DOI: 10.2298/CSIS100906003W

Integrating Instance-level and Attribute-level Knowledge into Document Clustering

Jinlong Wang[1,3], Shunyao Wu[1], Gang Li[2], Zhe Wei[4,5]

[1] School of Computer Engineering, Qingdao Technological University
266033 Qingdao, China
{wangjinlong,shunyaowu}@gmail.com
[2] School of Information Technology, Deakin University
3125, Victoria, Australia
gang.li@deakin.edu.au
[3] Medical College of Qingdao University,
266021 Qingdao, China
[4] State Key Laboratory of CAD&CG, Zhejiang University
310027 Hangzhou, China
[5] SANYHE International Holding Co., Ltd,
110027 Shenyang, China

Abstract. In this paper, we present a document clustering framework incorporating instance-level knowledge in the form of pairwise constraints and attribute-level knowledge in the form of keyphrases. Firstly, we initialize weights based on metric learning with pairwise constraints, then simultaneously learn two kinds of knowledge by combining the distance-based and the constraint-based approaches, finally evaluate and select clustering result based on the degree of users' satisfaction. The experimental results demonstrate the effectiveness and potential of the proposed method.

Keywords: document clustering, pairwise constraints, keyphrases.

1. Introduction

Document clustering is one of the paramount tasks in text analysis and mining, for a wide range of information retrieval tasks, such as documents classification, documents summarization and visualization, *etc*. The traditional document clustering is unsupervised exploratory learning process, assuming no training samples from the user, automatically grouping unlabeled similar documents into meaningful clusters while separating documents with different topics. However, the performance of document clustering is usually unsatisfactory. There are many reasons, such as (1) the bag of words (BOW) model which is usually used in document clustering is relatively weak [11]; (2) it is unsupervised and impossible to interact with people; (3) it is difficult to understand the meaning of partitions sometimes.

In practice, there is usually some prior knowledge available for use, which can improve the clustering quality. Recently, many researchers have employed these prior knowledge to assist unsupervised document clustering, becoming a

hot topic in data mining and machine learning communities [3,4,5,14,7,11,12]. [4] proposed a probabilistic semi-supervised framework combining constraint-based and distance-based approaches with instance-level knowledge in the form of pairwise constraints. [3] proposed an effective method to actively obtain pairwise constraints based on [4] and [5]. [11] utilized Wikipedia as background knowledge to construct bag of concepts (BOC) model, and partitioned documents with pairwise constraints obtained by active learning. [12] proposed a semi-supervised clustering framework that actively selects informative pairwise constraints for obtaining user feedback.

Indeed, the semi-supervised document clustering approaches make use of additional information to increase clustering quality and make the partition easy to understand. Nevertheless, a majority of existing work are overwhelmed by attribute-level knowledge side information, except [2], which extracts keyphrases from *Title* and *Keywords*, and sets large weights to the keyphrases. Experimental results demonstrate the effectiveness of their method on short articles such as News. In addition, keyphrases can be obtained by utilizing some methods of keyphrase extraction or keyphrase assignment [22,13].

However, almost all the aforementioned approaches only incorporated one kind of knowledge. The performance of clustering quality with both kinds of side information becomes an interesting problem. In text classification, Vikas Sindhwani *etc*. proposed two classification algorithms that supported dual supervision in the form of labels for both examples and features in 2008 [16], and designed two strategies for active dual supervision in 2009 [15,17]. Experimental results demonstrate the effectiveness and potential of their algorithms.

In this paper, we aim to integrate both the instance-level knowledge in the form of pairwise constraints and the attribute-level knowledge in the form of keyphrases to assist document clustering. Based on the semi-supervised method integrating pair-wise constraints and attribute preferences [20], we present a framework for document clustering analysis. Firstly, we utilize pairwise constraints to construct optimization so as to obtain initial weights, then, we add keyphrases and simultaneously learn the two knowledge, finally, we evaluate and select the result according to the degree of users' satisfaction.

The rest of the article is organized as follows. In section 2, we introduce the two knowledge incorporated by our method, pairwise constraints and keyphrases; in section 3, we propose our framework incorporating pairwise constraints and keyphrases; we demonstrate experimental results in section 4; finally we conclude the paper in Section 5.

2. Notations

Given a set of n documents $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ with d words, where $\mathbf{x}_i = [x_{i1}, \cdots, x_{id}]^t$ (t denotes the transpose operation), $\mathbf{x}_i \in \mathbb{R}^d$, the desired number of clusters k, "must-link" set $\mathcal{S}$, "cannot-link" set $\mathcal{D}$ and keyphrases set $\mathcal{P}$, the objective of clustering is to obtain a partition of $\mathcal{X}$. In addition, $|\mathcal{S}|$ stands for the number of constraints in set $\mathcal{S}$.

2.1. Pairwise Constraints

Instance-level knowledge utilized by constrained clustering includes labels, pairwise constraints, *etc*. Considering the definition of traditional clustering and our strategy to incorporate dual knowledge, this paper chooses pairwise constraints as instance-level knowledge.

The set of pairwise constraints comprises "must-link" set $\mathcal{S}$ and "cannot-link" set $\mathcal{D}$.

- $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}$ means $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same cluster.
- $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}$ means $\mathbf{x}_i$ and $\mathbf{x}_j$ are in different clusters.

2.2. Keyphrases

Keyphrases provide brief summaries of documents' content and reflect main topic of documents [22], such as words in title, keywords, MeSH (Medical Subject Headings) information in biomedical texts, *etc*. There are many different types of approaches to obtain keyphrases, such as keyphrase extraction, keyphrase assignment, and so on. In this paper, we extract keyphrases from *Title* and *Keywords* [2], and utilize attribute order preferences [18] to express keyphrases.

An attribute order preference (s, t, δ) ($\delta > 0$) stands for $\mathbf{w}_s - \mathbf{w}_t \geq \delta$. This means that the attribute s is more important than the attribute t. However, it is complicated to exactly specify how much term s is important than term t in document clustering. Thus, we define keyphrases as (s, δ) ($\mathbf{w}_s \geq \delta$) and set a large enough value for δ.

2.3. Bregman divergences

For the consideration of expansibility, we incorporate Bregman divergences into our framework. The Bregman divergences [1] include many useful distance metrics, such as squared Euclidean distance, Mahalanobis distance, KL divergence, generalized I-divergence, *etc*.

Definition 1. *Provided that $\phi : \mathcal{S} \to \mathbb{R}$ is a strictly convex function defined on a convex set $\mathcal{S} \subseteq \mathbb{R}^d$ so that ϕ is differentiable on $ri(\mathcal{S})$ (the relative interior of $\mathcal{S}$). The Bregman divergences d_ϕ is defined as*

$$d_\phi(\mathbf{x_i}, \mathbf{x}_j) = \phi(\mathbf{x}_i) - \phi(\mathbf{x}_j) - < \mathbf{x}_i - \mathbf{x}_j, \nabla\phi(\mathbf{x}_j) >$$

where $\nabla\phi$ is the gradient vector of ϕ.

We can obtain different divergences by setting a different function ϕ. Given $\phi(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$, we can have $d_\phi(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)A(\mathbf{x}_i - \mathbf{x}_j)$; or when given $\phi(\mathbf{x}) = \sum_{m=1}^{d} x_m log x_m$, we have $d_\phi(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{d} x_{im} log \frac{x_{im}}{x_{jm}} - \sum_{m=1}^{d}(x_{im} - x_{jm})$.

There are many types of distances for document clustering, such as cosine similarity, KL divergence, generalized I-divergence, *etc*. In order to facilitate solving optimization problem constructed based on metric learning, this

paper considers to use generalized I-divergence as the distance metric. Since I-divergence is not symmetric, we will modify it to "I-divergence to the mean", d_{IM} [4].

$$d_{IM}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{d} x_{im} log \frac{2x_{im}}{x_{im} + x_{jm}} + \sum_{m=1}^{d} x_{jm} log \frac{2x_{jm}}{x_{im} + x_{jm}}$$

Then, we parameterize the above distance by a vector of non-negative weights $\mathbf{w}$:

$$d_{IM_{\mathbf{w}}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{d} w_m x_{im} log \frac{2x_{im}}{x_{im} + x_{jm}} + \sum_{m=1}^{d} w_m x_{jm} log \frac{2x_{jm}}{x_{im} + x_{jm}}$$

3. A Semi-supervised Document Clustering Framework

In this section, we will propose the document clustering framework which incorporates pairwise constraints and keyphrases. Given a document repository and the two kinds of prior knowledge, our approach deals with the problem of effectively incorporating them with the appropriate distance learning. In general, the steps of our approach are as follows:

1. Incorporate pairwise constraints to initialize weights based on metric learning.
2. Add keyphrases to simultaneously learn the two knowledge combining constrained-based and distance-based approaches.
3. Evaluate and select clustering result according to the degree of users' satisfaction.

3.1. Initialize Weights Based on Metric Learning with Pairwise Constraints

Obtaining good initial weights is important to metric leaning, thus we initialize weights according to Halkidi's approach [8]. We construct optimization with pairwise constraints according to Xing's thought [24] so as to make sure must-link pair documents as similar as possible.

$$\min_{\mathbf{w}} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j) - \lambda H(\mathbf{w})$$

$$\textit{subject to:} \quad \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j) \geq 1$$

$$\mathbf{w} \in \mathbb{R}_{+}^{d} \tag{1}$$

3.2. Learn the Two Knowledge Combining Constraint-based and Distance-based Approaches Simultaneously

Through solving the optimization problem (1), we can obtain initial weights $\mathbf{w}_{initial}$. After that, we aim to simultaneously learn the two knowledge and the objective function is as follows:

$$min_{\mathbf{w},\mu,\pi} \frac{1}{n} \sum_{c=1}^{k} \sum_{x_i \in \pi_c} D_{\mathbf{w}}(\mathbf{x}_i, \mu_c) + \lambda_1 \sum_{(s,\delta)\in\mathcal{P}} max(\delta - w_s, 0) +$$

$$\lambda_2 \Phi_{pairwise\_constraints} - \lambda_3 H(\mathbf{w})$$

The first term is an objective clustering validation index, intra-cluster distortion of the clusters $\{\pi_c\}_{c=1}^{k}$; the second term is the penalty term of keyphrases which represents the satisfactory of attribute weights for keyphrases; the third term stands for the penalty term of pairwise constraints; the last term is the regularization term which guarantees the consistence of attribute weights.

The third term includes the penalty of must-link constraints and cannot-link constraints. According to [4], we set $(\sum \phi(\mathbf{x}_i \neq \mathbf{x}_j) D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j)/|\mathcal{S}_{unsat}|)^2$ for the penalty of must-link constraints and $(\sum \phi(\mathbf{x}_i = \mathbf{x}_j)(D_{\mathbf{w}_{max}} - D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j)/|\mathcal{D}_{unsat}|)^2$ for cannot-link constraints.

$$\Phi_{pairwise\_constraints} = (\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \phi(\mathbf{x}_i \neq \mathbf{x}_j) D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j)/|\mathcal{S}_{unsat}|)^2 +$$

$$(\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \phi(\mathbf{x}_i = \mathbf{x}_j)(D_{\mathbf{w}_{max}} - D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j))/|\mathcal{D}_{unsat}|)^2$$

Here, $\phi(True) = 1$ and $\phi(False) = 0$; $\mathbf{x}_i \neq \mathbf{x}_j$ stands for cluster index of $\mathbf{x}_i$ unequal to $\mathbf{x}_j$ ($\mathbf{x}_i \in \pi_c$ and $\mathbf{x}_j \notin \pi_c$), while $\mathbf{x}_i = \mathbf{x}_j$ stands for cluster index of $\mathbf{x}_i$ equal to $\mathbf{x}_j$ ($\mathbf{x}_i \in \pi_c$ and $\mathbf{x}_j \in \pi_c$); $D_{\mathbf{w}_{max}}$ stands for the maximum distance between two arbitrary points for the dataset; $|\mathcal{S}_{unsat}|$ stands for number of unsatisfied must-link constraints while $|\mathcal{D}_{unsat}|$ stands for number of unsatisfied cannot-link constraints. The higher the satisfaction level, the lower the penalty term.

In order to ensure that attribute weights are uniform, we use l_2 *entropy* as the regularization term and set $H(\mathbf{w}) = 1 - \mathbf{w}^T\mathbf{w}$.

There are three variables in the optimization problem, and it is impossible to solve it directly. Thus, we use EM framework to deal with the problem and design three steps. Firstly, given $\{\mu_c\}_{c=1}^{k}$ and $\mathbf{w}$, assign each data point to minimize objective function; then, given $\{\pi_c\}_{c=1}^{k}$, re-calculate cluster centroids $\{\mu_c\}_{c=1}^{k}$; finally, given $\{\pi_c\}_{c=1}^{k}$ and $\{\mu_c\}_{c=1}^{k}$, solve the optimization problem to obtain $\mathbf{w}$. Iterate until convergence.

E-step In simple *k*-means clustering, the E-step assigns each point to the nearest cluster centroid given a certain clustering distance metric. There are

Algorithm 1: The procedure of clustering with the two knowledge

Data: Dataset $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, number of output clusters k, initial weights $\mathbf{w_{initial}}$, must-link constraints $\mathcal{S}$, cannot-link constraints $\mathcal{D}$, and keyphrases $\mathcal{P}$.

Result: Clusters obtained with pairwise constraints and keyphrases.

begin

1. Initialize k cluster representatives $\{\mu_c\}_{c=1}^k$ and set $\mathbf{w} = \mathbf{w_{initial}}$;

2. repeat

2-a. **E-step**: Given $\{\mu_c\}_{c=1}^k$ and $\mathbf{w}$, re-assign data points to clusters to obtain $\{\pi_c\}_{c=1}^k$.

2-b. **M-step(A)**: Given $\{\pi_c\}_{c=1}^k$, re-calculate cluster centroids $\{\mu_c\}_{c=1}^k$.

2-c. **M-step(B)**: Given $\{\pi_c\}_{c=1}^k$ and $\{\mu_c\}_{c=1}^k$, re-estimate $\mathbf{w}$ by solving the optimization problem.

until convergence

3. return $\{\pi_c\}_{c=1}^k$.

also some other methods, such as iterated conditional models (ICM) in [4] that treated objective function as optimization problem to solve, evolutionary algorithm [9] and so on.

When given $\{\mu_c\}_{c=1}^k$ and $\mathbf{w}$, the objective function is transformed into:

$$J_\pi = min_\pi \frac{1}{n} \sum_{c=1}^{k} \sum_{x_i \in \pi_c} D_{\mathbf{w}}(\mathbf{x}_i, \mu_c) + \lambda_2 \Phi_{pairwise\_constraints}$$

Thus, this paper solves the optimization problem by ICM approach to obtain cluster assignments. Firstly, the ICM algorithm sets random order for each point; then, assign each point to the cluster centroid which minimizes the above objective function J_π. Iterate until convergence ($\{\pi_c\}_{c=1}^k$ does not change or J_π dose not obviously decrease between two sequential iterations).

M-step(A) The M-step(A) is one step of the M-step to re-estimate cluster centroids $\{\mu_c\}_{c=1}^k$. [1] has shown each cluster centroid re-estimated in M-step is the arithmetic mean of the points in that cluster. Thus, we calculated cluster centroids in *k*-means clustering with squared Euclidian distance as the formula: $\mu_k^{squared} = \frac{\sum_{\mathbf{x}_i \in \pi_c} \mathbf{x}_i}{|\pi_c|}$. Different from the squared Euclidian distance, given I-divergence, cluster centroids are re-estimated as follows:

$$\mu_k^{IM} = \alpha \frac{\sum_{\mathbf{x}_i \in \pi_c} \mathbf{x}_i}{|\pi_c|} + (1 - \alpha) \frac{1}{n}$$

Here, α ($0 < \alpha < 1$, such as $\alpha = 0.9$) is a smoothing factor to guarantee the denominator of $log \frac{2x_{im}}{x_{im} + \mu_{km}^{IM}}$ in $d_{I_{\mathbf{w}}}(\mathbf{x}_i, \mu_k^{IM})$ is unequal to 0.

M-step(B) The M-step(B) aims to compute weight by solving optimization constructed according to the objective function provided that $\{\pi_c\}_{c=1}^k$ and $\{\mu_c\}_{c=1}^k$ are given.

$$min_{\mathbf{w}} \frac{1}{n} \sum_{c=1}^{k} \sum_{x_i \in \pi_c} D_{\mathbf{w}}(\mathbf{x}_i, \mu_c) + \lambda_1 \sum_{(s,\delta)\in\mathcal{P}} max(\delta - w_s, 0) + \lambda_2 \Phi_{pairwise\_constraints} + \lambda_3 \mathbf{w}^T \mathbf{w}$$

$$\begin{aligned} \textit{subject to:} \quad & for\ each\ important\ word\ p \in P, w_s \geq \delta \\ & w_1 + ... + w_d = 1 \\ & \sum_{(\mathbf{x}_i, \mathbf{x}_j)\in\mathcal{D}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j) \geq 1 \\ & \mathbf{w} \in \mathbb{R}_+^d \end{aligned} \tag{2}$$

The problem (formula 2) is a convex optimization according to [4,18], and there are many effective algorithms to solve the optimization, such as newton method, homogeneous algorithm, active set method an so on [6]. We utilize MOSEK package [6] to solve optimization problems (formula 1 and formula 2).

3.3. Evaluate and Select Clustering Result Based on the Degree of Users' Satisfaction

Degree of users' satisfaction is the portion of knowledge that is satisfied in the clustering result. Many researchers utilize the degree of users' satisfaction to evaluate intermediate results and further improve clustering quality [8,21]. Generally, we think large degree of users' satisfaction can reflect good clustering quality. Thus, in this paper, we wish our approach can effectively incorporate the two knowledge so that the degree of users' satisfaction is satisfied.

$$\begin{aligned} accuracy &= accuracy_{pairwise\ constraints} + accuracy_{important\ words} \\ &= (|sat(\mathcal{S})| + |sat(\mathcal{D})|)/(|\mathcal{S}| + |\mathcal{D}|) + |sat(\mathcal{P})|/|\mathcal{P}| \end{aligned}$$

Here, $sat(*)$ means the satisfied constraints in the set *.

In this paper, the degree of users' satisfaction includes satisfaction of pairwise constraints and keyphrases. We utilize Sun's approach [18] to set parameters in objective function, and all the keyphrases information can be satisfied. Thus, we should lay stress on pairwise constraints. As we integrate two knowledge, our approach should be better than those only incorporating pairwise constraints, and our approach on satisfaction of pairwise constraints should be also better. Even if worse, it should not be much lower.

However, there are many complicated issues when incorporating these two kinds of knowledge, such as the conflicting information, suitable initial cluster centroids [21], *etc*. Thus, our approach is not always optimal on satisfaction of

[6] http://www.mosek.com/

pairwise constraints. As a heuristic, when decrease performance is observed (5% decrease is observed in this paper), we think it is inappropriate to incorporate the two knowledge, and utilize clustering result of metric learning method only with pairwise constraints in section 3.1 as final result.

3.4. Time Complexity

Let N be the number of documents in the collection. The first step includes two parts, constructing and solving the optimization problem with pairwise constraints to obtain new metric and partition documents by utilizing new metric. Time complexity of constructing and solving the optimization problem is related with number of pairwise constraints. Hence, its complexity is estimated to be $O(N)$ [8]. Then, given a clustering algorithm Alg, such as EM hard clustering algorithm utilized in this paper, we can partition the documents by new metric.

The main work of the second step is utilizing a variant EM clustering algorithm to partition documents. Different from the unsupervised version, we utilize ICM approach to assign each document in E-Step, and add M-Step(B) to optimize the distance, solving the optimization problem with two types of knowledge. Let $Complexity(ICM)$ be time complexity of ICM approach, and t be iterations of EM clustering algorithm. Thus, the time cost of the second step is estimated to be $O(Complexity(Alg) + t * (Complexity(ICM) + N))$.

The third step is just a simple comparison, and we can ignore its complexity.

According to the preceding analysis, the complexity of our approach is $O(Com plexity(Alg) + t * Complexity(ICM) + t * N)$. Usually, $t << N$ and the complexity of ICM approach is very low. Hence, the time complexity of our approach mainly depends on the complexity of the clustering algorithm.

4. Experimental Results

In this section, we demonstrate experimental results of our approach comparing with *k*-means, Xing's method [24] and CFP algorithm [18] on *20Newsgroups* collection.

1. *k*-means algorithm is unsupervised and only depends on objective criteria to partitions documents.
2. Xing's method [24] constructs optimization to learn pairwise constraints, and utilizes obtained new metric to partition documents. In this paper, we solve the optimization problem in section 3.1 to obtain new metric.
3. CFP algorithm [18] incorporates keyphrases (attribute order preferences) to assist document clustering. In this paper, we utilize EM framework in section 3.2 to integrate keyphrases while the objective function is as follows:

$$min_{\mathbf{w},\mu,\pi} \frac{1}{n} \sum_{c=1}^{k} \sum_{x_i \in \pi_c} D_{\mathbf{w}}(\mathbf{x}_i, \mu_c) + \lambda_1 \sum_{(s,\delta)\in\mathcal{P}} max(\delta - w_s, 0) - \lambda_3 H(\mathbf{w})$$

4. Our method integrates pairwise constraints and keyphrases into document clustering.

4.1. Datasets and Experimental Settings

We derive 3 datasets from *20Newsgroups* collection. We randomly select 100 documents for each category from original dataset, and derive 3 datasets with 3 categories, News_Different_3 (alt.atheism, rec.sport.baseball, and sci.space) including 3 newsgroup on different topics, News_Related_3 (talk.politics.misc, talk.politics.guns, and talk.politics.mideast) with relevant topics and News_Similar_3 (comp.graphics, comp.os.ms-windows, and comp.windows.x) with large overlap among each category.

We remove stop words, high-frequency and low-frequency words, and express each dataset by TFIDF weighting. Finally, we normalize each dataset so as to avoid impact of document length and make the dissimilarity among documents clearer [23]. Each text vector, $< tf_1 log(\frac{|D|}{df_1}), ..., tf_d log(\frac{|D|}{df_d}) >$, is normalized as follows:

$$< \frac{tf_1 log(\frac{|D|}{df_1})}{\sqrt{(tf_1 log(\frac{|D|}{df_1}))^2 + ... + (tf_d log(\frac{|D|}{df_d}))^2}}, ..., \frac{tf_d log(\frac{|D|}{df_d})}{\sqrt{(tf_1 log(\frac{|D|}{df_1}))^2 + ... + (tf_d log(\frac{|D|}{df_d}))^2}} >$$

As the keyphrases in each derived dataset are few and we want to provide enough keyphrases to assist document clustering, we treat the whole *20Newsgroups* collection as background knowledge, and extract keyphrases from categories that each derived dataset belongs to. In this way, we can obtain many keyphrases, and further select some keyphrases with high word frequency (we select $\lfloor \frac{d}{4} \rfloor$ keyphrases in experiments).

For reliability of experimental results, we make 2-fold cross-validation for each dataset [19,4,5]. We randomly select pairwise constraints from 50% of the dataset, and test methods on remaining 50%. For robustness of experimental results, clustering accuracy is averaged using 10 runs with randomly selected pairwise constraints.

In addition, we set $\lambda_1 = \frac{d}{|\mathcal{P}|}$, $\lambda_2 = 1$ and $\lambda = \lambda_3 = d$ so as to make sure three terms to contribute equally to the objective value [18], $\delta = 4/d$, and $d_{IM} = [\underbrace{1, .., 1}_{d}]$ for $D_{\mathbf{w}_{max}}$ (After normalizing dataset, the value range of attributes becomes [0,1]).

4.2. Evaluation Criteria

In this paper, we utilize two common indexes in document clustering, Purity and Normalized Mutual Information (NMI) to evaluate clustering quality.

Purity measures how close the cluster assignment versus underlying class labels by building one to one correspondence between the clusters and the classes.

$$Purity(\mathcal{C}, \mathcal{B}) = \frac{max_{Map(i) \in [1,..,k]} \sum_{i=1}^{k} n_{i,Map(i)}}{n}$$

Here, $\mathcal{C}$ stands for random variables denoting the clustering assignment while $\mathcal{B}$ presents random variables for the pre-specified class labels. The number of groups in $\mathcal{C}$ and $\mathcal{B}$ are both k. n stands for number of documents in the corpus, and i stands for the cluster index. $Map(i)$ is the class label corresponding to the cluster index i, and $n_{i,Map(i)}$ is the number of documents not only belonging to cluster i but class $Map(i)$.

Normalized Mutual Information (NMI) is an effective index based on information theory.

$$NMI(\mathcal{C},\mathcal{B}) = \frac{I(\mathcal{C};\mathcal{B})}{\sqrt{H(\mathcal{C})H(\mathcal{B})}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} log\frac{n\cdot n_{ij}}{n_i\cdot n'_j}}{\sqrt{\sum_{i=1}^{k} n_i log\frac{n_i}{n}\sum_{j=1}^{k} n'_j log\frac{n'_j}{n}}}$$

Here, n_i presents the document number in the ith cluster of $\mathcal{C}$, n'_j denotes the document number in the jth class of $\mathcal{B}$. n_{ij} denotes the item number included in ith cluster and jth class.

4.3. Results Comparison

Comparison to Other Methods Table 1 and Table 2 demonstrate the result comparisons under the Purity and NMI indexes. Overall, our approach is obviously better than other methods. Especially on the News_Similar_3 dataset, our approach increases 10% under Purity index with a small amount of prior knowledge.

Table 1. Our approach versus Competing methods under Purity index with 30 pairwise constraints (15 must-link constraints and 15 cannot-link constraints) and $\lfloor\frac{d}{4}\rfloor$ keyphrases

Datasets	*k*-means	Xing's method	CFP algorithm	Our method
News_Different_3	0.8160 ± 0.0872	0.9027 ± 0.0412	0.9260 ± 0.0438	0.9400 ± 0.0231
News_Related_3	0.6427 ± 0.0530	0.6800 ± 0.0514	0.6847 ± 0.0655	0.7467 ± 0.0916
News_Similar_3	0.4547 ± 0.0584	0.4747 ± 0.0603	0.5313 ± 0.0595	0.5847 ± 0.0784

Clustering Accuracy versus Constraints We keep the number of keyphrases as $\lfloor\frac{d}{4}\rfloor$ and get results with number of pairwise constraints increasing. The number of pairwise constraints in Fig. 1, 2 and 3, m stands for m must-link constraints and m cannot-link constraints.

Table 2. Our approach versus Competing methods under NMI index with 30 pairwise constraints (15 must-link constraints and 15 cannot-link constraints) and $\lfloor \frac{d}{4} \rfloor$ keyphrases

Datasets	*k*-means	Xing's method	CFP algorithm	Our method
News_Different_3	0.5591 ± 0.0965	0.6919 ± 0.1007	0.7680 ± 0.0765	0.7858 ± 0.0660
News_Related_3	0.3307 ± 0.0850	0.3651 ± 0.0662	0.4311 ± 0.0758	0.4923 ± 0.0829
News_Similar_3	0.0568 ± 0.0410	0.0859 ± 0.0621	0.1188 ± 0.0557	0.1779 ± 0.0744

k-means and CFP algorithm do not incorporate pairwise constraints, their clustering quality should not be affected by pairwise constraints. However, their performances are all unstable. It is mainly due to the initialization of cluster centroids. Even so, as shown in Fig. 1, 2 and 3, CFP algorithm is always much better than *k*-means. It illuminates that incorporating keyphrases extracted from *Title* and *Keywords* can increase document clustering quality.

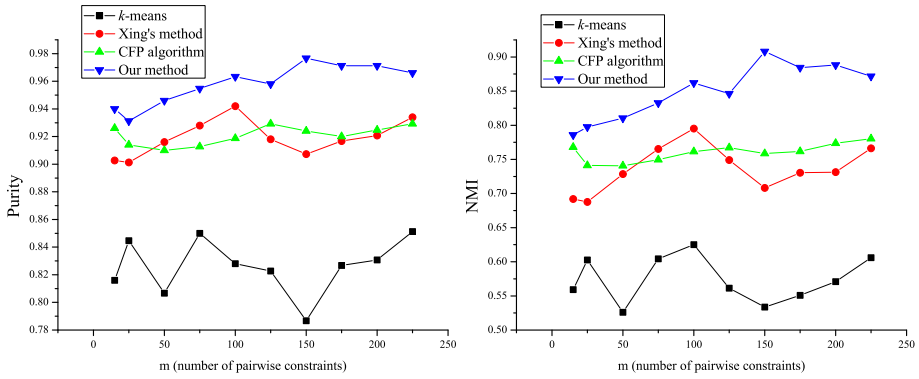

Fig. 1. Clustering accuracy on News_Different_3 with number of pairwise constraints increasing

In Fig. 1, the performance of our approach is obviously better than Xing's method, and even can improves about 10% under NMI index when number of pairwise constraints is few. It is mainly because topics in News_Different_3 dataset are easy to distinguish and keyphrases can effectively reflect topics. The topic of alt.atheism is religion, atheism, *etc.*, rec.sport.baseball is basketball, and sci.space is astrospace, universal gravitation, *etc.* As shown in Table 3, keyphrases of News_Different_3 can be directly matched with corresponding topics. For example, "atheists", "morality", "islamic", "christian", *etc.* should belong to alt.atheism, while "sky" and "moon" belong to sci.space. With the effec-

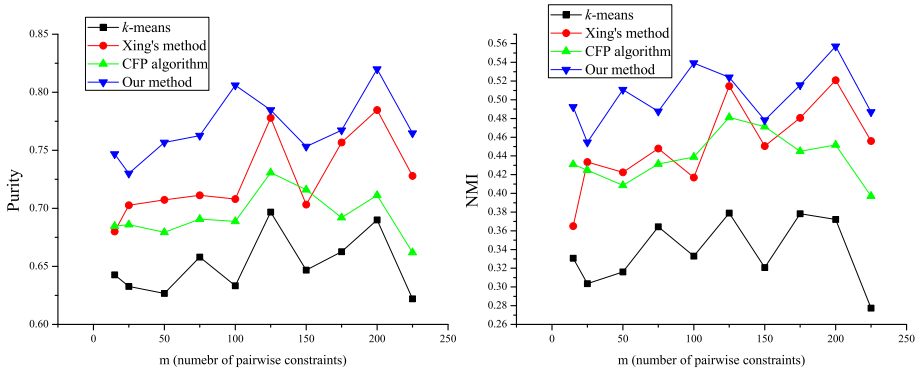

Fig. 2. Clustering accuracy on News_Related_3 with number of pairwise constraints increasing

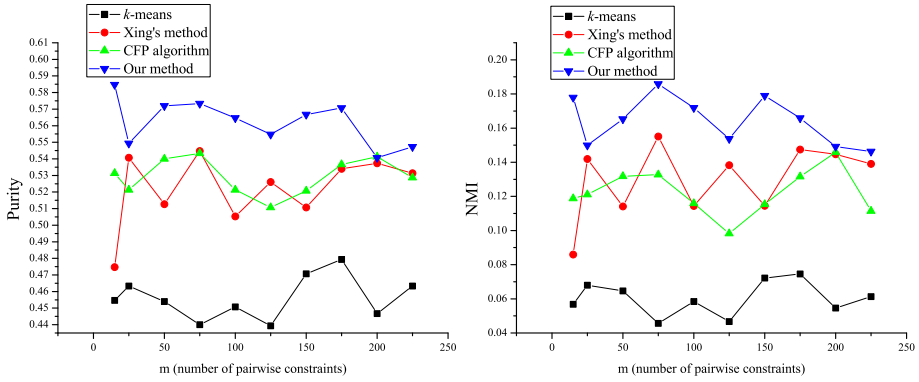

Fig. 3. Clustering accuracy on News_Similar_3 with number of pairwise constraints increasing

tive keyphrases, the performance of CFP algorithm is better than Xing's method under NMI index in most cases.

In Fig. 2 and Fig. 3, our method is slightly better than Xing's method. On the one hand, it is due to correlation and confused topics of the two datasets. There are many related and overlap among three categorization of News_Related_3 (talk.politics.misc, talk.politics.guns, and talk.politics.mideast), such as topic "gun" may appear in each categorization. Similar with News_Related_3, the documents of News_Similar_3 is mainly about computer help problems, and it is hardly to distinguish. On the other hand, keyphrases extracted from *Title* and *Keywords* is also hard to be assigned to corresponding topics. For example, as show in Table 3, keyphrases of News_Similar_3, "help", "do", "file", "problem", *etc*. belong to all the three categorization topics.

Table 3. Top 10 keyphrases (sorted by word frequency) of three datasets

News_Different_3	News_Related_3	News_Similar_3
atheists	waco	help
political	gun	do
morality	atf	dos
islamic	Clinton	window
baseball	burns	microsoft
update	ranch	file
sky	israel	win
moon	survivors	problem
players	gay	ms
christian	israeli	need

Table 4 and Table 5 show the *t*-Test [10] of our approach versus competing methods under Purity and NMI. When the probability is lower than 5%, it demonstrates the robustness of our approach is good and the performance of our method is obviously better than other method; However, when the probability is larger than 5%, it illuminates our approach is similar with other method. As shown in Table 4 and Table 5, our approach is obviously better than *k*-means, Xing's method and CFP algorithm on three datasets.

Table 4. *t*-Test: Our method versus Competing methods under Purity index

	k-means	Xing's method	CFP algorithm
News_Different_3	1.0723e-007	1.5592e-005	1.0886e-005
News_Related_3	1.4329e-007	6.4472e-004	1.1223e-005
News_Similar_3	3.4242e-008	0.0038	1.4409e-004
total	3.7362e-021	7.9132e-010	1.6270e-010

Table 5. *t*-Test: Our method versus Competing methods under NMI index

	k-means	Xing's method	CFP algorithm
News_Different_3	7.8849e-008	1.7325e-005	3.3633e-005
News_Related_3	2.7749e-008	0.0031	1.4376e-004
News_Similar_3	3.3062e-008	0.0062	8.1523e-005
total	4.6540e-013	6.2477e-008	5.5935e-011

Time Complexity Evaluation Fig. 4 shows the time complexity of our approach with respect to the size of collection and the number of pairwise constraints. In Fig. 4(a), we present results using a 3000-dimensional dataset with 100 randomly selected pairwise constraints (50 must-link constraints and 50 cannot-link constraints) and $\lfloor \frac{d}{4} \rfloor$ keyphrases. As shown in Fig. 4(a), the time complexity of our approach is nearly linear to the number of documents in the dataset. In addition, Fig. 4(b) shows the time cost increases linearly with number of pairwise constraints.

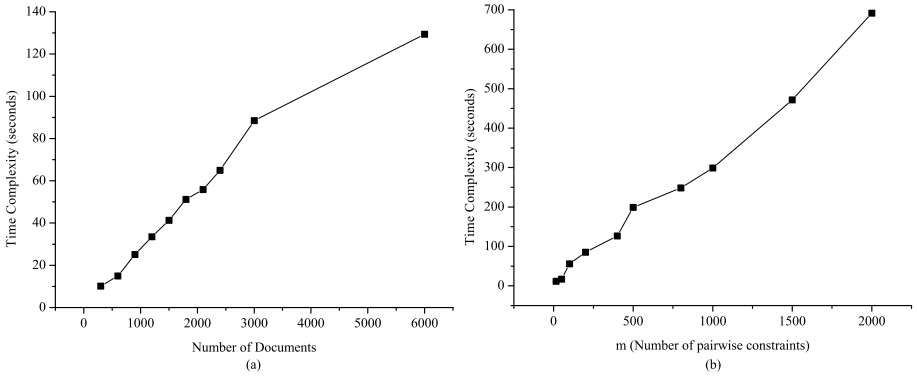

Fig. 4. Time complexity of our approach versus: (a) number of documents; (b) number of pairwise constraints.

5. Conclusion

This paper presents an effective semi-supervised document clustering framework for incorporating pairwise constraints and keyphrases. Our framework initializes attribute weights based on metric learning with pairwise constraints firstly, then simultaneously learn the two knowledge, finally evaluate and select clustering result according to the degree of users' satisfaction. The experimental results validate our method.

Our method can effectively integrate pairwise constraints and keyphrases into document clustering. It not only meets users' need but improve clustering quality. Even with few knowledge, the performance of our method is still satisfied. Moreover, document clustering with keyphrases should be paid much attention to, and its performance is better than clustering with pairwise constraints when keyphrases can effectively reflect document topics.

However, there are many parts to be improved. For simplicity, we set the same value ($\delta = \frac{4}{d}$) for all keyphrases, and it should treat keyphrases according to some criterions, such as word frequency. Secondly, how to solve the contradiction between keyphrases and pairwise constraints should be taken into

account. In addition, we should select suitable center centroids for CFP algorithm and our method, so as to improve the accuracy and robustness.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of P.R.China (No.60802066, No.51005202, No.61004104), the China Postdoctoral Science Foundation (No.20100471494), the Excellent Young Scientist Foundation of Shandong Province of China under Grant (No.2008BS01009), and Deakin CRGS Grant 2011.

References

1. Banerjee, A., Merugu, S., Dhillon, I., Ghosh, J.: Clustering with bregman divergences. The Journal of Machine Learning Research 6, 1705–1749 (2005)
2. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using wikipedia. In: SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 787–788 (2007)
3. Basu, S., Banjeree, A., Mooney, E., Banerjee, A., Mooney, R.J.: Active semi-supervision for pairwise constrained clustering. In: SDM' 04: Proceedings of the 4th SIAM International Conference on Data Mining. pp. 333–344 (2004)
4. Basu, S., Bilenko, M., Mooney, R.J.: A probabilistic framework for semi-supervised clustering. In: KDD '04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 59–68 (2004)
5. Bilenko, M., Basu, S., Mooney, R.J.: Integrating constraints and metric learning in semi-supervised clustering. In: ICML '04: Proceedings of the 21st International Conference on Machine learning. p. 11 (2004)
6. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge Univversity Press (2004)
7. Chen, Y., Rege, M., Dong, M., Hua, J.: Incorporating user provided constraints into document clustering. In: ICDM '07: Proceedings of the 7th IEEE International Conference on Data Mining. pp. 103–112 (2007)
8. Halkidi, M., Gunopulos, D., Vazirgiannis, M., Kumar, N., Domeniconi, C.: A clustering framework based on subjective and objective validity criteria. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(4), 4 (2008)
9. Handl, J., Knowles, J.: An evolutionary approach to multiobjective clustering. IEEE Transactions on Evolutionary Computation 11(1), 56–76 (2007)
10. Hogg, R., Craig, A., McKean, J.: Introduction to mathematical statistics. Macmillan New York (1959)
11. Huang, A., Milne, D., Frank, E., Witten, I.H.: Clustering documents with active learning using wikipedia. In: ICDM '08: Proceedings of the 8th IEEE International Conference on Data Mining. pp. 839–844 (2008)
12. Huang, R., Lam, W.: An active learning framework for semi-supervised document clustering with language modeling. Data and Knowledge Engineering 68(1), 49–67 (2009)
13. Hulth, A., Karlgren, J., Jonsson, A., Bostroem, H., Asker, L.: Automatic keyword extraction using domain knowledge. In: CICLing '01 Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing. pp. 472–482 (2001)

14. Ji, X., Xu, W.: Document clustering with prior knowledge. In: SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 405–412 (2006)
15. Melville, P., Sindhwani, V.: Active dual supervision: reducing the cost of annotating examples and features. In: HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing. pp. 49–57 (2009)
16. Sindhwani, V., Hu, J., Mojsilovic, A.: Regularized co-clustering with dual supervision. In: NIPS '08: Neural Information Processing Systems. vol. 21 (2008)
17. Sindhwani, V., Melville, P., Lawrence, R.D.: Uncertainty sampling and transductive experimental design for active dual supervision. In: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 953–960 (2009)
18. Sun, J., Zhao, W., Xue, J., Shen, Z., Shen, Y.: Clustering with feature order preferences. In: PRICAI '08: Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence. pp. 382–393 (2008)
19. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. In: ICML '01: Proceedings of the 18th International Conference on Machine Learning. pp. 577–584 (2001)
20. Wang, J., Wu, S., Li, G.: An effective semi-supervised clustering framework integrating pair-wise constraints and attribute preferences. Computing and Informatics (in press)
21. Wang, J., Wu, S., Vu, H.Q., Li, G.: Text document clustering with metric learning. In: SIGIR '10: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 783–784 (2010)
22. Witten, I., Paynter, G., Frank, E., Gutwin, C., Nevill-Manning, C.: KEA: Practical automatic keyphrase extraction. In: JCDL '99: Proceedings of the 4th ACM Conference on Digital Libraries. p. 255 (1999)
23. Wu, S., Wang, J., Vu, H.Q., Li, G.: Text clustering with important words using normalization. In: JCDL '10: Proceedings of the 10th Annual Joint Conference on Digital Libraries. pp. 393–394 (2010)
24. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.J.: Distance metric learning with application to clustering with side-information. In: NIPS '02: Neural Information Processing Systems. pp. 505–512 (2002)

Wang, Jinlong is an associate professor at School of Computer Engineering, Qingdao Technological University. He received the Diploma and Ph.D. degree at College of Computer Science and Technology from Zhejiang University, China in 2002 and 2007 respectively. His research interests include data mining, machine learning and artificial intelligence.

Wu, Shunyao is currently a master student at School of Computer Engineering, Qingdao Technological University.

Li, Gang is a lecturer in the School of Information Technology, Deakin University. He completed a PhD in 2005 at Deakin University in the area of data mining, and received the bachelor's degree in computer science from Xi'an Petroleum Institute in 1994, the master by research degree from Shanghai University of Science and technology in 1997. His research interests include data mining, wireless sensor networks and sentiment mining from multimedia.

Wei, Zhe is the ClO of SANYI Heavy Equipment CO. LTD. He received the Ph.D. degree at College of Mechanical Engineering from Zhejiang University, China in 2009. His research interests include mechanical design, product data management and so on.

Received: September 6, 2010; Accepted: February 14, 2011