) whether comparative elements are composed of only a single word; ii) how to distinguish SE, OE and CA that have similar POS tags.

Composition of Elements: comparative elements can be composed of one or more words. For instance, "better and more practical" is composed of multiple words in example 1. If we only extract one word "better" substituted for "better and more practical", some important information will be lost. We thus define that comparative elements can be composed of one or more words.

Distinction of Elements: Subject entity, object entity and aspect are mainly noun or noun phrase. So, we could not effectively distinguish them by only using the POS tags. Fortunately, we find that various elements commonly play different grammatical roles in comparative sentences. For instance, Subject entity is mainly as the subject of sentence, object entity acts as the object, and opinion phrase is as the predicate in the syntax function. Furthermore, we also find that subject entity is usually in the left of a keyword and object entity is in the right of a keyword. These linguistic clues are useful for distinction of SE, OE and CA elements.

Feature representation We introduce various linguistic-related features to extract comparative elements. Several preprocessing steps are executed towards comparative sentences, including word segmentation, POS tagging, phrase syntactic parsing. In this study, we use some basic linguistic features and more advanced ones as follows:

1) Words: A Word is the smallest linguistic unit that expresses natural language semantics. In western phonetic language, there is a clear delimiter between words. In Chinese, there is no obvious delimiter between words. Therefore, we first perform word segmentation for each sentence. Then each word in a sentence is used as a baseline feature of CRF model for Chinese information extraction work.

2) POS tags: part-of-speech tag is also a class of important features. Due to SE, OE and CA are mainly noun. Sentiment words are commonly adjective or verb. Comparative keywords are mainly preposition or adverb. Hence, POS tags are helpful for identifying different types of elements.

3) Chunks: Chunk division, also known as shallow parsing, is used to recognize independent components in a sentence whose structure is relatively simple, such as non-recursive noun phrases, verb phrases etc. The chunk labels are derived from syntactic parsing tree of a sentence. In a comparative sentence, SE and CA can be composed of noun or noun phrase. Keyword and OE usually form a preposition phrase. OP can be adjective or adjective phrase. So, chunk feature can contribute to identify comparative elements in phrase level.

4) Keywords: The keyword candidates in a comparative sentence are labeled by using a set of paired keywords e.g. *"与…不同(different...from)"*. A lexicon of 660 paired keywords is created by counting their co-occurrence frequency, and then pruning manually. The keyword candidates are useful for discriminating SEs from OEs in a comparative sentence.

5) Positions: Most of SEs are in the left of keywords and OEs are in the right of keywords in comparative sentences. By using the heuristic position information between entity and keyword can further distinguish SE from OE.

6) Contexts: The context of a word in a sentence can also affect the type of element. In this study, we use context within the radius of 3 of each target word in a sentence as feature. The context feature is set by feature template of CRF model.

The above linguistic features are automatically extracted by using Stanford segmenter, and Stanford parser.

Conditional random field model Conditional random fields (CRF) [10], which is an undirected probabilistic graphical model, has the following advantages for labeling and segmenting sequence data: i) CRF can effectively exploit the rich, global features of the inputs, and do not need to represent dependencies of the inputs. ii) Context information are taken into account by CRF, e.g., the linear chain CRF predicts sequences of labels for sequences of input samples in natural language processing. iii) Long-range dependencies between the inputs can be represented. The extraction of comparative elements involves multiple entities, rich features from the inputs, and long-range dependencies. Thus CRF is the very appropriate algorithm for modeling it.

In this paper, we adopt CRF++0.53 toolkit to execute training and labeling for model. The features extracted from the feature set are added to the model by setting feature template. Therefore, the feature selection problem is transformed into a feature template selection problem. This paper designs 6 feature templates based on the linguistic related features described above as shown in Table 1.

In Table 1, *w, t* denotes word and POS tag feature respectively. *c, l* represents comparative word candidate and heuristic position feature. *s* denotes shallow parsing feature. In order to verify the effective of syntactic and heuristic features, we build 6 feature templates in the experiments. Followed by the lexical level (baseline) feature template(T1), comparative word candidates are added to T1 template(T2), comparative word candidate and heuristic position and word features(T3), comparative word candidate and heuristic position features are added to T1 template (T4), shallow parsing features are added to T1 template(T5), All features (T6).

4. Experimental evaluation

We conduct various experiments to evaluate the performance of the proposed methods for comparative element extraction task.

4.1. Experiment data

The experiment data is derived from task 2 of the fourth Chinese Opinion Analysis Evaluation (COAE2012) [13]. It consists of consumer reviews of automotive and electronic products. The sentence distribution of the data is shown in Table 2. The ratio of training set, development set and test set is 4: 4: 1.

The COAE2012 task 2 is divided into two sub-tasks. Task 2.1: Identify which sentences are comparative sentences in a given set of sentences. Task 2.2: Extract comparative elements from the identified comparative sentences, including compared entity and compared aspect, and determine the opinion direction of compared entities.

Table 1. Feature Template

Template	Feature	Feature Template
T1	*w,t*	w_n ,t_n $n \in \{-3, ..., 3\}$
		$w_{n-1}w_n$,$t_{n-1}t_n$ $n \in \{0, 1, 2\}$
		w_nw_{n+1} ,t_nt_{n+1} $n \in \{-2, -1, 0\}$
		$t_{n-1}t_n$ t_{n+1} $n \in \{-1, 0, 1\}$
		w_nt_n $n = 0$
T2	*w,t,c*	w_n ,t_n ,c_n $n \in \{-3, ..., 3\}$
		$w_{n-1}w_n$,$t_{n-1}t_n$,$c_{n-1}c_n$ $n \in \{0, 1, 2\}$
		w_nw_{n+1} ,t_nt_{n+1}, c_nc_{n+1} $n \in \{-2, -1, 0\}$
		$t_{n-1}t_n$ t_{n+1}, $c_{n-1}c_n$ c_{n+1} $n \in \{-1, 0, 1\}$
		w_nt_n ,t_nc_n $n = 0$
T3	*w,c,l*	w_n ,c_n ,l_n $n \in \{-3, ..., 3\}$
		$w_{n-1}w_n$,$c_{n-1}c_n$,$l_{n-1}l_n$ $n \in \{0, 1, 2\}$
		w_nw_{n+1} ,c_nc_{n+1}, l_nl_{n+1} $n \in \{-2, -1, 0\}$
		$c_{n-1}c_n$ c_{n+1}, $l_{n-1}l_n$ l_{n+1} $n \in \{-1, 0, 1\}$
		w_nc_n ,c_nl_n $n = 0$
T4	*w,t,c,l*	w_n ,t_n, c_n ,l_n $n \in \{-3, ..., 3\}$
		$w_{n-1}w_n$,$t_{n-1}t_n$,$c_{n-1}c_n$,$l_{n-1}l_n$ $n \in \{0, 1, 2\}$
		w_nw_{n+1} ,t_nt_{n+1} ,c_nc_{n+1}, l_nl_{n+1} $n \in \{-2, -1, 0\}$
		$t_{n-1}t_n$ t_{n+1},$c_{n-1}c_n$ c_{n+1}, $l_{n-1}l_n$ l_{n+1} $n \in \{-1, 0, 1\}$
		w_nt_n ,t_nc_n ,t_nl_n ,c_nl_n $n = 0$
T5	*w,t,s*	w_n ,t_n ,s_n $n \in \{-3, ..., 3\}$
		$w_{n-1}w_n$,$t_{n-1}t_n$,$s_{n-1}s_n$ $n \in \{0, 1, 2\}$
		w_nw_{n+1} ,t_nt_{n+1}, s_ns_{n+1} $n \in \{-2, -1, 0\}$
		$t_{n-1}t_n$ t_{n+1}, $s_{n-1}s_n$ s_{n+1} $n \in \{-1, 0, 1\}$
		w_nt_n ,t_ns_n $n = 0$
T6	*w,t,c,l,s*	w_n ,t_n, c_n ,l_n ,s_n $n \in \{-3, ..., 3\}$
		$w_{n-1}w_n$,$t_{n-1}t_n$,$c_{n-1}c_n$,$l_{n-1}l_n$,$s_{n-1}l_n$ $n \in \{0, 1, 2\}$
		w_nw_{n+1} ,t_nt_{n+1} ,c_nc_{n+1}, l_nl_{n+1}, s_nl_{n+1} $n \in \{-2, -1, 0\}$
		$t_{n-1}t_n$ t_{n+1},$c_{n-1}c_n$ c_{n+1}, $l_{n-1}l_n$ l_{n+1} ,$s_{n-1}s_n$ s_{n+1} $n \in \{-1, 0, 1\}$
		w_nt_n ,t_nc_n ,t_nl_n ,t_ns_n ,c_nl_n,c_ns_n ,l_ns_n n=0

Table 2. Sentence distribution of the data

Type	Sentence distribution
Comparatives	1624(16.92%)
Non-comparatives	7976(83.08%)
Total	9600(100%)

Task 2.2 marks three parts: ProductName, FeatureName and Polarity. Our study is similar to task 2.2. In this task, the coverage is used to assess for the consistency. Set *x, y* are the results of different people annotation, coverage is defined as follows:

$$Coverage(x, y) = len(x \cap y)/len(x) * 100\% \quad (1)$$

Where *len(x)* represents the length of *x*, $x \cap y$ is the intersection of *x* and *y*. We set coverage is 0.2 in the experiment.

4.2. Evaluation methods

Task 2.2 is an information extraction task. It is difficult to determine the boundary of ProductName and FeatureName for task 2.2. Therefore, evaluation adopts two indicators: accurate evaluation and coverage evaluation.

Accurate evaluation: the extracted entity exactly matches with the answer. For example, when the answer is 'screen resolution', it is incorrect result to submit either 'screen' or 'resolution'.

Coverage evaluation: the extracted entity has overlap portion with the answer. In the above example, it is correct result if we submit 'screen' or 'resolution'.

4.3. Experimental results

Experimental results of comparative element extraction We use the comparative sentences in the automotive and electronic fields in COAE2012 task 2 to extract the comparative elements, a total of 1600 comparative sentences. Most of these sentences are typical comparative sentences, and a few implicit comparisons. The distribution of comparative elements is shown in Table 3. Stanford parser is used to perform phrase syntactic parsing. The experimental results are an average of 5 fold cross validation. We use two evaluation methods, accurate evaluation and coverage evaluation to measure the performance of system. The experiment results are shown in Table 4 where SUB represents subject entities, OBJ represents object entities, ATTR represents aspect names, OPIN represents evaluation words or phrases.

Table 3. The distribution of comparative elements in two fields

Field	Comparative Sentences	Subject Entity	Object Entity	Compared Aspect	Keyword	Opinion Phrase
Car	800	650	810	836	1421	831
Electronic	800	505	860	687	943	802

Table 4 shows the average result of element extraction in two fields. When introducing all features (T6 template), the results of element extraction are superior to other feature combinations (T1-T5 template). When using T1 template, the performance of system is poor, particularly recall.

Because T1 template contains only lexical level features, such as words and POS tags, these features can provide limited information for classification task, and the information

Table 4. The average results of 5-fold cross validation(%)

Element	Template	Accurate Evaluation			Coverage Evaluation		
		Precision	Recall	F1-score	Precision	Recall	F1-score
SUB	T1	67.43	39.03	48.78	74.91	41.53	53.44
	T2	68.47	41.57	50.99	73.29	47.35	57.53
	T3	73.12	32.00	43.83	76.41	36.44	49.35
	T4	70.25	41.81	51.51	75.66	48.01	61.29
	T5	66.08	37.94	48.21	72.25	42.19	53.12
	T6	71.61	41.36	51.54	80.44	50.31	61.90
OBJ	T1	81.60	66.93	73.36	83.00	69.11	75.42
	T2	81.57	69.83	74.99	84.72	72.02	77.86
	T3	78.05	70.63	73.82	78.77	72.71	75.62
	T4	80.75	73.77	76.90	87.88	76.89	82.02
	T5	81.78	66.13	73.02	83.86	68.25	75.25
	T6	82.22	73.03	77.18	91.69	77.21	83.24
ATTR	T1	72.80	48.38	58.13	78.17	50.04	61.02
	T2	74.43	52.83	61.80	79.96	55.57	65.57
	T3	76.51	39.69	52.27	80.66	42.84	55.96
	T4	73.88	52.11	61.11	78.88	55.31	65.03
	T5	71.36	49.71	58.12	75.06	51.73	61.25
	T6	73.70	51.74	60.80	81.95	55.91	66.47
OPIN	T1	87.12	61.67	72.17	89.15	62.05	73.17
	T2	87.38	64.55	74.19	88.98	66.48	76.10
	T3	88.69	64.26	74.44	88.77	66.10	75.78
	T4	87.45	68.47	76.74	90.51	68.98	78.29
	T5	86.34	62.95	72.70	88.45	64.97	74.91
	T6	87.08	68.85	76.86	89.30	71.15	79.20

obtained contains some noise. T2 template expands features from lexical level to heuristic information. It adds keyword candidate feature that makes the evaluation indicators to be significantly raised. The performance of the T3 template is polarized. On the one hand, the worst performance is gotten for SE and CA identification. Because the T3 template does not contain part of speech tag feature, it is the primary indication of the subject entities and attributes. On the other hand, keyword candidate and heuristic position features are added to T3 template, which improve the performance of OE and OP identification.

T4 template adds keyword candidate and heuristic information on the basis of T1 template, and provides the position information of other elements relative to candidate keywords in the sentence. The recall and F1-score are greatly improved, which show that keyword candidate and heuristic position features are very effective in the comparative element extraction problem.

However, T4 template has limited recognition ability for phrase level elements. Thus, T5 template expands features from lexical level to phrase level. It adds shallow parsing feature which improves the F1-score of CA and OP, but decrease the F1-score of other elements. T6 template, which includes all features, greatly increases the recall and F1-score of system. This proves that various features, such as lexical, syntactic, and heuristic features, are effectively for the comparative element extraction.

Table 4 compares the performance of the system for accurate evaluation and coverage evaluation. The best performance is obtained when we use coverage evaluation. This means that the system can correctly locate the comparative elements, but the ability to accurately identify the boundaries of the elements is limited. As we can see in Table 3, the precision is relatively high, while the recall is low in each of results. One possible reason is that multiple feature decisions improve the precision of system, while reduce the recall. The other reason is that domain knowledge is not introduced into the system. Experimental results show that the annotated results of OE and OP are better than those of SE and CA. Since the positions of OE and OP in the sentence are relatively fixed, they are commonly in the right of keyword or are degree adverbs. While a number of SEs to be omitted, and the positions of CA to be unfixed increase the difficulty of identifying them.

As the conditional random fields is a supervised learning method, there are domain adaptability problems. In order to verify the effectiveness of our method, for the car field, we use the electronic field corpus as training set. Similarly, for the electronic field, we use the car field corpus as training set. The average of these experiments is taken as the final experimental result.

Table 5. The results of domain cross annotation(%)

Element	Template	Accurate Evaluation			Coverage Evaluation		
		Precision	Recall	F1-score	Precision	Recall	F1-score
SUB	T6 Template	58.15	18.51	27.91	64.53	24.74	35.77
OBJ	T6 Template	79.00	58.42	66.65	85.60	61.70	71.71
ATTR	T6 Template	63.53	37.05	45.14	67.75	40.28	50.52
OPIN	T6 Template	80.43	60.27	68.78	82.67	62.59	71.24

By comparing table 5 with table 4, we find that, in Table 5, the model established by domain cross training has a substantial decrease in the performance of element extraction compared to table 4. Among them, the subject entities and compared aspects have the biggest decrease, and the comparative keywords and opinion words have a smaller decrease. The reason is that the subject entities, object entities and compared aspects are domain related. For example, a subject entity or object entity is usually a brand name or product name in a domain, and an aspect is a component or characteristic of a product. Thus, these three elements are domain related. On the other hand, the position of the subject and attribute varies greatly in the sentence, and aspect is not easily distinguished from subject entity. Therefore, domain cross annotation has the greatest impact on the subject entity and attribute. Since the position of object entity is relatively fixed, the recognition performance is better than that of the subject entity and attribute. Because of the small domain correlation of opinion words, and its recognition performance is relatively good.

Compare with COAE2012 evaluation results The average result of element extraction using T6 template in our experiments is compared with the max value of evaluation results in COAE2012. In contrast experiment, the average of 5-fold cross validation is adopted. The result is shown in Table 6, where PROD represents product name(SE and OE), ATTR represents aspect name.

Table 6. The result contrast on COAE2012 data (%)

Element	Method	Accurate Evaluation			Coverage Evaluation		
		Precision	Recall	F1-score	Precision	Recall	F1-score
ATTR	T6 Template	73.70	51.74	60.80	81.95	55.91	66.47
	Max value	66.05	62.52	60.78	77.94	67.51	65.69
PROD	T6 Template	76.92	57.20	64.36	86.07	63.76	72.57
	Max value	67.77	66.05	64.30	82.67	73.58	71.58
PROD+	T6 Template	75.84	55.38	63.17	84.69	61.14	70.54
ATTR	Max value	60.81	53.89	52.55	67.45	58.56	57.00

Table 6 shows that F1-scores of extracting entity, aspect, entity and aspect are higher than the max values of COAE2012, which indicate the proposed method in this paper is effective. Table 6 shows the precision is higher, and the recall is lower in each result. On the one hand, the positions of SEs and CAs in comparative sentences are too flexible to capture, and SEs are often omitted in comparative sentences, which can affect the mean recall of system. On the other hand, we do not introduce any domain knowledge, such as domain knowledge base, domain dictionary in the process of element extraction. If the domain dictionary is introduced in the model training phase, it will improve the recall of the extraction results. However, the cost of the artificial domain dictionary is relatively large and can not be applied to other fields. Therefore, in order to improve the recall of the system, it is necessary to find more effective features of the universal domain to solve the problem. In addition, the indices of all coverage matching of the system are higher than those of the accurate matching. It shows that the system can be more accurate to locate various elements, but the boundary identification is not accurate enough. The

reason is mainly from the accumulation of errors in the Natural Language Processing tools at the bottom, including word segmentation, part of speech tagging and syntactic analysis tools. Therefore, the improvement of low-level language processing technology is of great significance for improving the accuracy of information extraction.

Performance analysis of sentences with multiple comparative relations A comparative sentence can contain one or more comparative relations. In the car corpus, 25.4% of sentences contains more than one comparative relation. Therefore, It is necessary to analyze the element extraction performance of these sentences. The experimental results are shown in Table 7.

Table 7. The performance of element extraction in multi-relation sentences(%)

Element	Accurate Evaluation			Coverage Evaluation		
	Precision	Recall	F1-score	Precision	Recall	F1-score
SUB	65.97	43.75	52.61	70.93	44.01	54.32
OBJ	77.62	85.28	81.27	79.42	86.35	82.74
ATTR	76.73	58.54	66.41	78.38	58.90	67.26
OPIN	96.18	64.09	76.92	96.50	65.12	77.76

Table 7 shows that recall and F1-score of accurate evaluation are significantly improved in sentences with multiple comparisons. F1-score of coverage evaluation decrease significantly.

5. Conclusions and Future Work

This paper studies the problem of comparative element extraction in the comparative sentences. Conditional random fields model is employed to extract comparative elements, which fuses various lexical, syntactic and heuristic features. A comparative element extraction model is constructed by using the supervised method. The performance indices of the element extraction are improved. The experiment results show that the shallow syntactic features can effectively identify the phrase-level comparative elements. The comparative keyword candidate features can not only compensate for the lack of comparative words in the training samples, but also make a preliminary locating of other elements. Heuristic position features are helpful to distinguish between elements such as subject entities and object entities. All the features introduced in the model are domain-independent, so the method can be applied directly to other areas. In the future, we plan to find more effective features that represent a sentence to further improve the recall of our system. We also plan to summarize extracted information into an opinion summarization.

References

1. Ding, X., Liu, B., Zhang, L.: Entity discovery and assignment for opinion mining applications. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1125–1134. ACM, Paris, France (2009)

2. Dong, L., Wei, F., Tan, C., Tang, D.: Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. pp. 49–54. Association for Computational Linguistics, Baltimore, Maryland, USA (2014)
3. Feldman, R., Fresko, M., Goldenberg, J., Netzer, O., Ungar, L.: Extracting product comparisons from discussion boards. In: Proceedings of the Seventh IEEE International Conference on Data Mining. pp. 469–474. IEEE, Omaha, Nebraska, USA. (2007)
4. Ganapathibhotla, M., Liu, B.: Mining opinions in comparative sentences. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 241–248. Association for Computational Linguistics, Manchester, UK (2008)
5. He, H., Li, Z., Yao, C., Zhang, W.: Sentiment classification technology based on markov logic networks. New Review of Hypermedia and Multimedia 22(3), 243–256 (2016)
6. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceedings of the Nineteenth National Conference on Artificial Intelligence. pp. 755–760. Association for the Advancement of Artificial Intelligence, San Jose, California (2004)
7. Huang, G.H., Yao, T.F., Liu, Q.: Mining chinese comparative sentences and relations based on crf algorithm. Application Research of Computers 27(6), 2061–2064 (2010)
8. Jindal, N., Liu, B.: Mining comparative sentences and relations. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence. pp. 1331–1336. Association for the Advancement of Artificial Intelligence, Boston, Massachusetts (2006)
9. Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S.: Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 437–442. COLING, Dublin, Ireland (2014)
10. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning. pp. 282–289. ACM, Williamstown, MA, USA (2001)
11. Li, J.J.: Research on the identification of comparative sentences and relations and its application. Master's thesis, Chongqing University, China (2010)
12. Li, S., Lin, C.Y., Song, Y.I., Li, Z.: Comparable entity mining from comparative questions. IEEE Transactions on Knowledge and Data Engineering 25(7), 1498–1509 (2013)
13. Liu, K., Wang, S., Liao, X., Xu, H.: Overview of chinese opinion analysis evaluation. In: Proceedings of the 4st Chinese Opinion Analysis Evaluation. pp. 1–32. Chinese Informaiton Processing Society of China, Nanchang, Jiangxi, China (2012)
14. Qiu, G., Liu, B., Bu, J., Chen, C.: Opinion word expansion and target extraction through double propagation. Computational Linguistics 37(1), 9–27 (2011)
15. Shi, L., Li, S., Jiang, P., Liu, H.: Improving comparative sentence extraction of chinese product reviews by sentiment analysis. Journal of Engineering Science & Technology Review 9(6), 149–156 (2016)
16. Tang, D., Qin, B., Liu, T.: Aspect level sentiment classification with deep memory network. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 214–224. Association for Computational Linguistics, Austin, Texas, USA (2016)
17. Wang, S., Li, H., Song, X.: Automatic semantic role labeling for chinese comparative sentences based on hybrid patterns. In: Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence. pp. 378–382. IEEE, Sanya, China (2010)
18. Wang, W., Pan, S.J., Dahlmeier, D., Xiao, X.: Recursive neural conditional random fields for aspect-based sentiment analysis. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 616–626. Association for Computational Linguistics, Austin, Texas, USA (2016)
19. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 606–615. Association for Computational Linguistics, Austin, Texas, USA (2016)

20. Xing, L., Liu, L.: Chinese standard comparative sentence recognition and extraction research. In: Proceedings of the International Conference on Information Engineering and Applications. pp. 415–422. Springer, London, Chongqing, China (2013)
21. Xu, K., Liao, S.S., Li, J., Song, Y.: Mining comparative opinions from customer reviews for competitive intelligence. Decision Support Systems 50(4), 743–754 (2011)
22. Yang, S., KoJindal, Y.: Extracting comparative entities and predicates from texts using comparative type classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. pp. 1636–1644. Association for Computational Linguistics, Portland, Oregon (2011)
23. Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M.: Unsupervised word and dependency path embeddings for aspect term extraction. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. pp. 2979–2985. Association for the Advancement of Artificial Intelligence, New York, USA (2016)
24. Zhang, L., Liu, B., Lim, S.H., O' Brien-Strain, E.: Extracting and ranking product features in opinion documents. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 1462–1470. Association for Computational Linguistics, Uppsala, Sweden (2010)
25. Zhang, W., He, H., Cao, B.: Identifying and evaluating the internet opinion leader community based on k-clique clustering. Neural Computing and Applications 25(3-4), 595–602 (2014)

Wei Wang lecturer at Harbin Institute of Technology, her research interests cover natural language processing, sentiment analysis and information safety.

Guodong Xin lecturer at Harbin Institute of Technology, his research interests cover social networks and network security.

Bailing Wang professor at Harbin Institute of Technology, his research interests cover network security, computer network and social networks.

Junheng Huang associate professor at Harbin Institute of Technology, his research interests cover social networks, natural language processing and network security.

Yang Liu associate professor at Harbin Institute of Technology, his research interests cover network security, computer network and social networks.

Received: December 29, 2016; Accepted: August 25, 2017.

Computer Science and Information Systems 14(3):839–856 https://doi.org/10.2298/CSIS161230032K

Distinguishing Flooding Distributed Denial of Service from Flash Crowds Using Four Data Mining Approaches*

Bin Kong[1,2], Kun Yang[4,5], Degang Sun[4,5], Meimei Li*[3,4,5], and Zhixin Shi[4,5]

[1] School of Economics and Management, Beijing Jiaotong University
Beijing, China
pingpangfan@163.com
[2] National Secrecy Science and Technology Evaluation Center
Beijing, China
pingpangfan@163.com
[3] School of Computer and Information Technology, Beijing Jiaotong University
Beijing, China
limeimei@iie.ac.cn
[4] Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
{yangkun,sundegang,limeimei,shizhixin@iie.ac.cn}
[5] School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
{yangkun,sundegang,limeimei,shizhixin@iie.ac.cn}
Corresponding Author Email: limeimei@iie.ac.cn

Abstract. Flooding Distributed Denial of Service (DDoS) attacks can cause significant damage to Internet. These attacks have many similarities to Flash Crowds (FCs) and are always difficult to distinguish. To solve this issue, this paper first divides existing methods into two categories to clarify existing researches. Moreover, after conducting an extensive analysis, a new feature set is concluded to profile DDoS and FC. Along with this feature set, this paper proposes a new method that employs Data Mining approaches to discriminate between DDoS attacks and FCs. Experiments are conducted to evaluate the proposed method based on two real-world datasets. The results demonstrate that the proposed method could achieve a high accuracy (more than 98%). Additionally, compared with a traditional entropy method, the proposed method still demonstrates better performance.

Keywords: Flooding DDoS, Flash Crowds, Data Mining, Entropy.

1. Introduction

Distributed Denial of Service (DDoS) attacks have been wreaking havoc on the Internet, and these attacks show no signs of disappearing. In fact, attackers are constantly looking for new targets and new ways to deplete network performance [23], [21]. DDoS attacks are becoming more complex and sophisticated. Among all kinds of DDoS attacks, Flooding DDoS attacks are the most common and the most dangerous. Especially, when they

* This work is supported by the National Natural Science Foundation of China (61372062) and Major National Projects of China-The Core Electronic Components, High-end General Chips and Basic Software (CHBS) (2017ZX01045101).

happened under Flash Crowds (FCs) which have many similarities to the kind of DDoS attacks (more can be seen in Table 1) [5], [27], it usually render defense systems helpless.

FCs cause a large amount of traffic to surge simultaneously, causing dramatic stress on the server's network links and resulting in considerable loss of packets and network congestion at last. Although Flooding DDoS attacks are often launched by Botnets [33], [15], the master (attacker) in a Botnet orders compromised hosts (Bots) to simultaneously send packets to deplete the victims resources (e.g., memory, network bandwidth), eventually leaving the victim's system paralyzed [29].

Table 1. A Comparison Between DDoS And FC

Category	DDoS	FC
Network Status	Congested	Congested
Server Status	Overloaded	Overloaded
Traffic Type	Malicious	Genuine
Response to Traffic Control	Unresponsive	Responsive
Traffic Source	Any	Mostly Web
Flow Size	Any	Large Number of Flows
Predictability	Unpredictable	Mostly Predictable

Not only wired networks but also wireless networks [12] face resource constraints, such as limited bandwidth and less memory. Wireless networks also face other constraints, such as short communication ranges, less computational power, open channels, and short lifetimes. These characteristics of wireless networks make them vulnerable to anomalies. Any anomalies in a wireless ad hoc network degrade the overall performance of the network. DDoS attacks are one of these network anomalies, they are still severe attacks on wireless networks and may not be easily identifiable from FCs.

Due to plenty of similarities existed in DDoS and FC, all of them make the issue of discriminating DDoS and FC hardly to be tackled [29]. In the case of FCs, the high volume of traffic generated by legitimate users needs to be serviced by provisioning extra resources, whereas in the case of DDoS attacks, traffic generated by Bots needs be filtered as early as possible. Therefore, DDoS attacks and FCs should to be treated in different ways. To solve this discrimination issue, we extensively analyze these two phenomena and find that a few abnormal statistical features exist in DDoS attacks and FCs. With these features, we can translate the problem of differentiating DDoS attacks and FCs into ways to classify points in Euclidean n-spaces. As a result, we propose a method that employs Data Mining to discriminate between DDoS attacks and FCs.

This paper has been organized as follows: Section 2 reviews the currently available literatures. Section 3 concludes a new feature set and explains our proposed method in detail. Section 4 conducts experiments to evaluate this method and analyzes the results. Section 5 concludes our work.

2. Related Work

Since DDoS attacks first began in the early 2000s, considerable literatures have been published on detecting DDoS to avoid unnecessary economic loss, but little works related to the topic of distinguishing between DDoS attacks and FCs has been published [12], [31], [39]. Based on our understanding of the field, we simply divided the existing methods into two categories: Turing Test and Anomaly Behavior Analysis.

2.1. Turing Test

DDoS attacks are usually launched by Botnets, whereas FCs derive from legitimate clients; consequently, the problem of differentiating between DDoS attacks and FCs can be simplified to the problem of identifying whether the client is a human or a Bot. According to the client's responses, then distinguish whether the client is a normal user or not, that is Turing Test, which is the main and pervasive method for distinguishing DDoS attacks from FCs. The common Turing Test includes graphic puzzles, which display a slightly blurred or distorted picture or a puzzle and ask the user to type in the depicted symbols. This task is easy for humans yet hard for computers to answer.

CAPTCHAs (Completely Automated Public Turing Test to Tell Computers and Humans Apart) [36], [18] and AYAHs (Are You a Human) [1] are the most commonly used Turing tests. These methods, however, may cause some delays for normal users and usually annoy users with the increasingly difficult images employed. At the same time, various mechanisms, such as Reverse Turing Test, have also been developed by hacker communities to break these visual puzzles, which means that Turing Test will no longer completely defend against DDoS attacks or be able to distinguish DDoS attacks from FCs.

2.2. Anomaly Behavior Analysis

DDoS attacks mainly rely on Botnets, and Bots are usually executed by preprogrammed codes [14], whereas legitimate users are different individuals, and consequently, a few anomalies do exist between Bots and legitimate clients.

Jung et al. [17] first identified a few characteristics for discriminating DDoS attacks from FCs after analyzing various FC traces. The authors found that during FCs, most of the requests were generated either from those clients who had visited previously or from those clients who belonged to the same networks or administrative domains.

Xie et al. [37] proposed a novel method to detect anomaly events based on the hidden Markov model. This approach used the entropy of document popularity as the input feature to establish this model.

Ke et al. [20] proposed novel approaches using probability metrics to discriminate DDoS attacks from FCs. These methods efficiently identified the FC attacks from the DDoS attacks, reduced the number of false positives and false negatives, and also identified the attacks. Conversely, probability metric approaches failed to maintain the same accuracy for discriminating the FC attack from significant attack traffic.

Oikonomou et al. [25] tried to discriminate mimicked attacks from real FCs by modeling human behavior. The study is mainly based on the dynamic changes of requests and

the semantic meaning of requests and then builds the normal behavior model, which it used to distinguish Bots from normal visitors. This model is difficult to employ, however, for large-scale dynamic web pages because of the complicated process of establishing a transfer probability matrix.

Theerasak et al. [34] proposed a discrimination method based on packet arrival patterns. Pearsons correlation coefficient was used to measure packet patterns. These patterns are defined using the repeated properties observed from the traffic flow and are also calculated by the packet delay. Defining packet patterns is difficult, however.

Bhatia et al. [4], [5] proposed a technique combining the analysis of both network traffic features (e.g., incoming traffic volume, new source IP addresses, number of source IP addresses, and incoming traffic distribution) and server load characteristics (e.g., system-level CPU utilization, user-level CPU utilization, CPU load, and real memory utilization) to distinguish DDoS from FC. The computational complexity of this approach, however, is quite high.

Rabia et al. [19] reviewed the state-of-the-art detection mechanisms for the identified DDoS attacks in wireless body area networks (WBANs). The most serious threat to data availability is a DDoS attack that directly affects the all-time availability of a patients data. The existing solutions for standalone WBANs and sensor networks are not applicable in the cloud. Therefore, the purpose of this review was to identify the most threatening types of DDoS attacks affecting the availability of a cloud-assisted WBAN and review existing mechanisms to detect DDoS attacks.

Yu et al. [40], [42] employed flow similarities to discriminate DDoS attacks from FCs and achieved better results. The authors mainly used fixed thresholds, which required craft design and professional field knowledge.

Somani et al. [32] surveyed new environments for DDoS. The authors presented developments related to DDoS attack mitigation solutions in the cloud. In particular, this paper presented a comprehensive survey with detailed insights into the characterization, prevention, detection, and mitigation of these attacks. Additionally, it presented a comprehensive taxonomy to classify DDoS attack solutions.

Sachdeva et al. [29] combined multiclusters of source address entropy not only to detect various types of DDoS attacks against web services but also to distinguish DDoS attacks from FCs. Optimal thresholds for traffic cluster entropy were calibrated through receiver operating characteristic curves.

Saravanan et al. [30] found that during FCs, human users always tried to access hot pages, but Bots accessed pages randomly. To some extent, this finding could be helpful for differentiating between DDoS and FCs. The approach combined multiparameters with weights to discriminate DDoS from FC and achieved better results than when using a single parameter. The weights, however, were fixed and could not be updated automatically.

Gupta et al. [12] reviewed the researches on DDoS attacks on wireless networks and outlined various types of DDoS attacks. The impact of the attack occurs at various points along the network, most significantly on the routing mechanism, security goals, and protocol stack layer. The genuine nodes are kept busy by the malicious node while processing a large number of route requests or sending large data packets to other nodes. An ad hoc network must have a secured mechanism to evade the attacks.

DDoS is a spy-on-spy game between attackers and detectors, and these attacks have caused huge losses [2]. In particular, these attacks can mimic normal users, which look

like FCs, and can often evade the existing defense systems. As far as we know, Turing Test is the most popular and pervasive method used to distinguish between DDoS and FC; however, with the development of Reverse Turing Test [22], increasingly distorted and obscure images have been employed to defend against the reverse Turing test, which usually causes user annoyance and helplessness. Additionally, the existing anomaly analysis approaches are too sensitive to detect the thresholds needed to elaborate on the design and are usually not flexible. In this paper, we propose a Data Mining method to solve this problem, which may act as a complementary mechanism of existing defence systems.

3. Proposed Method

To solve this issue of discriminating between DDoS attacks and FCs, we conducted an extensive analysis of these two phenomena and identified a few abnormal statistical features in DDoS attacks and FCs, such as the number of packets sent by Bots and legitimate users is different, the number of new IPs appeared in DDoS and FC is different. With these features, we were able to translate the differentiating problem into a method for classifying points in Euclidean n-spaces. As a result, we propose a method to employ Data Mining to discriminate between DDoS attacks and FCs.

3.1. The Concluded Feature Set

Based on our understanding and analysis of Flooding DDoS attacks and FCs, we conclude the following:

1) Unique Source IPs' Or Clients' Number In Each Interval ($uniqueSrcIPs$): In FCs, users are interested in specified events only, such as flash news or interesting information. These users usually come from the whole Internet, so the distribution of source IP addresses in FCs may be largely dispersive. In DDoS, however, the attacker collects hosts that are vulnerable, so the distribution of IP addresses is relatively concentrated because the availability of Bots or zombies is limited. As a result, in each interval Δt, the unique source IPs' or clients' numbers for DDoS and FC is different, and this feature can be formally represented as follows:

$$uniqueSrcIPs = \{uniClients|\ different\ IPs\ number\ in\ each\ interval.\} \quad (1)$$

where, $uniClients$ is the number of different source IPs or clients in an interval.

2) New Increased IPs' Number In Adjacent Interval ($newIncresedSrcIPs$): In FCs, many more individuals care about a specific event than during a DDoS attack, and these individuals are usually more evenly distributed geographically than Bots in DDoS. Therefore, in the adjacent interval, the number of source IPs or clients in the FC has increased more than that of IPs or clients in DDoS.

$$newIncresedSrcIPs = \{x|x \in uniClients, \\ x \notin uniClientsPrevious\} \quad (2)$$

where, $uniClientsPrevious$ is the number of different source IPs or clients in the previous interval, which is adjacent to the current interval.

3) The Average Of The Number Of Packets Sent By Source IPs or Clients In Each Interval ($uSrcIPsSendPkts$): To attain the expected attack effect, such as network congestion, Bots usually have to send as many packets as possible, so the average number of packets sent by each Bot and legitimate client is different. In general, the average number of packets sent by Bots is larger than that of normal users in each interval.

$$uSrcIPsSendPkts = \left\{ \frac{\sum_{i=1}^{n} numPackets_i}{n} \right\} \tag{3}$$

where, $numPackets_i$ is the packets number sent by the i-th client, n is the unique number of clients in each interval Δt.

4) The Standard Of The Number Of Packets Sent By Source IPs In Each Interval ($stdSrcIPsSendPkts$): DDoS mainly rely on Botnets, and Bots are usually executed by preprogrammed codes. Each bot exhibits similar traffic behavior, whereas legitimate users are different individuals who exhibit varied behavior. Thus, the standard of the number of packets sent by each user in a DDoS attack is lower than that of in a FC.

$$stdSrcIPsSendPkts = \left\{ \sqrt[2]{\frac{\sum_{i=1}^{n} (numPackets_i - u)^2}{n-1}} \right\} \tag{4}$$

where, u is the uSrcIPsSendPkts.

To demonstrate our feature set, it is useful to distinguish between DDoS and FC. The following section will discuss some experiments we conducted.

3.2. Proposed Idea

In general, different datasets are created in different situations, such as different topologies and network bandwidths. For these reasons, different datasets cannot be mixed without any preprocessing. To address this problem and obtain the new feature set, we conducted the following preprocessing tasks:

1) Difference Topology: Different IP masks and different IP addresses exist in each dataset, so we apply relevant statistical features instead of using IPs directly, such as adopting the number of packets sent by each IP, the average packet size sent by each IP, and the standard deviation of the packet size sent by each source IP. With these preprocessing tasks, we believe that we could eliminate the effect caused by network topology.

2) Network Bandwidth: For various reasons, datasets are usually created at different bandwidths. To address this issue, we scaled the interval to ensure that different datasets have the same network bandwidth at each interval. For example, assume that the bandwidth of the first dataset is 10 M/s, and the second dataset is 100 M/s. To achieve the same traffic volume in each interval (2s), we enlarged the first dataset to be 100 M/s, which was an increase 100/10 (ten) times that of the interval. In this way, the impact of the network bandwidth was minimized.

With the new feature set, we proposed a method to employ several common Data Mining methods [13] [24], including logistic, multilayer perception, J48, and PART, to distinguish between DDoS attacks and FCs. The entire procedure is given in Algorithm 1. We labeled each interval with one class label: DDoS or FC.

Algorithm 1 Proposed Idea.

Input:
DDoS and FC dataset

Output:
The discriminating results of DDoS or FC

1: Calculate each feature with default interval Δt for each data separately, then label the class-DDoS or FC, and we could obtain 5 dimensional row vectors:

$$(uniqueSrcIPs, newIncresedSrcIPs, uSrcIPs-SendPkts, stdSrcIPsSendPkts, Class)$$

These vectors can be used as input. Consequently, we can translate the differentiating problem into a method to classify points in Euclidean n-spaces.

2: Mix the results in Step 1 and normalize the mixed results.

3: Take the mixed and normalized results in Step 2 as input for Data Mining methods, such as, Logistic, MultilayerPerceptron, J48 and PART, and estimate these results on public datasets.

4: return the final distinguishing results.

4. Experiments

In this section, we conducted additional experiments to verify our method on two public real-world datasets: CAIDA DDoS Attack 2007 Dataset (CAIDA2007) [7] used as the DDoS data and the World Cup 1998 Dataset (WorldCup1998) [11] used as the FC data.

CAIDA2007 contains approximately 1 hour of anonymized traffic traces from a DDoS attack on August 4, 2007 (Universal Time Coordinated, UTC). This attack attempted to block access to the targeted server. The 1-hour trace is split up into 5-minute pcap files. The total size of the dataset is 5.3 GB (compressed; 21 GB uncompressed). Only attack traffic directed to the victim and responses to the attack from the victim are included in the traces. We have removed as much of the nonattack traffic as possible. Traces in this dataset were anonymized using CryptoPAn prefix-preserving anonymization using a single key. The payload has been removed from all packets [7].

WorldCup1998 includes all requests made to the 1998 World Cup website between April 30, 1998, and July 26, 1998. During this 88-day period, the World Cup site received 1,352,804,107 requests. No information is available regarding how many requests were not logged, although it is believed no system outages occurred during the collection period [11].

We selected 1-minute of DDoS data (2007-08-05 05:30:00 to 2007-08-05 05:31:00) from CAIDA2007 and 1-hour of FC data (1998-06-10 16:00:00 to 1998-06-10 17:00:00) from WorldCup1998 for analysis (see Fig. 1). We ensured that different datasets would

produce the same traffic volume in each interval by scaling the interval to reduce the effects of bandwidth. After analyzing the selected data, we selected a scale interval rate of 1:100, which meant that the traffic produced in 1 DDoS interval was nearly equal to the traffic produced in 100 FCs intervals. After this process, we achieved the scaled traffic shown in Fig. 2. We found that the average amounts of traffic were almost the same, which significantly eliminated the effect of different bandwidths.

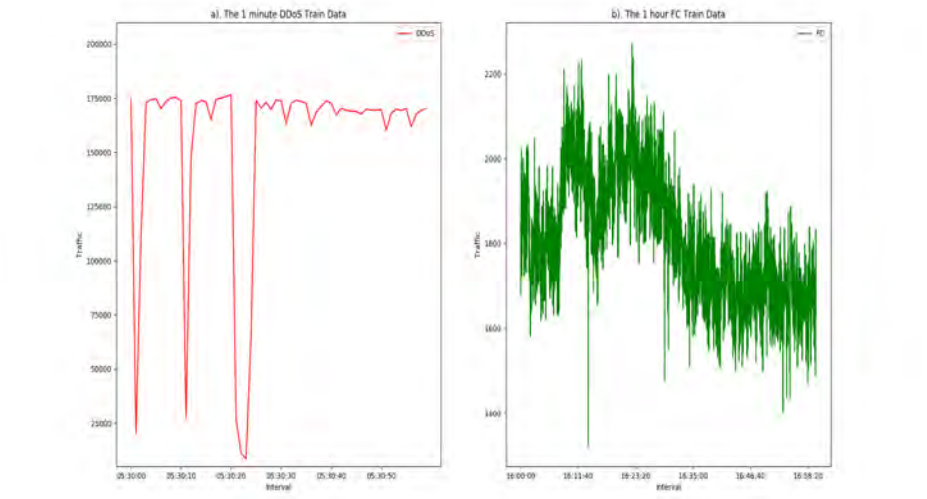


Fig. 1. a). The 1 minute DDoS Data with the interval-1s from 2007-08-05 05:30:00 to 2007-08-05 05:31:00. b). The 1 hour FC Data with the interval-1s from 1998-06-10 16:00:00 to 1998-06-10 17:00:00.

4.1. New Feature Set Evaluation

To evaluate the importance of each feature of our feature set, Correlation (Pearsons) based Method (CorrelationAttributeEval), Gain Ratio Method (GainRatioAttributeEval) and Information Gain Method (InfoGainAttributeEval) have been selected to do features

Tab. 2 shows the estimated results of the new features in detail. The results reveal that all three feature selection methods gave $uniqueSrcIPs$ the greatest priority; the second most important feature was $stdSrcIPsSendPkts$ and the next most important feature was $newIncresedSrcIPs$; the least vital feature was $uSrcIPsSendPkts$. These results are basically consistent with our expectation of real data.

Consequently, the importance of the entire feature set in descending order is as follows:

$$uniqueSrcIPs > stdSrcIPsSendPkts > newIncresedSrcIPs > uSrcIPsSendPkts$$

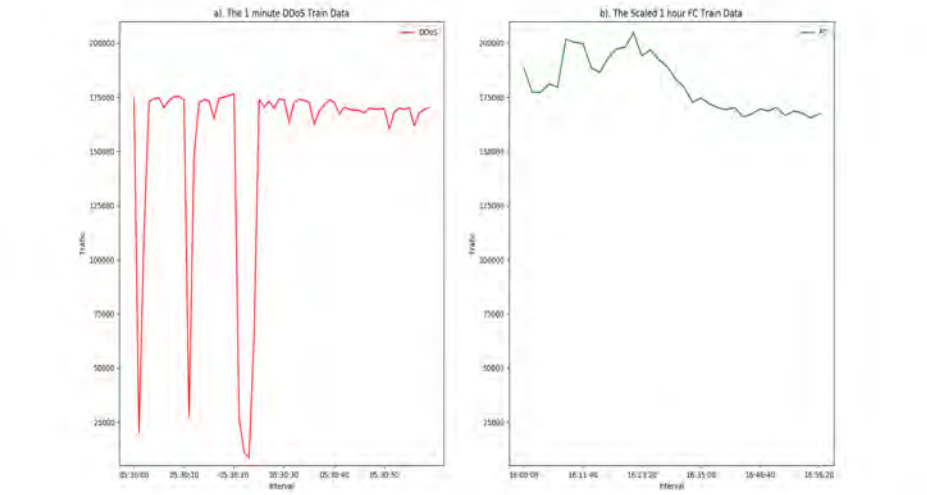


Fig. 2. a). The 1 minute DDoS Data with the interval-1s from 2007-08-05 05:30:00 to 2007-08-05 05:31:00. b). The scaled 1 hour FC Data with the intervel-100s from 1998-06-10 16:00:00 to 1998-06-10 17:00:00.

Table 2. Features Selection Estimated By 3 Common Features Selection Methods

Attribute Evaluator	GainRatio Attribute Eval		Correlation Attribute Eval		InfoGain Attribute Eval	
Search Method	Ranker					
Features	Rank	Average	Rank	Average	Rank	Average
unique SrcIPs	1	1.000	1	0.962	1	0.954
newIncreased SrcIPs	3	0.922	3	0.837	4	0.877
uSrcIPs SendPkts	4	0.747	4	0.387	3	0.954
stdSrcIPs SendPkts	2	1.000	2	0.890	2	0.954

4.2. Proposed Idea

In this section, we employ several common Data Mining methods included Logistic, MultilayerPerceptron, J48 and PART to distinguish DDoS and FC, combining with Confusion Matrix, Relative absolute error (RAE), Root relative squared error (RRSE), Accuracy, False Positive Rate (FPR) and False Negative Rate (FNR) all together as measurement standards.

The results are shown in Tab. 3. With these new features, we could distinguish between DDoS attacks and FCs with high accuracy, less than 5% RAE, no more than 30% RRSE, nearly 100% Accuracy, no more than 0.04% FNR, and nearly 0% FPR with 10-fold cross-validation for those 96 train samples. We found that different Data Mining methods have almost the same accuracy, and Logistic method may be the best discrimination method, with the lowest RAE and FNR and the highest Accuracy (98.9583%).

Table 3. Distinguished Results With 4 Data Mining Methods on Train Sets.

Methods		Logistic		Multilayer Perceptron		J48		PART	
Confusion Matrix		DDoS	FC	DDoS	FC	DDoS	FC	DDoS	FC
	DDoS	59	1	59	1	58	2	58	2
	FC	0	36	0	36	0	36	0	36
Relative Absolute Error (RAE)		2.2181 %		3.911 %		4.4361 %		4.4361 %	
Root Relative Squared Error (RRSE)		21.0712 %		20.9248 %		29.7991 %		29.7991 %	
Accuracy		98.9583%		98.9583 %		97.9167 %		97.9167 %	
False Positive Rate (FPR)		0%		0 %		0%		0 %	
False Negative Rate (FNR)		0.017%		0.017%		0.033%		0.033%	

To further verify the proposed idea, we conducted additional experiments on the test data. We selected another 1-minute DDoS data (2007-08-05 05:34:00 to 2007-08-05 05:35:00) from CAIDA2007 and a 1-hour FC data (1998-07-03 16:00:00 to 1998-07-03 17:00:00) from WorldCup1998 for analysis. Other than the scale rate, all of the processes used were the same as those used in the previous test. In this case, the scale interval rate was 1:80. After this process, we achieved the scaled traffic in each new interval (see Fig. 3). It was evident that the average traffic was basically the same, which could reduce the bandwidth effect of different datasets.

After completing the preprocessing tasks on the test data, we achieved the features and employed the trained models to estimate the new test set. The results in Tab. 4 show that our method has nearly 100% Accuracy, 0% FPR and FNR, less than 0.01% RAE, and no more than 0.01% RRSE, which indicates that the concluded features are useful, and that this idea could perform well in discriminating between DDoS attacks and FCs.

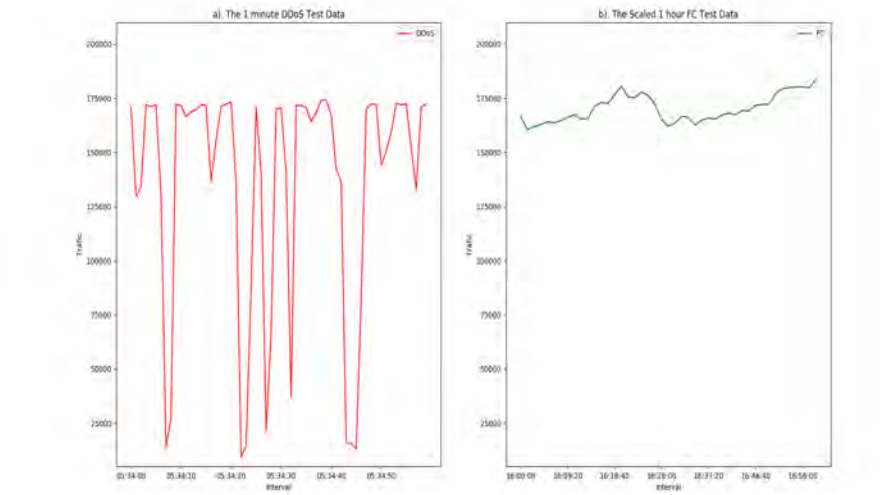


Fig. 3. a). The 1 minute DDoS Data with the interval-1s from 2007-08-05 05:34:00 to 2007-08-05 05:35:00. b). The scaled 1 hour FC Data with the intervel-80s from 1998-07-03 16:00:00 to 1998-07-03 17:00:00.

Table 4. Distinguished Results With 4 Data Mining Methods on Test Sets.

Methods		Logistic		Multilayer Perceptron		J48		PART	
Confusion Matrix		DDoS	FC	DDoS	FC	DDoS	FC	DDoS	FC
	DDoS	60	0	60	0	60	0	60	0
	FC	0	45	0	45	0	45	0	45
Relative Absolute Error (RAE)		0 %		0.007 %		0%		0 %	
Root Relative Squared Error (RRSE)		0 %		0.0075 %		0 %		0 %	
Accuracy		100%		100 %		100 %		100 %	
False Positive Rate (FPR)		0%		0 %		0%		0 %	
False Negative Rate (FNR)		0%		0%		0%		0%	

4.3. Experiments Analysis

In this section, we compare our proposed method with traditional methods. Yu et al. [41] made use of information distance, such as Sibson and Jeffrey distance measures, to discriminate between DDoS and FCs, and achieved only 65% accuracy. Bhatia et al. [5] proposed a few parameters (different from our parameters) to distinguish between DDoS and FCs and conducted experiments that did not achieve any accuracy. The methods used by Bhatia et al. [5] involved only simple statistics, and the study was totally different from our study. Saravanan et al. [30] employed a behavior-based detection method on Application Layer to distinguish between DDoS and FCs and achieved about 91% Accuracy.

To further compare our proposed method with traditional methods, such as the entropy method [35], we selected the Shannon Entropy of Source IPs (srcIPs) in each interval as a feature. The values of Shannon Entropy were calculated by the following formula:

$$Entropy = -\sum_{i=1}^{n} p_i * log_2(p_i) \tag{5}$$

where, p_i is the probability of each source IP or client in each interval.

The results are shown in Fig. 4. The red line represents the entropy of DDoS, and the green line represents the entropy of FCs, which indicates that it is not easy to discriminate DDoS and FCs using the entropy of srcIPs because their entropies are quite similar, which results in the discrimination thresholds rarely being selected.

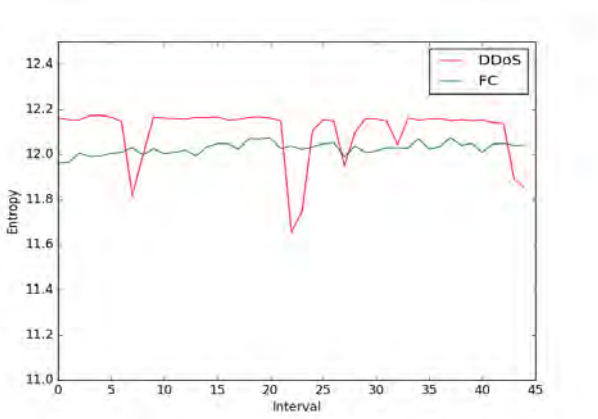


Fig. 4. Shannon Entropy of Source IPs In Each Interval.

Compared with the traditional methods-Entropy in Fig. 4, and review the results of our idea in Tab. 3 and Tab. 4. We find that our idea has a better accuracy to distinguish DDoS and FC with the new feature set on Train Set and Test Set than that of the traditional Entropy method.

Why could our idea achieve a better accuracy? On the basis of an extensive analysis of DDoS and FCs (see Fig. 5), we believe that this feature set plays a vital role in our

proposed method, achieving better accuracy. In Fig. 5, the red color represents FCs, and the blue color represents DDoS. This figure shows the distribution of the class for each feature dimension. In Fig. 5, we found that each concluded feature had a slightly better distinguished effect. In addition, traditional methods are primarily threshold-based and usually required crafted thresholds, which are difficult to obtain in reality. Through Data Mining, our proposed method was able to detect DDoS and FCs with fewer human interruptions and better accuracy. Our method is based on the analysis of end-victim. As a result, it is easier to deploy without modifying the existing network protocols, just to deploy at the front of end-victim.

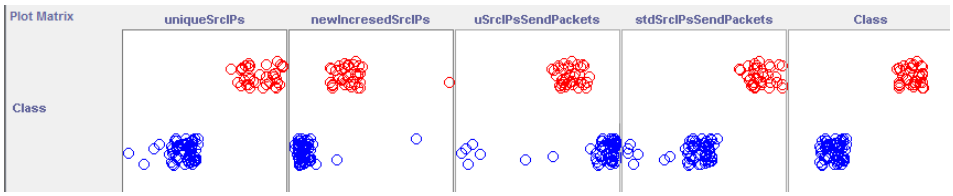


Fig. 5. The Correlation Between Each Feature And The Class. The Red Color Represents FC, While The Blue Color Represents DDoS

4.4. Issues and Limitations

The experiments demonstrated that our method could differentiate between DDoS and FCs well; however, this method does have some insufficiencies. Because this method is based on a few assumptions, the following corresponding issues existed.

1) We assumed that compromised machines do not use spoofing IPs. Although many methods can be used to handle spoofing (such as Ingress and Egress filtering [10], [9], HCF [16], Packet Marking [38], [43]), it is still a useful and potential technique for sophisticated users. According to the MIT Spoofer Project, which provides an aggregate view of ingress and egress filtering and IP spoofing on the Internet, 23% of autonomous systems and 16.8% of IP addresses are able to spoof, which means that an estimated 560 million out of 3.32 billion IP addresses still can be spoofed [28].

2) We also assumed that the number of Bots that can simultaneously launch attacks is limited, so for a Flooding attack, Bots have to send as many packets as possible [33], [8]]. Today, many other types of DDoS attacks occur. For example, low-rate DDoS (LDDoS), which our method cannot distinguish. And Botnets are becoming increasingly larger and more complex [14] with new techniques (such as Cloud Computing [31], Internet of Things [6], [26], SDN [39]) that have brought new challenges.

3) We also had a few other issues. In this paper, our proposed method relied mainly on time intervals, not individuals. As a result, our method could detect abnormalities but not identify attackers, which omitted the vulnerabilities for those mimicking attacks [40], [3].

For these reasons, we should not be overly optimistic. The battle to protect the Internet in a relatively secure environment is ongoing and much more research is required to solve these dilemmas.

5. Conclusion

To discriminate between DDoS attacks and FCs, we first categorized existing methods to clarify the issue. We conducted an extensive analysis of DDoS and FCs and identified a few features to profile DDoS attacks and FCs. Using these features, we translated the discrimination issue into a method to classify points in Euclidean n-spaces. As a result of this analysis, we proposed a method to employ Data Mining to discriminate between DDoS attacks and FCs. We evaluated the results of our experiments and found that the idea employed Data Mining techniques based on identified features can achieve high accuracy and reduced FPR and FNR. We further compared this method to a traditional method (i.e., entropy method), and the results indicated that our proposed method could have a better distinguished effect than that of the entropy method. At last, we discussed some shortcomings in this paper; for example, our method cannot detect LDDoS, and although it could detect abnormalities, it could not identify attackers.

Our future work will focus mainly on how to identify individuals. To do this work, more refined features that may better profile the traffic behavior of clients should be identified. Other researches are to find new datasets or real-world applications to further evaluate.

References

1. AYAHs: website:ayahs. http://areyouahuman.com
2. Behal, S., Kumar, K.: Trends in validation of ddos research. Procedia Computer Science 85, 7–15 (2016)
3. Behal, S., Kumar, K.: Detection of ddos attacks and flash events using novel information theory metrics. Computer Networks 116, 96–110 (2017)
4. Bhatia, S.: Detecting distributed denial-of-service attacks and flash events (2013)
5. Bhatia, S., Mohay, G., Tickle, A., Ahmed, E.: Parametric differences between a real-world distributed denial-of-service attack and a flash event. In: Availability, Reliability and Security (ARES), 2011 Sixth International Conference on. pp. 210–217. IEEE (2011)
6. Borgohain, T., Kumar, U., Sanyal, S.: Survey of security and privacy issues of internet of things. arXiv preprint arXiv:1501.02211 (2015)
7. DDoS: Caida ddos attack 2007 dataset. http://www.caida.org/data/passive/ddos-20070804_dataset.xml (2007)
8. Feily, M., Shahrestani, A., Ramadass, S.: A survey of botnet and botnet detection. In: Emerging Security Information, Systems and Technologies, 2009. SECURWARE'09. Third International Conference on. pp. 268–273. IEEE (2009)
9. Ferguson, P.: Network ingress filtering: Defeating denial of service attacks which employ ip source address spoofing (2000)
10. Ferguson, P., Senie, D.: Network ingress filtering: Defeating denial of service attacks which employ ip source address spoofing. Tech. rep. (1997)
11. FlashCrowds: World cup 1998 dataset. http://ita.ee.lbl.gov/html/contrib/WorldCup.html (1998)
12. Gupta, P., Bansal, P.: A survey of attacks and countermeasures for denial of services (dos) in wireless ad hoc networks. In: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. p. 25. ACM (2016)
13. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edn. (2011)
14. Hoque, N., Bhattacharyya, D.K., Kalita, J.K.: Botnet in ddos attacks: trends and challenges. IEEE Communications Surveys & Tutorials 17(4), 2242–2270 (2015)
15. Ismail, Z., Jantan, A.: A review of machine learning application in botnet detection system. Sindh University Research Journal-SURJ (Science Series) 48(4D) (2016)
16. Jin, C., Wang, H., Shin, K.G.: Hop-count filtering: an effective defense against spoofed ddos traffic. In: Proceedings of the 10th ACM conference on Computer and communications security. pp. 30–41. ACM (2003)
17. Jung, J., Krishnamurthy, B., Rabinovich, M.: Flash crowds and denial of service attacks: Characterization and implications for cdns and web sites. In: Proceedings of the 11th international conference on World Wide Web. pp. 293–304. ACM (2002)

18. Kandula, S., Katabi, D., Jacob, M., Berger, A.W.: Botz-4-sale: Surviving organized ddos attacks that mimic flash crowds (awarded best student paper). In: NSDI. USENIX (2005)
19. Latif, R., Abbas, H., Assar, S., Latif, S.: Analyzing Feasibility for Deploying Very Fast Decision Tree for DDoS Attack Detection in Cloud-Assisted WBAN. Springer International Publishing (2014)
20. Li, K., Zhou, W., Li, P., Hai, J., Liu, J.: Distinguishing ddos attacks from flash crowds using probability metrics. In: NSS. pp. 9–17. IEEE Computer Society (2009)
21. Mansfield-Devine, S.: The growth and evolution of ddos. Network Security 2015(10), 13–20 (2015)
22. Marcus, G., Rossi, F., Veloso, M.: Beyond the turing test. Ai Magazine (2016)
23. Networks, A.: Worldwide infrastructure security report (2016)
24. Ngo, T.: Data mining: Practical machine learning tools and technique, third edition by ian h. witten, eibe frank, mark a. hell. SIGSOFT Softw. Eng. Notes 36(5), 51–52 (Sep 2011), http://doi.acm.org/10.1145/2020976.2021004
25. Oikonomou, G., Mirkovic, J.: Modeling human behavior for defense against flash-crowd attacks. In: 2009 IEEE International Conference on Communications. pp. 1–6. IEEE (2009)
26. Patel, K., Thoke, A.: A details survey on black-hole and denial of service attack over manet environment (2016)
27. Prasad, K.M., Reddy, A.R.M., Rao, K.V.: Discriminating ddos attack traffic from flash crowds on internet threat monitors (itm) using entropy variations. African Journal of Computing & ICT 6(3) (2013)
28. Project, M.S.: http://spoofer.cmand.org/summary.php
29. Sachdeva, M., Kumar, K., Singh, G.: A comprehensive approach to discriminate ddos attacks from flash events. J. Inf. Sec. Appl. 26, 8–22 (2016)
30. Saravanan, R., Shanmuganathan, S., Palanichamy, Y.: Behavior-based detection of application layer distributed denial of service attacks during flash events. Turkish Journal of Electrical Engineering & Computer Sciences 24(2), 510–523 (2016)
31. Somani, G., Gaur, M.S., Sanghi, D., Conti, M., Buyya, R.: Ddos attacks in cloud computing: Issues, taxonomy, and future directions. arXiv preprint arXiv:1512.08187 (2015)
32. Somani, G., Gaur, M.S., Sanghi, D., Conti, M., Buyya, R.: Ddos attacks in cloud computing: issues, taxonomy, and future directions. Computer Communications (2017)
33. Thakar, B., Parekh, C.: Advance persistent threat: Botnet. In: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. p. 143. ACM (2016)
34. Thapngam, T., Yu, S., Zhou, W., Beliakov, G.: Discriminating ddos attack traffic from flash crowd through packet arrival patterns. In: Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. pp. 952–957. IEEE (2011)
35. Thomas M. Cover, J.A.T.: IEEE (1991)
36. Von Ahn, L., Blum, M., Langford, J.: Telling humans and computers apart automatically. Communications of the ACM 47(2), 56–60 (2004)

37. Xie, Y., Yu, S.Z.: A large-scale hidden semi-markov model for anomaly detection on user browsing behaviors. IEEE/ACM Transactions on Networking (TON) 17(1), 54–65 (2009)
38. Yaar, A., Perrig, A., Song, D.: Stackpi: New packet marking and filtering mechanisms for ddos and ip spoofing defense. IEEE Journal on Selected Areas in Communications 24(10), 1853–1863 (2006)
39. Yan, Q., Yu, F.R., Gong, Q., Li, J.: Software-defined networking (sdn) and distributed denial of service (ddos) attacks in cloud computing environments: A survey, some research issues, and challenges. IEEE Communications Surveys & Tutorials 18(1), 602–622 (2016)
40. Yu, S., Guo, S., Stojmenovic, I.: Fool me if you can: mimicking attacks and anti-attacks in cyberspace. IEEE Transactions on Computers 64(1), 139–151 (2015)
41. Yu, S., Thapngam, T., Liu, J., Wei, S., Zhou, W.: Discriminating ddos flows from flash crowds using information distance. In: NSS 2009: Proceedings of the third International Conference on Network and System Security. pp. 351–356. IEEE (2009)
42. Yu, S., Zhou, W., Jia, W., Guo, S., Xiang, Y., Tang, F.: Discriminating ddos attacks from flash crowds using flow correlation coefficient. IEEE Trans. Parallel Distrib. Syst. 23(6), 1073–1080 (2012)
43. Zhang, J., Liu, P., He, J., Zhang, Y.: A hadoop based analysis and detection model for ip spoofing typed ddos attack. In: Trustcom/BigDataSE/I? SPA, 2016 IEEE. pp. 1976–1983. IEEE (2016)

Bin Kong is currently a Ph.D student of Beijing Jiaotong University. He is currently the deputy director of National Secrecy Science and Technology Evaluation Center and Senior Engineer. His research interests include information security, risk assessment system, anomaly detection analysis, etc.

Kun Yang is currently a Ph.D student and study in Institute of Information Engineering, Chinese Academy of Science in Beijing from 2012 to the present. His research interests focus on DDoS attack detection, network traffic analysis and Machine Learning.

Degang Sun received his master degree from Beijing Jiaotong University. He has long engaged in Information Security Technology Research. His main research interests include electromagnetic leakage emission protection, wireless communication security and so on. He has published more than 40 academic papers, 6 books and more than 10 patents.

Meimei Li received her master degree from Peking University in 2007. She is currently a senior engineer of Institute of Information Engineering, Chinese Academy of Science, Beijing and an associate professor of University of Chinese Academy of Sciences. Her research interests include high security level system security, integration analysis, abnormal behavior detection and so on.

Shi Zhixin received the PhD degree in pattern recognition & intelligent systems from

Institute of Automation, Chinese Academy of Science, Beijing in 2014. He is currently a senior assistant professor at Institute of Information Engineering, Chinese Academy of Science. His research interests include the areas of DDoS attack detection, network monitoring, massive dataset mining. He is a member of the CCF.

Received: December 30, 2016; Accepted: May 10, 2017.

Computer Science and Information Systems 14(3):857–873 https://doi.org/10.2298/CSIS161230033X

Building a Lightweight Testbed Using Devices in Personal Area Networks

Qiaozhi Xu[1,2] and Junxing Zhang[1,#]

[1] College of Computer Science, Inner Mongolia University
Hohhot, China
ciecxqz@imnu.edu.cn,junxing@imu.edu.cn (#Corresponding author)
[2] College of Computer Science, Inner Mongolia Normal University
Hohhot, China

Abstract. Various networking applications and systems must be tested before the final deployment. Many of the tests are performed on network testbeds such as Emulab, PlanetLab, etc. These testbeds are large in scale and organize devices in relatively fixed ways. It is difficult for them to incorporate the latest personalized devices, such as smart watches, smart glasses and other emerging gadgets, so they tend to fall short in supporting personalized experiments using devices around users. Moreover, these testbeds commonly impose restrictions on users in terms of when and where to carry out experiments making them clumsy or inconvenient to use. The paper proposes to build a testbed utilizing users' devices in their own personal area networks (PANs). We have designed and implemented a prototype, which we call PANBED. Our experiments show that PANBED allows users to set up different scenes to test applications using a home router, PCs, mobile phones and other equipment. PANBED is light weighted with a size less than 16 KB and it has little impact to the other functions of the PAN. The experiment results also prove the realism, effectiveness, flexibility and convenience of PANBED.

Keywords: Testbed, Personal Area Network, PAN, Personal Network.

1. Introduction

Various networking applications and systems must be tested before the final deployment. Presently there are four commonly used test methods in networking and distributed system research: network simulation, overlay network, network emulation, and network testbed.

Network simulation is the method of simulating the operations of each network layer using a kind of software called simulator. Changes of system states caused by different events are recorded and modeled by the simulator. Network simulation is relatively simple, low-cost, controllable, repeatable, and relies on pure software environment to breaks the limitation of physical resources. Ns-2 [19], ns-3 [32], OMNet ++ [39], Atemu [30], TOSSIM [25], GloMoSim [8], SensorSim [28] are the most widely used simulation systems. However, a simulated environment can be quite different from the real physical environment, and it cannot capture changes of lower network layers in many cases, resulting in poor realism of experimental results.

A overlay network is a real network environment built on another existing network [4], which can test and evaluate the realistic performance of protocols and algorithms, such

as RON [7] and PlanetLab [12]. But overlay networks are high-cost, vulnerable to the impact of other network environments, and users are cumbersome to modify the network parameters, and unable to monitor the network behaviors which making the experiments and tests unrepeatable and difficult to control.

Network emulation is a trade-off method between network simulation and overlay network [44]. An emulation system achieves functions of a real system by using the software simulation and abstraction techniques on real devices and introducing the configurable parameters such as packet loss and link delay. Typical emulation systems are VMNet [45], Avrora [37], Dummynet [33], NSE [16], and ModelNet [38] and so on. An emulated environment is very close to a real one. It also possesses the repeatability of network simulation and realism of overlay network but it requires cumbersome manual configuration.

Current network testbeds typically incorporate simulation, emulation and overlay network into an integrated experimental platform of software and hardware. They are capable of producing repeatable and controllable scenes to reduce costs of setting up experiments. They are easy to use and offer different degrees of realism. The well-known network testbeds include Emulab [44], Kansei [14], MoteLab [43], GNOMES [42], GENI [9], Winlab [6], etc. However, these large-scale network testbeds tend to have some shortcomings: (i) A user must login to a testbed remotely to carry out his experiment, and the availability of devices is out of his control; (ii) it is difficult for these testbeds to incorporate the up-to-date or personalized devices, such as smart watches, smart glasses and other emerging gadgets, so they tend to fall short in supporting experiments aiming at the latest devices around users. (iii) The background traffic in both simulated and emulated systems is not realistic, which affects the realism of experimental results.

In order to overcome the shortcomings of large-scale network testbeds, this paper proposes to build a network testbed using users' devices in their own personal area networks (PANs). We have designed and implemented a prototype of this system, which we call PANBED. Our experiments show that PANBED allow users to set up different scenes to test applications using a home router, PCs, and mobile phones. The results demonstrate PANBED enable users to assess applications at their convenience using diverse, personal, up-to-date and low cost devices around them with little impact to existing PANs. As far as we know, PANBED is the first network testbed built on devices in a PAN.

To design and implement PANBED, we have identified and addressed the following six key challenges:

(1) Where to implement traffic shaping?

The traffic shaping in some testbeds is implemented by running DummyNet on the intermediate nodes (delay node). However, the same way cannot be adopted in the PANBED, because devices in a PAN are quite different from that in a testbed. Firstly, the number of devices in a PAN is limited. Most likely, there are only one home router and several devices. Secondly, not all devices in a PAN support DummyNet. Thirdly, these devices not only participate in experiments, but also complete original tasks for users, thus PANBED should change these devices as little as possible. For these reasons, we choose a home router to support the traffic shaping in the PANBED.

(2) How to implement traffic shaping?

There are two flow control modules, Netem and TC, in Linux that can shape the flow through a network card. OpenWRT [15] is based on the Linux kernel and often

used in embedded network devices. PANBED loads OpenWRT into the home router and implements the traffic shaping utilizing the Netem and TC modules.

(3) Is isolation of the control flow and data flow necessary on PANBED?

Some testbeds isolate the control flow from data flow by installing multiple network cards in experimental devices, and creating the control vlan and experimental vlans on switches, but devices in a PAN do not have such hardware conditions, so similar methods cannot be used in the PANBED. Devices in a PAN are all around users, and users know well about these devices. PANBED almost does not modify users' devices and only applies data flow control policies on the home router, so there is no isolation of the control flow from data flow.

(4) How to ensure the repeatability of experiments?

To ensure the repeatability of experiments and the consistency of experimental environments, many testbeds initialize devices with default or saved parameters before an experiment. However, in PANBED experimental devices are not dedicated. They need to complete their original tasks and cannot be frequently initialized, so it is difficult to ensure the repeatability of experiments in this situation. PANBED adopts two empirical approaches: (i) Since users know enough about the experimental devices, PANBED allows users to decide whether or not to initialize their devices, thereby improving the repeatability of experiments; (ii) before an experiment, PANBED collects and records system states of experimental devices, and provides secondary reference to users for the analysis of experimental results.

(5) How about realism?

In some testbeds there exist both real devices and simulated components such as NSE and DummyNet, but in PANBED, there are only real devices and it makes experimental results more realistic. In addition, the background traffic in many testbeds is not realistic, which affects the realism of experimental results. In PANBED, the background traffic is real, so the experimental results are more realistic.

(6) How does a user deploy her own application?

In most testbeds, a user remotely logins to experimental devices through SSH and deploys his application. In PANBED, all devices are around users, he can directly login and operate these devices.

We have designed and implemented PANBED based on the solutions to the above issues. Our experimental results show PANBED allows users to evaluate applications at their convenience, allows users to set up different scenes to test applications using a home router, PCs, and mobile phones and guarantees the realism of the experiment.

The remaining of the paper is organized as follows. The background and related work are given in Section 2. Section 3 and Section 4 describes the design and implementation of PANBED. The experiment and evaluation results are analyzed in Section 5. Finally, Section 6 concludes the paper.

2. Related Work

The goal of PANBED is to provide users with a low-cost and flexible network testbed utilizing devices in users' PAN. The following introduces the related research works about the network testbed, OpenWrt and PAN.

2.1. Network Testbed

Various networking applications and systems must be tested before the final deployment. Tests based on a real environment are high cost, long time, difficult to control and repeat. Test results based on a simulation are inaccurate compared with that based on a real environment. So in recent years, many researchers have begun to study and build network testbeds [12], [44], [38].

These testbeds are relatively large in scale and their operation mode generally are: (1) In one or several physical spaces, devices are organized in a fixed ways and some resource pools are formed; (2) users login the web server of a testbed through Internet after registration and being agreed by the administrator; (3) users submit their experimental requirements which specifying the device type, operating system version, connection topology of these devices and parameters of these links, such as bandwidth, delay, and loss; (4) testbed servers parse these experimental requirements, allocate resources, and build network environments for users according to their requirements; (5) users login the assigned devices via SSH or telnet remotely and begin their experiments.

These testbeds provide users with controllable and repeatable experiment environments without any input from them, but, their device types are limited and it is time-consuming to introduce emerging devices into testbeds which limits their flexibility. PANBED proposed in this paper can build a testbed for users using devices in their PAN and enables user to assess applications at their conveniences using diverse, personal, up-to-date and low cost devices around them with little impact to the existing PAN.

2.2. OpenWrt Routers

OpenWrt is an embedded system based on Linux kernel and often used in network devices such as industrial devices, telephones, small robots, smart homes and routers, etc. The sofeware architecture of OpenWrt is shown in Fig. 1.

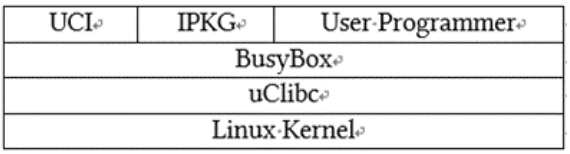


Fig. 1. Software architecture of OpenWrt [15] which shows OpenWrt is embedded a number of tools, such as uClibc, busybox and shell interpreter, etc. based on the basic Linux kernel.

Recently, OpenWrt is supported by more and more router vendors, such as 3Com, D-Link, TP-Link, Huawei, Netgear, etc. [3]. At the same time, many researchers also choose OpenWrt routers as a basic component in their researches for its openness and programmability. For example, Kai implemented a PPPoE traffic control system [46], Kim implemented the remotely intelligent management to an indoor lighting system so as to save energy [22], Kciuk realized the remotely control on robots in an intelligent buildings [21], Palazzi achieved the fast and smooth transmission of real-time flow and meanwhile ensured the high throughput of TCP applications [27], Lee designed a software-defined

wireless mesh network architecture SD-WMN [24], Serrano built a low-cost wireless network testbed [34], Reich designed a delay tolerance network testbed-MadNet [31].

2.3. Personal Area Network

PAN (Personal Area Network) refers to connecting personal terminal devices as a network by using variety of communication technologies, and nowadays, wireless personal area network (WPAN) is one of the main forms of PAN. Currently, researches on WPAN mainly focus on enhancements and improvements of the performance of WPAN such as throughput, energy consumption, coverage, data transmission rate and so on [35], [26], [11], [49], [50], [40], [23], some researchers also use WPAN to realize intelligent home, telemedicine and other purposes [17], [41], [29], [20], [13], [47], [48]. Up to now, there is no research to build a testing platform for users using devices in their PAN.

PANBED enables users to do testing with diverse, personal, latest devices around them and with little impact to existing PAN. As far as we know, PANBED is the first testbed to be built by utilizing devices within a PAN.

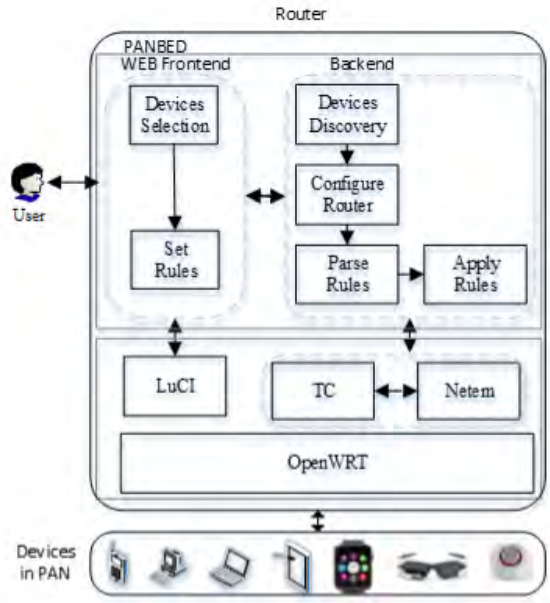


Fig. 2. PANBED architecture which is a light weight function based on the OpenWrt and is divided into frontend and backend.

3. PANBED Design

3.1. Overall Architecture

Usually there is a home router, several PCs, smart phones, tablets, sensors and other equipment in a PAN, and it is difficult and complex for users to do tests directly using these devices. PANBED provides users a convenient way to do testing under different network scenes using diverse, personal, and latest devices around them.

The architecture of PANBED is shown in Fig. 2. Users choose experimental devices and set experiment rules through the Web frontend. The backend completes functions such as discovering devices, configuring router, analyzing and applying rules, etc.

3.2. Frontend Design

PANBED provides a web access for users to facilitate their operations. Many testbeds set up a separate web server to deal with experimental requests. However, the number of devices within a PAN is limited, even no PC, so it is unrealistic to set up a separate web server in PANBED but to build the web service on the OpenWrt router.

Usually, users configure a router through a web page which is a web service based on uHTTPd [5]. uHTTPd is aimed towards being an efficient and stable server, suitable for lightweight tasks, commonly used with embedded devices and proper integration with OpenWrt's configuration framework [2].

For minimizing impact on the router and convenience of users, the web frontend of PANBED is embedded within the LuCI configuration page in a OpenWrt router with the template way, as shown in Fig. 3.

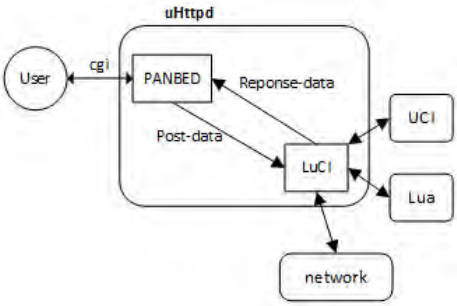


Fig. 3. Front-End Design of PANBED.

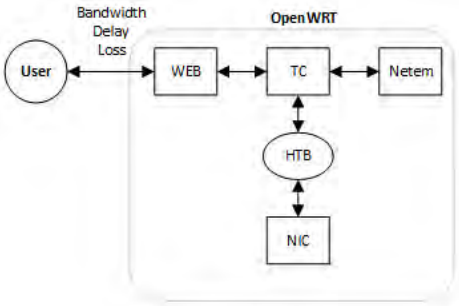


Fig. 4. Traffic Shaping of PANBED.

3.3. Backend Design

In PANBED, traffic control is the basis of backend's other functions, and repeatability of experiments is one of the important issues that PANBED needs to consider as a testbed.

Design of traffic control. In some testbeds, traffic shaping was implemented by DummyNet running on another Delay Node. DummyNet is a tool of FreeBSD system. In a PAN, there are a limited number of devices, PANBED cannot install and run DummyNet on all devices because these devices are not proprietary lab devices and may not support DummyNet; in addition, there may be no enough nodes to act as delay nodes; moreover, in order not to affect users' daily use, we should minimize the changes to these devices. For these reasons, traffic shaping in PANBED is implemented on the OpenWrt router.

OpenWrt supports two network emulation modules of Linux: netem [18] and TC (Traffic Controller)[10], [1]. Netem can simulate complex network transmission characteristics in a well-behaved LAN, such as variable bandwidth, delay, loss, repetition and reordering. TC controls the working mode of netem. TC supports classless and classful queue disciplines. The classless disciplines are relatively simple, and the date flow can be sorted, speed limited, and discarded, but cannot be differentiated fine-grained. The classful disciplines can implement fine-grained and differentiated traffic control by classifying packets with the classifier and filter.

In PANBED, a home router transmits experimental flow and non-experimental flow at the same time, but only control the experimental flow, so the classful disciplines are applied to achieve fine-grained traffic control. There are three types of classful queue disciplines: CBQ (based on class queuing), HTB (hierarchical token bucket) and PRIO (priority queue), but only HTB can control the flow fine-grained and easily, so we implement the traffic control using the HTB queue as shown in Fig. 4.

Repeatability of experiments. To ensure the repeatability of experiments, many testbeds initialized experimental devices before experiments starting. However, devices in a PAN are not special testing devices, and store a large amounts of users' data, it is not possible to initialize these devices frequently which poses challenges to the repeatability of experiments.

We take two empirical approaches to improve the repeatability: (i) Since users know enough about their experimental devices, PANBED allows them to decide whether or not to initialize these devices; (ii) PANBED records the system state of every experimental device before experiments starting, such as system version, CPU, memory usage, etc., and provides users a reference for analyzing of experiment results.

4. Implementation Details

4.1. Implementation of Frontend

The frontend is implemented by using the template technology through the OpenWrt LuCI configuration page. Users can view all devices connected to the router, select experimental devices and do their experiments through the web page.

4.2. Implementation of Backend

In order to enable users to easily build their personal testbed using devices around them, the backend needs to accomplish tasks such as device discovery, router configuration, rules parsing, rules applying and restore the PAN to the initial states after experiments.

Discovering devices. In home environments, most users use DHCP to allocate IP addresses, subnet masks, gateways, and DNS information to devices in a PAN. On a OpenWrt router, the information of devices connected to it is stored in dhcp.leases file and odhcpd file. PANBED discovers and displays all connected devices on the LuCI web page by parsing them.

Configuring the router. By default, a home router cannot directly control the flow using TC because for most of home routers, a LAN port doesn't be equipped a separate network interface adapter card (NIC). Usually, all wired LAN ports share a NIC, and all wireless devices share another NIC. Data exchanges among WAN, LAN and Wi-Fi go by bridge, and data exchanges among wired LAN ports don't pass through the NIC, as shown in Fig. 5 [2]. TC is a tool of network layer, and it can control the traffic only when the traffic passes through a physical NIC.

To solve the problem, vlan technology is used in PANBED. When users select N test devices, PANBED automatically creates N vlans on the router, and puts each device into a separate vlan so that data flow among these devices must pass through a physical NIC and make the TC take effect. In addition, PANBED creates routings for these vlans, because devices belonging to different vlans cannot communicate directly.

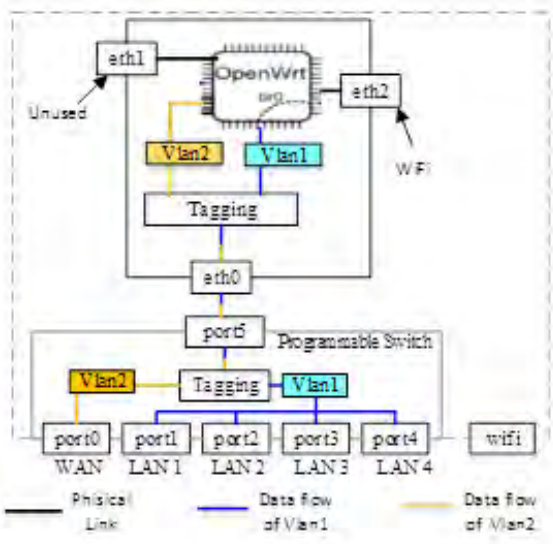


Fig. 5. Architecture of a Common AP which shows that all wired LAN ports share a NIC, and wireless devices share another NIC. The data exchanges among wired LAN ports don't pass through the NIC.

Parsing and applying rules. Links among testing devices specified by users are correspond to a series of rules. The format of each rule is as ¡source device, destination device,

bandwidth, delay, loss¿. PANBED parses and applies these rules by performing Algorithm 1.

```
Algorithm 1:
  (1)Creates N vlans on the Router;
  (2)Bonds every vlan with a specific interface,
      for example, vlan3 with eth0.3, vlan4 with eth0.4;
  (3)Allocates N experiment devices into N Vlans respectively;
  (4)Set IPs= {IPd1, IPd2 IPdn}; d1, d2,, and dn refers to
      experiment device 1, device 2, , and device n.
  (5)Read Rules
  (6)For each rule[i] in Rules
    Creates a htb sub-class with handle i and sets up the bandwidth limiter;
    Sets up netem for configuring delay and loss for sub-class i;
    Sets up the filter for filtering data flows of meeting conditions;
    Writes all above to a shell script file;
  (7)Run the script file;
```

Restoration of PAN environment. After finishing experiments, PANBED removes all experimental rules, vlans and routings on the router and restores the PAN to original states.

5. Evaluation

Users just need to upload several script files less than 16KB to a home router to utilize PANBED to do experiments. The following illustrates the realism, effectiveness and convenience of PANBED with a use case.

5.1. Use Case

With the enhancement of smart devices, mobile applications based on crowdsourcing are more and more. Assuming that a crowdsourcing application requires smart devices such as mobile phones, smart bracelets, watches and glasses to periodically report position, temperature, humidity and noise. For not influencing users' experience, the application can store the data in a file for a while until the file size exceeds a threshold, then upload it to the server.

A suitable threshold is important for better users' experience and it is related to many factors, among which the network condition is an important factor. When users are in a good Wi-Fi environment, the threshold can be set larger, but when users are in a mobile network environment with bad signal, too large threshold will increase the upload delay, even cause fail.

In order to improve users' experience, the developer wants to test a suitable threshold at different network conditions, and make the application adjust the threshold automatically according to the current network condition.

The current testbeds' supporting for smart devices are limited, some of them use virtual machine, and some of them provide real android smart phones, but their versions are old and their hardware condition is limited, so there exist some problems to complete similar testing tasks described above on these testbeds.

Certainly, the developer also can use devices around him to do the testing directly, but the processes are too complex. Firstly, the developer should be good at the Linux and network technology such as cross compile of OpenWrt, routing, vlan, traffic control and other technologies. Secondly, the developer should know the network configurations of each experimental device, configure the testing rules and restore them after experiments by manually.

PANBED builds a convenient experimental environment for testers utilizing devices around them. Using PANBED, testers can do testing easily and needn't to care about the configuration of each device and the implementation details of the underlying.

5.2. PANBED Setup

One of the important advantages of PANBED is portability and economy. We purchase an OEM OpenWrt router which works in 2.4GHz, with four 5dBi high omnidirectional antennas, supports IEEE 802.11b / g / n, IEEE 802.3 and IEEE 802.3u protocols, with a maximum wireless speed of 300Mbps, with one adaptive WAN port of 10/100M, four adaptive LAN port of 100/1000M, with wireless security basic features, with motherboard chipset of MT7620N, and costs about 65 Yuan RMB.

PANBED is light weighted. Firstly, it is very easy to install PANBED on an OpenWrt home router which only requires users to upload several script files to the router by SHH. The size of these files is less than 16KB, and they provide users all functions of PANBED. The total time of loading PANBED is less than 2 minutes.

Secondly, it is very simple for users to do testing on the PANBED. Users log on the router through the web browser such as http://192.168.1.1/. Then they can choose devices and do testing easily. Fig. 6 shows our devices in the PAN, including one PC, two Android phones, one Android smart glasses and one android smart bracelet. All devices are in network 192.168.1.0/24, connect to the router by Ethernet or Wi-Fi, can communicate with each other, and access the Internet.

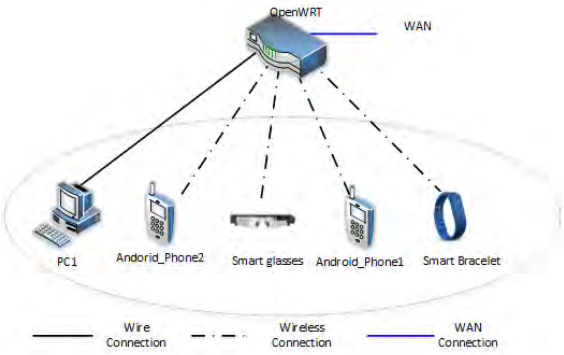


Fig. 6. Devices in our PAN including an OpenWrt home router, a PC, two android mobile phones, a smart glass and a smart bracelet, and they are in a same vlan initially.

5.3. Experiment Results

For completing the testing tasks described in the use case and verifying the effectiveness of PANBED, we implement a client app based on the android platform and a sever program of language C. The client periodically collects the position, temperature, humidity and noise of current environments and writes to a file. The client will send the file to a server when the file size is larger than the threshold set by the user. After uploading, the client returns the time taken by the uploading operation.

In following experiments, we choose android mobile phones as experimental devices for convenience, but PANBED is not confined to them, and those who support Wi-Fi and TCP/IP protocols can all be supported by PANBED.

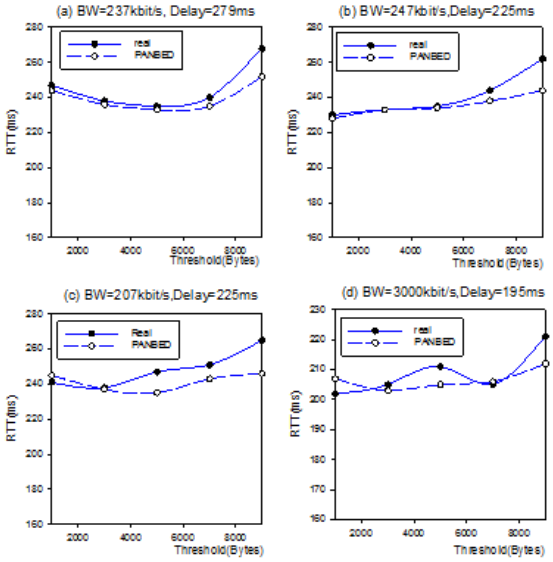


Fig. 7. Comparing of the results between PANBED and the real environment in four different network conditions which shows the values of RTT in PANBED are very similar with the real environment, and the difference is related to the dynamics of the network in some extent.

Realism of PANBED. Firstly we validate the realism of PANBED. We run the client on the android phone1 and run the server respectively on a remote PC and on the PC1 shown in Fig. 6. Then we measure the time taken by the uploading operation in two situation. In order to ensure the realism, we measure the network link between the android phone1 and the remote PC in real time with iperf [36] before the uploading, then emulate a same network link on the PANBED. The comparing results of PANBED and real situation in four different network conditions are shown in Fig. 7(a) - Fig. 7(d). We find that the experiment results on PANBED are essentially in agreement with that in the real environment and it proves the realism of PANBED.

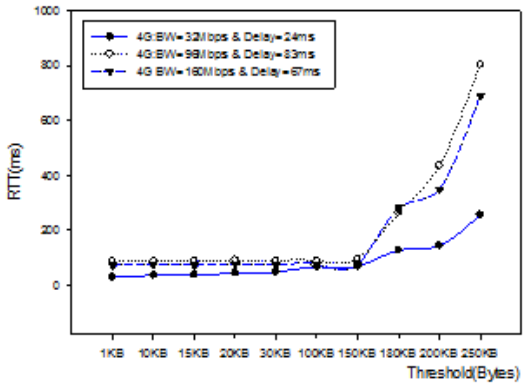


Fig. 8. RTT at Different Network and Different Threshold which shows the PANBED could simulate different network scene effectively.

Effectiveness of PANBED. To resolve the problems described in the use case and verify the effectiveness of PANBED, we emulate three different network links on PANBED and measure the uploading time at different threshold: (1) 4G network at 32Mbps bandwidth and 24ms delay; (2) 4G network at 96Mbps bandwidth and 83ms delay; (3) 4G network at 160Mbps bandwidth and 67ms delay. The results are shown in Fig. 8. We find, for the general mobile network, when the file size is less than 150K Bytes, the time spent on uploading file is relatively stable, but when the file size exceeds 180K Bytes, the time taken by uploading file will grow at speed of two times, three times, and even 4 times. It proves that PANBED can effectively solve the similar experimental requirements described in the use case.

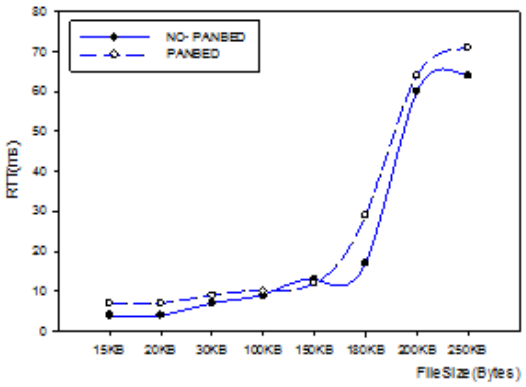


Fig. 9. Downloading time of android phone2 in two situations which shows the impact of doing experiment in the PAN is slight to other non-experimental devices.

Impact to other functions of PAN. The impact of PANBED to other non-experimental devices is very slight. We measure the time consumed by the android phone2 (shown in Fig. 6) which downloads files in case of not running PANBED and running PANBED on the router. When running PANBED, android phone1 and PC1 are selected as experimental devices, the link between them is set as BW=3Mbit/s and delay=19ms. The android phone1 sends data to PC1 continually. The results is shown in Fig. 9 which illustrates that the time taken by the android phone2 to download files is similar under two conditions and that proves it has little influence for other non-experiment devices to run PANBED on the router.

Fig. 10 displays the impact of PANBED on the memory and CPU of the router in three cases: (i) PANBED is not installed on the router; (ii) PANBED is installed but no testing task; (iii) PANBED is installed and works. The network scene is set as BW = 56 Kbit/s, Delay = 10ms, Loss = 0.10. The results are shown in Fig. 10 which prove that installing PANBED has little influence on the router and the PANBED is light weighted.

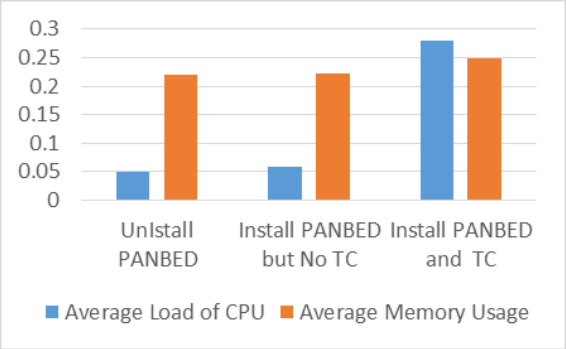


Fig. 10. Performance impact of PANBED to the router which shows the PANBED is a lightweight function embedded into the OpenWrt and its impact to the router is acceptable.

6. Conclusion

The paper describes the design and implementation of the prototype of PANBED, which build a small-scale personal testbed for users utilizing devices in their own personal area networks (PANs). The experiment results show that PANBED allows users to set up different network scenes to test applications easily using a home router, PCs, mobile phones and other devices. PANBED is light weighted with a size less than 16 KB and it has little impact to other functions of a PAN. The experiment results also prove the realism, effectiveness, flexibility and convenience of PANBED. PANBED can be used as a supplement to some existing testbeds and enable users to assess small applications at their convenience using diverse, personal, latest and low cost devices around them.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (Grant No.61261019), the Inner Mongolia Autonomous Region Natural Science Foundation

(Grant No.113113), the Program of Higher-level Talents of Inner Mongolia University, the Inner Mongolia Autonomous Region Natural Science Foundation (Grant No.2012MS0930), and the Inner Mongolia Autonomous Region Higher Education Institutions Scientific Research Project (Grant No.NJZY12032).

References

1. Network traffic control in openwrt. https://wiki.OpenWrt.org/doc/howto/packet.scheduler/packet.scheduler, accessed June, 2017
2. Openwrt network interfaces. http://wiki.openwrt.org/OpenWrtDocs/NetworkInterfaces?action=attachFile&do=get&target=ASUS-Internals-default-sm.png, accessed January, 2017
3. Openwrt supported devices. https://wiki.OpenWrt.org/toh/start, accessed June, 2017
4. Overlay network. http://en.wikipedia.org/wiki/Overlay_network, accessed June, 2017
5. Web server configuration (uhttpd). https://wiki.OpenWrt.org/doc/uci/uhttpd, accessed June, 2017
6. Wireless information network laboratory. http://www.winlab.rutgers.edu, accessed June, 2017
7. Andersen, D., Balakrishnan, H., Kaashoek, F., Morris, R.: Resilient overlay networks. ACM SIGCOMM Computer Communication Review 32(1), 66–66 (2002)
8. Bajaj, L., Takai, M., Ahuja, R., Tang, K., Bagrodia, R., Gerla, M.: Glomosim: A scalable network simulation environment. UCLA computer science department technical report 990027(1999), 213 (1999)
9. Berman, M., Chase, J.S., Landweber, L., Nakao, A., Ott, M., Raychaudhuri, D., Ricci, R., Seskar, I.: Geni: A federated testbed for innovative network experiments. Computer Networks 61, 5–23 (2014)
10. Brown, M.A.: Traffic control howto. Guide to IP Layer Network (2006)
11. Chillara, V.K., Liu, Y.H., Wang, B., Ba, A., Vidojkovic, M., Philips, K., de Groot, H., Staszewski, R.B.: 9.8 an 860μw 2.1-to-2.7 ghz all-digital pll-based frequency modulator with a dtc-assisted snapshot tdc for wpan (bluetooth smart and zigbee) applications. In: Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International. pp. 172–173. IEEE (2014)
12. Chun, B., Culler, D., Roscoe, T., Bavier, A., Peterson, L., Wawrzoniak, M., Bowman, M.: Planetlab: an overlay testbed for broad-coverage services. ACM SIGCOMM Computer Communication Review 33(3), 3–12 (2003)
13. Drake, J.D., Brian, J.M.: System and method for enabling a viewable pan id over a wireless personal area network (Jan 12 2016), uS Patent 9,237,511
14. Ertin, E., Arora, A., Ramnath, R., Naik, V., Bapat, S., Kulathumani, V., Sridharan, M., Zhang, H., Cao, H., Nesterenko, M.: Kansei: a testbed for sensing at scale. In: Proceedings of the 5th international conference on Information processing in sensor networks. pp. 399–406. ACM (2006)
15. Fainelli, F.: The openwrt embedded development framework. In: Proceedings of the Free and Open Source Software Developers European Meeting (2008)
16. Fall, K.: Network emulation in the vint/ns simulator. In: Computers and Communications, 1999. Proceedings. IEEE International Symposium on. pp. 244–250. IEEE (1999)
17. Gutirrez J, Villa-Medina J F, N.G.A.e.a.: Automated irrigation system using a wireless sensor network and gprs module. IEEE transactions on instrumentation and measurement 63(1), 166–176 (2014)

18. Hemminger, S., et al.: Network emulation with netem. In: Linux conf au. pp. 18–23 (2005)
19. Issariyakul, T., Hossain, E.: Introduction to network simulator NS2. Springer Science & Business Media (2011)
20. Katsaounis, G., Tsilomitrou, O., Manesis, S.: A wireless sensors and controllers network in automation a laboratory-scale implementation for students training. In: Control and Automation (MED), 2014 22nd Mediterranean Conference of. pp. 1067–1073. IEEE (2014)
21. Kciuk, M.: Openwrt operating system based controllers for mobile robot and building automation system students projects realization. In: Research and Education in Mechatronics (REM), 2014 15th International Workshop on. pp. 1–4. IEEE (2014)
22. Kim, C.G., Kim, K.J.: Implementation of a cost-effective home lighting control system on embedded linux with openwrt. Personal and ubiquitous computing 18(3), 535–542 (2014)
23. Kim, D.H., Bae, K.: System level approach for low energy consumption in wireless personal area networks. In: Consumer Electronics (ICCE), 2016 IEEE International Conference on. pp. 520–521. IEEE (2016)
24. Lee, W.J., Shin, J.W., Lee, H.Y., Chung, M.Y.: Testbed implementation for routing wlan traffic in software defined wireless mesh network. In: Ubiquitous and Future Networks (ICUFN), 2016 Eighth International Conference on. pp. 1052–1055. IEEE (2016)
25. Levis, P., Lee, N., Welsh, M., Culler, D.: Tossim: Accurate and scalable simulation of entire tinyos applications. In: Proceedings of the 1st international conference on Embedded networked sensor systems. pp. 126–137. ACM (2003)
26. Noh, J.Y., Kim, M.K., Yim, C.H., Han, K.S., Choi, K.H., Yu, J.H., Kim, M.S.: Packet transmission system based on wireless personal area network and method thereof (Oct 7 2014), uS Patent 8,855,090
27. Palazzi, C.E., Brunati, M., Roccetti, M.: An openwrt solution for future wireless homes. In: Multimedia and Expo (ICME), 2010 IEEE International Conference on. pp. 1701–1706. IEEE (2010)
28. Park, S., Savvides, A., Srivastava, M.B.: Sensorsim: A simulation framework for sensor networks. In: Proceedings of the 3rd ACM international workshop on Modeling, analysis and simulation of wireless and mobile systems. pp. 104–111. ACM (2000)
29. Paul, B., Marcombes, S., David, A., Struijk, L.N.A., Le Moullec, Y.: A context-aware user interface for wireless personal-area network assistive environments. Wireless Personal Communications 69(1), 427–447 (2013)
30. Polley, J., Blazakis, D., McGee, J., Rusk, D., Baras, J.S.: Atemu: a fine-grained sensor network simulator. In: Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004. 2004 First Annual IEEE Communications Society Conference on. pp. 145–152. IEEE (2004)
31. Reich, J., Misra, V., Rubenstein, D.: Roomba madnet: a mobile ad-hoc delay tolerant network testbed. ACM SIGMOBILE Mobile Computing and Communications Review 12(1), 68–70 (2008)
32. Riley, G.F., Henderson, T.R.: The ns-3 network simulator. Modeling and tools for network simulation pp. 15–34 (2010)
33. Rizzo, L.: Dummynet: a simple approach to the evaluation of network protocols. ACM SIGCOMM Computer Communication Review 27(1), 31–41 (1997)
34. Serrano, P., Bernardos, C.J., de La Oliva, A., Banchs, A., Soto, I., Zink, M.: Floornet: deployment and evaluation of a multihop wireless 802.11 testbed. EURASIP Journal on Wireless Communications and Networking 2010, 8 (2010)
35. Shrestha, B., Hossain, A.Z.E., Camorlinga, S.G., Krishnamoorthy, R., Niyato, D.: Method and system for allocation guaranteed time slots for efficient transmission of time-critical data in ieee 802.15. 4 wireless personal area networks (Mar 10 2015), uS Patent 8,976,763
36. Tirumala, A., Qin, F., Dugan, J., Ferguson, J., Gibbs, K.: Iperf: The tcp/udp bandwidth measurement tool. http://dast. nlanr. net/Projects (2005)

37. Titzer, B.L., Lee, D.K., Palsberg, J.: Avrora: Scalable sensor network simulation with precise timing. In: Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on. pp. 477–482. IEEE (2005)
38. Vahdat, A., Yocum, K., Walsh, K., Mahadevan, P., Kostić, D., Chase, J., Becker, D.: Scalability and accuracy in a large-scale network emulator. ACM SIGOPS Operating Systems Review 36(SI), 271–284 (2002)
39. Varga, A.: Omnet++. Modeling and Tools for Network Simulation pp. 35–59 (2010)
40. Vatti, R.A., Gaikwad, A.N.: Throughput improvement of high density wireless personal area networks. In: Computational Intelligence and Communication Networks (CICN), 2014 International Conference on. pp. 506–509. IEEE (2014)
41. Wang, Y., Wang, Q., Zheng, G., Zeng, Z., Zheng, R., Zhang, Q.: Wicop: Engineering wifi temporal white-spaces for safe operations of wireless personal area networks in medical applications. IEEE transactions on mobile computing 13(5), 1145–1158 (2014)
42. Welsh, E., Fish, W., Frantz, J.P.: Gnomes: A testbed for low power heterogeneous wireless sensor networks. In: Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on. vol. 4, pp. IV–IV. IEEE (2003)
43. Werner-Allen, G., Swieskowski, P., Welsh, M.: Motelab: A wireless sensor network testbed. In: Proceedings of the 4th international symposium on Information processing in sensor networks. p. 68. IEEE Press (2005)
44. White, B., Lepreau, J., Stoller, L., Ricci, R., Guruprasad, S., Newbold, M., Hibler, M., Barb, C., Joglekar, A.: An integrated experimental environment for distributed systems and networks. ACM SIGOPS Operating Systems Review 36(SI), 255–270 (2002)
45. Wu, H., Luo, Q., Zheng, P., Ni, L.M.: Vmnet: Realistic emulation of wireless sensor networks. IEEE Transactions on Parallel and Distributed Systems 18(2), 277–288 (2007)
46. Zhang, K.: The Design and Implementation of An OpenWrtłbased PPPoE Traffic Control System. Ph.D. thesis, Nankai University,Chian (2014)
47. Zhang, W., Han, S., He, H., Chen, H.: Network-aware virtual machine migration in an overcommitted cloud. Future Generation Computer Systems (2016)
48. Zhang, W., Li, X., Xiong, N., Vasilakos, A.V.: Android platform-based individual privacy information protection system. Personal and Ubiquitous Computing 20(6), 875–884 (2016)
49. Zheng, G., Hua, C., Zheng, R., Wang, Q.: Toward robust relay placement in 60 ghz mmwave wireless personal area networks with directional antenna. IEEE Transactions on Mobile Computing 15(3), 762–773 (2016)
50. Zhu, Y.H., Chi, K., Tian, X., Leung, V.C.: Network coding-based reliable ipv6 packet delivery over ieee 802.15. 4 wireless personal area networks. IEEE Transactions on Vehicular Technology 65(4), 2219–2230 (2016)

Qiaozhi Xu received the B.S. degrees in computer science and techonology from Inner Mongolia Normal University of China in 2000 and the M.S degree from Nanjing Normal University of China in 2005.She joined Inner Mongolia Normal University of China since 2005. She is currently working towards her Ph.D. in Inner Mongolia University since 2013. Her research interests include network testbed,computer network and cloud computing.

Junxing Zhang is a professor in the College of Computer Science at the Inner Mongolia University. He is also the Director of the Inner Mongolia Key Laboratory of Wireless Networking and Mobile Computing. He received a B.S. degree in Computer Engineering from the Beijing University of Posts and Telecommunications, a M.S. degree in Computer Science from the Colorado State University, and a Ph.D. degree from the University

of Utah. His research interests include network measurement and modeling, mobile and wireless networking, network security and verification, etc. Prof. Zhang was awarded the title of "Grassland Talent" by the government of the Inner Mongolia Autonomous Region in 2010. He has published over 40 papers in various internationally recognized journals and conferences, and led several national and provincial research projects. He also served as a peer reviewer for several international journals and conferences, such as IEEE Transactions on Mobile Computing, Wireless Networks, and ICNP.

Received: December 30, 2016; Accepted: August 8, 2017.

Computer Science and Information Systems 14(3):875–912 https://doi.org/10.2298/CSIS170113034K

Supporting the platform extensibility for the model-driven development of agent systems by the interoperability between domain-specific modeling languages of multi-agent systems

Geylani Kardas[1], Emine Bircan[2], and Moharram Challenger[3]

International Computer Institute, Ege University, 35100, Bornova, Izmir, Turkey
[1]geylani.kardas@ege.edu.tr, [2]eminebircanbircan@gmail.com,
[3]moharram.challenger@ege.edu.tr

Abstract. The conventional approach currently followed in the development of domain-specific modeling languages (DSMLs) for multi-agent systems (MASs) requires the definition and implementation of new model-to-model and model-to-text transformations from scratch in order to make the DSMLs functional for each different agent execution platforms. In this paper, we present an alternative approach which considers the construction of the interoperability between MAS DSMLs for a more efficient way of platform support extension. The feasibility of using this new interoperability approach instead of the conventional approach is exhibited by discussing and evaluating the model-driven engineering required for the application of both approaches. Use of the approaches is also exemplified with a case study which covers the model-driven development of an agent-based stock exchange system. In comparison to the conventional approach, evaluation results show that the interoperability approach requires both less development time and effort considering design and implementation of all required transformations.

Keywords: Metamodel, Model transformation, Model-driven development, Domain-specific Modeling Language, Multi-agent System, Interoperability.

1. Introduction

Software agents in a Multi-agent system (MAS) interact with each other to solve problems in a competitive or collaborative manner within an environment. In a MAS, software agents are expected to be autonomous, mostly through a set of reactive/proactive behaviors designed for addressing situations likely to happen in particular domains [1-3]. Both internal agent behavior model and interactions within a MAS become even more complex and hard to implement when taking into account the varying requirements of different agent environments [4]. Hence, working in a higher abstraction level is of critical importance for the development of MASs since it is almost impossible to observe code level details of MASs due to their internal complexity, distributedness and openness [5].

Agent-oriented software engineering (AOSE) [6] researchers define various agent metamodels (e.g. [7-11]), which include fundamental entities and relations of agent

systems in order to master the abovementioned problems of developing MASs. In addition, many model-driven agent development approaches are provided such as [12-15] and researchers also propose domain-specific languages (DSLs) / domain-specific modeling languages (DSMLs) (e.g. [16-24]) for facilitating the development of MASs by enriching MAS metamodels with some defined syntax and semantics (usually translational semantics [25]). DSLs / DSMLs [26-28] have notations and constructs tailored toward a particular application domain (e.g. MAS) and help to the model-driven development (MDD) of MASs. MDD aims to change the focus of software development from code to models [29], and hence many AOSE researchers believe that this paradigm shift introduced by MDD may also provide the desired abstraction level and simplify the development of complex MAS software [5].

In AOSE, perhaps the most popular way of applying model-driven engineering (MDE) for MASs is based on providing DSMLs specific to agent domain with including appropriate integrated development environments (IDEs) in which both modelling and code generation for system-to-be-developed can be performed properly. Proposed MAS DSMLs such as [17, 21, 23] usually support modelling both the static and the dynamic aspects of agent software from different MAS viewpoints including agent internal behaviour model, interaction with other agents, use of other environment entities, etc. Within this context, abstract syntaxes of the languages are represented with metamodels covering those aspects and required viewpoints to some extent. Following the construction of abstract and concrete syntaxes based on the MAS metamodels, the operational semantics of the languages are provided in the current MAS DSML proposals by defining and implementing entity mappings and model-to-model (M2M) transformations between the related DSML's metamodel and the metamodel(s) of popular agent implementation and execution platform(s) such as JACK [30], JADE [31] and JADEX [32]. Finally, a series of model-to-text (M2T) transformations are implemented and applied on the outputs of the previous M2M transformations which are the MAS models conforming to the related agent execution platforms. Hence, agent software codes, MAS configuration files, etc. pertaining to the implementation and deployment of the modeled agent systems on the target MAS platform are generated automatically.

When we take into account the different abstractions covered by the metamodels of MAS DSMLs and the underlying agent execution platforms, DSML metamodels can be accepted as the platform-independent metamodels (PIMMs) of agent systems while metamodels of the agent execution platforms are platform-specific metamodels (PSMMs) according to the Object Management Group (OMG)'s well-known Model-driven Architecture (MDA) [33] as also indicated in [9] and [14].

Above described approach (which we can refer as the conventional or classical approach) applied in the current MAS DSML development studies, unfortunately requires the definition and implementation of new M2M and M2T transformations from scratch in order to make the DSMLs functional for different agent execution platforms. In other words, for each new target agent execution platform, MAS DSML designers should repeat all the time-consuming and mostly troublesome steps of preparing the vertical transformations [34] between the related DSML and this new agent platform.

Motivated by the similarity encountered in the abstract syntaxes of the available MAS DSMLs, we are quite convinced that both the definition and the implementation of M2M transformations between the PIMMs of MAS DSMLs would be more convenient and

less laborious comparing with the transformations required between MAS PIMMs and PSMMs in the way of enriching the support of MAS DSMLs for various agent execution platforms. Hence, in this paper, we present our approach which aims at improving the mechanism of constructing language semantics over the interoperability of MAS DSMLs and providing a more efficient way of extension for the executability of modeled agent systems on various underlying agent platforms. Differentiating from the existing MAS DSML studies (e.g. [17, 20, 21, 23, 24]), our proposal is based on determining entity mappings and building horizontal M2M transformations between the metamodels of MAS DSMLs which are in the same abstraction level. In this paper, we also investigate the feasibility of using this new interoperability approach instead of the conventional approach of platform support for current MAS DSMLs by first discussing the MDE required for the application of both approaches, and then conducting a comparative evaluation of two approaches according to an evaluation framework [35] which is specific for the assessment of MAS DSMLs. For this purpose, application of the proposed interoperability approach is demonstrated by constructing horizontal transformations between two full-fledged agent DSMLs called SEA_ML [23] and DSML4MAS [9] respectively. In order to provide the related comparison, we also discuss the application of the classical approach on SEA_ML instance models. Use of both approaches is exemplified with a case study which covers the MDD of an agent-based stock exchange system. Finally, development costs of two approaches are evaluated.

This paper is an extended version of the paper [36]. It differs from the latter by including: 1) a completely new discussion on design and implementation of M2M and M2T transformations required for extending the platform support of a MAS DSML according to the conventional approach 2) an improved case study in which MDD of an agent-based stock exchange is realized by using both the proposed interoperability approach and the conventional approach currently in-use, and finally 3) another new section which covers the comparative evaluation of applying two approaches.

The rest of the paper is organized as follows: In Sect. 2, the approach considering the MAS DSML interoperability is presented. Two agent DSMLs used in this study are briefly discussed in Sect. 3. Sect. 4 discusses how the interoperability can be built between MAS DSMLs while Sect. 5 demonstrates the application of the conventional way of platform support for MAS DSMLs. In Sect. 6, a case study on the development of an agent-based stock exchange system with using both the proposed approach and following the conventional way is given. Sect. 7 includes the comparative evaluation of the two approaches. Related work is given in Sect. 8. Finally, Sect. 9 concludes the paper and states the future work.

2. Proposed approach for the interoperability of MAS DSMLs

As indicated in [36], support of current MAS DSMLs for each agent execution platform is enabled by repetitively defining and implementing a chain of vertical M2M and M2T transformations. Available M2M and M2T transformations are specific for each different agent platform (such as JADE [31, 37], JACK [30, 38], JADEX [32, 39]) and almost all of them cannot be re-used while extending the executability of the MAS

models for a new agent platform. Due to the difficulty encountered on repeating those vertical model transformation steps, current MAS DSML proposals mostly support the execution of modeled agents on just one agent platform (e.g. [17-19, 21, 23]). Very few of them enable the execution of models on two different agent platforms (e.g. [9, 14]) and, as far as we know, there is no any MAS DSML which provides the execution of modeled agents on more than two different agent platforms. In order to increase the platform variety, we propose benefiting from the vertical transformations already existing between the syntax of a MAS DSML (let us call $DSML_1$) and metamodels of various agent platforms for enabling model instances of another MAS DSML (let us call $DSML_2$) executable on the same agent platforms by just constructing horizontal transformations between the PIMMs of the MAS DSMLs in question. Therefore, instead of defining and implementing N different M2M and M2T transformations for N different agent platforms, creation of only one single set of M2M transformations between $DSML_1$ and $DSML_2$ can be enough for the execution of $DSML_2$'s model instances on these N different agent platforms. Taking into account the MDE of software systems in general, our approach also fits into the theory of modeling spaces [40] where model transformations are proposed to bridge two conceptual spaces. In here, the metamodels of different MAS DSMLs can be considered as representing different conceptual spaces and our aim is to bridge these metamodels to support agent platform extensibility for the related DSMLs.

The construction of model transformations between MAS DSMLs and hence re-use of already existing transformations between those DSMLs and agent platforms are depicted in Fig. 1. Let the abstract syntaxes of $DSML_1$, $DSML_2$ and $DSML_3$ be the metamodels MM_1, MM_2 and MM_3 respectively. Horizontal lines between these MAS DSMLs represent the M2M transformations between these metamodels while each vertical line between a DSML and the MM of an agent platform represents the M2M transformations between this DSML and the agent platform. According to the figure, agent systems modeled in $DSML_1$ are already executable on the agent platforms A and B (due to the existing vertical transformations for these platforms), while $DSML_2$ model instances are executable on the agent platforms X, Y and Z. Similarly, M2M and M2T transformations were already provided for the execution of $DSML_3$ model instances on the agent platforms α, β, θ respectively. If $DSML_1$ is required to support X and Y agent platforms, developers should prepare new model transformations separately for those agent platforms (shown with dotted arrows in Fig. 1) in case of the absence of horizontal transformations between MM_1 and MM_2. Hence, construction of only one set of horizontal M2M transformations between $DSML_1$ and $DSML_2$ enables $DSML_1$'s automatic support on agent platforms X, Y (and also Z). Conversely, same is also valid for extending the $DSML_2$'s support for agent execution platforms. Interoperability between $DSML_1$ and $DSML_2$ over these newly defined horizontal transformations also makes transformation and code generation of $DSML_2$ model instances for the agent platforms A and B. In addition to the important decrease in the number of transformations, construction of horizontal model transformations between the PIMMs of MAS DSMLs will be more feasible and easier than the vertical transformations since the DSMLs are in the same abstraction level according to MDA [33].

In this paper, we discuss the applicability of the above proposed approach by taking into account the construction of the interoperability between two MAS DSMLs called SEA_ML [23] and DSML4MAS [17]. Both DSMLs enable the modeling of agent

systems according to various agent internal and MAS organizational viewpoints. They provide a clear visual syntax for MAS modeling and code generation for agent implementation and execution platforms. Moreover, both languages are equipped with Eclipse-based IDEs in which modeling and automatic generation of MAS components are possible. These features of the languages led us to choose them in this study. Before discussing the details of how the approach is applied over these DSMLs, SEA_ML and DSML4MAS are briefly introduced in the following section.

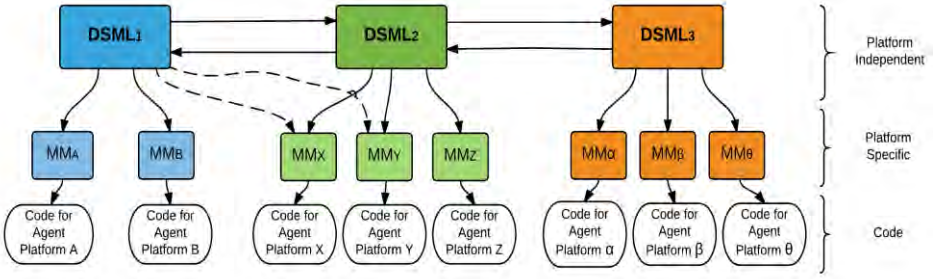


Fig. 1. Interoperability of MAS DSMLs via horizontal model transformations

3. Two agent DSMLs: SEA_ML and DSML4MAS

In the following subsections, main language features and metamodels of SEA_ML and DSML4MAS are briefly discussed before constructing the MAS DSML interoperability between them as proposed in this study.

3.1. SEA_ML

SEA_ML [23] is a MAS modeling language which enables the developers to model the agent systems in a platform independent level and then automatically achieve codes and related documents required for the execution of the modeled MAS on target MAS implementation platforms. It provides a convenient and handy environment for agent developers to construct and implement software agent systems working on various application domains. In order to support MAS experts when programming their own systems, and to be able to fine-tune them visually, SEA_ML covers all aspects of an agent system from the internal view of a single agent to the complex MAS organization. In addition to these capabilities, SEA_ML also supports the model-driven design and implementation of autonomous agents who can evaluate semantic data and collaborate with semantically-defined entities of the Semantic Web [41], like Semantic Web Services (SWS) [42]. That feature exactly differentiates SEA_ML and makes unique regarding any other MAS DSML currently available. Within this context, it includes new viewpoints which specifically pave the way for the development of software agents working on the Semantic Web environment [41]. Modeling agents, agent knowledge-

bases, platform ontologies, SWS and interactions between agents and SWS are all possible in SEA_ML.

SEA_ML's metamodel is divided into eight viewpoints, each of which represents a different aspect for developing Semantic Web enabled MASs [23, 43]. *Agent's Internal Viewpoint* is related to the internal structures of semantic web agents (SWAs) and defines entities and their relations required for the construction of agents. It covers both reactive and Belief-Desire-Intention (BDI) [44] agent architectures. *Interaction Viewpoint* expresses the interactions and the communications in a MAS by taking messages and message sequences into account. *MAS Viewpoint* solely deals with the construction of a MAS as a whole. It includes the main blocks which compose the complex system as an organization. *Role Viewpoint* delves into the complex controlling structure of the agents and addresses role types. *Environmental Viewpoint* addresses the use of resources and interaction between agents with their surroundings. *Plan Viewpoint* deals with an agent Plan's internal structure, which is composed of Tasks and atomic elements such as Actions. *Ontology Viewpoint* addresses the ontological concepts which constitute agent's knowledgebase (such as belief and fact). *Agent - SWS Interaction Viewpoint* defines the interaction of agents with SWS including the definition of entities and relations for service discovery, agreement and execution. A SWA executes the semantic service finder Plan (SS_FinderPlan) to discover the appropriate services with the help of a special type of agent called SSMatchMakerAgent who executes the service registration plan (SS_RegisterPlan) for registering the new SWS for the agents. After finding the necessary service, one SWA executes an agreement plan (SS_AgreementPlan) to negotiate with the service. After negotiation, a plan for service execution (SS_ExecutorPlan) is applied for invoking the service.

The collection of SEA_ML viewpoints constitutes an extensive and all-embracing model of the MAS domain. SEA_ML's abstract syntax combines the generally accepted aspects of MAS (such as MAS, Agent Internal, Role and Environment) and introduces two new viewpoints (Agent-SWS Interaction and Ontology) for supporting the development of software agents working within the Semantic Web environment [4].

SEA_ML can be used for both modeling MASs and generation of code from the defined models. SEA_ML instances are given as inputs to a series of M2M and M2T transformations to achieve executable artifacts of the system-to-be-built for JADEX [32] agent platform and semantic web service description documents conforming to *Web Ontology Language for Services (OWL-S)* ontology [45]. It is also possible to automatically check the integrity and validity of SEA_ML models [46]. A complete discussion on SEA_ML can be found in [23]. The language and its supporting tool are available in [47].

3.2. DSML4MAS

DSML4MAS [9, 17] is perhaps one of the first complete MAS DSMLs in which a PIMM, called PIM4Agents, provides an abstract syntax for different aspects of agent systems. Similar to SEA_ML's viewpoints, both internal behavior model of agents and agent interactions in a MAS are covered by PIM4Agents views / aspects. *Multi-agent view* contains all the main concepts in a MAS such as Agent, Cooperation, Capability, Interaction, Role and Environment. *Agent view* focuses on the single autonomous entity

(agent), the roles it plays within the MAS and the capabilities it has to solve tasks and to reach the environment resources. *Behavioural view* describes how plans are composed by complex control structures and simple atomic tasks like sending a message and how information flows between those constructs. In here, a plan is a specialized version of behavior composed of activities and flows. Activities and tasks are minimized parts of the work and flows provide the communication between these parts. *Organization view* describes how single autonomous entities cooperate within the MAS and how complex organizational structures can be defined. Social structure in the system is defined with cooperation entity where agents and organizations take part in. The structure has its own protocol defining how the entities interact in a cooperation. Agents have "domainRoles" for the interaction and these roles are attached to the actors by "actorBinding" entities where actors are representative entities within the corresponding interaction protocol. *Role view* examines the behaviour of an agent entity in an organization or cooperation. An agent's role covers the capabilities and information to have access to a set of resources. *Interaction view* describes how the interaction in the form of interaction protocols takes place between autonomous entities or organizations. Agents communicate over the PIM4Agents Protocol which refers to actors and "messageFlows" between these actors. Finally, *Environment view* contains the resources accessed and shared by agents and organizations. Agents can communicate with the environment indirectly via using resources. Resources can store knowledge from BDI agents for changing beliefs by using Messages and Information flows.

Grouping modelling concepts in DSML4MAS allows the metamodel evolution by adding new modelling concepts in the existing aspects, extending existing modelling concepts in the defined aspects, or defining new modelling concepts for describing additional aspects of agent systems [9]. For instance, SWS integration into the system models conforming to DSML4MAS is provided via introducing the SOAEnvironment entity [48] which extends the Environment entity and contains service descriptions. Agents use service descriptions to specify the Services they are searching for and then service interaction is realized by InvokeWS and ReceiveWS tasks which are inherited from Send and Receive task entities described in PIM4Agents.

Similar to SEA_ML, DSML4MAS also enables the MDD of MAS including a concrete graphical syntax [49] based on the aforementioned PIMM (PIM4Agents) and an operational semantics for the execution of modeled agent systems on JACK [30] or JADE [31] agent platforms. Extensions to the language introduced in [48] provide the description of the services inside an agent environment according to specifications such as *Web Services Modeling Language (WSML)* [50] or *Semantic Annotation of WSDL and XML Schema (SAWSDL)* [51]. Interested readers may refer to [48] and [9] for an extensive discussion on DSML4MAS. The language is available with its modeling tools in [52].

4. Building the interoperability between SEA_ML and DSML4MAS with horizontal model transformations

We have applied the horizontal transformability approach described in Sect. 2 for establishing the interoperability between SEA_ML and DSML4MAS. As shown in Fig.

2, SEA_ML currently supports the MAS implementation for JADEX BDI architecture [32] and SWS generation according to the OWL-S ontology [45]. In order to extend its platform support capability, new M2M and M2T transformations should be prepared for each new implementation platform. Let us consider extending execution platforms for SEA_ML agents with another well-known and widely-used MAS execution and deployment platform called JADE [31]. In order to make SEA_ML instances also executable on the JADE platform, definition and implementation of M2M transformations are needed between the abstract syntax of SEA_ML and PSMM of JADE framework. It is worth indicating that definition and application of M2T transformations are also required for the code generation from the outputs of the previous SEA_ML to JADE transformations (as will be discussed in Sect. 5). The methodology described above is currently the dominating MAS DSML engineering approach and also the most preferred way of model-driven agent development in AOSE [4, 5, 53]. Instead, we can follow the approach introduced in Sect. 2 by just writing the horizontal transformation rules between the metamodels of SEA_ML and DSML4MAS and running those transformations on SEA_ML instances for the same purpose: making SEA_ML models executable also on JADE platform. That is possible since DSML4MAS has already support on JADE and JACK [38] agent platforms and SAWSDL [51] and WSML semantic service ontologies [50] via vertical transformations between its metamodel and metamodels of the corresponding system implementation platforms. Realization of horizontal transformations between SEA_ML and DSML4MAS has extra benefits such as the execution of SEA_ML instances also on JACK platform and/or implementation of the modeled SWS according to SAWSDL or WSML specifications (Fig. 2).

Before deriving the rules of transformations, we should determine the entity mappings between both languages since the transformations are definitely based on these entity mappings. Comparing with the mappings we previously provided in [22] or [23] for the transformability of SEA_ML instances to MAS execution platforms, we have experienced that the determination of the entity mappings in this study was easier and took less time. We believe that the reason of this efficiency originates from the fact that metamodels of SEA_ML and DSML4MAS are in the same abstraction level and provide close entities and relations in similar viewpoints for MAS modeling (as will be discussed in Sect. 7 of this paper).

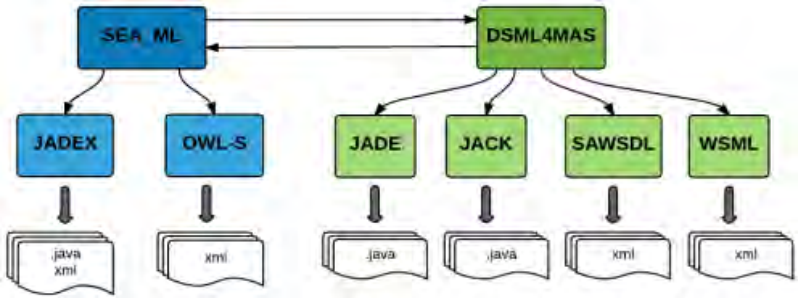


Fig. 2. Interoperability of SEA_ML and DSML4MAS

Table 1 lists some of the important mappings constructed between first-class entities of these two languages. For instance, two agent types (SWA and SSMatchmakerAgent) defined in SEA_ML are mapped onto the autonomous entity Agent defined in DSML4MAS. Likewise, meta-entities pertaining to agent plan types (SS_RegisterPlan, SS_FinderPlan, SS_AgreementPlan and SS_ExecutorPlan) required for the interaction between the semantic services are mapped with the Plan concept of DSML4MAS. Since Actor entity in DSML4MAS has access to resources and owns capabilities needed for agent interactions, SEA_ML's Role entity is mapped onto Actor entity.

Table 1. Entity mappings between the metamodels of SEA_ML and DSML4MAS.

SEA_ML MM Entity	DSML MM Entity
SemanticWebAgent (SWA)	Agent
SSMatchmakerAgent	Agent
Role	Actor
SemanticWebService (SWS)	SOAEnvironment
Environment	Environment
WebService	Service
Interface	Functionals
Process	Functionals
Grounding	InvokeWS
Input	Input
Output	Output
Precondition	Precondition
SS_RegisterPlan	Plan
SS_FinderPlan	Plan
SS_AgreementPlan	Plan
SS_ExcecutorPlan	Plan
SemanticWebOrganization (SWO)	Organization

One interesting mapping is encountered between SEA_ML's SWS entity and DSML4MAS's SOAEnvironment since it enables the representation of SEA_ML semantic services in DSML4MAS model instances. On DSML4MAS side, SOAEnvironment entity, which is extended from Environment entity, includes services in general. Hence, SEA_ML SWS entity is mapped onto SOAEnvironment entity and SEA_ML WebService entities are mapped onto Service entities. In SEA_ML WebService definition, every service has Interface, Process and Grounding. Interface entity represents the information about service inputs, outputs and any other necessary information. Process entity has internal information about the service and finally Grounding entity defines the invocation protocol of the web service [23]. DSML4MAS services are described with Blackbox and Glassbox entities [48]. BlackBox is used to define a service's functional and non-functional parameters while Glassbox includes the description of the internal service process. The Functionals are described in terms of

service signature that are input and output parameters, and specifications that are preconditions and effects. The NonFunctionals are defined in terms of price, service name and developer. Hence, Interface and Process entities of services defined in SEA_ML are mapped onto DSML4MAS Functionals which have input and output definitions. On DSML4MAS side, agent interactions with services are provided by InvokeWS and ReceiveWS tasks. Therefore, SEA_ML Grounding, which represents the physical structure of the underlying web service executed for the corresponding SWS, is mapped to InvokeWS. Remaining mappings listed in Table 1 (e.g. SEA_ML SWO to DSML4MAS Organization, SEA_ML Environment to DSML4MAS Environment) are very simple to determine since the related entities on both sides have similar or almost same functionality within the syntaxes of the languages.

After determining the entity mappings between SEA_ML and DSML4MAS, it is necessary to provide model transformation rules which are applied at runtime on SEA_ML instances to generate DSML4MAS counterparts of these instances. For that purpose, transformation rules should be formally defined and written according to a model transformation language ([34, 54]). In this study, we preferred to use ATL Transformation Language (ATL) [55] to define the model transformations between SEA_ML and DSML4MAS. ATL is one of the well-known model transformation languages, specified as both metamodel and textual concrete syntax. An ATL transformation program is composed of rules that define how the source model elements are matched and navigated to create and initialize the elements of the target models. In addition, ATL can define an additional model querying facility which enables specifying the requests onto models. ATL also allows code factorization through the definition of ATL libraries. Finally, ATL has a transformation engine and an IDE [56] that can be used as a plug-in on an Eclipse platform. These features of ATL caused us to prefer it as the implementation language for the horizontal transformations from SEA_ML to DSML4MAS.

ATL is composed of four fundamental elements. The first one is the header section defining attributes relative to the transformation module. The next element is the import section which is optional and enables the importing of some existing ATL libraries. The third element is a set of helpers that can be viewed as the ATL equivalents to the Java methods. The last element is a set of rules that defines the way target models are generated from source models.

Following listing include an excerpt from the written ATL rules in order to give some flavor of M2M transformations provided in this study. To this end, the rule in Listing 1 enables the transformation of the elements covered by the Agent-SWS Interaction viewpoint of SEA_ML to their counterparts included in the Multi-agent viewpoint of DSML4MAS. In line 1, the rule is named uniquely. In line 2, the source metamodel is chosen and renamed as swsinteractionvp with “from” keyword. The target metamodel is indicated and renamed as pim4agents with “to” keyword (Line 3). In the following lines (between 4 and 14), instances of SEA_ML SWA and SSMatchmakerAgent entities are selected and transformed to DSML4MAS Agent instances. Transformation of agent roles and plans are also realized by using “Set” and “allInstances” functions. It is worth indicating that types of Plan instances seem to be transformed to DSML4MAS behavior in the given listing although all SEA_ML Plan types are semantically mapped to DSML4MAS Plan as listed in Table 1. That is because some of the DSML4MAS meta-entities are collected with tag definitions in Ecore representations which take the same

name with the related viewpoint. For instance, plans are not defined solely with their names; instead they are collected in behavior definitions. Hence, in order to provide the full transformations of the plans with all their attributes, ATL rule is written here as mapping SEA_ML plan instances to the DSML4MAS behaviors. Inside another helper rule, those behaviors are separated into the corresponding plans and so exact transformation of SEA_ML plan instances to DSML4MAS plans are realized. More examples of the ATL rules written for the required transformations can be found in [36].

```
01 rule SWSInteractionVP2MultiagentSystem {
02   from   swsinteractionvp:
         SWSInteraction!SWSInteractionViewpoint
03       to          pim4agent: PIM4Agents!MultiagentSystem (
04         agent <- Set
          {SWSInteraction!SemanticWebAgent.allInstances()},
05         agent <- Set {SWSInteraction!SSMatchmakerAgent.allInstances()},
06         role <- Set {SWSInteraction!Role.allInstances()},
07         role <- Set {SWSInteraction!RegistrationRole.allInstances()},
08         behavior <- Set {SWSInteraction!SS_AgreementPlan.allInstances()},
09         behavior <- Set {SWSInteraction!SS_ExecutorPlan.allInstances()},
10         behavior <- Set {SWSInteraction!SS_FinderPlan.allInstances()},
11         behavior <- Set {SWSInteraction!SS_RegisterPlan.allInstances()},
12         environment <-Set {SWSInteraction!SWS.allInstances()},
13         environment <-Set {SWSInteraction!Grounding.allInstances()} )
14   }
```

Listing 1. An excerpt from the SWSInteractionVP2MultiagentSystem rule

As it will be demonstrated in Sect. 6, the application of these horizontal model transformations on SEA_ML model instances automatically produces the counterparts of these MAS models according to DSML4MAS specifications. The ATL engine uses the Ecore representation of a SEA_ML MAS model as the input, executes the transformation rules and outputs the corresponding DSML4MAS model again in Ecore format. Produced MAS model is ready to be processed inside the IDE of DSML4MAS. The model can be opened and/or directly utilized in this IDE for the generation of executable codes for JADE or JACK agent platforms.

5. Following the conventional way: Execution of SEA_ML Models on JADE Platform via vertical M2M and M2T transformations

In order to provide a comparison between the new proposed approach and the classical way of platform support for MAS DSMLs, we also designed and implemented direct transformations from SEA_ML instances into the JADE counterparts and realized code generation from the output agent models. This section discusses how the new platform extensibility for SEA_ML can be enabled by a series of vertical M2M and M2T transformations according to the well-known MDA principles [33, 57] and hence it gives some flavor of applying MDD methodology which is currently followed by most

of the agent developers to design and implement a DSML with including an operational semantics from scratch.

Execution of any MAS model conforming to an agent DSML requires first a M2M transformation to prepare the counterpart of the model in the targeted agent execution platform. Then a series of M2T transformations are applied on this platform specific model to generate executable software codes and/or files (e.g. [5, 9, 13, 23]). Hence, the first subsection describes how the transformations between SEA_ML and JADE are built while the second subsection discusses code generation from the output JADE model instances.

5.1. M2M Transformations between SEA_ML and JADE platform

Taking into consideration the MDA and its abstraction layers, SEA_ML resides on the platform independent model (PIM) layer and its abstract syntax (discussed in Sect. 3) can be utilized in this work as a PIMM while JADE platform locates on the platform specific model (PSM) layer and its metamodel represents a PSMM. JADE [31, 37] is one of the widely used agent development and execution platforms. It provides an open source Java API, currently distributed by Telecom Italia [58]. The API can be used to implement agents as Java objects. In JADE, agent internals including agent behaviours can be developed according to the IEEE Foundation for Intelligent Physical Agents (FIPA) standards [59]. Moreover, interactions between the software agents can be programmed based on the FIPA Agent Communication Language specifications [60] and MAS platform is supported with agent management and directory facilitator services which are all defined in FIPA standards to manage agents and provide yellow page services for agents to find and communicate with other agents.

After in-depth examination of the JADE API, a general metamodel of JADE platform has been derived and prepared in Ecore format which can be used as a PSMM. Fig. 3 gives an excerpt from this metamodel which reflects the main JADE entities for agent behaviours and messages. As its name denotes, *Agent* is any JADE entity which will be programmed as Java class for platform agents. In addition to one shot behaviours, the *Behaviour*s of agents can be in many types such as *CompositeBehaviour* (a series of behaviours bounded each other with input/output chains), *ParallelBehaviour* (hence concurrent actions of the agent can be modeled) and *FSMBehaviour* (tasks and actions of the agent can be modeled as a finite-state machine (FSM)). Each message between the platform agents can be modeled as *ACLMessage* instances which includes the information on the performative (e.g. INFORM, QUERY, PROPOSE), sender agent, receiver agent, applied conversation protocol, content language, used ontology, etc. All modeled agents, their behaviours and other related entities are covered in a *MASmodel* PIMM entity.

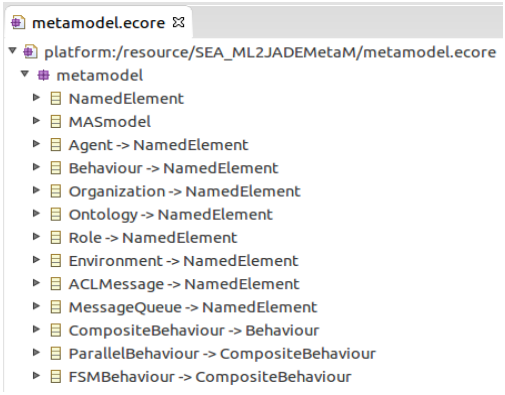


Fig. 3. An excerpt from the Ecore representation of the derived JADE metamodel

Following the derivation of the JADE metamodel as the PSMM in our study, we needed to construct the model transformation rules between SEA_ML PIMM and JADE PSMM. Similar to the method given in Sect. 4, entity mappings between these two metamodels are provided. Table 2 lists some of these entity mappings.

Table 2. Entity mappings between the metamodels of SEA_ML and JADE.

SEA_ML MM Entity	JADE MM Entity
SemanticWebAgent	Agent
SSMatchMakerAgent	DFAgent
SS_AgrementPlan	FSMBehaviour
SS_ExecutorPlan	ParallelBehaviour
SS_FinderPlan	OneShotBehaviour
SS_RegisterPlan	CyclicBehaviour
Role	SimpleBehaviour
OntologyMediatorRole	CompositeBehaviour
RegistrationRole	AgentManagementSystem (AMS)
RoleOntology	Ontology
OrganizationOntology	Ontology
ServiceOntology	Ontology
SemanticWebOrganization	Agent Platform
Environment	Environment

As seen from the table, SEA_ML SemanticWebAgent is mapped to JADE Agent as expected. SSMatchmakerAgent is mapped to DFAgent. SSMatchMaker agent is responsible for finding appropriate services which is needed by other agents in a MAS system, similar to DFAgent (Directory Facilitator Agent) on the JADE platform. Also SEA_ML SS_AgreementPlan, SS_ExecutorPlan, SS_FinderPlan and SS_RegisterPlan are mapped to FSMBehaviour, ParallelBehaviour, OneShotBehaviour and CyclicBehaviour respectively based on the characteristics of these plans on semantic service discovery, engagement and execution features of agents in SEA_ML and similarity of tasks defined in the corresponding JADE entities.

Moreover, Agents have roles and use ontologies to accomplish their duties in SEA_ML. A SEA_ML Role can be simply described as a JADE SimpleBehaviour while a specialization of role, SEA_ML OntologyMediatorRole is mapped to JADE CompositeBehaviour since it handles the ontologies with a more complex duty. SEA_ML RegistrationRole, which registers agents and their services, is mapped to JADE AgentManagementSystem (AMS) entity since AMS is responsible for managing a JADE agent platform with including the determination of agent statuses and registering/deregistering of agents. SEA_ML RoleOntology, ServiceOntology and OrganizationOntology are mapped to JADE Ontology entity as expected. In both sides, Ontology holds necessary information as a knowledgebase for the environment. SEA_ML Environment holds the non-agent resources for the agents and this can be mapped to a configuration file on JADE side, where access configurations for external sources are stated. Finally, SEA_ML SemanticWebOrganization entity models a MAS with including all SWAs, their goals and plans, so it matches well with JADE Agent Platform entity which possesses the similar features for platform agents living together.

```
01 rule SWSInteractionVP2MASmodel{
02   from
03     swsinteractionvp: SWSInteraction!SWSInteractionViewpoint
04   to
05     jademm: metamodel!MASmodel(
06       hasAgent<- Set{SWSInteraction!SemanticWebAgent.allInstances()},
07       hasDFAgent<- Set{SWSInteraction!SSMatchmakerAgent.allInstances()},
08       hasFSMBehaviour <- Set{SWSInteraction!SS_AgreementPlan.allInstances()},
09       hasParallelBehaviour <- Set{SWSInteraction!SS_ExecutorPlan.allInstances()},
10       hasOneShotBehaviour <- Set{SWSInteraction!SS_FinderPlan.allInstances()},
11       hasCyclicBehaviour <- Set{SWSInteraction!SS_RegisterPlan.allInstances()},
12       hasSimpleBehaviour <- Set{Ontology!Role.allInstances()},
13       hasCompositeBehaviour<-Set{Ontology!OntologyMediatorRole.allInstances()},
14       hasOntology <- Set{Ontology!RoleOntology.allInstances()},
15       hasOntology <- Set{Ontology!OrganizationOntology.allInstances()},
16       hasOntology <- Set{Ontology!ServiceOntology.allInstances()},
17       hasAgentPlatform <-Set{MASandOrg!SemanticWebOrganization.allInstances()},
18       hasEnvironment <- Set{MASandOrg!Environment.allInstances()},
19       hasAgentManagementSystem<-Set{SWSInteraction!RegistrationRole.allInstances()}
20     )
21 }
```

Listing 2. An excerpt from the SWSInteractionVP2MASmodel rule

After determination of the entity mappings, M2M transformation rules according to these mappings are written in ATL. This time the source metamodel which is used by the ATL engine will be SEA_ML metamodel while the target metamodel is JADE metamodel. Rules are executed on SEA_ML model instances to generate PSMs conforming to JADE specifications. In the following, some examples from the prepared model transformation rules are given. The first example is an excerpt which demonstrates a union rule (see Listing 2). Since SEA_ML is designed with multiple viewpoints, all necessary elements are united under a MAS viewpoint on target JADE

model. The rule transforms all SemanticWebAgents instances encountered in a SEA_ML model into JADE Agents with including all its plans and ontologies.

An excerpt from the ATL rule which provides the transformation of SEA_ML agent-SWS agreement plan into a JADE FSM behavior is shown in Listing 3. Following this rule, a helper rule which is used by this rule is also given in Listing 4. setAgreementPlanName helper rule is called by the AgreementPlan2FSMBehaviour ATL rule to control "name" attribute of the SEA_ML SSAgreementPlan instance and set the default name for this attribute in case of it is not specified in the source model.

```
01 rule AgreementPlan2FSMBehaviour {
02   from
03     Agreementplan: SWSInteraction!SS_AgreementPlan
04   to
05     jBehaviour: metamodel!FSMBehaviour (
06       name <- Agreementplan.setAgreementPlanName()
07     )
08 }
```

Listing 3. An excerpt from the AgreementPlan2FSMBehaviour

```
01 helper context SWSInteraction!SS_AgreementPlan def:
       setAgreementPlanName():  String =
02   if (self.name = thisModule.controlString or self.name.oclIsUndefined() ) then
03     'SS_AGREEMENT_PLAN_NAME_IS_EMPTY'
04   else
05     self.name
06   endif;
```

Listing 4. An excerpt from the setAgreementPlanName helper rule

5.2. M2T Transformations for code generation from JADE PSMs

In the interoperability approach we followed in Sect. 4, we did not need to worry about constructing the way of producing executables of the SEA_ML model instances on JADE platform since we benefited from ready-to-use code generation features already provided inside the DSML4MAS environment. However, this time, we should design and implement all M2T transformations for JADE model instances to generate artifacts executable inside the JADE platform. For this purpose, we prepared a series of M2T transformations by using Xpand language [61] which enables code generation from EMF models. Since the metamodel we derived for JADE is already encoded in Ecore (see Sect. 5.1), we can apply our Xpand rules on model instances conforming this metamodel to generate JADE Java classes. That completes the MDA we designed for the execution of SEA_ML models on JADE platform: SEA_ML instance models can be transformed into their JADE counterparts by executing the ATL rules discussed in Sect. 5.1 and then output of this transformation, which is a JADE model instance, can be processed with the prepared Xpand M2T rules to generate Java classes for this JADE

model instance. Following Xpand snippets give some flavor of the implemented M2T rules. As seen in Listing 5, the reference model is imported first. Here, it is indicated which element from the JADE instance model is going to be used to create which target element in the text part. For example, if instance model has *hasAgent* element than it is going to be defined by counter element Agent Java class. In Listing 6, an excerpt from the template for definition of Java class for each Jade Agent element is given. Attribute values of the corresponding class are set and some parts of the methods are generated as the result of executing the related Xpand rule on the proper JADE model instance.

```
«IMPORT metamodel»
«DEFINE main FOR MASmodel»
«EXPAND agent FOREACH hasAgent»
«EXPAND dfagent FOREACH hasDFAgent»
«EXPAND parallelbehaviour FOREACH hasParallelBehaviour»
«EXPAND oneshotbehaviour FOREACH hasOneShotBehaviour»
«EXPAND cyclicbehaviour FOREACH hasCyclicBehaviour»
«EXPAND compositebehaviour FOREACH hasCompositeBehaviour»
«EXPAND ontology FOREACH hasOntology»
«EXPAND ams FOREACH hasAgentManagementSystem»
«EXPAND agentplatform FOREACH hasAgentPlatform»
«ENDDEFILE»
```

Listing 5. Xpand code snippet to parse a JADE model instance to set target elements in the text

```
«DEFINE agent FOR Agent»
«FILE name + ".java"»
  import jade.core.*;
  public class «name» extends Agent {
     /*constructor definiton*/
     public «name»(DataStore ds) {
        super ();
        this.ds=ds;
     }
     /*constructor definition*/
     public «name» () {
        super();
        this.ds=new DataStore ();
     }
     /* Agent initializations */
     protected void setup () {  }
}
```

Listing 6. An excerpt from the Java class template for Jade Agent elements

6. Case Study

In this section, the use of the proposed MAS DSML interoperability approach is demonstrated for the development of an agent-based stock exchange system which will be deployed on the JADE platform. The system-to-be-used is modeled in SEA_ML and transformed to a DSML4MAS instance by applying the method described in Sect. 4 in order to use the generation power of DSML4MAS language. In this way, the implementation of this system's agents on JADE (or JACK) platform can be possible by using the operational semantics of DSML4MAS which is already provided for the execution of agents and the generation of semantic web services (see Fig. 2). In the second part of the case study, we also exemplified the use of the direct transformations constructed between SEA_ML and JADE within the scope of applying the conventional approach (discussed in Sect. 5). Hence, it will be possible to evaluate and compare the new interoperability approach with the traditional model-driven agent development. In the first subsection, the general architecture of the agent-based stock exchange systems and their modeling with SEA_ML are briefly introduced. Following subsections discuss the development of the MAS with the interoperability between SEA_ML and DSML4MAS and direct transformations from SEA_ML to JADE respectively.

6.1. Agent-based Stock Exchange Systems

Stock trading is one of the key items in economy and estimating its behavior and taking the best decision in it are among the most challenging issues. Agents in a MAS can share a common goal or they can pursue their own interests. That nature of MASs exactly fits to the requirements of free market economy. Moreover, Stock Exchange Market has lots of services which are offered for Investors (Buyer or Seller), Brokers, and Stock Managers. These services can be represented with semantic web services to achieve more accurate service finding and service matching.

When considering the structure of the system, the semantic web agents work within a semantic web organization for Stock System including sub-organizations for Stock Users where the Investor and Broker agents reside, and the Stock Market where the system's internal agents, e.g. Trade Managers (SSMatchmaker agent instances) work. The Stock Market organization also has two sub-organizations, the Trading Floor and the Stock Information System. These organizations and sub-organizations have their own organizational roles. These organizations also need to access some resources in other environments. Therefore, they have interactions with the required environments to gain access permissions. For example, agents in the Stock Market sub-organization need to access bank accounts and some security features, so that they can interact with the Banking & Security environment. All of the user agents including Investors and Brokers cooperate with Trade Manager to access the Stock Market. Also, the user agents interact with each other. For instance, Investor Agents can cooperate with Brokers to exchange stock for which Brokers are expert. More information on developing such stock trading agents can be found in [62].

To model the system in SEA_ML, Agent-SWS Interaction viewpoint is considered as the representative for SEA_ML viewpoints. This viewpoint is the most important aspect

of MASs working in semantic web environments. Fig. 4 shows a screenshot from the SEA_ML's modeling environment in which instances of both the semantic services and the agent plans required for the stock exchange are modeled, including their relations according to Agent-SWS interaction viewpoint of SEA_ML. Investor and Broker agents can be modeled with appropriate plan instances in order to find, make the agreement with and execute the services. The services can also be modeled for the interaction between the semantic web service's internal components (such as Process, Grounding, and Interface), and the SWA's plans. It is important to indicate that the stock exchange system given in here was already modeled in the SEA_ML environment before this study and instead of re-modeling the whole system (e.g. in DSML4MAS), the existing model is intentionally adopted in here to examine the applicability of the proposed approach. In fact, the model in question is much more complicated and we can only consider the agent-SWS interaction aspect due to the limits and scope of this paper. Discussion on the whole model can be found in [23] and the sources of the model pertaining to the case study are all available at the SEA_ML's distribution website [47].

We can see from the instance model given in Fig. 4 that an investor agent (e.g. InvestorA) plays the Buying role and applies its StockFinder plan for finding an appropriate Trading service interface of one TradingService SWS in order to buy some stocks. This plan enables the discovery by interacting with the TradeManager SSMatchmakerAgent which registers the services by applying the StockRecorder plan. InvestorA cooperates with Broker1 in order to receive some expert advice for its investment. At the next step, the Broker1 agent applies its StockBargaining plan for negotiating with the already discovered services. This negotiation is made through the Trade interface of the SWS. Finally, if the result of the negotiation is positive, the agent applies the StockOrder plan to call the TradingFloor of the SWS by executing its Exchange process and using its TradeAccess grounding with which the service is realized. In a similar way, Investor agents can cooperate with Brokers and interact with the TradeManager in order to collect some information about the market, e.g. the rate of exchange for a currency or the fluctuation rate for a specific stock.

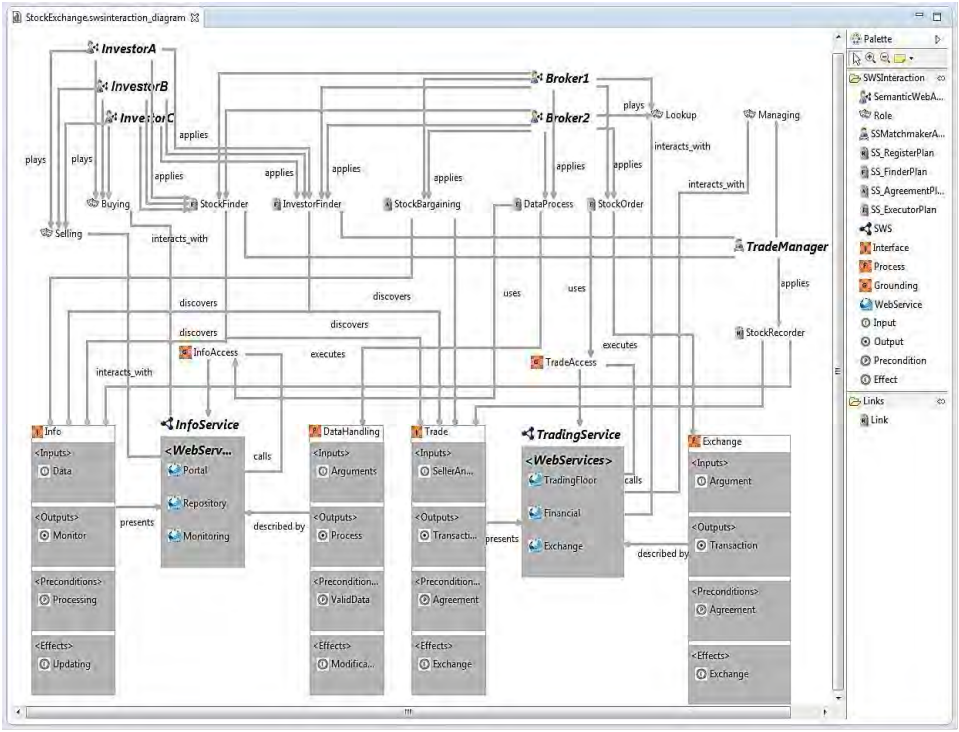


Fig. 4. Instance model of the multi-agent stock exchange system in SEA_ML with including the agents, semantic web services and their relations

6.2. Use of the interoperability between SEA_ML to DSML4MAS

The designed instance MAS model described in the previous subsection is controlled based on the provided constraint rules in SEA_ML tool to check its validity. Now, we can benefit from the interoperability provided between SEA_ML and DSML4MAS to enable the modeled MAS executable on the JADE platform. The horizontal model transformations discussed in Sect. 4 are executed on this SEA_ML instance model and as result; we succeed to automatically achieve the counterpart models conforming to DSML4MAS. To realize the transformation, the SEA_ML metamodel, the SEA_ML instance models for this case study, and the DSML4MAS metamodel are given to the ATL engine as input and the instance models of the case study in DSML4MAS are generated by the engine with executing our transformation rules.

The generated model conforms to the specification of DSML4MAS's abstract syntax, so it can be handled with DSML4MAS's graphical editor [49]. To visualize the instance model in DSML4MAS, the only thing needed is to add the related graphical notations to the generated instance model. The screenshot given in Fig. 5 shows the appearance of the output instance model in the concrete syntax of DSML4MAS. We can examine from the figure that the agents and their relations we modeled in SEA_ML are exactly

reflected to a DSML4MAS model after execution of the M2M transformations proposed in this study. From now on, it is straightforward to automatically achieve platform-specific executables and documents of this MAS model for JADE or JACK agent platforms since DSML4MAS already owns a chain of M2M and M2T transformations for these agent execution platforms and service ontologies as discussed in Sect. 3.2.

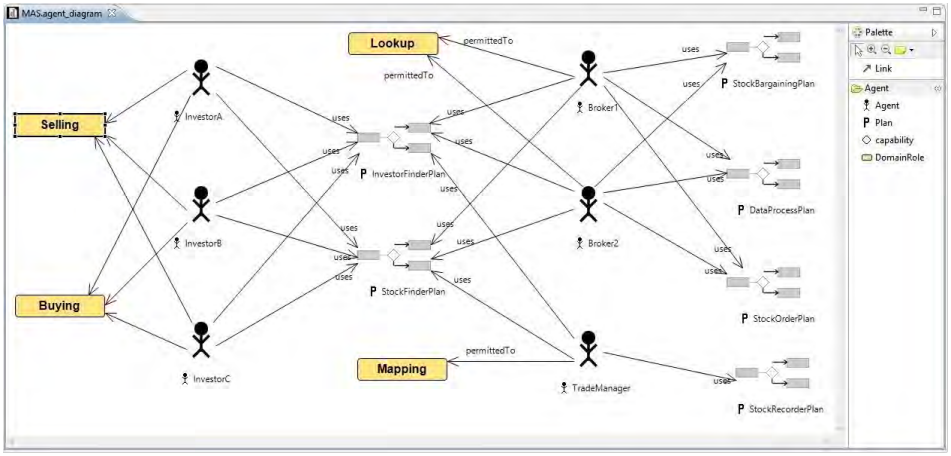


Fig. 5. Partial instance model of the agent-based stock exchange system in DSML4MAS achieved after application of the defined M2M transformations

6.3. Use of direct transformations between SEA_ML and JADE

The second way of implementing the same modeled agent-based stock exchange system on the JADE platform is to employ the vertical M2M and M2T transformations introduced in Sect. 5. At first, we need to automatically produce the corresponding JADE model of the same MAS currently modeled in SEA_ML. For this purpose, Ecore representations of SEA_ML, JADE and the instance MAS model of the stock exchange system are all given to the ATL engine and M2M transformation rules are executed on the SEA_ML instance model to achieve the counterpart instance model conforming to the JADE metamodel. Hence, the JADE model of all investor and broker agents in the stock exchange system with including their knowledgebases and other attributes can be automatically produced. An excerpt from the output XMI model is given in Listing 7.

The output of the above M2M transformation will be the input of the next vertical transformation which is a M2T transformation providing the automatic code generation for the JADE platform. As discussed in Sect. 5.2, M2T rules, we defined with using Xpand, can be executed on a JADE instance model to generate executable artifacts. When we apply those transformations on the JADE model of our stock exchange system, template codes for the Java classes for each agent in the system are automatically generated by parsing the instance model, determining each JADE model entity instance and producing Java codes described in the appropriate Xpand rule. For example, Listing

8 includes a code snippet from the Java class generated for the investor agent, called Investor A.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<metamodel:MASmodel xmi:version="2.0"
          xmlns:xmi="http://www.omg.org/XMI"
   xmlns:metamodel="http://JADEmetamodel">
      <hasAgent name="InvestorC"/>
      <hasAgent name="Broker2"/>
  <hasAgent name="InvestorB"/>
  <hasAgent name="InvestorA"/>
  <hasAgent name="Broker1"/>
  <hasOntology name="SellingAndBuyingRolesOnto"/>
  <hasOntology name="StockUserOrgOnt"/>
  <hasOntology name="SearchServiceOnt"/>
   …
```

Listing 7. An excerpt from the target JADE instance model

```
01 import jade.core.*;
02 public class InvestorA extends
          Agent {
03    /*constructor definiton*/
04    public InvestorA(DataStore ds){
05       super ();
06        this.ds=ds;
07    }
08    /*constructor definition*/
09    public InvestorA(){
10       super ();
11       this.ds=new DataStore();
12    }
13    /* Agent initializations */
14    protected void setup () { }
15 }
```

Listing 8. A snippet from the template JADE code generated for a system agent

7. Evaluation

An evaluation of the proposed MAS DSML interoperability approach was performed in this study by taking into account the language developers' perspective. Although the discussion given in the previous sections shows the applicability and effectiveness of the interoperability approach in the way of extending the execution platform support of

MAS DSMLs, we believe that some kind of comparative evaluation may help clarifying the feasibility of choosing the interoperability approach instead of the conventional one, i.e. design and implementation of separate M2M and M2T transformations for each new agent execution platform. For this purpose, we adopted the evaluation framework proposed in [35] which provides the systematic assessment of both the language constructs and the use of agent DSMLs according to various dimensions and criteria. To the best of our knowledge, the work in [35] presents the current unique evaluation framework which is specific to the MAS DSMLs and guides the assessment of model-driven agent development methodologies in general. However, since the scope of our evaluation in this study is limited mainly with model transformations in different abstractions levels and does not cover the evaluation of a full-fledged MAS DSML, only the dimensions called *model transformations* and *development* and the evaluation criteria pertaining to these dimensions called *M2M*, *M2T*, *Overall (output) performance ratio* and *development time* defined in [35] are taken into consideration. Furthermore, we revisited these dimensions and criteria in order to make them more meaningful and appropriate for our quantitative evaluation; and we separated our evaluation into two parts in which the related dimensions and metrics are included, namely Time Evaluation and Development Effort Evaluation.

A group of four software developers was employed during this evaluation. All of the evaluators were graduate students in computer science: one of them was a PhD candidate while remaining three evaluators were M.Sc. students. All group members had experience on software modeling and development of agent systems ranging from 2 to 4 years. In addition, two of the evaluators were also working as software engineer in industry for 2 years on average at the time of this study was realized. All group members passed related graduate courses in their master or PhD program, including Agent-oriented Software Development, Multi-agent Systems and Model-driven Software Engineering which are taught in Computer Engineering Department and International Computer Institute of Ege University, Turkey. All evaluators were familiar with Eclipse environment and skilled on Java programming language. They also had knowledge and practical experience on using MDD technologies like ATL, MOFScript, Xpand, Acceleo earned from above listed courses and previous projects.

The evaluation was performed both for 1) the interoperability provided between SEA_ML and DSML4MAS via horizontal model transformations and 2) definition and implementation of vertical M2M and M2T transformations directly between SEA_ML and JADE. In the remaining of the discussion, the former approach is shortly referred as the interoperability approach while the latter is referred as the conventional approach. It is worth stating that the evaluators worked individually for the application of these two approaches and both elapsed times and development throughputs (e.g. number of written M2M rules) were recorded for each evaluator and for each approach separately. Average times and throughputs were calculated for each phase of the development required during the application of interoperability or conventional approach by considering all evaluators' development processes. These average results gained from abovementioned time and development effort evaluations are reported in Sect. 7.1 and 7.2 respectively. Discussion on these results of the conducted evaluation is given in Sect. 7.3.

7.1. Time Evaluation

Time evaluation consists of measuring, analyzing and comparing the time elapsed for the design and the implementation of transformations required for each approach.

As it is discussed in Sect. 4, building the interoperability between two MAS DSMLs, SEA_ML and DSML4MAS includes the determination of entity mappings between two DSMLs' metamodels (which are in the same abstraction level) and writing the horizontal model transformations according to these mappings. Hence, the horizontal transformations between SEA_ML PIMM and DSML4MAS PIMM are realized by using ATL transformation language in four steps:

1. *Analyzing the source PIMM, SEA_ML metamodel*: This analysis is performed by the developer of the transformations to comprehend and infer on the source metamodel by considering the MAS features and elements.
2. *Analyzing the target PIMM, DSML4MAS metamodel*: Similar to step 1, the language developer also needs to analyze the metamodel of the target language to be able to determine main language entities and their relations.
3. *Mapping*: In this step, the language developer maps the meta elements in the source PIMM (SEA_ML) with the meta elements of the target PIMM (DSML4MAS) in a way that the semantics of mapped elements are representing similar concepts and associations in the domain (see Table 1). Mappings can be in m:n manner.
4. *Implementation*: Based on the defined entity mappings, M2M transformation rules and supporting helper rules are written by using ATL. Hence source models conforming to SEA_ML PIMM can be transformed into target DSML4MAS instance models by executing these transformations on ATL engine. This step also contains the test procedure of all written rules.

Average times spent by the evaluators for each of the abovementioned steps are calculated for time evaluation of the interoperability approach as shown in Table 3.

Table 3. Cost of building horizontal transformations for the interoperability approach

Step	Analysis for source PIMM (SEA_ML)	Analysis for target PIMM (DSML4MAS)	Determination of entity mappings	Implementation of M2M transformations	Total
Average Elapsed Time (in hours)	8	7	2	4	21

On the other hand, as discussed in Sect. 5, each developer (evaluator) needs to design and implement two types of vertical transformations for extending the execution support of SEA_ML on JADE platform in case of following the conventional approach:

A) PIM to PSM transformation between SEA_ML PIMM and JADE PSMM.

B) PSM to Code transformation for code generation from instance MAS models conforming to JADE PSMM.

The steps of preparing the vertical M2M transformations for type A are similar to the steps of providing horizontal transformations of the interoperability approach:

A) PIM to PSM transformation between SEA_ML PIMM and JADE PSMM

1. *Analyzing the source PIMM, SEA_ML metamodel*: This analysis is performed by the developer of the transformations to comprehend and infer on the source metamodel by considering the MAS features and elements.
2. *Analyzing JADE platform and derivation of JADE metamodel as a PSMM*: Unlike SEA_ML and DSML4MAS, the JADE MM needs to be prepared from scratch.
3. *Mapping*: In this step, the language developer maps the meta elements of the source PIMM (SEA_ML) with the meta elements of the target PSMM (JADE) in a way that the semantics of mapped elements are representing similar concepts and associations in the domain (see Table 2). Mappings can be in m:n manner.
4. *Implementation*: Based on the defined entity mappings, M2M transformation rules and some supporting helper rules are written by using ATL. Hence source models conforming to SEA_ML PIMM can be transformed into target JADE instance models by executing these transformations on ATL engine. This step also contains the test procedure of all written rules.

Average times spent by the evaluators for each of the abovementioned steps are shown in Table 4.

Table 4. Cost of building vertical transformations between SEA_ML PIMM and JADE PSMM

Step	Analysis for source PIMM (SEA_ML)	Derivation of target PSMM (JADE)	Determination of entity mappings	Implementation of M2M transformations	Total
Average Elapsed Time (hours)	8	16	3	6	33

For the conventional approach, the evaluators should also provide the M2T transformations for code generation from JADE instance models. Followings are the steps required for this transformation.

B) PSM to Code transformation for code generation from instance MAS models conforming to JADE PSMM

1. Analyzing JADE API for required Java class structures
2. Design of code templates for JADE PSMM meta-entities
3. *Implementation*: A series of M2T transformations are written by using Xpand (see Sect. 5.2).

Average times spent for each of the abovementioned steps are shown in Table 5.

Table 5. Cost of building vertical transformations for code generation from instance MAS models conforming to JADE PSMM

Step	Analysis for the agent platform API	Design of code templates	Implementation of M2T transformations	Total
Average Elapsed Time (in hours)	2	3	5	10

The figures presented in the abovementioned tables will be used in Sect. 7.3 to compare the interoperability approach with the conventional one.

7.2. Development Effort Evaluation

In this part, the development required both for the interoperability approach and the conventional approach is evaluated by comparing the number of rules, helper rules and templates as the main building blocks of the transformations including PIM to PIM, PIM to PSM and PSM to Code.

In the interoperability approach, each evaluator only needed to write horizontal M2M transformation rules in ATL which are required for the transformation between SEA_ML and DSML4MAS. The related figures for average numbers of rules and helper rules and average total number of line of codes (LoC) pertaining to these transformation rules are given in Table 6.

Table 6. Specification of the horizontal M2M transformations for the interoperability approach

Item	M2M Rules	M2M Helper Rules	Total LoC
Average Quantity	8	11	200

In the conventional approach, each evaluator needed to write vertical M2M transformation rules in ATL which are required for the transformation between SEA_ML and JADE. Moreover, templates for the generation of codes from JADE instance models were also needed to be written in Xpand. The related figures showing average quantities of rules, helper rules, templates and LoC are shown in Table 7.

Table 7. Specification of the vertical M2M and M2T transformations for conventional approach

Phase	PIM to PSM (SEA_ML to JADE)			Code Generation	
Item	M2M Rules	M2M Helper Rules	Total LoC	Templates	Total LoC
Average Quantity	15	17	240	12	316

The analysis and comparison of these figures are discussed in the next section.

7.3. Discussion

In this section, average results gained from time and development effort evaluations (Sect. 7.1 and 7.2 respectively) are discussed. To ease analysis, time evaluation results are shown in a bar chart (see Fig. 6) and development effort evaluation results are shown in another bar chart (see Fig. 7).

Based on the result figures given in Fig. 6, the followings can be deduced:

- Average times elapsed for the analysis of source PIMMs in both approaches are equal. This is expected since this step consists of the efforts for analyzing the same PIMM (SEA_ML MM).
- Analyzing the target PIMM (DSML4MAS) took a bit less time than the first step in the interoperability approach since abstract syntax of SEA_ML is more complicated with including a detailed Agent-SWS interaction viewpoint comparing with DSML4MAS. However, analyzing the target PSMM (JADE MM) in the conventional approach took more than two times (16 hours) comparing with the corresponding PIMM analysis in the interoperability approach. Main reason of this extra cost encountered in the conventional approach is each evaluator's need for examining the whole JADE platform first and then derive its metamodel in Ecore format to be used during the model transformations whereas each evaluator just needed to analyze an already available PIMM (metamodel of DSML4MAS) in the interoperability approach. Another reason of this cost in the conventional approach is the necessity to work in different abstraction levels according to MDA while working in the platform independent level is sufficient in the proposed interoperability approach.
- Mapping and Implementation steps of the interoperability approach took also less time than the corresponding steps in the conventional way on average. Average time elapsed for the MAS entity mappings between SEA_ML and DSML4MAS in the interoperability approach (2 hours) is less than the average time needed for setting the mappings between SEA_ML and JADE (3 hours). That difference is also another result of working in different abstraction levels in the conventional approach. Since they are in the same abstraction level, concepts and relations defined in SEA_ML and DSML4MAS are closer to each other and it is relatively easy for evaluators to set mapping between these concepts. However, each evaluator should deal with setting mappings between SEA_ML and JADE which are in different abstraction levels.

Moreover, there is an additional cost of the conventional approach comparing with the interoperability approach: allocating time for building vertical transformations for code generation from instance MAS models conforming to JADE PSMM. That process includes the analysis for JADE API, design of code templates and the implementation of M2T rules.

Considering the average total cost of vertical transformations constructed in the conventional approach, about 76.7% of the cost comes from PIMM to PSMM

transformations (33 hours) and the rest, about 23.3% comes from PSMM to Code transformations (10 hours) which is not required in the interoperability approach.

When we compare the time cost of developing the horizontal transformations in the interoperability approach with the vertical ones in the conventional approach to provide the platform extensibility of a MAS DSML, we can see that the average grand total of time needed for the interoperability (21 hours) is approximately half (about 48.8%) of the average grand total time needed for the conventional approach (43 hours). That is because the proposed interoperability approach benefits from the already provided M2T transformations and only needs the construction of M2M transformations between two MAS DSMLs (in our case, SEA_ML and DSML4MAS) while in the conventional approach, it is required to prepare both 1) M2M transformations between a MAS PIMM and agent platform PSMM (in our case SEA_ML and JADE) and 2) M2T transformations for code generation.

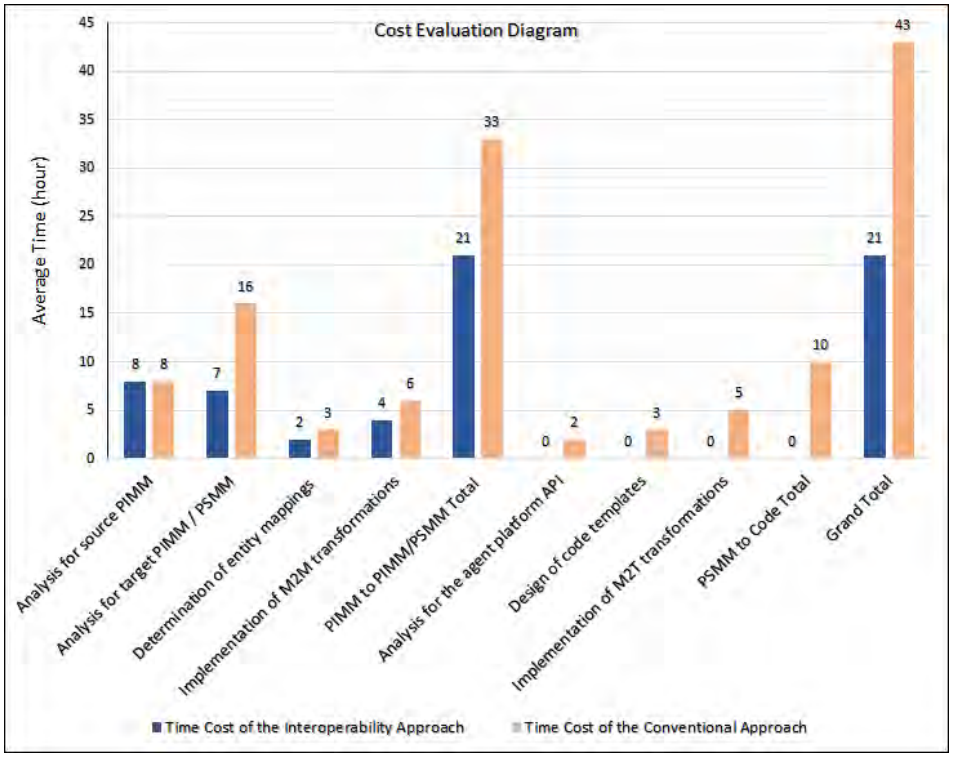


Fig. 6. Demonstration of average time evaluation results for both the interoperability and the conventional approaches

When we consider comparing only the M2M transformations required for both approaches instead of comparing the grand total efforts, we can see that horizontal M2M transformations in the interoperability needs 21 hours on average which is about 36.4% less than the time required for providing the vertical M2M transformations of the conventional approach (33 hours). Therefore, even in the case that the conventional approach has no need for M2T transformations to be developed, the interoperability

approach would be still advantageous. The reason of this difference is clear: working in the same abstraction level for MAS modeling (in the interoperability approach) takes less development time for M2M transformations comparing with the overhead of preparing M2M between agent models residing at the different abstraction levels (as in the case of the conventional approach).

Considering the evaluation of the development effort given for the application of each approach, the average cost figures for both the interoperability and the conventional approaches are shown inside a bar chart (Fig. 7).

According to the figures given in Fig. 7, the average number of M2M rules and helper rules for the interoperability approach are much less than those of the conventional approach. The reason is that the horizontal transformations written by the evaluators for the interoperability are only between the metamodels of two MAS DSMLs while the vertical transformations of the conventional approach consist of both PIM to PSM and PSM to Code transformations.

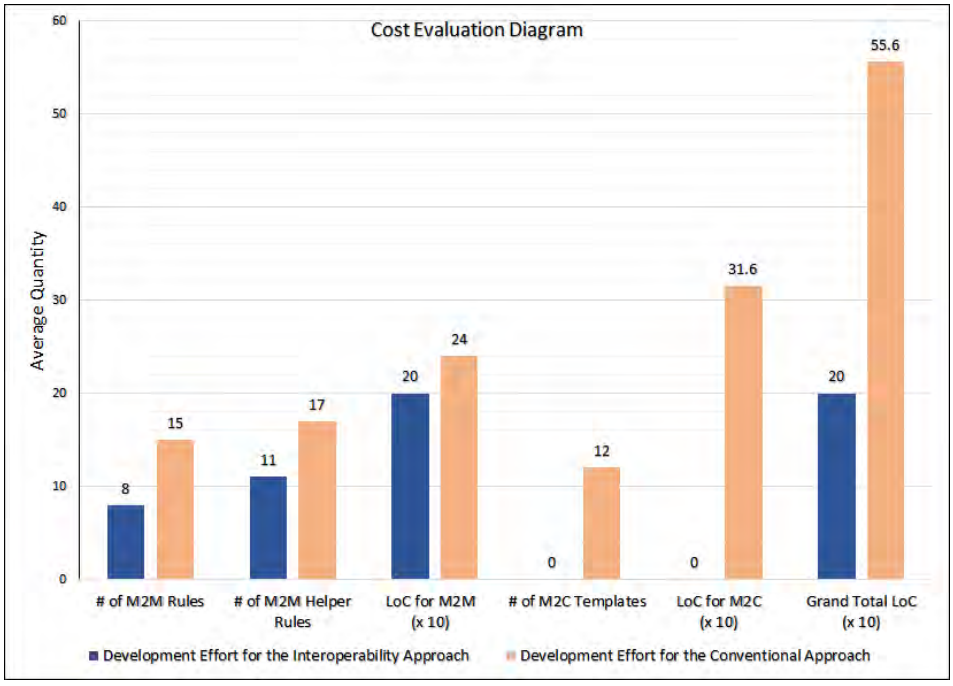


Fig. 7. Demonstration of the development effort evaluation results on average for both the interoperability and the conventional approaches

As we discussed earlier in this section, the conventional approach has an additional transformation phase of M2T. Comparing the total LoC developed for both approaches, we can see that the application of interoperability approach with 200 LoC on average needs about 64% less coding effort for transformations against the conventional approach where 556 LoC are written on average. The main reason is that M2T transformation for code generation is very close to the target agent platform, so it has lots of domain details that need to be coded in the templates. However, the

interoperability approach does not need this part of transformation since it is already available for the PIMM of the target MAS DSML.

Taking into consideration the threats to validity of this study, the first threat can be the number of the evaluators which may have an effect upon the generalization of the findings. However, software developers especially with practical experience on the application of both agents and MDD comparing with most of the other domains are very rare. In addition, the evaluators need to have knowledge and experience of languages, tools and frameworks for constructing transformation rules both for the interoperability and the conventional approaches. Although these limitations caused us to have a relatively small size of evaluators, we think that our sample was still sufficient for measuring the variability in such a dedicated field. Moreover, the conducted study aimed at performing evaluation according to the perspective of MAS DSML developers instead of MAS DSML users (agent developers) since the work herein mostly includes language implementation rather than language use. Within this context, the number of the evaluators employed in this study is again satisfactory since the number of MAS DSML developers constitutes a small amount inside the total number of MAS DSML users currently available.

In the execution phase of the evaluation, application of both approaches (interoperability and conventional) can be realized with a combination of single group/two groups and one problem domain/two different domains. In this study, we preferred to use a single group for the evaluation of both approaches instead of employing two groups with different evaluator profiles, in which, e.g. the first group experiences the application of the interoperability approach while the second group deals with applying the conventional approach. Choosing one single evaluator group may cause another threat to the validity. Using two different evaluator groups can be a good option where methods and implementation technologies required for each assessment case vary for the groups and comparison of these variations present the key point of the conducted evaluation. One such example of utilizing two separate groups for the assessment of a MAS DSML can be found in our previous work [35]. However, implementation methods and used language frameworks / technologies are so similar for the comparative evaluation of interoperability and conventional approaches in this study. For instance, the way of developing model transformations is almost same for both approaches in the conducted evaluation: evaluators implemented horizontal transformations in the interoperability approach while vertical transformations were constructed in the conventional approach with using the same Eclipse framework and ATL. Hence, we benefited from using a single group of evaluators having the same level of knowledge and experience in fair comparison of two approaches.

Finally, one may find the demonstration of applying the proposed horizontal approach with a single case study (discussed in Sect. 6) as an additional threat to validity since generation throughputs of employing both interoperability and conventional approaches probably differ in developing real agent systems for various business domains. However, the evaluation performed in this study mainly considers the language implementation and we investigate how an interoperability between MAS DSMLs facilitates the construction of both a model-driven MAS development process and its supporting tools for DSML developers. The evaluators in our work play a MAS DSML developer role more than a MAS DSML user role. Hence, we believe that size, complexity or type of the case study selected for exemplifying the use of the proposed

approach is not critical and does not directly affect the achieved results of evaluating interoperability and conventional approaches within the scope of implementing DSML-based development processes as given in this study.

8. Related Work

In the last decade, AOSE researchers have significant efforts on using model-driven approaches for agent development and the derivation of DSLs / DSMLs for MAS. For instance, Agent-DSL [16] was used to specify the agency properties that an agent needs to accomplish its tasks. However, the proposed DSL was presented only with its metamodel and provided just a visual modeling of the agent systems according to agent features, like knowledge, interaction, adaptation, autonomy and collaboration. Likewise, in [63], the authors introduced two dedicated modeling languages and called these languages as DSMLs. These languages were described by metamodels which can be seen as the representations of main concepts and relationships identified for each of the particular domains again introduced in [63]. The study included only the abstract syntaxes of the related DSMLs and did not give the concrete syntaxes or semantics. In fact, the study only defined generic agent metamodels for MDD of MASs. The work in [14] presented a methodology based on OMG's MDA [33] for modeling and implementing agent and service interactions on the Semantic Web. A PIMM for MAS and model transformations from instances of this PIMM to two different MAS deployment platforms were discussed in this study. But neither a DSML approach nor semantics of service execution was covered in the study.

As previously discussed in this paper, Hahn [17] introduced a DSML for MAS called DSML4MAS. The abstract syntax of the DSML was derived from a platform independent metamodel [9] which was structured into several aspects, each focusing on a specific viewpoint of a MAS. In order to provide a concrete syntax, the appropriate graphical notations for the concepts and relations were defined [49]. Furthermore, DSML4MAS supports the deployment of modeled MASs both in JACK and JADE agent platforms by providing an operational semantics over model transformations. Combination of these studies [9, 17, 49] are important because they provided the construction of probably the first complete DSML for agents with all of its specifications and guided MDD of agent applications. For instance, Ayala et al. [64] used DSML4MAS for the development of agent-based ambient intelligence systems. The metamodel of DSML4MAS was employed as a source metamodel to support the modeling of context aware systems and conforming models were transformed into target models which are instances of an aspect-oriented agent metamodel called Malaca. Code generation enabled the implementation of Malaca models to run in the ambient intelligence devices.

Another DSML was provided for MASs in [21]. The abstract syntax was presented using the Meta-object Facility (MOF) [65], the concrete syntax and its tool was provided with Eclipse Graphical Modeling Framework (GMF) [66], and finally the code generation for the JACK agent platform was realized with model transformations using Eclipse JET [67]. However, the developed modeling language was not generic since it was based on only the metamodel of one of the specific MAS methodologies called

Prometheus [68]. A similar study was performed in [18] which proposes a technique for the definition of agent-oriented engineering process models and can be used to define processes for creating both hardware and software agents. This study also offered a related MDD tool based on a specific MAS development methodology called INGENIAS [69].

Originating from a well-formalized syntax and semantics, Ciobanu and Juravle defined and implemented a language for mobile agents in [20]. They generated a text editor with auto-completion and error signaling features and presented a way of code generation for agent systems starting from their textual description. The work conducted in [24] aimed at creating a UML-based agent modeling language, called MAS-ML, which is able to model the well-known types of agent internal architectures, namely simple reflex agent, model-based agent, reflex agent, goal-based agent and utility-based agent. Representation and exemplification of all supported agent architectures in the concrete syntax of the introduced language were given. MAS-ML is also accompanied with a graphical tool which enables agent modeling. However, the current version of MAS-ML does not support any code generation for MAS frameworks which prevents the execution of the modeled agent systems.

Wautelet and Kolp [70] investigated how a model-driven framework can be constructed to develop agent-oriented software by proposing strategic, tactical and operational views. Within this context, they introduced a Strategic Services Model in which strategic agent services can be modeled and then transformed into the dependencies modeled according to the well-known i* early phase system modeling language [71] for a problem domain. In addition, generated i* dependencies can be converted to BDI agents to be executable on appropriate agent platforms such as JACK [30] and JADEX [32]. However, implementation of the required transformations and code generation were not included in this study. Another work for model-driven development of BDI agents [72] introduced a metamodel for the definition of entities and relations pertaining to Jason BDI architecture [73]. The work only consisted of a metamodel and a graphical concrete syntax for this metamodel. Generation of executable artifacts was not included in the study.

In a recent work [74], a metamodel, describing some modelling units and constraints, was introduced in order to identify the real time requirements of a MAS during the analysis phase of the development. Hence, the requirement analysis was supported with a model-driven approach to determine real-time tasks. Bergenti et al. [75] proposed a DSL, called JADEL, for the MDD of agents on JADE platform. Instead of covering all features of JADE, JADEL only provided high-level agent-oriented abstractions, namely agents, behaviours, communication ontologies, and interaction protocols. JADEL was supported with a compiler which enabled source code generation for implementing agents on JADE platform. However, the related code generation feature of JADEL is not functional enough to fully implement JADE agents as also indicated by the authors in [75].

Finally, by considering our previous studies, in [19] and [22], we showed the derivation of a DSL for the MDE of agent systems working on the Semantic Web. That initial version of the language was refined and enriched with a graphical concrete syntax in [23]. This new language, called SEA_ML, covered an enhanced version of agent-SWS interaction viewpoint in which modeling those interactions can be elaborated as much as possible for the exact implementation of agent's service discovery, agreement

and execution dynamics. We also presented the formal semantics of the language [46] and discussed how the applied methodology can pave the way of evolutionary language development for MAS DSLs [4]. Moreover, qualitative evaluation and quantitative analysis of SEA_ML have been recently performed over a multi-case study protocol [35].

The work presented in this paper contributes to the abovementioned MAS DSL/DSML studies by introducing the interoperability of the languages and hence the proposed MDE technique helps to facilitate the platform support of the MAS DSMLs comparing with the existing agent platform extensibility approaches which deal with the definition and the implementation of new M2M and M2T transformations for each execution platform. To the best of our knowledge, the work herein is the first effort on the interoperability of the MAS DSMLs and it is the first study in AOSE which employs horizontal model transformations to enable this interoperability. It is worth indicating that only the work conducted in [15] considers the application of horizontal transformations for agent domain apart from our proposal. However, that study just provides the transformation between the metamodels of two specific AOSE methodologies (Prometheus [68] and INGENIAS [69]) to realize MAS implementation on exactly one agent deployment platform and does not support MAS DSML interoperability or language extensibility on various agent platforms.

Taking into account the interoperability of software systems within the context of MDE, various noteworthy studies also exist for enabling these systems to work together [76]. For instance, an MDE platform was used in [77] both for representing various software bug tracing tools and executing transformations among their conceptual models to enhance the interoperability between these tools. Likewise, Sun et al. [78] benefited from MDE to address tool interoperability for supporting different data formats among similar tools. Kern [79] introduced an interoperability interface for the exchange of metamodels and models between MetaEdit+ and Eclipse EMF tools based on the mappings specified at meta-metamodel level. That bridging approach was extended to construct the interoperability between modeling tools such as ARIS, EMF, MetaEdit+ and Microsoft Visio in [80]. BPM-X-Change tool, introduced in [81], provided the interchange of data models and their visual diagrams between different enterprise management tools and repositories for the interoperability of them. In [82], a comprehensive analysis of modeling tools was performed by considering both the degree of supported interoperability and the variety of approaches for realizing the interoperability. Horizontal transformations, defined and implemented between MAS DSMLs in our study, directly support the interoperability of agent modeling tools owned by these DSMLs. Hence, the work herein can also be considered inside above software tool interoperability studies with emphasizing MDD of agent systems.

9. Conclusion

We presented an approach for extending the execution platform support of MAS DSMLs over language interoperability in this paper. The interoperability is provided by defining and implementing horizontal M2M transformations between the agent metamodels which constitute the syntaxes of MAS DSMLs. Extending the platform

support with applying the conventional way which is widely in-use for MAS DSMLs was also demonstrated in the paper to provide a comparison for the new interoperability approach. In comparison to the conventional approach, evaluation results showed that the interoperability approach requires both less development time and effort considering the quantity of transformation rules, code generation templates and total LoC required for all transformations. Due to being at the same abstraction level, both mapping the model entities and implementing the model transformations were more convenient and less laborious comparing with M2M and M2T transformation chain required in the way of enriching the support of DSMLs for various agent execution platforms in the conventional approach.

As the future work, we plan to extend the applicability of this interoperability approach for some other MAS DSMLs. For instance, modeling SEA_ML agents can be improved by constructing a similar interoperability with another MAS DSML, called MAS-ML [24]. MAS-ML owns a built-in modeling for agent architectures such as reflex agent, model-based agent, or utility-based agent which are not currently supported in SEA_ML. Hence, instead of constructing all required components for such agent architecture support in SEA_ML from scratch, an interoperability with MAS-ML can automatically improve the features of SEA_ML within this context.

Acknowledgment. This study is funded by the Scientific and Technological Research Council of Turkey (TUBITAK) Electric, Electronic and Informatics Research Group (EEEAG) under grant 115E591 and the Scientific Research Projects Directorate of Ege University under grant 16-UBE-001.

References

1. Wooldridge, M., Jennings, N. R.: Intelligent Agents - Theory and Practice. Knowledge Engineering Review, Vol. 10, No. 2, 115-152. (1995)
2. Badica, C., Budimac, Z., Burkhard, H. D., Ivanovic, M.: Software agents: Languages, tools, platforms. Computer Science and Information Systems, Vol. 8, No. 2, 255-298. (2011)
3. Paprzycki, M.: Editorial: Agent-oriented computing for distributed systems and networks. Journal of Network and Computer Applications, Vol. 37, 45–46. (2014)
4. Challenger, M., Mernik, M., Kardas, G., Kosar, T.: Declarative specifications for the development of multi-agent systems. Computer Standards & Interfaces, Vol. 43, 91-115. (2016)
5. Kardas, G.: Model-driven development of multiagent systems: a survey and evaluation. The Knowledge Engineering Review, Vol. 28, No. 4, 479-503. (2013)
6. Shehory, O., Sturm, A.: Agent-Oriented Software Engineering: Reflections on Architectures, Methodologies, Languages, and Frameworks. Springer-Verlag Berlin Heidelberg. (2014)
7. Bernon, C., Cossentino, M., Gleizes, M.-P., Turci, P., Zambonelli, F.: A study of some multi-agent meta-models. Agent-Oriented Software Engineering V, Lecture Notes in Computer Science, Vol. 3382, 62-77. (2005)
8. Omicini, A., Ricci, A., Viroli, M.: Artifacts in the A&A meta-model for multi-agent systems. Autonomous Agents and Multi-Agent Systems, Vol. 17, No. 3, 432-456. (2008)
9. Hahn, C., Madrigal-Mora, C., Fischer, K.: A Platform-Independent Metamodel for Multiagent Systems. Autonomous Agents and Multi-Agent Systems, Vol. 18, No. 2, 239-266. (2009)

10. Beydoun, G., Low, G., Henderson-Sellers, B., Mouratidis, H., Gomez-Sanz, J. J., Pavon, J., Gonzalez-Perez, C.: FAML: A Generic Metamodel for MAS Development. IEEE Transactions on Software Engineering, Vol. 35, No. 6, 841-863. (2009)
11. Garcia-Magarino, I.: Towards the integration of the agent-oriented modeling diversity with a powertype-based language. Computer Standards & Interfaces, Vol. 36, 941–952. (2014)
12. Bauer, B., Odell, J.: UML 2.0 and agents: how to build agent-based systems with the new UML standard. Engineering Applications of Artificial Intelligence, Vol. 18, No. 2, 141-157. (2005)
13. Pavon, J., Gomez-Sanz, J., Fuentes, R.: Model driven development of multi-agent systems. Lecture Notes in Computer. Science, Vol. 4066, 284–298. (2006)
14. Kardas, G., Goknil, A., Dikenelli, O., Topaloglu, N. Y.: Model driven development of semantic web enabled multi-agent systems. International Journal of Cooperative Information Systems, Vol. 18, No. 2, 261-308. (2009)
15. Gascuena, J. M., Navarro, E., Fernandez-Caballero, A., Martínez-Tomas, R.: Model-to-model and model-to-text: looking for the automation of VigilAgent. Expert Systems, Vol. 31, No. 3, 199-212. (2014)
16. Kulesza, U., Garcia, A., Lucena, C., Alencar, P.: A generative approach for multi-agent system development. Lecture Notes in Computer Science, Vol. 3390, 52–69. (2005)
17. Hahn, C.: A Domain Specific Modeling Language for Multiagent Systems. In Proceedings of 7th International Conference on Autonomous Agents and Multi-Agent Systems, Estoril, Portugal, 233-240. (2008)
18. Fuentes-Fernandez, R., Garcia-Magarino, L., Gomez-Rodriguez, A. M., Gonzalez-Moreno, J. C.: A technique for defining agent-oriented engineering processes with tool support. Engineering Applications of Artificial Intelligence, Vol. 23, No. 3, 432-444. (2010)
19. Demirkol, S., Challenger, M., Getir, S., Kosar, T., Kardas, G., Mernik, M.: SEA_L: A Domain-specific Language for Semantic Web enabled Multi-agent Systems. In Proceedings of the 2nd Workshop on Model Driven Approaches in System Development, held in conjunction with 2012 Federated Conference on Computer Science and Information Systems, IEEE Conference Publications, Wrocław, Poland, 1373-1380. (2012)
20. Ciobanu, G., Juravle, C.: Flexible Software Architecture and Language for Mobile Agents. Concurrency and Computation-Practice & Experience, Vol. 24, No. 6, 559-571. (2012)
21. Gascuena, J. M., Navarro, E., Fernandez-Caballero, A.: Model-Driven Engineering Techniques for the Development of Multi-agent Systems. Engineering Applications of Artificial Intelligence, Vol. 25, No. 1, 159-173. (2012)
22. Demirkol, S., Challenger, M., Getir, S., Kosar, T., Kardas, G., Mernik, M.: A DSL for the development of software agents working within a semantic web environment. Computer Science and Information Systems, Vol. 10, No. 4, 1525-1556. (2013)
23. Challenger, M, Demirkol, S., Getir, S., Mernik, M., Kardas, G., Kosar, T.: On the use of a domain-specific modeling language in the development of multiagent systems. Engineering Applications of Artificial Intelligence, vol. 28, 111-141. (2014)
24. Goncalves, E. J. T., Cortes, M. I., Campos, G. A. L., Lopes, Y. S., Freire, E. S. S., da Silva, V. T., de Oliveira, K. S. F., de Oliveira, M. A.: MAS-ML2.0: Supporting the modelling of multi-agent systems with different agent architectures. Journal of Systems and Software, Vol. 108, 77-109. (2015)
25. Bryant, B.R., Gray, J., Mernik, M., Clarke, P. J., France, R. B., Karsai, G.: Challenges and Directions in Formalizing the Semantics of Modeling Languages. Computer Science and Information Systems, Vol. 8, No. 2, 225-253. (2011)
26. Mernik, M., Heering, J., Sloane, A.: When and how to develop domain-specific languages. ACM Computing Surveys, Vol. 37, No. 4, 316-344. (2015)
27. Varanda Pereira, J. M., Mernik, M., da Cruz, D., Henriques, P. R.: Program Comprehension for Domain-specific Languages. Computer Science and Information Systems, Vol. 5, No. 2, 1-17. (2008)

28. Lukovic, I., Varanda Pereira, J. M., Oliveira, N., da Cruz, D., Henriques, P. R.: A DSL for PIM specifications: Design and attribute grammar based implementation, Computer Science and Information Systems, Vol. 8, No. 2, 379-403. (2011)
29. Selic, B.: The pragmatics of model-driven development. IEEE Software, Vol. 20, 19-25. (2003)
30. Howden, N., Rönnquist, R., Hodgson, A., Lucas, A.: JACK intelligent agents-summary of an agent infrastructure. In Proceedings of the 5th International Conference on Autonomous Agents, Montreal, Canada, 1-6. (2001)
31. Bellifemine, F. L., Caire, G., Greenwood, D.: Developing Multi-Agent Systems with JADE, Wiley & Sons. (2007)
32. Pokahr, A., Braubach, L., Walczak, A., Lamersdorf, W.: Jadex-engineering goal-oriented agents. Developing Multi-Agent Systems with JADE. Bellifemine et al.(Eds)., Wiley & Sons, 254-258. (2007)
33. Object Management Group (OMG). Model Driven Architecture (MDA) Specification. (2003). [Online]. Available: http://www.omg.org/mda/ (current June 2017)
34. Mens, T., Van Gorp, P.: A Taxonomy of Model Transformation. Electronic Notes in Theoretical Computer Science, Vol. 152, issue 27, 125-142. (2006)
35. Challenger, M., Kardas, G., Tekinerdogan, B.: A systematic approach to evaluating domain-specific modeling language environments for multi-agent systems. Software Quality Journal, vol. 24, no. 3, pp. 755-795. (2016)
36. Bircan, E., Challenger, M., Kardas, G.: Interoperability of MAS DSMLs via Horizontal Model Transformations. In Proceedings of the 4th Workshop on Model Driven Approaches in System Development, held in conjunction with 2016 Federated Conference on Computer Science and Information Systems, IEEE Conference Publications, Gdansk, Poland, 1555-1564. (2016)
37. Bellifemine, F., Poggi, A., Rimassa, G.: Developing multi-agent systems with a FIPA-compliant agent framework. Software-Practice & Experience, Vol. 31, No. 2, 103-128. (2001)
38. Agent Oriented Software (AOS). Agent Oriented Software Inc., JACK Intelligent Agents. (2001). [Online]. Available: http://www.aosgrp.com/products/jack/ (current June 2017).
39. Pokahr, A., Braubach, L., Lamersdorf, W.: Jadex: A BDI reasoning engine. Multi-agent programming, Springer, 149-174. (2005)
40. Djuric, D., Gasevic, D., Devedzic, V.: The Tao of Modeling Spaces. Journal of Object Technology, Vol. 5, No. 8, 125-147. (2006)
41. Shadbolt, N., Hall, W., Berners-Lee, T.: The Semantic Web revisited. IEEE Intelligent Systems, Vol. 21, No. 3, 96-101. (2006)
42. Sycara, K., Paolucci, M., Ankolekar, A., Srinivasan, N.: Automated discovery, interaction and composition of Semantic Web Services. Journal of Web Semantics, Vol. 1, No. 1, 27-46. (2003)
43. Challenger, M., Getir, S., Demirkol, S., Kardas, G.: A Domain Specific Metamodel for Semantic Web Enabled Multi-Agent Systems. Advanced Information Systems Engineering Workshops, Lecture Notes in Business Information Processing, Vol. 83, 177-186. (2011)
44. Rao, A. S., Georgeff, M. P.: BDI-agents: From Theory to Practice. In Proceedings of the 1st International Conference on Multiagent Systems, San Francisco, USA, 312-319. (1995)
45. World Wide Web Consortium (W3C). OWL-S: Semantic markup for web services. (2004). [Online]. Available: http://www.w3.org/Submission/OWL-S/ (current June 2017).
46. Getir, S., Challenger, M., Kardas, G.: The formal semantics of a domain-specific modeling language for semantic web enabled multi-agent systems. International Journal of Cooperative Information Systems, Vol. 23, No. 3, 1-53. (2014)
47. SEA_ML: A domain-specific modeling language for MAS. (2016). [Online]. Available: http://serlab.ube.ege.edu.tr/resources.html#SEA_ML (current June 2017)

48. Hahn, C., Nesbigall, S., Warwas, S., Zinnikus, I., Fischer, K., Klusch, M.: Integration of Multiagent Systems and Semantic Web Services on a Platform Independent Level. In Proceedings of 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, 200-206. (2008)
49. Warwas, S., Hahn, C.: The concrete syntax of the platform independent modeling language for multiagent systems. In Proceedings of the 2nd Workshop on Agent-based Technologies and Applications for Enterprise Interoperability, Estoril, Portugal, 94-105. (2008)
50. Web Service Modeling Ontology (WSMO). (2005). [Online]. Available: https://www.w3.org/Submission/WSMO/ (current June 2017)
51. Kopecky, J., Vitvar, T., Bournez, C., Farrell, J.: Sawsdl: Semantic annotations for wsdl and xml schema. IEEE Internet Computing, Vol. 11, No. 6, 60-67. (2007)
52. DSML4MAS Development Environment. (2008). [Online]. Available: https://sourceforge.net/projects/dsml4mas/ (current June 2017)
53. Gomez-Sanz, J. J., Fuentes-Fernandez, R.: Understanding Agent-Oriented Software Engineering methodologies. The Knowledge Engineering Review, Vol. 30, Issue 4, 375–393. (2015)
54. Milanovic, M., Gasevic, D., Giurca, A., Wagner, G., Lukichev, S., Devedzic, V.: Model Transformations to Bridge Concrete and Abstract Syntax of Web Rule Languages. Computer Science and Information Systems, Vol. 6, No. 2, 47-85. (2009)
55. Jouault, F., Allilaire, F., Bezivin, J., Kurtev, I.: ATL: A model transformation tool. Science of Computer Programming, Vol. 72, No. 1-2, 31-39. (2008)
56. Eclipse. ATL Model Transformation Language and Toolkit. (2007). [Online]. Available: http://www.eclipse.org/atl/ (current June 2017)
57. Frankel, D.: Model Driven Architecture: Applying MDA to Enterprise Computing: The Complete Book. Wiley Publishing, US. (2008)
58. JAVA Agent DEvelopment Framework (JADE). (2015). [Online]. Available: http://jade.tilab.com/ (current June 2017)
59. Foundation for Intelligent Physical Agents (FIPA). (2002b). "IEEE Foundation for Intelligent Physical Agents (FIPA), FIPA Standards." http://www.fipa.org (current June 2017)
60. Foundation for Intelligent Physical Agents (FIPA).: The Foundation for Intelligent Physical Agents, Agent Communication Language Message Structure Specification 00061. (2005). [Online]. Available: http://www.fipa.org/specs/fipa00061/ (current June 2017)
61. Xpand. (2016). [Online]. Available: http://wiki.eclipse.org/Xpand (current June 2017)
62. Kardas, G., Challenger, M., Yildirim, S., Yamuc. A.: Design and implementation of a multiagent stock trading system. Software: Practice and Experience, Vol. 42, No. 10, 1247-1273. (2012)
63. Rougemaille, S., Migeon, F., Maurel, C., Gleizes, M-P.: Model Driven Engineering for Designing Adaptive Multi-agent Systems, Lecture Notes in Artificial Intelligence, Vol. 4995, 318-333. (2007)
64. Ayala, I., Amor, M., Fuentes, L.: A model driven engineering process of platform neutral agents for ambient intelligence devices. Autonomous Agents and Multi-agent Systems, Vol. 28, 214-255. (2014)
65. Object Management Group (OMG). Meta Object Facility (MOF). (2002). [Online]. Available: http://www.omg.org/spec/MOF (current June 2017)
66. Eclipse. Graphical Modeling Framework (GMF). (2006). [Online]. Available: http://www.eclipse.org/modeling/gmp/ (current June 2017)
67. Eclipse. The Eclipse Modeling project, Model to Text (M2T) transformation, JET code generator (2007). [Online]. Available: http://www.eclipse.org/modeling/m2t/?project=jet#jet (current June 2017)
68. Padgham, L., Winikoff, M.: Prometheus: A practical agent-oriented methodology. Agent-oriented methodologies, Henderson-Sellers and Giorgini (Eds), 107-135. (2005)

69. Pavón, J., Gómez-Sanz, J. J., Fuentes, R.: The INGENIAS methodology and tools. Agent-oriented methodologies, Henderson-Sellers and Giorgini (Eds), Vol. 9, 236-276. (2005)
70. Wautelet, Y., Kolp, M.: Business and model-driven development of BDI multi-agent systems. Neurocomputing, Vol. 182, 304–321. (2016)
71. Yu, E., Giorgini, P., Maiden, N., Mylopoulos, J.: Social Modeling for Requirements Engineering: The Complete Book. MIT Press, Cambridge, Massachusetts. (2011)
72. Tezel, B. T., Challenger, M., Kardas, G.: A Metamodel for Jason BDI Agents. In Proceedings of the 5th Symposium on Languages, Applications and Technologies, Maribor, Slovenia, OpenAccess Series in Informatics, Vol. 51, 8:1-8:9. (2016)
73. Bordini, R. H., Hübner, J. F., Wooldridge, M.: Programming Multi-Agent Systems in AgentSpeak Using Jason, John Wiley & Sons. (2007)
74. Ashamalla, A., Beydoun, G., Low, G.: Model Driven Approach for Real-time Requirement Analysis of Multi-Agent Systems. Computer Languages, Systems & Structures, DOI: 10.1016/j.cl.2017.05.006. (2017)
75. Bergenti, F., Iotti, E., Monica, S., Poggi, A.: Agent-Oriented Model-Driven Development for JADE with the JADEL Programming Language. Computer Languages, Systems & Structures, DOI: 10.1016/j.cl.2017.06.001. (2017)
76. Chen, D.: Practices, principles and patterns for interoperability. Research report of INTEROP NoE, FP6 - Network of Excellence - Contract no: 508011, Deliverable 6.1 (2005)
77. Bezivin, J., Bruneliere, H., Jouault, F., Kurtev, I: Model Engineering Support for Tool Interoperability. In Proceedings of the 4th UML Workshop in Software Model Engineering, Montego Bay, Jamaica, 2-16. (2005)
78. Sun, Y., Demirezen, Z., Jouault, F., Tairas, R., Gray, J.: A Model Engineering Approach to Tool Interoperability. Lecture Notes in Computer Science, Vol. 5452, 178-187. (2009)
79. Kern, H.: The Interchange of (Meta)Models between MetaEdit+ and Eclipse EMF Using M3-Level-Based Bridges. In Proceedings of 8th OOPSLA Workshop on Domain-Specific Modeling, Birmingham, USA, 14-19. (2008)
80. Kern, H.: Model Interoperability between Meta-Modeling Environments by using M3-Level-Based Bridges. Ph.D Thesis, University of Leipzig, Leipzig (2016)
81. Kammermeier, F., Rautenberg, V., Scherer, H.-J.: Pattern-based model transformations: Software for the interoperability of enterprise management tools and model interchange. (2011). [Online]. Available: http://www.bpm-x.com/ (current June 2017)
82. Kern, H.: Study of Interoperability between Meta-Modeling Tools. In Proceedings of 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, 1629–1637 (2014)

Geylani Kardas received his B.Sc. in computer engineering and both M.Sc., and Ph.D. degrees in information technologies from Ege University in 2001, 2003 and 2008. He is currently an associate professor at Ege University, International Computer Institute (ICI) and the head of Software Engineering Research Laboratory (Ege-SERLab) at ICI. His research interests mainly include agent-oriented software engineering, model-driven software development, and domain-specific (modeling) languages. He has authored or co-authored over 60 peer-reviewed papers in these research areas. Dr. Kardas worked and is still working as the principle investigator, researcher or consultant in various R&D projects funded by governments, agencies and private corporations. He is a member of the ACM.

Emine Bircan received her B.Sc. in computer engineering from Izmir Institute of Technology in 2013 and M.Sc. in information technologies from Ege University, International Computer Institute (ICI) in 2017. During her graduate study, she is a student member of Software Engineering Research Laboratory (Ege-SERLab) at ICI. She is currently a researcher in the Scientific and Technological Research Council of Turkey (TUBITAK). Her interests include model-driven software engineering, multi-agent systems and recently data security & privacy.

Moharram Challenger received his B.Sc., and M.Sc. degrees in computer engineering from IAU-Shabestar and IAU-Arak Universities (Iran) in 2001 and 2005 respectively. He also received his Ph.D. in Information Technologies from Ege University (Turkey), in Feb 2016. After PhD, he has worked as an external Postdoc researcher in IT group, Wageningen University of Research (the Netherlands) about 1 year. Since Jan 2017, he is working as an assistant professor at International Computer Institute, Ege University. His research interests include domain-specific (modeling) languages, multi-agent systems, and Internet of Things.

Received: January 13, 2017; Accepted: June 24, 2017

Computer Science and Information Systems 14(3):913–937 DOI: https://doi.org/10.2298/CSIS170109035R

Towards OntoUML for Software Engineering: Transformation of Kinds and Subkinds into Relational Databases

Zdeněk Rybola and Robert Pergl

Faculty of Information Technology
Czech Technical University in Prague
Thákurova 9, 16000 Praha 6
{zdenek.rybola, robert.pergl}@fit.cvut.cz

Abstract. OntoUML is an ontologically well-founded conceptual modelling language that distinguishes various types of classifiers and relations providing precise meaning to the modelled entities. While Model-Driven Development is a well-established approach, OntoUML has been overlooked so far as a conceptual modelling language for the PIM of application data. This paper is an extension of the paper presented at MDASD 2016, where we outlined the transformation of Rigid Sortal Types – Kinds and Subkinds. In this paper, we discuss the details of various variants of the transformation of these types and the rigid generalization sets. The result of our effort is a complete method for preserving high-level ontological constraints during the transformations, specifically special multiplicities and generalization set meta-properties in a relational database using views, CHECK constraints and triggers.

Keywords: OntoUML, UML, transformation, relational database, Kind, Subkind, generalization set.

1. Introduction

Software engineering is a demanding discipline that deals with complex systems [8]. The goal of software engineering is to ensure high-quality software implementation of these complex systems. To achieve this, various software development approaches have been developed. One of these approaches is Model-Driven Development (MDD), which is based on elaborating models and transformations between them [20]. The most usual part of the MDD approach used in the practice is the process of *forward engineering*: transformations of more abstract models (e.g. a PIM (platform independent model)) into more specific ones (e.g. PSM (platform specific model) or ISM (implementation specific model)). The most common use-case of such a process is the development of conceptual data models and their transformation into source codes or database scripts.

To achieve a high-quality software system, high-quality expressive models are necessary to define the requirements for the system [8]. To successfully use them in the MDD approach, the models should define all requirements and all constraints of the system. Moreover, it should hold that more specific models persist the constraints defined in the more abstract models [10]. OntoUML seems to be a very suitable language for

modelling platform-independent data models. As it is based on Unified Foundational Ontology (UFO), it is domain-agnostic and it provides mechanisms to create ontologically well-founded conceptual models [10].

In this paper, we discuss introduction of OntoUML into MDD, specifically we formulate two research questions:

1. Are there benefits of using OntoUML Kinds and Subkinds in PIM?
2. Is it possible to preserve generalization sets constraints of Kinds and Subkinds in a relational database?

The first research question is motivated by success of OntoUML for making precise ontological models and thus exploring the possibility of incorporating OntoUML into the MDD as a primary language for conceptual data modelling. As relational databases are still the most popular type of data storage[1], we focus on modelling OntoUML PIMs and their transformation into their realization in relational databases. One of the key challenges in MDD is preserving model constraints during transformations (the second research question).

To elevate the current knowledge of model transformation and database modelling, the transformation of the OntoUML PIMs is divided into three consecutive steps:

1. transformation of the OntoUML PIM into UML PIM;
2. transformation of the UML PIM into RDB PSM (PSM of relational database);
3. transformation of the RDB PSM into SQL ISM (SQL scripts for database schema creation) [28].

This paper is part of a series discussing the transformation of various types of universals used in OntoUML into their correct and complete realization in the RDB ([28], [29], [30]). In particular, this paper is an extended version of the paper presented at MDASD 2016 [30], where the transformation of Rigid Sortal types – Kinds and Subkinds – was introduced. In this paper, we elaborate deeply on various possibilities of the realization during the second and third step of the transformation, focusing on the realization of meta-properties of the generalization sets. We illustrate the possibilities on the running example discussed in section 4.

The structure of the paper is as follows: in section 2, the work related to our approach is discussed; in section 3, the relevant constructs and principles of OntoUML and UFO for this contribution are introduced; in section 4, the running example used for the demonstration of our approach is introduced; in section 5, our approach is discussed and illustrated on the running example; in section 6, discussion to our approach is provided; finally, in section 7, the conclusion of the paper results is provided.

2. Background

2.1. Previous work

Our approach was introduced in [28], where we presented the idea of the three-step transformation using UML and OCL constraints as the intermediate steps. We also discussed

[1] According to ranking published on `https://db-engines.com/en/ranking`, 7 of 10 most popular database systems are relational.

the possibility to use the approach discussed in [32] for the realization of the OCL constraints derived during the transformation of the OntoUML universal types. In [29], we presented more details about the transformation of anti-rigid Sortal universal types (Roles and Phases), discussing the options for the realization of the anti-rigid relations including special OCL constraints and their realization in the RDB using database views.

Finally, in [30], we discussed in more details the transformation of rigid Sortal universal types – Kinds and Subkinds, describing the possible realizations, the derived OCL constraints realizing the meta-properties of the generalization sets in the RDB PSM and their realization in the RDB using database views, CHECK constraints triggers. This paper is an extended version of the paper [30], discussing details of possible realizations and their consequences and providing more examples (especially for the triggers missing in the initial paper).

2.2. Related work

We may distinguish several efforts dealing with the Impedance Mismatch Problem (IMP) [37] of conceptual models and the relational model. These are:

1. Transformation of traditional Entity-Relationship (ER) models into the relational model.
2. Using the UML notation to express relational models.
3. Object-relational mapping technologies (ORM).
4. Transformation from (Onto)UML into a relational model: UML PIM into an RDB PSM

Ad 1 is a long-studied and well-established approach documented e.g. in [27]. We use these approaches in our work. There are also approaches for transformation of the Extended ER models (EER), e.g. [5]. However, the traditional approaches neglect checking certain constraints. As discussed later, when realizing generalization sets and their meta-properties in RDB, the existence of referencing records in other tables must be checked. Similar problem is also addressed in [1] as *inverse referential integrity constraints* (IRICs), where the authors present an approach to the automated implementation of the IRICs by database triggers in a tool called IIS*Case. This tool was designed to provide a complete support for developing database schemas including the check of the consistency of constraints embedded into the DB [18] and the integration of subschemas into a relational DB schema [17]. The transformation of constraints specifically has been elaborated e.g. in [18] and [23]. Next, Rybola and Richta discuss implementation of special multiplicity constraints in [32], which we also use in our work.

There are several approaches to ad 2, but we do not refer to them further here, as they actually do not deal with a conceptual transformation and are mentioned just for the sake of completeness. Ad 3 are technologies offered by libraries of various object-oriented languages (such as Java, Smalltalk, Ruby and C#) to overcome the object-relational Impedance Mismatch Problem (IMP). The library routines perform an automatic run-time transformation between the object model and the relational model. An extensive study of the current leading ORM solutions is presented by Torres et al. in [37]. Nevertheless, ORM works at an application level, while our goal is to push richer semantics to the database level. Also, as noted by Torres et al., ORM provides tools to work with the IMP, but not a complete methodology to solve them.

Ad 4 are approaches most similar to our effort. In [16], the authors describe a tool called Dresden OCL Toolkit [6] which is able to validate a model instance against defined OCL constraints. The tool can be also used to generate SQL code from the model and attached OCL constraints, realizing the constraints using database views. We inspired by this approach in one of our proposed realizations of the generalization set constraints (as well as other types of constraints not discussed here) derived from the initial OntoUML PIM. Another related work can be found in [7], where the authors transform OCL constraints into stored procedures. However, we prefer using other techniques, as procedures must be invoked explicitly by the application, while the suggested CHECK constraints and triggers are executed automatically when performing standard DML operations. Also, we present the realization specially for Oracle Database 12g, while the authors of [7] discusses MariaDB, PostgreSQL and SQL server. Another approach to the realization of UML PIM is discussed e.g. in [22] or [38], where the authors present transformation of a UML PIM into an Object-Relational database. Also, they do not discuss realization of the meta-properties of generalization sets nor other types of constraints, which can be derived from an OntoUML PIM, which is the focus of our research.

As for the transformation of OntoUML specifically, based on the literature review and personally confirmed by the author of OntoUML Dr. Giancarlo Guizzardi, there is no published method for transformation into UML apart from our previous work so far. Instead, there are works dealing with transformation of OntoUML into different languages such as OWL [39] and Alloy [4].

2.3. UML

As OntoUML is based on UML and UML is used in the intermediate steps of the transformation, certain aspects of the language require to be outlined. UML (Unified Modeling Language) [25] is a popular modelling language for creating and maintaining variety of models using diagrams and additional components [34]. In context of the data modelling, UML Class Diagram is the notation mostly used to define conceptual models of application data [2]. Also, to describe the structure of a relational database schema, UML Data Modelling profile as an extension to the UML Class Diagrams may be used [35].

The main elements of a UML Class Diagram are classes, which serve to classify various types of objects of the domain and specify the features and behaviour of their instances. The classes can be related by various types of connectors to define relations between the classes or their instances. In context of this paper, the generalization/specialization relation is important. It is used between the classes to inherit features form a superclass (more abstract concept) to the subclasses (more specific concepts) [25]. As UML is designed following the object-oriented programming approach, an object can be an instance of only one class [2], although according to *Liskov substitution principle*, an instance of a subclass can be used on any place where instance of its superclass is expected [19].

The subclasses of the same superclass may form a *generalization set* to define a partition of subclasses with common meaning [25]. For each generalization set, two meta-properties should be set to restrict the relation of an instance to the individual subclasses: *isCovering* – expressing whether each instance of the superclass must be also an instance of some subclass in the generalization set – and *isDisjoint* – expressing whether an object can be an instance of multiple subclasses in the set at the same time. The default setting of

these properties differ in the individual versions of UML: UML 2.4.1and older define the {incomplete, disjoint} as default, while UML 2.5 defines the {incomplete, overlapping} as default. However, as each object is an instance of exactly one class in the most current programming languages, the concept of generalization sets can only be used in conceptual models and it must be transformed before its actual realization.

To define additional constraints in the UML models, Object Constraint Language (OCL) [24] is used. OCL is a specification language, which is part of the UML standard. In OCL, it is possible to define various conditions, which must be satisfied by all instances of contextual classes, and many other types of constraints. In our approach, we are using OCL to define constraints in the UML PIM and PSM of the relational database.

3. OntoUML

OntoUML is a conceptual modelling language focused on building ontologically well-founded models. It was formulated in Guizzardi's PhD Thesis [10] as a light-weight extension of UML based on UML profiles.

The language is based on *Unified Foundational Ontology* (UFO) [14], which is based on the cognitive science and modal logic and related mathematical foundations such as sets and relations. Thanks to this fact, it provides expressive and precise constructs for modellers to capture the domain of interest. Unlike other extensions of UML, OntoUML does not build on the UML's ontologically vague "class" notion, but builds on the notion of *universals* and *individuals*. It uses the basic notation of UML Class Diagram like classes, associations and generalization/specialization together with stereotypes and meta-attributes to define the nature of individual elements more specifically. On the other hand, it omits a set of other problematic concepts (for instance aggregation and composition) and replaces them with its own ontologically correct concepts.

UFO and OntoUML address many problems in conceptual modelling, such as part-whole relations [12] or roles and the counting problem [11]. The language has been successfully applied in different domains such as interoperability for medical protocols in electrophysiology [9] and the evaluation of an ITU-T standard for transport networks [3].

However, being domain-agnostic, we believe that it may be suitable even for conceptual modelling of application data in the context of MDD. Using OntoUML, we can create very precise and expressive models of application data. These models can be later transformed into relational database schema containing various domain-specific constraints to maintain consistency according to the OntoUML model.

3.1. Universals and Individuals

UFO distinguishes two types of things. *Universals* are general classifiers of various objects and they are represented as classes in OntoUML (e.g. Person). There are various types of universals according to their properties and constraints as discussed later. *Individuals*, on the other hand, are the individual objects instantiating the universals (e.g. Mark, Dan, Kate) [10].

The fact that an individual is an instance of a universal means that – in the given context – we perceive the object *to be* the Universal (e.g. Mark is a Person). Important feature of UFO is the fact that an individual may instantiate multiple universals at the same

time but all the universals must have a common ancestor providing the identity principle (e.g. `Mark` is a `Person` and he is a `Student` as well) [14].

3.2. Identity and Identity Principle

Identity is one of the key features of UFO. Identity is the fact of being what an individual is, enabling distinction of different individuals. Identity principle, on the other hand, defines the methods to determine, if two apparent instances of the same concept are actually the very same individual. Various universals define different identity principles and thus different ways how to distinguish their individuals (e.g. a `Person` is something else than a `University`); different individuals of the same universal have different identities (e.g. `Mark` is not `Kate` even when both are `Persons`).

Each individual always needs to have a single specific identity, otherwise there is a clash of identities (e.g. `Mark` is a `Person` and therefore it can never be confused with another concept such as a `University`). The identity of an individual is determined at the time the individual comes to existence and it is immutable – it can never be changed (e.g. `Mark` will always be `Mark` and he will always be a `Person`).

The types of universals that provide the identity principle for their instances are called *Sortal universals* (e.g. `Person`, `Student`). The types of universals not providing the identity principle are called *Non-Sortal universals* [14] or *Dispersive universals* [13] (e.g. a `Customer` may be a `Person` or a `Company`). In this paper, we discuss only the transformations of the Sortal types of universals, as they form the basis of models.

3.3. Rigidity

UFO and OntoUML are built on the notion of worlds coming from Modal Logic – various configurations of the individuals in various circumstances and contexts of time and space. *Rigidity* is, then, the meta-property of universals that defines the fact if the extension of the universal (i.e. the set of all instances of the universal) is world invariant [15]. UFO distinguishes *rigid* (instances of rigid universals are their instances in all worlds), *anti-rigid* (each instance of an anti-rigid universal from any world is not its instance in certain other world(s)) and *semi-rigid* (some instances of semi-rigid universals are always their instances, other instances may not be their instances in certain worlds) universals [10].

In this paper, we discuss only on the Rigid Sortal types of universals – Kinds and Subkinds – and we discuss the details of the transformation of such universals into the relational databases.

3.4. Generalization and Specialization

In contrast to UML, in UFO and OntoUML, the generalization relation defines the inheritance of the *identity principle*. According to that, an individual which is an instance of the subtype is also an instance of the supertype automatically by following the same identity principle, and thus it also receives the properties defined by the supertype. Additionally, the relation is rigid in UML – an instance of the subclass can always be used as an instance of the superclass unless it ceases to exist – while in OntoUML, the relation may be non-rigid: a single individual may be an instance of both the superclass and subclass in one world and it may be an instance of only the superclass in another world [10].

Although not very common in UML models, the generalization sets are crucial in OntoUML models as they define the required identity for various universal types. Unless altered, {*incomplete, non-disjoint*} is considered the default value of the meta-properties, which is in contrast to UML 2.4.1 and earlier.

3.5. Kinds and Subkinds

The backbone of an OntoUML model is created by Kinds. *Kind* is a Rigid Sortal type of universals that defines the identity principle for its instances, thus defining the way how we are able to distinguish individual instances of that universal [14]. In OntoUML, the Kind universals are depicted as classes with the $\ll Kind \gg$ stereotype [15].

Subkind is a Rigid Sortal universal type that does not define its own identity principle, but it inherits it from its ancestor and provides it to its instances. Therefore, Subkind universals form generalization sets of other Kind or Subkind universals; they form inheritance hierarchies with the root in a Kind universal. In other words, each instance of a Subkind universal is automatically – through the transitive generalization relation – also an instance of all the ancestral Kind and Subkind universals, receiving the identity principle from the root Kind universal. The inheritance may have any combination of values of the *isDisjoint* and *isCovering* meta-properties [14]. In OntoUML, the Kind universals are depicted as classes with the $\ll Subkind \gg$ stereotype [15].

4. Running Example

Our approach to the transformation of Kinds, Subkinds and their generalization sets from the OntoUML PIM into SQL ISM is illustrated on the running example shown in Figure 1. The model shows an excerpt of the domain of transportation company. The company uses various vehicles for transportation of various types of load, ranging from persons to documents to heavy cargo. Therefore, the main entity of such model is the type `Vehicle`. As it is the type defining the identity principle for its instances, it is classified as *Kind*. For each vehicle used by the company, its manufacturer, model and plate number is needed, represented by the respective attributes of the `Vehicle` type.

Furthermore, the company distinguishes three special types of vehicles – trucks, cars and motorcycles – for which additional attributes are important. Each of these types of vehicles is represented by its own type in the OntoUML PIM with the appropriate attributes. As all these types represent the specialization of the general concept of a vehicle, their are classified as *Subkinds* forming a generalization set specializing the type `Vehicle`. Moreover, as the types of the vehicles are disjoint – clearly, one vehicle cannot be a truck and a motorcycle at the same time – the generalization set is defined `disjoint`. On the other hand, there might be other types of vehicles, for which no special attributes need to be recorded. Therefore, the generalization set is not defined `complete`, but rather left `incomplete`.

5. Our Approach

Our approach to the transformation of a PIM in OntoUML into its realization in a relational database consists of three steps which are discussed in the following sections:

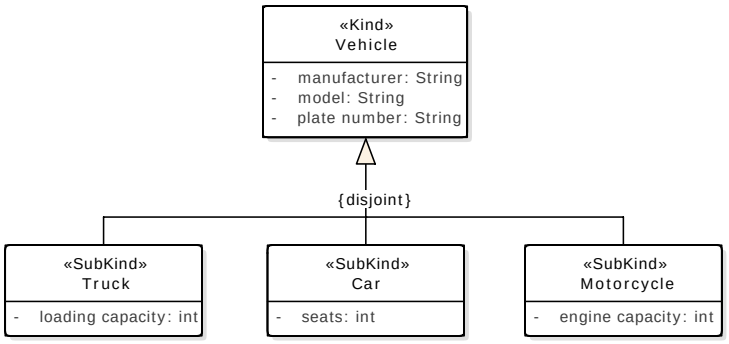


Fig. 1. Example of an OntoUML model with Kinds and Subkinds

1. subsection 5.1 discusses the transformation of an OntoUML PIM into a UML PIM,
2. subsection 5.2 discusses the transformation of the UML PIM into an RDB PSM,
3. subsection 5.3 discusses the transformation of the RDB PSM into an SQL ISM.

As mentioned in the introduction, it should hold that no information should be lost when transforming from a more abstract model into a more specific one. As OntoUML applies constraints for meta-properties of generalization sets, these constraints should be carried over to the other models. However, as generalization sets are not very common in UML models (although defined by the standard), these constraints are not addressed in the common transformation approaches. Therefore we focus on correct and complete realization of these generalization sets including their meta-properties. Whenever not possible to express a constraint directly in the diagrams, we use OCL to define such constraint.

Although we may formulate a direct transformation from OntoUML into the relational database, the transformation via an auxiliary UML model enables to leverage all the available knowledge (e.g. [27,32] and tools for transformation of a UML PIM into database models such as Enterprise Architect), while the direct transformation would have to be built from the scratch. Also, various optimizations and refactoring may be applied whenever possible (e.g. when the subclasses hold no attributes and are used only to distinguish various subtypes of the superclass entity, they can be expressed by values of a single attribute of the superclass).

To demonstrate our approach, we refer to the technical report [31], where all the OCL constraints and SQL scripts are defined for the running example discussed in section 4. We use Oracle Database 12c SQL dialect [26] in all the examples in the SQL ISM.

In the approach presented here, we assume the (most common) situation where all attributes of the model classes have multiplicities `[1..1]`. In the discussion in section 6, we discuss how the situation changes for different multiplicities.

5.1. Transformation of OntoUML PIM into UML PIM

In the first step of the transformation, the OntoUML model is transformed into pure UML model, transforming the types in the OntoUML PIM into the classes in the UML PIM while preserving all the semantics defined by the particular universals of the types.

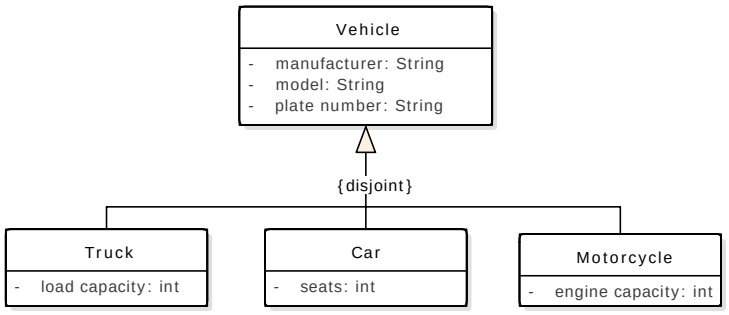


Fig. 2. UML PIM with the transformed Kinds and Subkinds

As various OntoUML universal types define different semantics, they are also transformed in a different manner. However, we discuss only the transformation of Kinds and Subkinds in this paper.

Kinds and Subkinds. As both Kind and Subkind universals in OntoUML are rigid, their instances cannot cease to be their instances without ceasing to exists. The same applies in UML for the classes and their instances. Therefore, the representation of Kinds and Subkinds in UML may stay the same: each $\ll Kind \gg$ and $\ll Subkind \gg$ class from the OntoUML PIM is transformed into a standard UML class in the UML PIM keeping all its features – attributes and relations.

The resulting transformed UML PIM of the running example is shown in Figure 2. Each of the $\ll Kind \gg$ and $\ll Subkind \gg$ classes has been transformed into standard UML class.

Generalization sets. A Subkind in OntoUML represents a special case of a Kind or other Subkind, forming a generalization set together with other Subkinds. As both Kinds and Subkinds are rigid, also the generalization set is rigid: when an object is an instance of the Subkind, it is also an instance of its rigid ancestor – a Kind or another Subkind – because of the inherited identity principle and it cannot cease to be the instance of any of them without loosing its identity.

Thanks to this rigidity, the generalization sets of $\ll Subkind \gg$ classes in the OntoUML PIM can be transformed into standard UML generalization sets in the UML PIM. Also, the meta-properties `isDisjoint` and `isCovering` of the generalization set remain the same – the only difference may be the need to show these properties, depending on the version of UML as discussed in section 2. The example of this transformation can be seen in Figure 2.

5.2. Transformation of PIM into PSM

The second step is the transformation of the UML PIM into RDB PSM. The UML Data Model profile – an extension to the UML class diagrams – is used in the examples to define the structure of the relational database in UML [35]. Additional constraints required

to preserve the semantics derived during the transformation of the OntoUML model are defined as OCL invariants, as OCL is part of the UML standard and there are tools supporting the transformation of OCL constraints into database constructs such as Dresden-OCL [6].

In general, when performing transformation from a UML PIM into a PSM of a relational database, classes are transformed into database tables, class's attributes are transformed into table columns and associations are transformed into references restricted by FOREIGN KEY constraints. Also, PRIMARY KEY constraints are defined for unique identification of individual records in the tables [32]. Therefore, the transformation of classes representing various Kind universals is straightforward – the class with its attributes is transformed into a table with its columns.

However, the concept of generalization is not present in relational databases. Therefore, the generalization sets must be transformed into into tables and references. In general, there are three approached commonly used for the realization of the generalization in relational database [27]:

- by a single table containing columns for all the attributes of the superclass and all the subclasses;
- by individual tables for each of the subclasses, containing the columns for the attributes of the respective subclass and the superclass;
- by a table for the superclass and individual tables for all the subclasses referencing the superclass table.

Each of these variants brings certain limitations and consequences. However, in any case, the generalization set constraints *isCovering* and *isDisjoint* defined in the UML PIM, as well as the multiplicity constraints of the attributes, should be realized as well to preserve the constraints explicitly defined in the model. However, these properties are usually not truly realized in the database. Therefore, we focus on the possibilities for their realization in the RDB PSM to truly preserve the semantics defined in the original OntoUML PIM.

In the following subsections, the details of each of the possible realizations are discussed in context of the generalization set constraints and their realization, as well as other limitations and consequences.

Single Table. In this variant, the superclass and all its subclasses are realized by a single table, defining the columns for all the attributes of the superclass and all its subclasses. Instances of the superclass are represented by rows with NULL values in the columns of the subclasses. Instances of a subclass contain values only in the superclass columns and the columns of that subclass – the columns of the other subclasses contain NULL values. Additionally, adding an extra *discriminator* column to discriminate the subclasses may be convenient [27]. An example of this realization of the generalization set in the running example is shown in Figure 3.

In this realization, the mandatory values of the supeclass's attributes can be easily realized by NOT NULL constraints, as all instances always have values for such mandatory attributes. However, the mandatory values of the subclasses's attributes cannot be realized by such simple constraints, as they are dependent on the actual type of the stored instance.

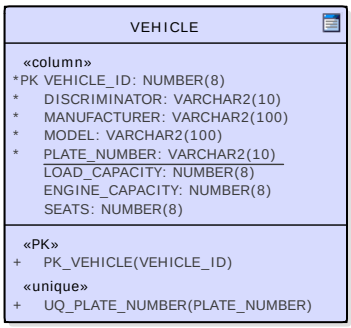


Fig. 3. RDB PSM with the generalization set realized by a single table

Instead, these constraints must be defined with the relation to the value in the *discriminator* column, which in turn can contain only values representing the appropriate classes of the original generalization set according to its meta-properties. Therefore, the following constraints should be realized:

- The *discriminator* value must match one of the subclasses. In the case of an `incomplete` generalization set, also the superclass is allowed. In the case of an `overlapping` generalization set, also all the combinations of the subclasses are allowed.
- All mandatory columns of the subclass(es) identified by the *discriminator* values must contain NOT NULL values, while the columns of all the other subclasses must contain NULL values. In the case of identifying the superclass, all columns of all the subclasses must contain NULL values.

As discussed above, the NOT NULL constraints for the columns representing the attributes of the superclass can be defined directly on the column level, as all instances will always provide values for these columns. However, the constraints for the columns representing the attributes of the subclasses follow complex rules, which cannot be defined on the column level. Instead, they must be defined on the table level by additional constraints. In our approach, we use OCL invariants to define such constraints in the RDB PSM. An example of the OCL constraint realizing the {`disjoint,incomplete`} generalization set from the running example realized by a single table is shown in Constraint 1. Examples of the realization of the other variants of the generalization set meta-properties can be found in [31].

In case that values of some of the attributes of the superclass or some of the subclasses should be unique, the realization of such constraint in the RDB PSM is simple. As all data are stored in the same table, the column representing the particular attribute can be simply restricted by the UNIQUE constraint, as shown in Figure 3 where the uniqueness is defined by the `UQ_PLATE_NUMBER` constraint.

Individual Tables. In this variant, a table is created for each possible type of instance from the generalization set. Usually, it means creating a table for each of the subclasses,

Constraint 1 OCL invariant for the {`disjoint,incomplete`} generalization set realized by a single table

```
context v:VEHICLE inv GS_Vehicle_Types:
def Vehicle_Instance: Boolean =
  v.DISCRIMINATOR = 'Vehicle' AND v.LOAD_CAPACITY = OclVoid
    AND v.SEATS = OclVoid AND v.ENGINE_CAPACITY = OclVoid
def Truck_Instance: Boolean =
  v.DISCRIMINATOR = 'Truck' AND v.LOAD_CAPACITY <> OclVoid
    AND v.SEATS = OclVoid AND v.ENGINE_CAPACITY = OclVoid
def Car_Instance: Boolean =
  v.DISCRIMINATOR = 'Vehicle' AND v.LOAD_CAPACITY = OclVoid
    AND v.SEATS <> OclVoid AND v.ENGINE_CAPACITY = OclVoid
def Motorcycle_Instance: Boolean =
  v.DISCRIMINATOR = 'Motorcycle' AND v.LOAD_CAPACITY = OclVoid
    AND v.SEATS = OclVoid AND v.ENGINE_CAPACITY <> OclVoid

Vehicle_Instance OR Truck_Instance OR Car_Instance OR Motorcycle_Instance
```

containing columns for the attributes of the superclass and the particular subclass [27]. However, to correctly realize the meta-properties of the generalization set, a table for the superclass should be also created in the case of an `incomplete` generalization set and a table for each of the combinations of the subclasses in the case of an `overlapping` generalization set. An example of such realization of the running example is shown in Figure 4.

With this realization, in each of the tables, the mandatory attributes can be easily realized by NOT NULL constraints defined for all the columns representing the mandatory attributes of both the superclass and the particular subclass(es), which the table represents, as only instances of the same combination of classes are stored in a single table, and thus they always have values for those attributes.

However, on the other hand, it is more complicated to ensure the unique values for attributes of the superclass, when needed. The same also applies for attributes of subclasses in the case of an `overlapping` generalization set. It is because the values of such attributes are distributed among multiple tables representing the particular combination of classes from the generalization set. For instance, in the running example, the values of the `plate number` attribute of all vehicles should be unique, regardless of its actual type. As it is realized by the `PLATE_NUMBER` column in all the tables shown in Figure 4, it is not possible to simply restrict the columns only by a simple UNIQUE constraint. Such constraint – named `UQ_PLATE_NUMBER` in each of the tables – only ensures the uniqueness of the values in that particular table. Instead, an additional constraint must be defined to ensure the uniqueness of the values across all the tables with the following properties:

- A constraint is generated for each of the tables realizing the generalization set.
- Each constraint checks, that there is no record in all the other tables with the same value of the unique column as in the constrained table.

As such constraint cannot be defined directly in the table, we use OCL to define them. An example of such OCL constraint for the table `VEHICLE` realiting the unique constraint of the `PLATE_NUMBER` column distributed across the tables shown in Figure 4 is shown in Constraint 2. The constraints for the other tables can be found in [31].

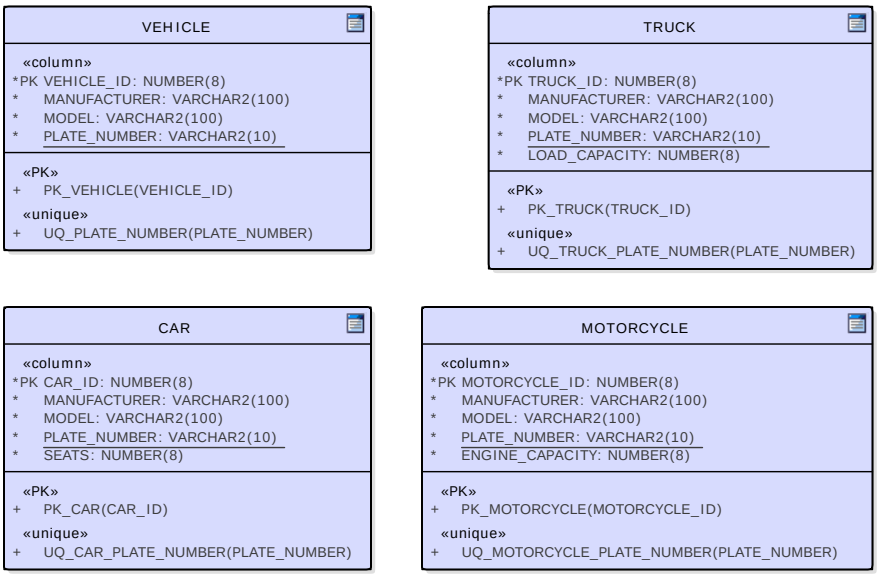


Fig. 4. RDB PSM with the generalization set realized by individual tables

Constraint 2 OCL invariants for distributed unique column `PLATE_NUMBER`

```
context v:VEHICLE inv UQ_Vehicle_Plate_number:
NOT(TRUCK.allInstances()->exists (t|t.PLATE_NUMBER = t.PLATE_NUMBER))
 AND
  NOT(CAR.allInstances()->exists (c|c.PLATE_NUMBER = t.PLATE_NUMBER))
 AND
  NOT(MOTORCYCLE.allInstances()->exists (m|m.PLATE_NUMBER = t.PLATE_NUMBER))
```

Related Tables. In this variant, all the classes are transformed into their own respective tables. Each table contains only attributes of the particular class and the PRIMARY KEY. Additionally, the subclass tables also contain reference to the superclass table restricted by the FOREIGN KEY constraint. Also, a special *discriminator* column in the superclass table to discriminate the actual type of the instance is convenient. The instances of the classes from the UML PIM are then stored in the appropriate combination of tables. Moreover, as the references realize generalization, the references are always one-to-one. Therefore, the reference should be unique and can be combined with the PRIMARY KEY column [27]. An example of this realization of the generalization set in the running example is shown in Figure 5.

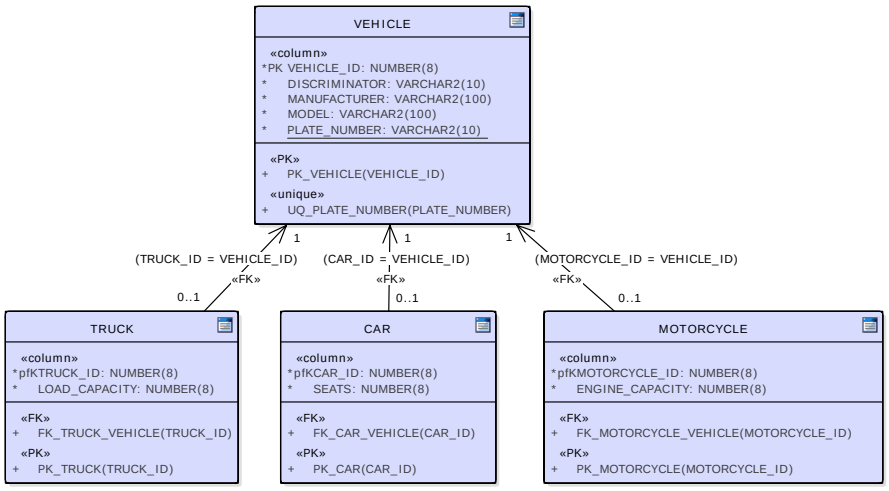


Fig. 5. RDB PSM with the generalization set realized by related tables

The mandatory attributes of the superclass and the subclasses can be simply realized by NOT NULL constraints, as each table always contains the same portion of data of instances of the same type. Also, the UNIQUE constraints for the individual attributes of the superclass and subclasses can be easily realized by UNIQUE constraints, as each attribute is realized only by a single column in a single table – see the `UQ_PLATE_NUMBER` constraint defined in the `VEHICLE` table.

To correctly realize the meta-properties of the generalization set, the existence of referencing records in the subclass tables should be checked in context of the value in the *discriminator* column of the superclass table, which in turn can contain only values representing the appropriate classes or their combination from the original generalization set according to its meta-properties. Therefore, the following constraints should be realized:

– The *discriminator* value in the superclass table must match one of the subclasses. In the case of an `incomplete` generalization set, also the superclass is allowed. In the case of an `overlapping` generalization set, also all the combinations of the subclasses are allowed.

Constraint 3 OCL invariant for the {disjoint,incomplete} generalization set realized by related tables

```
context v:VEHICLE inv GS_Vehicle_Types:
def Vehicle_Instance: Boolean = v.DISCRIMINATOR = 'Vehicle'
    AND NOT (TRUCK.allInstances()->exists (t|t.TRUCK_ID = v.VEHICLE_ID))
    AND NOT (CAR.allInstances()->exists (c|c.CAR_ID = v.VEHICLE_ID))
    AND NOT (MOTORCYCLE.allInstances()->exists (m|m.MOTORCYCLE_ID = v.VEHICLE_ID))
def Truck_Instance: Boolean = v.DISCRIMINATOR = 'Truck'
    AND TRUCK.allInstances()->exists (t|t.TRUCK_ID = v.VEHICLE_ID)
    AND NOT (CAR.allInstances()->exists (c|c.CAR_ID = v.VEHICLE_ID))
    AND NOT (MOTORCYCLE.allInstances()->exists (m|m.MOTORCYCLE_ID = v.VEHICLE_ID))
def Car_Instance: Boolean = v.DISCRIMINATOR = 'Car'
    AND ...
def Motorcycle_Instance: Boolean = v.DISCRIMINATOR = 'Motorcycle'
    AND ...

Vehicle_Instance OR Truck_Instance OR Car_Instance OR Motorcycle_Instance
```

– For each record in the superclass table, there is a single referencing record in each of the tables identified by the *discriminator* value and no referencing record in all the other tables. In the case of identifying the superclass, there are no referencing records in all the subclass tables.

As such constraints restrict data in multiple tables, they cannot be defined on the column level like other database constraints. Instead, a special constraint on the table level must be defined. We use OCL in our approach. Furthermore, as the references are based on the value of the *discriminator* column, covering all its possible values, both constraints can be realized by a single OCL invariant. An example of such constraint for the {disjoint,incomplete} generalization set from the running example in Constraint 3. Examples of the realization of the other variants of the generalization set meta-properties can be found in [31].

5.3. Transformation of PSM into ISM

The last step is the transformation of the RDB PSM into SQL ISM. This model consists of database scripts for the creation of the database tables, their constraints and other constructs.

As we have the PSM of the relational database, the transformation is quite easy. Most of the current CASE tools such as Enterprise Architect [36] can be used to generate SQL DDL scripts. These scripts usually include the CREATE statements for the tables, their columns, NOT NULL constraints and PRIMARY and FOREIGN KEY constraints. However, the OCL invariants defined for the additional constraints require special transformation. Only a few tools currently seem to offer transformation of such constraints – e.g. DresdenOCL [6].

Our approach to the realization of the OCL constraints derived from the OntoUML universal types is inspired by the approach for special multiplicity constraints discussed in [32]. Based on that approach, the following constructs may be used to prevent violating the derived constraints:

- *Database views* can be used to query only the valid data meeting the constraints. They do not slow down the DML operations when inserting, updating or removing data, however, they do not actually prevent inserting data violating the constraints. It is still possible to get invalid data into the database and query them using standard SELECT statements over the original tables. It is necessary to ensure the correct usage of the views on the application level.
- *Updatable database views with CHECK option* can be used to query only the valid data. Moreover, they can be used to manipulate the data by DML operations like INSERT, UPDATE and DELETE, as well. The CHECK option prevents creating a record which will not be accessible by the view, however, it does not prevent changing data in other tables which might violate the constraint. Moreover, the updatable views are restricted by several constraints for the query expression in their definition, as discussed in [32].
- *CHECK constraints* can be used to check the values inserted to various columns of the table. Any DML operation affecting the value in such a column is then validated against the CHECK constraint and rolled back, if the constraint is violated. Unfortunately, the common contemporary database engines (e.g. Oracle 11g) do not support subqueries in the CHECK constraint statements. Therefore, they can be effectively used to realize constraints restricting data in a single table, but not for the relational constraints restricting the records in related tables.
- *Triggers* can be defined to perform complex data validations and manipulations when various DML operations are executed on a table. Being able to define complex queries in the trigger body, they are capable to deal with almost every possible constraint and prevent any operation which would violate them. The constraint checks in the triggers slow down each constrained DML operation, however as shown in [32], the time increase is typically not substantial. On the other hand, it is possible to entirely prevent creating invalid data in the database and save a lot of checking implementation on the application level.

In [32], the research was focused on the realization of special multiplicity constraints. The same approach, however, may be used also for the realization of the constraints derived from the Rigid Sortal universal types and their generalization sets in OntoUML as discussed in the following sections. As the generalization set in the running example is {`disjoint,incomplete`}, only realization of such OCL constraints shown in subsection 5.2 is discussed. However, the other variants would be realized similarly.

Single table. When the generalization set is transformed into a single database table, a special OCL constraint is defined to ensure the meta-properties of the generalization set by checking the values in the columns of the particular subclasses according to the *discriminator* value. An example of the constraint is shown in Constraint 1.

As the constraint simply restricts values of columns in a single table to be empty or non-empty, it can be easily realized in the SQL ISM by a *database view* with a query selecting only records with NULL and NOT NULL values in the appropriate columns according to the *discriminator* value. Such *database view* can be used to query only valid records, ignoring all records violating the constraint. Moreover, as the view meets the condition to be defined as *updatable view* WITH CHECK OPTION, it can also be used for the DML operations, preventing creating invalid data in the original table. Still, the

SQL 1 Updatable database view to query valid data from the combined `Vehicle` table

```
CREATE VIEW GS_VEHICLE_TYPES_VIEW AS
SELECT * FROM VEHICLE v WHERE
  (v.DISCRIMINATOR = 'Vehicle' AND v.LOAD_CAPACITY IS NULL
    AND v.SEATS IS NULL AND v.CONTENT IS NULL)
  OR
  (v.DISCRIMINATOR = 'Truck' AND v.LOAD_CAPACITY IS NOT NULL
    AND v.SEATS IS NULL AND v.CONTENT IS NULL)
  OR
  (v.DISCRIMINATOR = 'Car' AND v.LOAD_CAPACITY IS NULL
    AND v.SEATS IS NOT NULL AND v.CONTENT IS NULL)
  OR
  (v.DISCRIMINATOR = 'Motorcycle' AND v.LOAD_CAPACITY IS NULL
    AND v.SEATS IS NULL AND v.CONTENT IS NOT NULL)
WITH CHECK OPTION;
```

SQL 2 CHECK constraint for the combined `Vehicle` table

```
ALTER TABLE VEHICLE ADD CONSTRAINT GS_VEHICLE_TYPES_CHECK CHECK (
  (DISCRIMINATOR = 'Vehicle' AND LOAD_CAPACITY IS NULL
    AND SEATS IS NULL AND CONTENT IS NULL)
  OR
  (DISCRIMINATOR = 'Truck' AND LOAD_CAPACITY IS NOT NULL
    AND SEATS IS NULL AND CONTENT IS NULL)
  OR
  (DISCRIMINATOR = 'Car' AND LOAD_CAPACITY IS NULL
    AND SEATS IS NOT NULL AND CONTENT IS NULL)
  OR
  (DISCRIMINATOR = 'Motorcycle' AND LOAD_CAPACITY IS NULL
    AND SEATS IS NULL AND CONTENT IS NOT NULL));
```

original table can be directly accessed by the DML and query operations, and therefore such view cannot guarantee the consistency. An example of such database view for the running example is shown in SQL 1.

Similarly, the constraint can also be realized by a *CHECK constraint*, which is checked after each operation on the table. Thanks to that, it is able to completely ensure the data consistency according to the constraint. As the constraint condition checks only data of a single record in a single table, there is no problem with its implementation in the common relational database engines. An example of such CHECK constraint for the running example is shown in SQL 2.

Finally, the constraint might also be realized by a trigger. As the constraint only restricts data in a single table, only a single trigger would be needed. This trigger would be defined to be executed BEFORE INSERT and UPDATE operations, as only these operations can create invalid records. Moreover, as the constraint restricts values of only isolated records, the trigger can be executed for each row, checking only the affected row instead of all the data in the table. Such trigger would check the new values of the affected row in the individual columns and roll back the operation, if the values do not match the condition of the constraint. However, as the constraint can be realized more conveniently by the CHECK constraint or a checked updatable database view, using the trigger is not recommended. Still, an example of such trigger can be found in [31].

SQL 3 Database views to query valid data from the individual tables

```
CREATE VIEW UQ_VEHICLE_PLATE_NUMBER_VIEW AS
SELECT * FROM VEHICLE v WHERE (
  NOT EXISTS (SELECT 1 FROM TRUCK t WHERE t.PLATE_NUMBER = v.PLATE_NUMBER)
  AND NOT EXISTS (SELECT 1 FROM CAR c WHERE c.PLATE_NUMBER = v.PLATE_NUMBER)
  AND NOT EXISTS (SELECT 1 FROM MOTORCYCLE m WHERE m.PLATE_NUMBER = v.PLATE_NUMBER)
) WITH CHECK OPTION;
```

Individual tables. When the generalization set is transformed using the approach of individual tables, no special constraint is needed to realize the meta-properties of the generalization set. However, when a uniqueness of a superclass attribute (or even a subclass attribute in case of an `overlapping` generalization set) must be ensured, special distributed uniqueness OCL constraints such as shown in Constraint 2 must be realized.

Each of such constraints can be transformed into a *database view* querying records from the particular table meeting the condition of the OCL constraint – having such value of the restricted unique column, which does not exist in any of the other tables. This view can be used to query only valid record, ignoring all records in that particular table containing a non-unique value. An example of the view realizing the constraint shown in Constraint 2 is shown in SQL 3. Moreover, as the view meets the conditions for an updatable view, it is defined with the WITH CHECK OPTION and it can be also used for the DML operations while preventing violation of the constraint. As similar views are also defined for the other tables containing the constrained attribute, their combination can completely prevent creation of invalid data in the tables. Still, the original tables can be accessed directly, and therefore such views cannot guarantee the entire database consistency.

Similarly, the constraint could be also realized by CHECK constraints, checking the value of the affected record does not exist in the other tables. However, such CHECK constraint would require subqueries for checking the data in the other tables, and although valid according to the SQL:1999 specification, the contemporary database engines do not support such CHECK constraints. Therefore, this realization is not possible.

Finally, a trigger can also be used to entirely prevent creating invalid data in the database. As the constraint restricts using a value already used in another table, only the INSERT and UPDATE operations executed on each of the tables can create data violating the constraints. Therefore, a trigger is defined FOR EACH ROW on each of the tables after INSERT and UPDATE operations, trying to find the new value defined for the constrained column in the other tables and rolling back the operation, if a record in the other tables is found. An example of the trigger for the table `VEHICLE` in the running example is shown in SQL 3. The other triggers for the other tables can be found in [31].

Related tables. When the generalization set is transformed using the approach of related tables, the constraint is defined in context of the superclass table checking the existence of referencing records in the appropriate subclass tables, such as shown in Constraint 3. However, the realization of this constraint in the SQL ISM is mutually exclusive with the FOREIGN KEY constraint – the generalization set constraint requires referencing records in the subclass tables and the FOREIGN KEY constraint requires existence of the referenced record, but the records must be inserted one-by-one. Therefore, the FOREIGN

SQL 4 Triggers for the individual tables

```
CREATE OR REPLACE TRIGGER UQ_VEHICLE_PLATE_NUMBER_TRIGGER
AFTER INSERT OR UPDATE ON VEHICLE
FOR EACH ROW
DECLARE
  l_count NUMBER := 0;
BEGIN
  SELECT count(1) INTO l_count FROM DUAL WHERE (
    EXISTS (SELECT 1 FROM TRUCK t WHERE t.PLATE_NUMBER = :new.PLATE_NUMBER)
    OR EXISTS (SELECT 1 FROM CAR c WHERE c.PLATE_NUMBER = :new.PLATE_NUMBER)
    OR EXISTS (SELECT 1 FROM MOTORCYCLE m WHERE m.PLATE_NUMBER = :new.PLATE_NUMBER));

  IF l_count > 0 THEN raise_application_error
    (-20101, 'OCL constraint UQ_Vehicle_Plate_number violated!');
  END IF;
END;
```

SQL 5 Database view to query only valid data from the superclass of the related tables

```
CREATE OR REPLACE VIEW GS_VEHICLE_TYPES_VIEW AS
SELECT * FROM VEHICLE v WHERE
  (v.DISCRIMINATOR = 'Vehicle'
    AND NOT EXISTS (SELECT 1 FROM TRUCK t WHERE t.TRUCK_ID = v.VEHICLE_ID)
    AND NOT EXISTS (SELECT 1 FROM CAR c WHERE c.CAR_ID = v.VEHICLE_ID)
    AND NOT EXISTS (SELECT 1 FROM MOTORCYCLE m WHERE m.MOTORCYCLE_ID = v.VEHICLE_ID))
  OR (v.DISCRIMINATOR = 'Truck'
    AND EXISTS (SELECT 1 FROM TRUCK t WHERE t.TRUCK_ID = v.VEHICLE_ID)
    AND NOT EXISTS (SELECT 1 FROM CAR c WHERE c.CAR_ID = v.VEHICLE_ID)
    AND NOT EXISTS (SELECT 1 FROM MOTORCYCLE m WHERE m.MOTORCYCLE_ID = v.VEHICLE_ID))
  OR (v.DISCRIMINATOR = 'Car' AND ...)
  OR (v.DISCRIMINATOR = 'Motorcycle' AND ...))
WITH CHECK OPTION;
```

KEY should be defined DEFERRABLE to be checked at the end of the transaction instead of at the time of each operation [21].

The transformation of the OCL constraint into a *database view* is simple. The view queries only data from the table representing the superclass meeting the condition of the constraint – records which have referencing records in the appropriate subclass tables according to the *discriminator* value. This view can be used to query valid instances of the superclass (in case of `incomplete` generalization set) and it is also used in the JOIN queries to query complete data of valid instances of the individual subclasses (or their combination in case of `overlapping` generalization sets). An example of such view for the running example and the constraint shown in Constraint 3 is shown in SQL 5. Moreover, as the view meets the criteria for being updatable, it is defined with the WITH CHECK OPTION clause and can be used for the DML operations instead of the original superclass table, preventing creation of invalid record by such operations. However, the constraint can be also violated by DML operations on the subclass tables. Although some of these operations can be checked by additional updatable views, it is not possible to prevent all possible violating operations (e.g. inserting a referencing record into inappropriate subclass table), and also the original tables can still be accessed directly. Therefore, the realization by the views cannot guarantee the data consistency entirely.

Similarly, a CHECK constraints could be also defined with the same conditions, automatically preventing creation of invalid data. However, such CHECK constraints would

require subqueries to check data in other tables, which, although valid according to the SQL:1999 specification, is not supported by the contemporary database engines. Therefore, this realization is not applicable.

The constraint can also be realized by a set of triggers checking the individual DML operations on the individual table able to cause violation of the constraint. The triggers verify the appropriate condition and rolls the operation back in the case of violation. In total, the following triggers are needed for the realization of such OCL constraint:

- **INSERT OR UPDATE ON the superclass table**: When inserting the data into the superclass table or updating them, referencing records in the appropriate subclass tables must exist according to the new *discriminator* value.
- **INSERT ON each subclass table**: When inserting a record into a subclass table, it should reference a non-existent record in the superclass table (thanks to the deferred FOREIGN KEY constraint, it should be inserted later while checking existence of appropriate referencing records). Otherwise, it necessarily must be an inappropriate or duplicate record.
- **UPDATE ON each subclass table**: When updating the data in a subclass table and changing its reference value, there must be no record in the superclass table referenced by the old reference value (such record would lose the required referencing record) nor the new reference value (such record should be inserted later to satisfy the insertion constraint discussed above).
- **DELETE ON each subclass table**: When deleting the data from a subclass tables, the referenced record in the superclass should not exist, otherwise it would lose the required referencing record.

All of these triggers can be defined BEFORE the particular operations, as it is possible to detect the constraint violation before really changing the data. Also, defining the triggers FOR EACH ROW is more efficient, as only the affected row can violate the constraint.

As presented above, a lot of triggers must be defined to realize the constraint. Also, the DML operations are slowed down by their execution as they query data from other tables for each of the affected rows. However, in contrast to the other realizations, they entirely prevent violation of the constraint, and thus ensuring the data consistency. An example of the trigger for the table `VEHICLE` from the running example is shown in SQL 6. The other triggers can be found in [31].

6. Discussion

This paper is a part of research on OntoUML-based MDE, however, specifically the topic of generalization sets is applicable in pure UML, as well, as the UML standard also incorporates them.

6.1. Efficiency

As mentioned above, our approach to the realization of the constraints derived from the meta-properties of rigid generalization sets is based on the approach discussed in [32]. In

SQL 6 Trigger for the INSERT and UPDATE operation on the superclass of the related tables realization

```
CREATE OR REPLACE TRIGGER GS_VEHICLE_TYPES_TRIGGER_VEHICLE
BEFORE INSERT OR UPDATE ON VEHICLE
FOR EACH ROW
DECLARE
  l_count NUMBER(1);
BEGIN
  SELECT COUNT(*) INTO l_count FROM DUAL WHERE (
    (:new.DISCRIMINATOR = 'Vehicle'
      AND NOT EXISTS (SELECT 1 FROM TRUCK t WHERE t.TRUCK_ID = :new.VEHICLE_ID)
      AND NOT EXISTS (SELECT 1 FROM CAR c WHERE c.CAR_ID = :new.VEHICLE_ID)
      AND NOT EXISTS (SELECT 1 FROM MOTORCYCLE m
        WHERE m.MOTORCYCLE_ID = :new.VEHICLE_ID))
    OR (:new.DISCRIMINATOR = 'Truck'
      AND EXISTS (SELECT 1 FROM TRUCK t WHERE ...)
      AND NOT EXISTS (SELECT 1 FROM CAR c WHERE ...)
      AND NOT EXISTS (SELECT 1 FROM MOTORCYCLE m WHERE ...))
    OR (:new.DISCRIMINATOR = 'Car' AND ...)
    OR (:new.DISCRIMINATOR = 'Motorcycle' AND ...));

  IF l_count = 0 THEN raise_application_error
      (-20101, 'OCL_constraint_GS_Vehicle_Types_violated!');
  END IF;
END;
```

the paper, the authors discuss possible ways to realize constraints for special multiplicity values using database views and triggers. The authors also provide results of experiments, proving that their realization guarantees database consistency in context of the multiplicity constraints with just a slight decrease in efficiency.

The OCL constraints derived from the meta-properties of generalization sets in OntoUML have the same structure – they are based on multiplicities of related records or their exclusivity. Therefore, also their realization using the views and triggers is very similar. Based on this, we can expect the same impact on the efficiency of the DML operations and queries. However, as our research is not yet fully concluded, experiments are yet to be done to prove that.

6.2. Transformation of UML PIM into RDB PSM

As discussed in subsection 5.2, generalization sets can be transformed into a *single table*, *individual tables* or *related tables*. Each of them has advantages and disadvantages.

The *single table* realization is suitable in situations when we transform subclasses with few of attributes. Otherwise, the constraints for the mandatory attributes of the subclasses are getting complicated. Also, this solution is suitable in the case of `overlapping` generalization sets, as all the data are stored in a single table, preventing data and structure duplicities.

In contrast to that, the *related tables* realization is suitable in situations when the subclasses have many attributes, as the realization of their multiplicity constraints is much easier. On the other hand, it requires joining data from multiple tables to query data of instances of the subclasses. Moreover, it is more complicated to preserve the meta-properties of the generalization set, as queries into other tables will be needed in the validation checks, which increases the execution time of the queries and DML operations.

We consider the possible realization by the *individual tables* not suitable generally, as it leads to structural duplicities and data distribution, which is even more obvious in the case of `overlapping` and `incomplete` generalization sets.

6.3. Transformation of RDB PSM into SQL ISM

As discussed in subsection 5.3, the OCL constraints defined as the result of the transformation of the UML PIM into the RDB PSM can be realized by *database views*, *CHECK constraints* and *triggers*.

The *database views* can be used to query only valid data from the database. The *updatable database views* can be even used to manipulate with the data while checking the constraints. However, still, it is possible to create invalid data in the database violating the constraints by using the tables directly. Therefore, the responsibility is transferred to the application level.

The realization by the *CHECK constraints* is able to ensure the data consistency and prevent creating invalid data violating the constraints. However, as current common database engines do not support subqueries in the CHECK constraint statements, they can be used only in the case of the generalization sets realized by the *single table* approach.

The most reliable and universal realization is the realization by the *triggers*. Using the triggers, it is possible to completely validate the manipulated data and entirely prevent creating data violating the constraints. On the other hand, the triggers increase the time to execute the DML operations. However, as discussed in [32], this time increase is not substantial, unless manipulating with very large database.

6.4. Attribute Multiplicity

In this paper, we focused on the most common situation of mandatory attributes (attribute multiplicity `[1..1]`), as in OntoUML, optional attributes (minimal multiplicity `0`) are considered anti-patterns and they typically signal missing anti-rigid types like Roles or Phases [33]. Still, our approach is applicable even for the case of optional attributes. Some of the constraints will even simplify – e.g. the NOT NULL constraints for individual columns representing the attributes of the subclasses (Constraint 1).

Also, the collection attributes (attributes with the maximal multiplicity $*$) can be simply realized by our approach. They just need to be transformed into separate classes and *one-to-many* relations, leading to the realization in the form of references and FOREIGN KEY constraints.

7. Conclusions

In this paper, we introduced our approach to transformation of an OntoUML PIM of application data into an ISM of a relational database. This transformation is divided into three consecutive steps: the transformation of the OntoUML PIM into UML PIM, the transformation of the UML PIM into RDB PSM and the transformation of the RDB PSM into SQL ISM.

In the transformations, various options are available and additional constraints should be defined and realized to preserve the semantics defined by the OntoUML universal

types. In this paper, we discussed the details of the transformation of Rigid Sortal universal types – Kinds and Subkinds and their generalization sets – discussing various possible realizations of the constraints derived from the semantics of these OntoUML constructs. All the variants are described using a running example of a simple OntoUML PIM of vehicle types.

As for the research questions formulated in section 1:

- Are there benefits of using OntoUML Kinds and Subkinds in PIM? — Yes, when following the proposed transformation, the resulting RDB is able to ensure the constraints, thus enabling better alignment with the problem domain.
- Is it possible to preserve generalization sets constraints of Kinds- and Subkinds in a relational database? — Yes, our transformation covered them.

OntoUML specifies numerous entity types and relation types. There are relatively a lot of rules and constraints (compared to e.g. UML) posed on them. As for the future research, a similar work as presented here should be elaborated for other important OntoUML constructs – the Non-sortal universal types – e.g. Category, Mixin, RoleMixin – and relational constructs – part-whole relations, Relators, etc. As for the continuation of the presented method, combinations of multiple generalization sets of a single universal with various combinations of the meta-properties should be investigated. Finally, experiments should be carried out to study the finer points of individual variants of the constraints realization, mostly with the respect to quantitative measures concerning time and space efficiency in various database backends. Last but not least, a (semi)automated tooling needs to be developed to be able to use the method efficiently in the MDD lifecycle.

Acknowledgments. This research was supported by CTU grant No. SGS17/211/OHK3/3T/18 and contributes to the CTU's ELIXIR CZ Service provision plan.

References

1. Aleksić, S., Ristić, S., Luković, I.: An approach to generating server implementation of the inverse referential integrity constraints. In: Proceedings. AL-Zaytoonah University of Jordan, Amman, Jordan (May 2011)
2. Arlow, J., Neustadt, I.: UML 2.0 and the Unified Process: Practical Object-Oriented Analysis and Design (2nd Edition). Addison-Wesley Professional (2005)
3. Barcelos, P.P.F., Guizzardi, G., Garcia, A.S., Monteiro, M.: Ontological evaluation of the ITU-T recommendation g.805. vol. 18. IEEE Press, Cyprus (2011)
4. Benevides, A.B., Guizzardi, G., Braga, B.F.B., Almeida, J.P.A.: Assessing Modal Aspects of OntoUML Conceptual Models in Alloy. In: Advances in Conceptual Modeling - Challenging Perspectives, vol. 5833, pp. 55–64. Springer Berlin Heidelberg, Berlin, Heidelberg (2009), DOI: 10.1007/978-3-642-04947-7_8
5. Dimitrieski, V., elikovic, M., Aleksic, S., Risti, S., Alargt, A., Lukovi, I.: Concepts and evaluation of the extended entity-relationship approach to database design in a multi-paradigm information system modeling tool. Computer Languages, Systems & Structures 44, 299–318 (Dec 2015)
6. DresdenOCL 3.4.0. `https://github.com/dresden-ocl/dresdenocl` (Aug 2014), accessed: 2016-02-11

7. Egea, M., Dania, C.: SQL-PL4ocl: an automatic code generator from OCL to SQL procedural language. Software & Systems Modeling pp. 1–23 (May 2017), https://link.springer.com/article/10.1007/s10270-017-0597-6
8. Ghezzi, C., Jazayeri, M., Mandrioli, D.: Fundamentals of Software Engineering. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edn. (2002)
9. Goncalves, B., Guizzardi, G., Pereira Filho, J.G.: Using an ECG reference ontology for semantic interoperability of ECG data. Special Issue on Ontologies for Clinical and Translational Research (2011)
10. Guizzardi, G.: Ontological Foundations for Structural Conceptual Models, vol. 015. University of Twente, Enschede (2005)
11. Guizzardi, G.: Agent roles, qua individuals and the counting problem. Software Engineering of Multi-Agent Systems (IV) (2006)
12. Guizzardi, G.: The Problem of Transitivity of Part-Whole Relations in Conceptual Modeling Revisited. In: Proccedings of 21st International Conference on Advanced Information Systems Engineering (CAISE09). Amsterdam, The Netherlands (2009), 00000
13. Guizzardi, G.: Ontological Meta-properties of Derived Object Types. In: Advanced Information Systems Engineering, pp. 318–333. No. 7328 in Lecture Notes in Computer Science, Springer Berlin Heidelberg (Jun 2012), http://link.springer.com/chapter/10.1007/978-3-642-31095-9_21, dOI: 10.1007/978-3-642-31095-9_21
14. Guizzardi, G., Wagner, G.: A unified foundational ontology and some applications of it in business modeling. In: In CAiSE Workshops (3. pp. 129–143 (2004)
15. Guizzardi, G., Wagner, G., Guarino, N., Sinderen, M.v.: An Ontologically Well-Founded Profile for UML Conceptual Models. In: Advanced Information Systems Engineering, pp. 112–126. No. 3084 in Lecture Notes in Computer Science, Springer Berlin Heidelberg (Jun 2004), http://dx.doi.org/10.1007/978-3-540-25975-6_10
16. Heidenreich, F., Wende, C., Demuth, B.: A Framework for Generating Query Language Code from OCL Invariants. Electronic Communications of the EASST 9(0) (Nov 2007), https://journal.ub.tu-berlin.de/eceasst/article/view/108
17. Luković, I., Mogin, P., Pavićević, J., Ristić, S.: An approach to developing complex database schemas using form types. Software: Practice and Experience 37(15), 16211656 (Dec 2007), http://dx.doi.org/10.1002/spe.v37:15
18. Luković, I., Popović, A., Mostić, J., Ristić, S.: A tool for modeling form type check constraints and complex functionalities of business applications. Computer Science and Information Systems 7(2), 359–385 (2010)
19. Martin, R.C.: Design principles and design patterns. Object Mentor 1, 34 (2000)
20. Mellor, S.J., Clark, A.N., Futagami, T.: Model-driven development. IEEE Software 20(5), 14 (Sep 2003)
21. Melton, J.: Advanced SQL:1999. Morgan Kaufmann Publishers (2003)
22. Mok, W.Y., Paper, D.P.: On transformations from UML models to object-relational databases. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences. pp. 10 pp.– (Jan 2001)
23. Obrenović, N., Popović, A., Aleksić, S., Luković, I.: Transformations of check constraint PIM specifications. Computing and Informatics 31(5), 1045–1079 (2012)
24. OMG: Object constraint language (OCL), version 2.4. http://www.omg.org/spec/OCL/2.4/ (Feb 2014), accessed: 2016-02-23
25. OMG: UML 2.5. http://www.omg.org/spec/UML/2.5/ (Mar 2015), accessed: 2016-02-08
26. Oracle: Oracle SQL. http://www.oracle.com/technetwork/database/database-technologies/sql/overview/index.html, accessed: 2017-01-19
27. Ramakrishnan, R., Gehrke, J.: Database Management Systems, 3rd Edition. McGraw-Hill, Boston, 3rd edition edn. (Aug 2002)

28. Rybola, Z., Pergl, R.: Towards OntoUML for Software Engineering: Introduction to the Transformation of OntoUML into Relational Databases. In: Enterprise and Organizational Modeling and Simulation. LNBIP, Springer, CAiSE 2016, Ljubljana, Slovenia (June 2016), in press.
29. Rybola, Z., Pergl, R.: Towards OntoUML for Software Engineering: Transformation of Anti-Rigid Sortal Types into Relational Databases. In: Model and Data Engineering. LNCS, vol. 9893, pp. 1–15. Springer, MEDI 2016, Almria, Spain (Sep 2016)
30. Rybola, Z., Pergl, R.: Towards OntoUML for Software Engineering: Transformation of Rigid Sortal Types into Relational Databases. In: Proceedings of the 2016 Federated Conference on Computer Science and Information Systems. p. 15811591. No. 8 in ACSIS, Gdansk, Poland (2016), doi: 10.15439/2016F250
31. Rybola, Z., Pergl, R.: Transformation of Kinds and Subkinds into Relational Databases: A Running Example. Technical Report TR-FIT-2017, Czech Technical University in Prague (2017)
32. Rybola, Z., Richta, K.: Possible Realizations of Multiplicity Constraints. Computer Science and Information Systems 10(4), 1621–1646 (Oct 2013), wOS:000327912000006
33. Sales, T.P., Guizzardi, G.: Anti-patterns in Ontology-driven Conceptual Modeling: The Case of Role Modeling in OntoUML. In: Ontology Engineering with Ontology Design Patterns: Foundations and Applications. IOS Press (2016)
34. da Silva, A.R.: Model-driven engineering: A survey supported by the unified conceptual model. Computer Languages, Systems & Structures 43, 139 – 155 (2015)
35. Sparks, G.: Database Modeling in UML, `http://www.eetimes.com/document.asp?doc_id=1255046`, accessed: 2016-02-02
36. Sparx Systems: Enterprise architect 13. `http://www.sparxsystems.com.au/products/ea/index.html`, accessed: 2017-01-02
37. Torres, A., Galante, R., Pimenta, M.S., Martins, A.J.B.: Twenty years of object-relational mapping: A survey on patterns, solutions, and their implications on application design. Information and Software Technology 82, 1–18 (Feb 2017)
38. Vara, J.M., Vela, B., Bollati, V.A., Marcos, E.: Supporting Model-Driven Development of Object-Relational Database Schemas: A Case Study. In: Theory and Practice of Model Transformations, vol. 5563, pp. 181–196. Springer Berlin Heidelberg, Berlin, Heidelberg (2009), dOI: 10.1007/978-3-642-02408-5_13
39. Zamborlini, V., Guizzardi, G.: On the representation of temporally changing information in OWL. In: Enterprise Distributed Object Computing Conference Workshops (EDOCW), 2010 14th IEEE International. pp. 283–292. IEEE (2010)

Zdeněk Rybola is an assistant professor at the Department of Software Engineering at the Faculty of Information Technology, Czech Technical University in Prague, teaching software engineering courses. He also just submitted his PhD thesis in August 2017. His area of interest includes OntoUML, UML, Model-Driven Development and Relational databases.

Robert Pergl is an assistant professor at the Department of Software Engineering at the Faculty of Information Technology, Czech Technical University in Prague and the head of the research group Centre for Conceptual Modelling and Implementation (CCMi). He focuses on ontologies, conceptual modelling, enterprise engineering, programming languages and paradigms.

Received: January 9, 2017; Accepted: August 17, 2017.

Computer Science and Information Systems 14(3):939–958 DOI: https://doi.org/10.2298/CSIS170116036C

Development of Custom Notation for XML-based Language: a Model-Driven Approach

Sergej Chodarev and Jaroslav Porubän

Technical University of Košice, Department of Computers and Informatics,
Letná 9, Košice, Slovakia
{sergej.chodarev, jaroslav.poruban}@tuke.sk

Abstract. In spite of its popularity, XML provides poor user experience and a lot of domain-specific languages can be improved by introducing custom, more human-friendly notation. This paper presents an approach for design and development of the custom notation for existing XML-based language together with a translator between the new notation and XML. The approach supports iterative design of the language concrete syntax, allowing its modification based on users feedback. The translator is developed using a model-driven approach. It is based on explicit representation of language abstract syntax (metamodel) that can be augmented with mappings to both XML and the custom notation. We provide recommendations for application of the approach and demonstrate them on a case study of a language for definition of graphs.

Keywords: domain-specific languages, human-computer interaction, iterative design, model-driven development, translator, XML.

1. Introduction

XML is very common and easy to parse generic language. It is well supported by existing tools and technologies and therefore it is a popular basis for domain-specific languages (DSLs). While XML is appropriate choice in many cases, especially for program-to-program communication, it is not well suited for cases, where humans need to manipulate documents. Although they are able to create, modify and read XML documents, it is not a pleasurable experience, because of uniformity and syntactic noise that makes it difficult to find useful information visually [28].

While a more appropriate syntax can be chosen for the development of new languages, a lot of languages was already implemented based on XML and their reimplementation would be complicated and time-consuming. One of the possible ways to solve this problem is to develop a translator that would read documents written in a specialized human-friendly notation and output them in the XML for further processing using existing tools. Ideally, the new notation would be specifically tailored to the domain of the language as is usual for DSLs [22].

In this paper we present an approach that supports iterative design of the notation. It is possible to evaluate the notation by automatically converting samples of existing documents from XML in each iteration. This makes it easier to experiment with different syntax alternatives and choose the most appropriate one.

To make such iterative process possible, we propose to use model-driven approach. This means that instead of traditional language development approach driven by *concrete*

syntax definition, we need to base development of the translator on *the abstract syntax* that is common for all notations of the language [16]. Abstract syntax should be expressed in a format that can be easily augmented with the definition of both notations and allow automatic generation of corresponding language processors.

For example, Java classes representing the structure of an XML-based language can be generated automatically from the XML Schema using JAXB[1]. The generated classes are already annotated in a way that allows automatic marshalling and unmarshalling their instances in the XML form. Additional annotations can be added to the classes that define their mapping to a different textual notation. In the next step an annotation based parser generator, like YAJCo [30], can be used to generate a parser and pretty-printer for the new notation. Connecting them with the XML marshaller and unmarshaller one would get a complete translator from the custom human-friendly notation to the original XML-based and back.

Rest of the paper is structured based on the main topics and contributions of the paper, that are the following:

1. The process of *iterative design* of new notation for existing language (Section 2). This process is in contrast with traditional approach, where concrete syntax is completely defined before the development of language processor.
2. The approach to language translator development that is based on explicit representation of language *abstract syntax* in a format that allows attaching definitions of different concrete notations (Section 3). This allows to develop a round-trip translator based on the specification of the abstract syntax.
3. Demonstration of the approach on *a case study* of a language for specification of graphs (Section 4). The case study shows possible challenges of the approach and can be used as a guide to develop similar translators.
4. Summary of *recommendations* for application of the approach that was generalized from the case study (Section 5).

The approach was originally presented in our conference paper [6]. In this extended version of the paper larger emphasis was given to the iterative process of notation design, which is now explained in greater detail.

This paper also presents completely new case study. In the previous paper a language for graphical user interface specification was used (examples from the original case study are provided in Appendix A). While the language demonstrated that the new notation could be much shorter and compendious compared to XML, it did not have complete specification in a form of XML Schema making the development more complicated. The new case study uses the GraphML language with proper XML Schema. The study also includes discussion of alternative solutions and describes testing of the translator.

In addition, the recommendations was extracted from both case studies, that summarize most important points in a tool-neutral form.

2. Iterative Process of Language Notation Design

Notation of a formal language defines a way how it is presented to its users and how they interact with code written in the language. Therefore, notation is a user interface

[1] Java Architecture for XML Binding: https://docs.oracle.com/javase/tutorial/jaxb/

of the language and design of the notation should follow the principles of user interface design [2]. This means it requires iterative evaluation of the design and its modification based on the evaluation results [25]. Development of the translator between the new notation and the original one should follow the same iterations, so the translator would allow to test the design in conditions similar to real life.

On the other hand, classical approach to language development [1] assumes that concrete syntax of the language is designed upfront. A complete specification of the grammar is then augmented with semantic actions and processed to generate a parser. Therefore, changes in the syntax often require modification of semantic rules, making the process laborious.

We propose an alternative process for development of the custom notation for existing XML-based languages together with the round-trip translator:

1. Extract the language abstract syntax from the XML Schema.
2. Augment the abstract syntax with initial definition of the new concrete syntax.
3. Generate a pretty-printer based on the definition.
4. Convert examples of existing XML documents to the new notation.
5. Evaluate the new notation on examples of converted documents.
6. If the notation is not satisfactory, modify the concrete syntax definition and go back to the step 3.
7. If the notation is satisfactory, complete the syntax definition and generate the parser.

In the first step, the central piece of the translator — language abstract syntax is defined based on the existing language definition. The first iteration then starts with the design of the initial version of the notation. Definition of the notation must support generation of a pretty-printer based on it. More detailed discussion of the implementation is provided in the Section 3.

A set of existing documents in the XML-based notation is converted to the new one and manually evaluated. Based on the results of the evaluation, the notation is either redesigned and reimplemented based on the feedback, or it is finalized to obtain full round-trip translator between the notations.

This process allows to easily use existing documents for testing the new notation instead of some artificial examples. Complete real-life documents in the new notation can be generated automatically immediately after the definition of the syntax has changed. This allows very fast evaluation and modification cycles, so problems in the notation can be spotted and resolved, even if they occur only in complex documents.

The evaluation can be done in different ways depending on the needs of the project. In the simplest case it consists of visual checking of comprehensibility of converted documents. Usual usability testing methods can be used as well. This includes testing with potential users of the language, in which they would be given realistic tasks. For example, *discount usability evaluation method* can be successfully applied to software languages [19]. Quantitative evaluation methods can be used as well [2], although they require larger number of participants to obtain statistically significant results.

This approach also provides a simple method for testing correctness of the developed translator, i.e. that no information is lost or corrupted during the translation. A set of example XML documents can be automatically converted to the new notation and then back to the XML. Result of the conversion can be compared with the original XML documents

to reveal missing support for some language features or other errors. If the translator is correct, no data is lost and documents are identical (except of differences in formatting that can be removed using normalization before the comparison).

The approach is not limited to XML-based languages. With some modifications it can be used for development of alternative notation for any software language.

3. Model-driven Development of Language Translator

To support described process, it is needed to use approach similar to model-driven software development [33], where the development of the language translator is driven by the model of the language — metamodel[2]. The metamodel defines language concepts with their properties and relations to other concepts. Definition of the structure is annotated with additional information about concrete syntax of the language that needs to be translated.

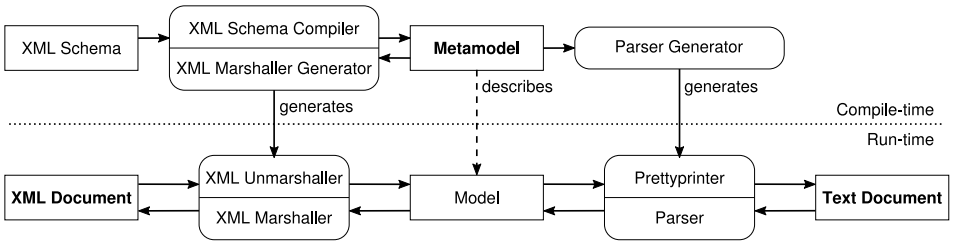


Fig. 1. Model-driven language translator development (arrows represent data-flow)

Figure 1 shows the whole architecture of the model-driven language translator development in the case of translating XML to textual notation and vice versa. The metamodel augmented with definition of concrete notations is the central element. It is used as an input to generate parser and pretty-printer for both the textual notation (using parser generator) and XML (using XML marshaller generator). The generated tools can be connected into a pipeline that handles the translation of one notation to the other with the internal representation of the model (defined by the metamodel) as an intermediate format.

What is important, the first version of the metamodel itself can be retrieved from the existing description of XML-based language — XML Schema. This allows to significantly shorten the development process, because large part of the language definition — its abstract syntax specification — is derived automatically.

This style of development also follows the "Single Point of Truth" principle [31], because the structure of the language is defined only once and its mappings to concrete notations are attached to it. Therefore, it is easier to keep concrete syntax definitions in sync.

In the case of evolution of the language, the changes should be expressed in the metamodel, so other artifacts, including the XML Schema, could be updated automatically. If

[2] If we consider documents written in a language to be *models*, then a model of the language itself is *a metamodel*.

it is not possible, it would be required to manually synchronize changes of the language with the metamodel definition.

The described approach does not depend on concrete tools. It, however, requires an XML marshaller and a parser/pretty-printer generator that both use the same format for the metamodel specification. In the case study in Section 4 Java classes are used to represent the metamodel. They are augmented using annotations, JAXB is used as an XML marshaller and YAJCo as a parser generator. Alternative solution can use Ecore from EMF [34] to represent the metamodel and Xtext [8] as a parser generator.

The approach is based on interconnecting different technological spaces (TS) [18]. Original language and its infrastructure is defined in the XML technological space, but for definition of the new textual notation it is appropriate to use the programming languages syntax TS. Both spaces are interconnected using the abstract syntax definition, that by itself can be placed in a different technological space. In our case it is object-oriented programming TS, but it can be Model-Driven Architecture TS if different representation of metamodel would be chosen (e.g. Ecore). The choice of technological space would substantially influence approaches and technologies used to solve the task of language translation. For example, while in the MDA TS some model transformation language (e.g. ATL [13]) would be used, in OOP TS the same task would be solved using methods of the classes representing metamodel or using the Visitor design pattern.

From this point of view, presented case study also demonstrates that object-oriented programming language like Java can be successfully used as a format for abstract syntax description, provided that it allows attaching structured metadata [26] (known as annotations or attributes) to program elements. This allows to use numerous existing tools and lowers barrier of learning new technologies for industrial programmers.

4. Case Study

The approach is demonstrated on the development of a new textual notation for the GraphML language. Graph Markup Language (GraphML) is a format for storage and exchange of graphs and associated metadata used by some graph drawing tools [5].

The translator was implemented using two tools: JAXB and YAJCo. JAXB is a standard solution for marshalling and unmarshalling Java objects to XML. YAJCo[3] (Yet Another Java Compiler Compiler) is a parser generator for Java that allows to specify language syntax using a metamodel in a form of annotated Java classes [30]. This allows declarative specification of the language and its mapping to Java objects [20]. In addition to the parser, YAJCo is able to generate a pretty-printer and other tools from the same specification [21].

This section describes the process of development of the translator using the chosen tools. It also explains challenges that arise during the implementation and their solutions. Readers can use it as a guide to develop their own translator[4].

As an additional illustration of the custom notation for an existing XML-based language, Appendix A provides example from our previous paper [6] — GtkBuilder language used to define graphical user interfaces.

[3] Available at https://github.com/kpi-tuke/yajco

[4] Complete source code of the translator is available at http://hron.fei.tuke.sk/~chodarev/graphl/

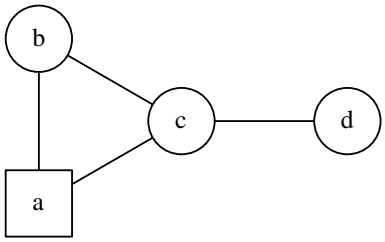


Fig. 2. Example graph

Listing 1. Example graph definition using XML notation

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <graphml xmlns="http://graphml.graphdrawing.org/xmlns">
3    <key id="ds" for="node" attr.name="shape" attr.type="string">
4      <default>circle</default>
5    </key>
6    <graph id="G" edgedefault="undirected">
7      <node id="a">
8        <data key="ds">square</data>
9      </node>
10     <node id="b"/>
11     <node id="c"/>
12     <node id="d"/>
13     <edge source="a" target="b"/>
14     <edge source="a" target="c"/>
15     <edge source="b" target="c"/>
16     <edge source="c" target="d"/>
17   </graph>
18 </graphml>
```

4.1. Graph Markup Language

A GraphML document contains one or more definitions of graphs and each graph consists from nodes and edges. In addition, nodes and edges can have user defined attributes attached. All possible attributes must be declared in the beginning of the document using a *key* element. GraphML also supports advanced concepts, like hyperedges (edges connecting more then two nodes), nodes with named ports (locations for edges to connect) and nested graphs.

An example definition for simple graph (depicted in Fig. 2) is presented in Listing 1. In addition to nodes and edges it defines one custom attribute with the name “shape” and default value “circle” (lines 3–5). The attribute is used to change shape of the node *a* to “square” (line 8).

The same definition in custom textual notation is presented in Listing 2. The notation is inspired by the DOT language [11]. Notation is much cleaner as it uses plain identifiers to define nodes and pseudo-graphical symbol “−−” to define edges. Additional properties of nodes or edges are specified in square brackets and may include port definitions or

Listing 2. Example graph definition using custom textual notation

```
key shape [id: ds, for: NODE, type: STRING, default: "circle"];
undirected graph G {
    a [ shape = "square" ] ;
    b ;
    c ;
    d ;
    a -- b ;
    a -- c ;
    b -- c ;
    c -- d ; }
```

attribute values. In contrast to XML notation, attributes are referenced by their name instead of ID, because custom name resolution strategy based on attribute names is used instead of default XML ID mechanism.

4.2. Project Layout

We recommend to split the project into two separate modules or subprojects:

1. the metamodel definition and the code generated based on it (parsers, pretty-printers),
2. the language translator that uses the metamodel and generated code from the first module.

This layout explicitly divides generated code and the code that depends on it, therefore simplifying build process. We used Apache Maven[5] to manage building and testing of the project and to configure multi-module project.

The first module contains only the classes representing metamodel and definitions of two concrete syntaxes. In our case, the concrete syntax definitions are attached directly to the metamodel classes in the form of Java annotations.

The second module contains implementation of the translator between the notations of the language. It instantiates JAXB marshaller and unmarshaller and also YAJCo generated parser and pretty-printer and uses them to read internal model of a document from one notation and write it in the other notation. This module also contains tests for automatic verifying of the translator and its parts (see section 4.6).

4.3. Metamodel Extraction

As was mentioned earlier, the metamodel represented by Java classes was generated based on the existing XML Schema using the XML binding compiler (`xjc`) that is a part of the JAXB. It generates Java classes corresponding to elements of XML-based language. Generated classes contain annotations that define mapping of classes and their fields to XML elements and attributes. JAXB uses these annotations to create instances of the

[5] Available at https://maven.apache.org/

Listing 3. NodeType class constructor with YAJCo annotations

```
@After(";") @NewLine
public NodeType(String id,
                @Before("[") @After("]") @Separator(",")
                List<NodeElement> dataOrPort) {
    this.id = id;
    this.dataOrPort = dataOrPort;
}
```

Listing 4. EBNF grammar rule generated from the NodeType class constructor

```
NodeType ::= <ID> <[> (NodeElement (<,> NodeElement)*)? <]> <;>
```

classes and set their properties based on XML document contents. The same annotations are used to serialize objects to the XML form.

This means that after the metamodel was extracted it is possible to use JAXB to read an existing graph definition from the XML notation to an internal representation defined by the metamodel and also to marshall the internal model back to the XML form.

In the case study, extracted classes directly corresponded to elements of the XML-based language. Therefore, they included classes like *GraphType*, *NodeType*, *EdgeType*, *KeyType* (declaration of data attribute), *DataType* (value of data attribute), etc. In total, 13 classes and 7 enum types was generated by JAXB.

4.4. Custom Syntax Definition

Definition of the new concrete syntax is also provided in the form of annotations added to the metamodel classes. This means that the metamodel generated using JAXB needs to be modified to include YAJCo-specific annotations.

While JAXB annotations are placed at classes and fields, in YAJCo most of the annotations are attached to constructors and their parameters. Each constructor is transformed into a grammar rule and parameters of the constructor determine the right-hand side of the rule. This allows to define syntactic alternatives for the same language concept and also explicitly specifying order of elements based on the order of parameters. In addition, YAJCo infers relations between language concepts from the inheritance relations between the metamodel classes.

For example, Listing 3 presents one of the constructors of the *NodeType* class. It defines that a graph node can be constructed from a string representing its identifier and a list of ports or data attributes (e.g. line 3 in Listing 2). The node definition would start with the *ID* token representing the identifier, followed by a sequence of elements enclosed in brackets and separated by comma. Grammar rule generated based on the constructor is presented in Listing 4. Annotations also contain hints on indentation and new-line placement that are used by the pretty-printer, but ignored by the parser.

Each variation of the element concrete syntax requires its own constructor. For example, the *node* can be defined with additional elements specified, or without them (for

example lines 3 and 4 in Listing 2) and therefore it needs at least two constructors. In addition to the constructors, factory methods can be used as an annotation target. This makes it possible to define different syntaxes even if they have the same types of parameters in Java.

Each class also needs a non-parametrized constructor required by JAXB. This constructor must be marked using the YAJCo `@Exclude` annotation so it would be ignored by the YAJCo tool.

4.5. Development Process

After defining initial subset of the concrete syntax, YAJCo was used to generate parser and pretty-printer. They were used to implement the translator according to the schema described in Section 3.

The translator was used to convert example XML documents taken from the official GraphML documentation[6] to the new syntax. Translated samples of the documents were manually checked by developers of the translator to evaluate the syntax. This process was repeated after each change of the syntax definition, therefore providing very short development cycles for experimenting with different notations for implemented language features.

The new notation was developed incrementally. Support for language concepts was gradually added and different variants of concrete syntax was considered and immediately evaluated by the authors of the translator. For example, several different notations for node attributes and ports was considered, before the final one was chosen.

4.6. Testing Translator Completeness

Development of the translator requires automatic testing of its completeness. This is done by executing round-trip translation — convert an XML document to the custom notation and then convert it back to XML. After the translation, contents of the document should not change, except of formatting.

To realize the testing, usual unit tests were implemented for each tested document. Comparison of XML documents was handled using XMLUnit[7] library that allows to perform XML comparisons for the purpose of application testing.

The same documents that were used to test the notation, was also used to test completeness of the translator. In addition to comparing original and resulting XML documents, unit tests printed intermediate form in the custom notation to help with notation checking.

4.7. Completing the Metamodel

The metamodel extracted from XML Schema can be incomplete in several ways. First of all, the XML Schema itself may be incomplete — missing definition of some language concepts or properties. On the other hand, the extraction tool can leave out some language properties or express them in a form that parser generator cannot understand. Language can also intentionally leave specification of some elements to extensions.

[6] GraphML Primer, available at http://graphml.graphdrawing.org/primer/graphml-primer.html
[7] Available at http://www.xmlunit.org/

Listing 5. Alternative types of values as defined by JAXB

```
public class GraphType {
    @XmlElements({
        @XmlElement(name = "data", type = DataType.class),
        @XmlElement(name = "node", type = NodeType.class),
        @XmlElement(name = "edge", type = EdgeType.class),
        @XmlElement(name = "hyperedge", type=HyperedgeType.class)
    })
    protected List<Object> dataOrNodeOrEdge;
    ...
}
```

Incomplete XML schema. Incompleteness of the schema can be easily solved by manual modification of the schema itself or extracted metamodel. For example, the schema of GraphML does not specify references between elements using identifiers. In most cases resolution of references is not required in a translator as it just passes identifiers from one notation to another without change. In the case of GraphML data attributes, however, we needed to access referenced data key. Therefore, generated classes was modified to include JAXB annotations `@XmlID` and `@XmlIDREF`.

Incomplete translation of the schema. This type of incompleteness is again solved by manual modification of the metamodel classes.

This problem appeared in cases, where several alternative values of different types are expected in the same context. For example, *graph* definition contains a sequence of nodes, edges, hyperedges or data attributes. In object-oriented model this situation can be expressed by inheritance. JAXB, however, does not use this technique in generated metamodel classes. Instead, it uses *Object* type in the container and adds `@XMLElements` annotation to specify all possible concrete types that can be used as is shown in Listing 5.

On the other hand, YAJCo requires the use of inheritance or implementation relations in these situations. So a new marker interface was created and classes of all elements that can appear in specific context are marked to implement it. The container class is then modified to reference the marker interface instead of the *Object*. Result of these modifications is presented in Listing 6.

Extensible languages. While first two cases can be easily fixed by manual or semi-automatic modification of the metamodel, the last one represents more complex problem. If all used extensions are known, it is possible to incorporate them into the metamodel. The other possibility is to include generic element type in the metamodel, that would represent all elements that are not defined explicitly. Then it would be possible to define also some generic notation for them in the custom form of the language that would be equivalent to XML.

Listing 6. Alternative types of values defined using inheritance

```
public class GraphType {
    @XmlElements( ... )
    protected List<GraphElement> dataOrNodeOrEdge;
    ...
}

public interface GraphElement {}

public class NodeType implements GraphElement { ... }

public class EdgeType implements GraphElement { ... }

public class HyperedgeType implements GraphElement { ... }
```

4.8. Model Transformations

In some cases it is useful to slightly modify abstract syntax for the purpose of custom notation. This can be done by defining separate metamodel and then transforming models from one metamodel to another. In simple cases, however, a single metamodel can be extended to include notation-specific properties.

Notation specific properties. Representation of the metamodel using Java classes allows to implement simple model transformations using constructors. Constructors of the metamodel classes can transform their parameters before storing to object fields. It makes it possible to define helper classes with own syntax rules, but store parsed information in a form expected by XML marshaller.

For example, hyperedges in XML notation contain endpoints and data attributes in arbitrary order. In textual notation, however, we need to separate them, because we decided to represent them differently: endpoints separated by "`--`" symbol and data attributes enclosed in brackets and separated by comma. In the result, hyperedge definition may look like this: `hyperedge: a -- b -- c [color="red", width="2.0"];`

To implement it we need to introduce separate constructor parameters for endpoints and data attributes that are used by YAJCo to generate parser. In addition, pretty-printer requires getters corresponding to these parameters.

As you can see in Listing 7, these separate lists are not even stored in the object fields. Instead, they are combined into existing field *dataOrEntrypoint*. The lists in the getters are constructed by filtering corresponding elements from the combined list. This means that constructor itself and getters implement this simple transformation.

Transforming visitors. Another way to implement transformation of the model is to introduce separate transformation step between reading the model in one notation and writing in the other notation. In object-oriented languages like Java, the Visitor design pattern can be used for this purpose.

Listing 7. Separate lists for endpoints and data attributes of hyperedge

```
public class HyperedgeType implements GraphElement {
    @XmlElements({
        @XmlElement(name = "data", type = DataType.class),
        @XmlElement(name = "endpoint", type = EndpointType.class)
    })
    protected List<Object> dataOrEndpoint;
    ...

    @Before({"hyperedge", ":"}) @After(";") @NewLine
    public HyperedgeType(
            @Separator("--") @Range(minOccurs = 1)
            List<EndpointType> endpoint,
            @Before("[") @After("]") @Separator(",")
            List<DataType> data) {
        this.dataOrEndpoint = new ArrayList<>(endpoint);
        this.dataOrEndpoint.addAll(data);
    }

    public List<EndpointType> getEndpoint() {
        return filterByType(dataOrEndpoint, EndpointType.class);
    }

    public List<DataType> getData() {
        return filterByType(dataOrEndpoint, DataType.class);
    }
    ...
}
```

This technique is especially useful for cases where values of fields used by both notations need to be modified. In our case it is used to change representation of strings and identifiers between notations (see next section).

4.9. Different Representations of Identifiers

An important problem arises from different treatment of keywords and other tokens in different notations. XML uses special syntax for language elements (tags delimited by angle brackets) and therefore it can allow to use language keywords as identifiers inside XML attributes and text fragments. For example, a graph can be named simply "graph": `<graph id="graph">...</graph>`

On the other hand, if element names like "`graph`" or "`hyperedge`" become reserved keywords in the custom notation, they could not be used as identifiers anymore, because standard lexical analyzer would not be able to distinguish them. Therefore equivalent expression "`graph graph {...}`" could not be parsed.

An ideal solution for the problem would be the use of scannerless parser [15]. It does not have separate lexical analysis step and therefore can distinguish identifiers and key-

words based on parsing context. In a case where it is not possible due to technological constraints, the conflict can be resolved in several ways:

1. by selecting language keywords with some special symbols that are not allowed in identifiers (for example, "`%graph`" instead of "`graph`"),
2. by requiring special notation for user defined identifiers (for example, beginning with the dollar sign "`$`"),
3. by modifying only conflicting identifiers using model transformation or in pretty-printer (for example, by appending underscore "`graph_`").

Another problem is in the definition of characters that are allowed in an identifier. If they are different, then illegal characters need to be replaced or escaped. This problem needs to be solved not only for identifiers, but also for other types of tokens, like strings.

Automatic sanitization of identifiers can be done in model transformation step described in the previous section. An alternative solution is to combine transformation in class constructor with customizing generated pretty-printer. In case of YAJCo, the pretty-printer is based on the Visitor pattern, so it can be easily customized by overriding needed methods in a subclass.

5. Recommendations

Experience from the case study can be summarized in several recommendations for application of the presented translator development approach. We suppose that most of these recommendations are not limited to used technologies and are applicable for development of any translator between two notations of the same language.

1. Use such representation of the language metamodel, that can be mapped to both translated notations and allows to automatically generate parser and pretty-printer from these mappings.
2. Use separate modules for the metamodel with generated code on one side and translator that uses it on the other side. This allows to define explicit dependencies between generated and handwritten code.
3. If it is possible, extract initial definition of the metamodel from specification of existing language notation. If resulting metamodel is incomplete, it can be completed manually.
4. Use notation-specific properties of model concepts to represent structural differences between notations. Simple transformations should be inserted in the translation process to convert values between these properties.
5. Take care of different representations of identifiers, strings and other types of tokens in different notations. This may require replacement or escaping of tokens during the transformation.
6. Use round-trip transformation of existing documents to test completeness of the translator and suitability of the new notation. This allows to see documents in the new notation without manually writing them.

6. Related Work

Domain-specific languages. Domain-specific languages are successfully used in different areas, for example specification of static structure of database applications [7], development of kiosk applications [40], or development of some specific aspects of applications like user interface [29], logging mechanisms [39], or acceptance tests [35]. It was also shown that DSLs improve comprehension of programs compared to general-purpose languages [17]. Therefore tools and methods for development of DSLs are active research topics.

Development of domain-specific languages can be guided by numerous patterns as described by Mernik et al. [22]. Since the approach described in this paper does not deal with design of a completely new language, not all patterns are applicable there. From the implementation patterns, the *Preprocessor* pattern clearly applies to our work. A language processor developed using the presented method does not perform complete static analysis of the code, so it is actually a preprocessor from the new concrete syntax to the old one.

Karsai et al. [14] provide guidelines for design of domain-specific languages. These guidelines are independent from concrete development approach and tools, so the guidelines from the *Concrete Syntax* category are fully applicable to the design of concrete syntax using our approach.

Application of usability testing methods to evaluation of DSLs is described in the work of Barišic et al. [2]. Kurtev et al. [19] also demonstrate that even low budget studies with small number of participants (Discount Usability Evaluation method) can be successfully used for this purpose. All these methods can be utilized during the design of the custom notation for DSL.

DSL development methods. A complete approach for the systematic development of domain-specific languages was presented by Strembeck et al. [36]. They define a model-driven development process for DSL development. Similarly to our approach, they suggest starting the process with the definition of the language model. In our case the language model is not developed based on domain analysis, but is extracted from the existing language specification. Behavior of the language and its integration with target platform are not defined, because they are provided by the existing implementation. Definition of the concrete syntax, however, can be done according to the process described in their work.

Villanueva Del Pozo in her thesis [38] defined an agile model-driven method for involving end-users in DSL development. The method proposes several concrete mechanisms to involve users in design and testing of the language based on questionnaires and specification of usage scenarios. Our approach supports similar ways of user involvement. In addition, in our case it is possible to use existing samples of DSL documents instead of usage scenarios.

To conclude, our approach differs from general-purpose DSL development methods in a fact that it does not cover design and implementation of a complete new language, but only design of the new concrete syntax for an existing DSL and implementation of the translator. Therefore we focus on aspects that are specific to this task and use the fact, that existing language definition and existing documents can be used to aid design, implementation and testing of the syntax and translator.

Alternative solutions. An alternative to development of the custom notation is the use of different generic language instead of the XML. YAML (Yet Another Markup Language) is a popular choice, for example Shearer [32] used it to provide textual representation for ontologies. YAML was specially designed as a human-friendly notation for expressing data structures [4]. Its syntax is readable, but the use of generic language does not allow to use specialized short-hand notations tailored for a developed language. While the basic structure of our example language may be expressed similar to the custom notation, problems start in the details. For example, the custom syntax uses infix notation for graph edges, that is not supported by YAML.

Similar solution is the use of OMG HUTN (Human-Usable Textual Notation) which specifies generic textual notation for MOF (Meta-Object Facility) based metamodels [23], again without possibility to customize concrete syntax.

XML-based language can be also replaced with an internal DSL that is embedded in a general-purpose language. For example, Nosal' and Porubän showed patterns for mapping XML to source-code annotations in case of configuration languages [27]. This approach, however, is limited to the specific type of languages that express application configuration and its mapping to elements of source code.

Another possibility is to derive textual notation automatically based on the metamodel. This approach was implemented for languages defined using Meta-Object Facility (MOF) [12]. Automatic derivation, however, does not allow to fine-tune the notation for the needs of users.

To conclude, all solutions based on some generic or automatically derived concrete syntax greatly simplify development process at the cost of restricted customizability of the syntax. Therefore, in cases where these limitations are acceptable, these methods may be more appropriate compared to the approach described in this paper. However, in cases where custom syntax is desirable, our approach can improve design and development process.

Alternative technologies. The approach presented in this paper does not depend on concrete tools, therefore it is possible to implement it using alternative technologies.

Neubauer et al. had shown in their work [24], that it is possible to use Ecore from the Eclipse Modeling Framework (EMF) [34] for representing metamodels and Eclipse Xtext [8] for generating parser, pretty-printer (*serializer* in the Xtext terminology), and editing support based on the Eclipse integrated development environment. They developed a tool, called XMLText, that automates development of round-trip transformation from XML-based languages defined by XML Schema to textual notation. Their tool also generates syntax definition for the language. This definition can be used as a starting point for customization using the process described in section 2, so our contribution compared to their work is in defining a tool-neutral development approach.

Another real-life example of migrating UML and XML based modeling languages to textual and graphical languages using EMF and Xtext was presented by Eysholdt and Rupprecht [9]. They, however, did not use a single metamodel for different notations. Instead, they used model-to-model transformations to migrate models.

The main difference compared to technologies presented in this paper is the fact that EMF and Xtext use specialized language for defining metamodel (Ecore), while JAXB and YAJCo rely on Java for this purpose. This allows to lower the entry barrier by mini-

mizing the amount of new technologies needed to be learned. It also allows to implement model transformations in Java using the techniques well-known by industrial programmers. On the other hand, EMF promises independence on concrete programming language. Together with Xtext they also provide a more mature platform for the development of language processors and editing environments. The approach itself, hovewer, is fully applicable using these tools as well.

Different language notations. The approach presented in this paper can be modified for other types of notations. The approach can be used, for example, in development of alternative notations for ontologies instead of XML-based languages, similarly to some existing tools [37, 10].

The new notation is also not required to be textual. As was shown in the work of Bačíková et al. [3], it is possible to use the same metamodel definition to generate a graphical user interface. This interface would consist of forms allowing to edit language sentences as an alternative to writing the model in textual form.

7. Conclusion

Presented case study showed the applicability of the model-driven translator development approach and therefore possibility of iterative design of language notation together with its translator to the original notation. It also allowed to formulate several recommendations for practical use of the approach. Most of them are not specific to the tools used in the study and should be applicable to other tools as well.

An advantage of the model-driven approach compared to grammar-driven approaches is in the fact that it allows to define concrete syntax variants as simple mappings to the abstract syntax and therefore to freely experiment with the concrete syntax, without the need to reimplement the whole translator.

Common representation of the model shared by several existing tools also allows to use them to construct complete translator with little effort. In our case study, Java classes was used as such common representation, therefore our work also showed that object-oriented programming language with support for annotations provide adequate foundation for expressing metamodels. This allows light-weight model-driven software development, that lowers barrier for adoption by allowing to use tools and knowledge from object-oriented programming.

The use of a custom notation, of course, have several disadvantages compared to the standard and well-supported notation such as the XML. It disables possibility to use existing tools like editors, code browsers and so on. If such tools are needed, they need to be developed by authors of the new notation, although this process can be supported by language development tools such as Xtext. Overall, benefits of custom textual notation compared to XML should be considered for each language individually based on possible improvements of the readability and environment in which the language is used.

Development of the case study also exposed several deficiencies and potential improvements in the YAJCo tool. Therefore, the future work would be devoted to its improvement. For example, built-in support for different types of tokens would greatly simplify language implementation, and generation of supporting tools, like editor, would improve experience of language users.

Acknowledgments. This work was supported by projects KEGA 047TUKE-4/2016 “Integrating software processes into the teaching of programming” and FEI-2015-23 “Pattern based domain-specific language development”.

References

1. Aho, A.V., Lam, M.S., Sethi, R., Ullman, J.D.: Compilers: Principles, Techniques, and Tools (2nd Edition). Addison-Wesley, Boston, USA (2006)
2. Barišic, A., Amaral, V., Goulão, M., Barroca, B.: Evaluating the Usability of Domain-Specific Languages. In: Software Design and Development, pp. 2120–2141. IGI Global (2012)
3. Bačíková, M., Lakatoš, D., Nosáľ, M.: Automatized generating of GUIs for domain-specific languages. In: CEUR Workshop Proceedings. vol. 935, pp. 27–35 (2012)
4. Ben-Kiki, O., Evans, C., Ingerson, B.: YAML Ain’t Markup Language. Version 1.2. Tech. rep. (2009), http://yaml.org/
5. Brandes, U., Eiglsperger, M., Lerner, J., Pich, C.: Graph Markup Language (GraphML). In: Roberto Tamassia (ed.) Handbook of Graph Drawing and Visualization, pp. 517–541. CRC Press (2013)
6. Chodarev, S.: Development of human-friendly notation for XML-based languages. In: Federated Conference on Computer Science and Information Systems (FedCSIS). pp. 1565–1571. IEEE (2016)
7. Dejanović, I., Milosavljević, G., Perišić, B., Tumbas, M.: A domain-specific language for defining static structure of database applications. Computer Science and Information Systems (ComSIS) 7(3), 409–440 (2010)
8. Efftinge, S., Völter, M.: oAW xText: A framework for textual DSLs. In: Workshop on Modeling Symposium at Eclipse Summit. vol. 32, p. 118 (2006)
9. Eysholdt, M., Rupprecht, J.: Migrating a Large Modeling Environment from XML/UML to Xtext/GMF. In: ACM International Conference Companion on Object Oriented Programming Systems Languages and Applications Companion (OOPSLA). pp. 97–104. ACM, New York, USA (2010)
10. Fonseca, J.M.S., Pereira, M.J.V., Henriques, P.R.: Converting Ontologies into DSLs. In: 3rd Symposium on Languages, Applications and Technologies. OpenAccess Series in Informatics (OASIcs), vol. 38, pp. 85–92. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2014)
11. Gansner, E.R., North, S.C.: An open graph visualization system and its applications to software engineering. Software: Practice and Experience 30(11), 1203–1233 (sep 2000)
12. Gargantini, A., Riccobene, E., Scandurra, P.: Deriving a textual notation from a metamodel: an experience on bridging modelware and grammarware. Milestones, Models and Mappings for Model-Driven Architecture p. 33 (2006)
13. Jouault, F., Allilaire, F., Bézivin, J., Kurtev, I.: ATL: A model transformation tool. Science of Computer Programming 72(1), 31 – 39 (2008)
14. Karsai, G., Krahn, H., Pinkernell, C., Rumpe, B., Schindler, M., Völkel, S.: Design Guidelines for Domain Specific Languages. In: 9th OOPSLA Workshop on Domain-Specific Modeling (DSM’ 09). p. 7. No. October (2009)
15. Kats, L.C., Visser, E., Wachsmuth, G.: Pure and declarative syntax definition. ACM SIGPLAN Notices 45(10), 918 (oct 2010)
16. Kleppe, A.: A Language Description is More than a Metamodel. In: Fourth International Workshop on Software Language Engineering. Grenoble, France (2007), http://doc.utwente.nl/64546/
17. Kosar, T., Oliveira, N., Mernik, M., Pereira, V.J.M., Črepinšek, M., Da Cruz, D., Henriques, R.P.: Comparing general-purpose and domain-specific languages: An empirical study. Computer Science and Information Systems (ComSIS) 7(2), 247–264 (2010)

18. Kurtev, I., Bézivin, J., Aksit, M.: Technological Spaces: An Initial Appraisal. In: International Symposium on Distributed Objects and Applications, DOA 2002 (2002), http://doc.utwente.nl/55814/
19. Kurtev, S., Christensen, T.A., Thomsen, B.: Discount method for programming language evaluation. In: 7th International Workshop on Evaluation and Usability of Programming Languages and Tools (PLATEAU). pp. 1–8. ACM Press, New York, New York, USA (2016)
20. Lakatoš, D., Porubän, J., Bačíková, M.: Declarative specification of references in DSLs. In: 2013 Federated Conference on Computer Science and Information Systems (FedCSIS). pp. 1527–1534. IEEE (2013)
21. Lakatoš, D., Porubän, J.: Generating tools from a computer language definition. In: International Scientific conference on Computer Science and Engineering (CSE 2010). pp. 76–83 (September 2010)
22. Mernik, M., Heering, J., Sloane, A.M.: When and how to develop domain-specific languages. ACM Computing Surveys 37(4), 316–344 (dec 2005)
23. Muller, P.A., Hassenforder, M.: HUTN as a Bridge between ModelWare and GrammarWare - An Experience Report. WISME Workshop, MODELS/UML pp. 1–10 (2005)
24. Neubauer, P., Bergmayr, A., Mayerhofer, T., Troya, J., Wimmer, M.: XMLText: from XML schema to Xtext. In: 2015 ACM SIGPLAN International Conference on Software Language Engineering. pp. 71–76. ACM (oct 2015)
25. Nielsen, J.: Iterative user-interface design. IEEE Computer 26(11), 32–41 (Nov 1993)
26. Nosáľ, M., Sulír, M., Juhár, J.: Source code annotations as formal languages. In: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS). pp. 953–964 (Sept 2015)
27. Nosáľ, M., Porubän, J.: XML to annotations mapping definition with patterns. Computer Science and Information Systems (ComSIS) 11(4), 1455–1477 (2014)
28. Parr, T.: Humans should not have to grok XML (8 2001), http://www.ibm.com/developerworks/library/x-sbxml/index.html
29. Perisic, B., Milosavljevic, G., Dejanovic, I., Milosavljevic, B.: UML profile for specifying user interfaces of business applications. Computer Science and Information Systems (ComSIS) 8(2), 405–426 (2011)
30. Porubän, J., Forgáč, M., Sabo, M., Běhálek, M.: Annotation based parser generator. Computer Science and Information Systems (ComSIS) 7(2), 291–307 (2010)
31. Raymond, E.S.: The art of Unix programming. Addison-Wesley (2004), http://www.catb.org/esr/writings/taoup/
32. Shearer, R.: Structured ontology format. In: Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions (2007)
33. Stahl, T., Voelter, M., Czarnecki, K.: Model-Driven Software Development: Technology, Engineering, Management. John Wiley & Sons (2006)
34. Steinberg, D., Budinsky, F., Merks, E., Paternostro, M.: EMF: eclipse modeling framework. Pearson Education (2008)
35. Straszak, T., Śmiałek, M.: Model-driven acceptance test automation based on use cases. Computer Science and Information Systems (ComSIS) 12(2), 707–728 (2015)
36. Strembeck, M., Zdun, U.: An approach for the systematic development of domain-specific languages. Software: Practice and Experience 39(15), 1253–1292 (oct 2009)
37. Čeh, I., Črepinšek, M., Kosar, T., Mernik, M.: Ontology driven development of domain-specific languages. Computer Science and Information Systems (ComSIS) 8(2), 317–342 (2011)
38. Villanueva Del Pozo, M.J.: An agile model-driven method for involving end-users in DSL development. Ph.D. thesis, Universitat Politècnica de València, Valencia (Spain) (jan 2016), https://riunet.upv.es/handle/10251/60156
39. Zawoad, S., Mernik, M., Hasan, R.: Towards building a forensics aware language for secure logging. Computer Science and Information Systems (ComSIS) 11(4), 1291–1314 (2014)

40. Živanov, Ž., Rakić, P., Hajduković, M.: Using code generation approach in developing kiosk applications. Computer Science and Information Systems (ComSIS) 5(1), 41–59 (2008)

A. GtkBuilder Language Example

This appendix provides an example of the concrete syntax of the GtkBuilder language from our previous case study. GtkBuilder is a part of the GTK+ GUI toolkit that allows to declaratively specify layout of a user interface using an XML-based language[8]. Details of the implementation of the translator can be found in the original paper [6].

The GtkBuilder UI definition language allows to specify a layout of widgets forming a user interface and their properties using an XML notation. Each instance of a widget is defined using an *object* element, which contains its type, identifier, properties, signal bindings, and child objects. Listing 8 presents an example UI definition in the XML notation.

Listing 8. Example of user interface definition using XML notation

```
<interface>
  <object class="GtkDialog" id="dialog1">
    <child internal-child="vbox">
      <object class="GtkVBox" id="vbox1">
        <property name="border-width">10</property>
        <child internal-child="action_area">
          <object class="GtkHButtonBox" id="hbuttonbox1">
            <property name="border-width">20</property>
            <child>
              <object class="GtkButton" id="save_button">
                <property name="label" translatable="yes">Save
                </property>
                <signal name="clicked"
                        handler="save_button_clicked"/>
              </object>
            </child>
          </object>
        </child>
      </object>
    </child>
  </object>
</interface>
```

The same definition can be expressed using a custom notation as shown in Listing 9. The notation uses special symbols to provide concise representation for language elements. For example, *object* is expressed using "`[ Class id ... ]`" notation (e.g. line 1), properties are written simply as pairs in a form "`name : value`" (e.g. line 4),

[8] Specified at https://developer.gnome.org/gtk3/stable/GtkBuilder.html

Listing 9. Example of user interface definition using custom textual notation

```
1  [ GtkDialog dialog1
2  %child vbox :
3     [ GtkVBox vbox1
4     border-width : 10
5     %child action_area :
6        [ GtkHButtonBox hbuttonbox1
7        border-width : 20
8        %child :
9           [ GtkButton save_button
10          label : _ Save
11          clicked -> save_button_clicked ]]]]
```

signal binding is expressed as "`signal_name -> handler`" (line 11), and strings that should be translated in localized versions of UI are marked with underscore (line 10). The notation is short and quite intuitive at the same time.

Sergej Chodarev is Assistant professor at the Department of Computers and Informatics, Technical university of Košice, Slovakia. He received his PhD. in Computer Science in 2012. The main areas of his current research are design and implementation of domain specific languages, meta-programming and user interfaces.

Jaroslav Porubän is Associate professor and the Head of Department of Computers and Informatics, Technical university of Košice, Slovakia. He received his MSc. in Computer Science in 2000 and his PhD. in Computer Science in 2004. Since 2003 he is a member of the Department of Computers and Informatics at Technical University of Košice. Currently the main subject of his research is the computer language engineering concentrating on design and implementation of domain specific languages and computer language composition and evolution.

Received: January 16, 2017; Accepted: September 21, 2017.

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

004

COMPUTER Science and Information Systems : the International journal /
Editor-in-Chief Mirjana Ivanović. – Vol. 14, No 3 (2017) - . – Novi Sad (Trg D. Obradovića 3): ComSIS Consortium, 2017 - (Belgrade : Sigra star). –30 cm

Polugodišnje. – Tekst na engleskom jeziku

ISSN 1820-0214 (Print) 2406-1018 (Online) = Computer Science and Information Systems
COBISS.SR-ID 112261644

Cover design: V. Štavljanin
Printed by: Sigra star, Belgrade

ComSIS Computer Science and Information Systems
Volume 14, Number 3, Special Issue, September 2017

Contents

Editorial
Guest editorial: Advances in Distributed Computing and Data Analysis
Guest editorial: Advances in Information Technology
Guest editorial: Model Driven Approaches in System Development

Special Section: Advances in Distributed Computing and Data Analysis

579 Imbalanced Data Classification Based on Hybrid Resampling and Twin Support Vector Machine
Lu Cao, Hong Shen

597 Promising Techniques for Anomaly Detection on Network Traffic
Hui Tian, Jingtian Liu, Meimei Ding

611 BHyberCube: a MapReduce Aware Heterogeneous Architecture for Data Center
Tao Jiang, Huaxi Gu, Kun Wang, Xiaoshan Yu, Yunfeng Lu

629 Click-Boosted Graph Ranking for Image Retrieval
Jun Wu, Yu He, Xiaohong Qin, Na Zhao, Yingpeng Sang

643 A Weighted Mutual Information Biclustering Algorithm for Gene Expression Data
Yidong Li, Wenhua Liu, Yankun Jia, Hairong Dong

661 An Optimization Scheme for Routing and Scheduling of Concurrent User Requests in Wireless Mesh Networks
Zhanmao Cao, Chase Q. Wu, Mark L. Berry

Special Section: Advances in Information Technology

685 Construction of Affective Education in Mobile Learning: The Study Based on Learner's Interest and Emotion Recognition
Haijian Chen, Yonghui Dai, Yanjie Feng, Bo Jiang, Jun Xiao, Ben You

703 A Retrieval Algorithm of Encrypted Speech based on Syllable-level Perceptual Hashing
Shaofang He, Huan Zhao

719 A Novel Link Quality Prediction Algorithm for Wireless Sensor Networks
Chenhao Jia, Linlan Liu, Xiaole Gu, Manlan Liu

735 Connected Model for Opportunistic Sensor Network Based on Katz Centrality
Jian Shu, Lei Xu, Shandong Jiang, Lingchong Meng

751 An Improved Artificial Bee Colony Algorithm with Elite-Guided Search Equations
Zhenxin Du, Dezhi Han, Guangzhong Liu, Kun Bi, Jianxin Jia

769 A DDoS Attack Detection System Based on Spark Framework
Dezhi Han, Kun Bi, Han Liu, Jianxin Jia

789 A kernel based true online Sarsa(λ) for continuous space control problems
Fei Zhu, Haijun Zhu, Yuchen Fu, Donghuo Chen, Xiaoke Zhou

805 Social evaluation of innovative drugs: A method based on big data analytics
Genghui Dai, Xinshuang Fu, Weihui Dai, Shengqi Lu

823 Sentiment information Extraction of comparative sentences based on CRF model
Wei Wang, Guodong Xin, Bailing Wang, Junheng Huang, Yang Liu

839 Distinguishing Flooding Distributed Denial of Service from Flash Crowds Using Four Data Mining Approaches
Bin Kong, Kun Yang, Degang Sun, Meimei Li, Zhixin Shi

857 Building a Lightweight Testbed Using Devices in Personal Area Networks
Qiaozhi Xu, Junxing Zhang

Special Section: Model Driven Approaches in System Development

875 Supporting the platform extensibility for the model-driven development of agent systems by the interoperability between domain-specific modeling languages of multi-agent systems
Geylani Kardas, Emine Bircan, Moharram Challenger

913 Towards OntoUML for Software Engineering: Transformation of Kinds and Subkinds into Relational Databases
Zdeněk Rybola, Robert Perg

939 Development of Custom Notation for XML-based Language: a Model-Driven Approach
Sergej Chodarev, Jaroslav Porubän

ISSN: 1820-0214 (Print) 2406-1018 (Online)