Orthogonal Nonnegative Matrix Tri-factorization for Semi-supervised Document Co-clustering

Huifang Ma, Weizhong Zhao, Qing Tan, and Zhongzhi Shi

Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100190 Beijing, China
Graduate University of the Chinese Academy of Sciences, 100049 Beijing, China
mahf@ics.ict.ac.cn
http://www.intsci.ac.cn/users/mahuifang/index.html

Abstract. Semi-supervised clustering is often viewed as using labeled data to aid the clustering process. However, existing algorithms fail to consider dual constraints between data points (e.g. documents) and features (e.g. words). To address this problem, in this paper, we propose a novel semi-supervised document co-clustering model OSS-NMF via orthogonal nonnegative matrix tri-factorization. Our model incorporates prior knowledge both on document and word side to aid the new word-category and document-cluster matrices construction. Besides, we prove the correctness and convergence of our model to demonstrate its mathematical rigorous. Our experimental evaluations show that the proposed document clustering model presents remarkable performance improvements with certain constraints.

Keywords: Semi-supervised Clustering, Pairwise Constraints, Word-Level Constraints, Nonnegative Matrix tri-Factorization.

1 Introduction

Providing a meaningful cluster hierarchy to a document corpus has always been a major goal for the data mining community. Approaches to solve this problem have focused on document clustering algorithms, which are widely used in a number of different areas of text mining and information retrieval. One of a latest presented approach for obtaining document cluster is Non-negative Matrix Factorization (NMF) [1], which aimed to provide a minimum error non-negative representation of the term-document matrix. This technique can be considered as co-clustering [2], which aimed to cluster both the rows and columns of the original data simultaneously by making efficient use of the duality between data points (e.g. documents) and features (e.g. words). Put it another way, document clustering and word clustering are performed in a reinforcing manner.

However, traditional clustering algorithms fail to take advantage of knowledge from domain experts. Incorporating the additional information can greatly enhance the performance of clustering algorithms. In recent years, a great amount of effort has been made for clustering document corpus in a semi-supervised way, aiming to cluster the document set under the guidance of some supervisory information.

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 189–200, 2010.

Unfortunately, traditional approaches to semi-supervised document clustering inherently strongly depend on constraints within document themselves while ignore the useful semantic correlation information hidden within the words of the document corpus. We believe that adding word semantic information (such as word clusters indicating word semantics) as additional constraints can definitely improve document clustering performance. Thereafter how to effectively combine both document-level and word-level constraints to guide the process of document clustering is a problem that is definitely worthy of researching.

Based on the above considerations, in this paper, we propose a novel semi-supervised document co-clustering method via non-negative factorization of the term-document matrix for the given document corpus. We have extended the classical NMF approach by introducing both document-level and word-level constraints based on some prior knowledge. Our clustering model encodes the user's prior knowledge with a set of constraints to the objective function, and the document clustering task is carried out by solving a constrained optimization problem. Specifically, we propose a semi-supervised co-clustering framework to cluster the words and documents simultaneously. Meanwhile, we derive iterative algorithms to perform orthogonal non-negative tri-factorization. The correctness and convergence of these algorithms are proved by showing that the solution satisfied the KKT optimality and these algorithms are guaranteed to converge. Experiments performed on various publicly available document datasets demonstrate the superior performance of the proposed work.

The basic outline of this paper is as follows: Section 2 introduces related works. Section 3 presents the semi-supervised orthogonal nonnegative matrix tri-factorization. The experiments and results are given in Section 4. Lastly, we conclude our paper in Section 5.

2 Related Work

This section briefly reviews related work about NMF and semi-supervised document clustering.

The classical NMF algorithms [3] aim to find two matrix factors for a matrix X such that $X \approx WH^T$, where $W^{m\times k}$ and $H^{n\times k}$ are both nonnegative matrices. Ding et al.[4] made systematical analysis of NMF and introduced 3-factor NMF. They demonstrated that the orthogonality constraint leads to rigorous clustering interpretation. When 3-factor NMF is applied to the term-document matrix X, each column X_j of X is an encoding of an original document and each entry x_{ij} of vector X_j is the significance of term i with respect to the semantics of X_j, where i ranges across the terms in the dictionary. Thereafter, Orthogonal NMF factorizes X into three non-negative matrices

$$X = FSG^T, \tag{1}$$

where G is the cluster indicator matrix for clustering of documents of X and F is the word cluster indicator matrix for clustering of rows of X. The simultaneous row/column clustering can be solved by optimizing

$$J = \min_{F\geq 0, S\geq 0, G\geq 0} \left\|X - FSG^T\right\|_F^2 \quad s.t \quad F^TF = I, G^TG = I. \tag{2}$$

The Frobenius norm is often used to measure the error between the original matrix X and its low rank approximation FSG^T. The rank of the approximation, k, is a parameter that must be set by users.

Several formulations of co-clustering problem are proposed in the past decade and they are superior to traditional one-side clustering. Dhillon [2] proposed a bipartite spectral graph partitioning approach to co-cluster words and documents. Long et al.[5] presented a general principled model, called relation summary network to co-cluster the heterogeneous data on a k-partite graph. As for semi-supervised co-clustering algorithms, Chen et al.[6] presented a semi-supervised document clustering model with simultaneous text representation and categorization. Fei et al.[7] proposed a semi-supervised clustering algorithm via matrix factorization. Li et al.[8] presented an interesting word-constrained clustering algorithm. The way of incorporating word constraints is very appealing and sets a good foundation for our model formulation. Even though these semi-supervised algorithms have shown to be superior to traditional clustering method, very little is known about the combination of constraints on both documents and words. One recent work came from Li et al.[9]. They have demonstrated a non-negative matrix tri-factorization approach to sentiment classification with prior knowledge about sentiment words in the lexicon and partial labels on documents.

3 Semi-supervised Orthogonal Nonnegative Matrix Tri-factorization for Co-clustering

In this section, we first describe how we integrate two different constraints in our model in Sect. 3.1. We then derive the OSS-NMF model, prove the correctness and convergence of the algorithm in Sect. 3.2 and Sect. 3.3 respectively.

3.1 Incorporating Document-Level Constraints

Our model treats the prior knowledge on the word side as categorization of words, represented by a complete specification F_0 for F. The prior knowledge on document-level is provided in the form of two sets of pairwise constraints on documents: two documents are known to be highly related and must be grouped into the same document cluster; or two documents that are irrelevant and can not be grouped into the same cluster.

We make use of set A_{ml} to denote that must-link document pairs (d_{i_1}, d_{j_1}) are similar and must be clustered into the same document cluster:

$$A_{ml} = \{(i_1; j_1); \ldots; (i_a; j_a)\}; a = |A_{ml}|. \tag{3}$$

It is easy to demonstrate that the must-link constraints represent equivalence relation. Therefore, we can compute a collection of transitive closures from A_{ml}. Each pair of documents in the same transitive closure must be in the same cluster in the clustering result.

Meanwhile, cannot-link document pairs are collected into another set:

$$B_{cl} = \{(i_1; j_1); \ldots; (i_b; j_b)\}; b = |B_{cl}|, \tag{4}$$

where each pair of documents are considered dissimilar and ought not to be clustered into the same document cluster.

We then encode the must-link document pairs as a symmetric matrix A whose diagonal entries are all equal to one and the cannot-link document pairs as another matrix B.

Suppose each document in the corpus either completely belongs to a particular topic, or is more or less related to several topics. We can then regard these constraints as the document class posterior probability on G. A must-link pair $(i_1; j_1)$ implies that the overlap $g_{i_1 k} g_{j_1 k} > 0$ for some class k, and therefore $\sum_k g_{i_1 k} g_{j_1 k} = (GG^T)_{i_1 j_1}$ should be maximized. The must-link condition can be presented as

$$\max_G \sum_{i,j \in A} (GG^T)_{ij} = \sum_{ij} A_{ij} (GG^T)_{ij} = TrG^T AG. \tag{5}$$

In terms of cannot-link pairs $(i_2; j_2)$, $g_{i_2 k} g_{j_2 k} = 0$ for all k. Likewise, we take the cannot-link constraints and minimize $\sum_k g_{i_2 k} g_{j_2 k} = (G^T G)_{i_2 j_2}$. Since g_{ik} are nonnegative, we write this condition as:

$$\sum_{i,j \in B} (GG^T)_{ij} = TrBGG^T = 0, \quad or \quad \min_G TrG^T BG. \tag{6}$$

3.2 Algorithm Derivation

Combining the above constraints together, we define the objective function of OSS-NMF as:

$$J = \min_{F \geq 0, S \geq 0, G \geq 0} ||X - FSG^T|| \; + \; \alpha \left\| F - F_0 \right\|_F^2 + Tr(-\beta GAG^T + \gamma GBG^T),$$
$$s.t.\ FF^T = I, GG^T = I, \tag{7}$$

where α, β and γ are positive trade-off parameters that control the degree of enforcement of the user's prior knowledge. The larger value the parameters take, the stronger enforcement of the users prior knowledge we will have; vise versa.

An iterative procedure to solve the optimization problem in Eq.(7) can be summarized as follows.

Computation of S. Optimizing Eq.(7) with respect to S is equivalent to optimizing

$$J_1 = \min_{F \geq 0, S \geq 0, G \geq 0} \left\| X - FSG^T \right\|_F^2. \tag{8}$$

Setting $\frac{\partial J_1}{\partial S} = 0$ leads to the following updating formula:

$$S_{ik} = S_{ik} \sqrt{\frac{(F^T XG)_{ik}}{(F^T FSG^T G)_{ik}}}. \tag{9}$$

Computation of F. Optimizing Eq.(7) with respect to F is equivalent to optimizing

$$J_2 = \min_{F\geq 0, S\geq 0, G\geq 0} \left\|X - FSG^T\right\|_F^2 + \alpha \left\|F - F_0\right\|_F^2, \quad s.t. \quad FF^T = I. \tag{10}$$

We present an iterative multiplicative updating solution. After introducing the Lagrangian multiplier, the Lagrangian function is stated as

$$L(F) = \left\|X - FSG^T\right\|_F^2 + \alpha \left\|F - F_0\right\|_F^2 + Tr[\lambda_1(F^TF - I)]. \tag{11}$$

This takes the exact form as Li demonstrated in [8], thereby we can update F as follows:

$$F_{ik} = F_{ik}\sqrt{\frac{(XGS^T + \alpha F_0)_{ik}}{(FF^TXGS^T + \alpha FF^TF_0)_{ik}}}. \tag{12}$$

Computation of G. Optimizing Eq.(7) with respect to G is equivalent to optimizing

$$J_3 = \min_{F\geq 0, S\geq 0, G\geq 0} \left\|X - FSG^T\right\|_F^2 + Tr(-\beta G^TAG + \gamma G^TBG), \quad s.t. \quad GG^T = I. \tag{13}$$

Similar with the computation of F, we introduce the Lagrangian multiplier, thus the Lagrangian function is

$$L(G) = \left\|X - FSG^T\right\|^2 + Tr(-\beta G^TAG + \gamma G^TBG) + Tr[\lambda_2(G^TG - I)]. \tag{14}$$

We show that G can be iterated as:

$$G_{ik} = G_{ik}\sqrt{\frac{(X^TFS + \beta AG)_{ik}}{(G(SF^TFS^T + \lambda_2) + \gamma BG)_{ik}}}. \tag{15}$$

The detailed analysis of computation of G is shown in the optimization section. When the iteration starts, we update one factor with others fixed.

3.3 Algorithm Correctness and Convergence

To prove the correctness and convergence of our algorithm, we will make use of optimization theory, matrix inequalities and auxiliary functions that used in [3].

Correctness

Theorem 1. *If the update rule of S, F and G in Eq.(9), Eq.(12) and Eq.(15) converge, then the final solution satisfies the KKT optimality condition, i.e., the algorithm converges correctly to a local optima.*

Proof: Following the standard theory of constrained optimization, we introduce the Lagrangian multipliers λ_1, λ_2 and construct the following Lagrangian function:

$$\begin{aligned} L &= \|X - FSG^T\| + \alpha\|F - F_0\| + Tr[\lambda_1(F^TF - I)] \\ &\quad + Tr[-\beta GAG^T + \gamma GBG^T + \lambda_2(G^TG - I)] \\ &= Tr[X^TX - 2G^TX^TFS + G^TGSF^TFS + \alpha(FF^T - 2FF_0^T + F_0F_0^T)] \\ &\quad - \beta G^TAG + \gamma G^TBG + \lambda_1(F^TF - I) + \lambda_2(G^TG - I)]. \end{aligned} \tag{16}$$

The correctness of updating rules for S in Eq.(9) and F in Eq.(12) have been proved in [8]. Therefore, we only need to proof the correctness of updating rules for G. Fixing F, S, we can get that the KKT complementary condition for the non-negativity of G

$$[-2X^TFS + 2G(SF^TFS^T + \lambda_2) - 2\beta AG + 2\gamma BG]_{ik}G_{ik} = 0. \tag{17}$$

We then obtain the Lagrangian multiplier, it is obvious that at convergence the solution satisfy

$$[-2X^TFS + 2G(SF^TFS^T + \lambda_2) - 2\beta AG + 2\gamma BG]_{ik}G_{ik}^2 = 0. \tag{18}$$

We can see that this is identical to the KKT condition. The above equation denotes that either the first factor equals to zero, or G_{ik} is zero. If the first factor is zero, the two equations are identical. If G_{ik} is zero, then G_{ik}^2 is zero as well, vice versa. Thus, we have proved that if the iteration converges, the converged solution satisfies the KKT condition, i.e., it converges correctly to a local minima.

Proof is completed.

Convergence. We demonstrate that the above objective function decreased monotonically under these three updating rules. Before we proof the convergence of the algorithm, we need to construct the auxiliary function similar to that used in Lee and Seung [3]. We first introduce the definition of auxiliary function.

Definition 1. *A function $Z(H, H')$ is called an auxiliary function of $L(H)$ if it satisfies*

$$Z(H, H') \geq L(H), \quad Z(H, H) = L(H). \tag{19}$$

Lemma 1. *If $Z(H, H')$ is an auxiliary function, then L is non-increasing under the update*

$$H^{(t+1)} = \arg\min_H Z(H, H^t). \tag{20}$$

By construction $L(H^{(t)}) = Z(H^{(t)}, H^{(t)}) \geq Z(H^{(t+1)}, H^{(t)}) \geq L(H^{(t+1)})$, $L(H^{(t+1)})$ is monotonic decreasing (non-increasing).

Lemma 2. *For any nonnegative matrices $A \in R^{n\times n}$,$B \in R^{k\times k}$,$S \in R^{n\times k}$,$S' \in R^{n\times k}$, A, B are symmetric, the following inequality holds[10]:*

$$\sum_{i=1}^{n}\sum_{p=1}^{k}\frac{(AS'B)_{ip}S_{ip}^2}{S'_{ip}} \geq tr(S^TASB). \tag{21}$$

Theorem 2. *The above iterative algorithms converge.*

Proof: To proof the algorithm converges, the key step is to find an appropriate auxiliary function $Z(G, G')$ of L(G) in Eq.(14). We show that the following function

$$Z(G,G') = \sum_{ik} [-2G'_{ik}(1+\log\frac{G_{ik}}{G'_{ik}})(X^TFS)_{ik} + \frac{[G'(SF^TFS+\lambda_2)]_{ik}G^2_{ik}}{G'_{ik}} - \beta G'_{ik}(AG')_{ik}(1+\log\frac{G^2_{ik}}{G'_{ik}}) + \gamma\frac{(BG')_{ik}G^2_{ik}}{G'_{ik}}]. \quad (22)$$

is its corresponding auxiliary function.

First, it is obvious that when $G = G'$, the equality holds. Second, the inequality holds $Z(G, G') \geq L'(G)$. This is based on the following: a) The first term and third term in $Z(G, G')$ are always smaller than the corresponding terms in $L'(G)$ because of the inequality $z \geq 1 + \log(z) \quad \forall z > 0$; b) The second and last term in Eq.(24) are always bigger than the corresponding terms in $L'(G)$, due to Lemma 2. Putting these together, we can guarantee that $Z(G, G') \geq L'(G)$.

To find the minimum of $Z(G, G')$, we take

$$\frac{\partial Z(G,G')}{\partial G_{ik}} = \sum_{ik} [-2\frac{G'_{ik}}{G_{ik}}(X^TFS)_{ik} + 2\frac{[G'(SF^TFS+\lambda_2)]_{ik}G_{ik}}{G'_{ik}} - 2\beta\frac{G'_{ik}(AG')_{ik}}{G_{ik}} + 2\gamma\frac{(BG')_{ik}G_{ik}}{G'_{ik}}] \quad (23)$$

and the Hessian matrix of $Z(G, G')$

$$\frac{\partial^2 Z(G,G')}{\partial G_{ik}\partial G_{jl}} = \sum_{ik} [2\frac{G'_{ik}}{G^2_{ik}}(X^TFS)_{ik} + 2\frac{[G'(SF^TFS+\lambda_2)]_{ik}}{G'_{ik}} + 2\beta\frac{G'_{ik}(AG')_{ik}}{G^2_{ik}} + 2\gamma\frac{(BG')_{ik}}{G'_{ik}}]\delta_{ij}\delta_{kl} \quad (24)$$

is a diagonal matrix with positive diagonal elements.

Thus $Z(G, G')$ is a convex function of G. Therefore, we can obtain the global minimum of Z. The minimum value is obtained by setting $\frac{\partial Z(G,G')}{\partial G_{ik}} = 0$, we get

$$\frac{G'_{ik}}{G_{ik}}(X^TFS+\beta AG)_{ik} = \frac{G_{ik}}{G'_{ik}}(G'(SF^TFS^T)+\lambda_2+\gamma BG')_{ik}. \quad (25)$$

We can thereafter derive the updating rule of Eq.(16)

$$G_{ik} = G_{ik}\sqrt{\frac{(X^TFS+\beta AG)_{ik}}{(G(SF^TFS^T+\lambda_2)+\gamma BG)_{ik}}}. \quad (26)$$

Under this updating rule, $L'(G)$ decreases monotonically, where the Lagrangian multiplier k-by-k matrix λ_2 for enforcing the orthogonality and $G^TG = I$ is given by

$$\lambda_2 = G^TX^TFS + \beta G^TAG - \gamma G^TBG - SF^TFS^T. \quad (27)$$

Proof is completed.

4 Experiments

This section provides empirical evidence to show the benefits of our model OSS-NMF. We compared our method with Constrained-Kmeans[11], Information-Theoretic Co-clustering, which is referred to as IT-Co-clustering[12], ONMF-W denoting Orthogonal NMF with word-level constraints[8], ONMF-D representing Orthogonal NMF with document-level constraints. Constrained K-means is the representative semi-supervised data clustering method; Information-Theoretic Co-clustering is one of the most popular co-clustering method; ONMF-W and ONMF-D are two derived algorithms from our approach.

The requirement of word constraints is the specification of word categorization. Similar with Li [8], we took advantage of the ACM term taxonomy, which come naturally and strictly decide the taxonomy of computer society. The document-level constraints were generated by randomly selecting pairs of documents. If the labels of this document pair are the same, then we generated a must link. In contrast, if the labels are different, a cannot link is generated. The amounts of constraints were determined by the size of input data. Incorporating dual constraints on our model, we believe that our approach should perform better given reasonable amount of labeled data.

4.1 Datasets

Three different datasets widely used as benchmark data sets in clustering literature were used.

Citeseer dataset: Citeseer collection was made publicly available by Lise Getoor's research group at University of Maryland. We end up with a sampling of Citeseer data containing 3312 documents. These data are classified into one of the following six classes: Agents, Artificial Intelligence, Data Base, Information Retrieval, Machine Learning, Human Computer Interaction.

DBLP Dataset: This dataset is downloaded from DBLP Computer Science Bibliography spanning from 1999 to 2004. We extract the paper titles to form our dataset from 5 categories, which contains 2170 documents.

URCS Technical Reports: This dataset is composed of abstracts of technical reports published in the Department of Computer Science at Rochester University. There are altogether 512 reports abstracts grouped according to 4 categories.

We pre-processed each document by tokenizing the text into bag-of-words. Then we applied stopwords removing and stemmed words. In particular, words that occur in less than three documents are removed. We used the weighted term-frequency vector to represent each document.

4.2 Evaluation Metrics

We adopt the clustering accuracy and normalized mutual information as our performance measures. These performance measures are standard measures widely

used for clustering. Clustering accuracy measures the cluster performance from the one-to-one relationship between clusters and classes point of view, which is defined as:

$$Acc = \max \frac{\sum_{i=1}^{N} \delta(map(r_i), d_i)}{N}, \tag{28}$$

where r_i denotes the cluster label of a document and d_i denotes the true class label, N is the total number of documents, $\delta(x, y)$ is a function which equals one if $x = y$ and equals zero otherwise, $map(r_i)$ is the permutation function which maps each cluster label to the corresponding label of the data set.

NMI measures how closely the clustering algorithm could reconstruct the underlying label distribution in the data. It is defined as:

$$NMI = \frac{I(Z'; Z)}{(H(Z') + H(Z))/2}, \tag{29}$$

where $I(Z'; Z) = H(Z) - H(Z|Z')$ is the mutual information between the random variables Z' and Z, $H(Z)$ is the Shannon entropy of Z, and $H(Z|Z')$ is the conditional entropy of Z given Z'. In general, the larger the NMI value is, the better the clustering quality is.

4.3 Clustering Results

Considering the document constraints are generated randomly, we run each algorithm 20 times for each dataset and took the average as statistical results. To give these algorithms some advantage, we set the number of clusters equal to the real number of all the document clusters and word clusters.

Overall Evaluation. Table 1 shows the cluster accuracy and normalized mutual information of all the algorithms on all the data sets. From the experimental comparisons, we observe that our proposed method OO-SNMF effectively combined prior knowledge from the word side with constraints on the document side for improving clustering results. Moreover, our model outperforms most of the clustering methods on all the data sets. In summary, the experimental results match favorably with our hypotheses and encouraged us to further explore the reasons.

The superiority of our model arises in the following three aspects: (1) the mechanism of tri-factorization for term-document matrix allows setting different classes of terms and documents, which is in line with the real applications; (2) co-clustering the terms and documents with both constraints leads to improvement in the clustering of documents; (3) last but not least, the constraints on word-level are quite different from that of document-level, which means our model can incorporate distinguished semantic information on both sides for clustering.

Effect of the Size of Words. In this section, we describe the effect of the size of words on clustering. These words can be used to represent the underlying

Table 1. Comparison of four algorithms on different datasets

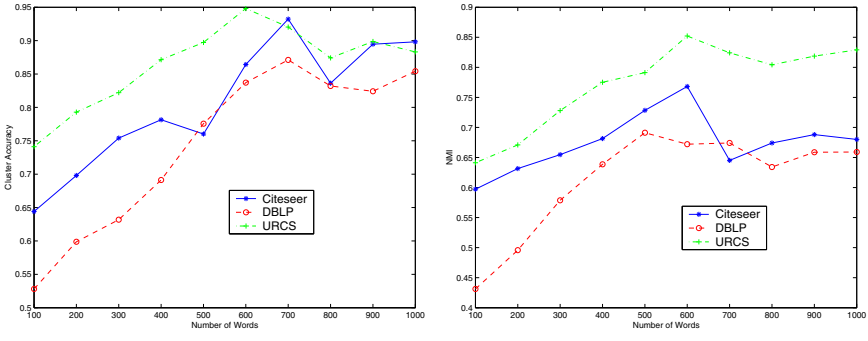
(a) Clustering Accuracy

Data Sets	Citeseer	DBLP	URCS
Constrained-Kmeans	0.5124	0.4215	0.5923
IT-Co-clustering	0.5765	0.4873	0.6214
ONMF-W	0.5514	0.4812	0.6052
ONMF-D	0.6142	0.5321	0.6812
OSS-NMF	**0.7235**	**0.6823**	**0.8368**

(b) Normalized Mutual Information

Data Sets	Citeseer	DBLP	URCS
Constrained-Kmeans	0.5813	0.5312	0.6358
IT-Co-clustering	0.6521	0.5821	0.7389
ONMF-W	0.6722	0.6312	0.7548
ONMF-D	0.7214	0.6523	0.7964
OSS-NMF	**0.8345**	**0.7643**	**0.9124**

'concept' of the corresponding category cluster. We follow the term frequency criteria to select word. The performance results with different numbers of words on all of the datasets are demonstrated.



(a) Cluster Accuracy with different numbers of words on 3 dataset.

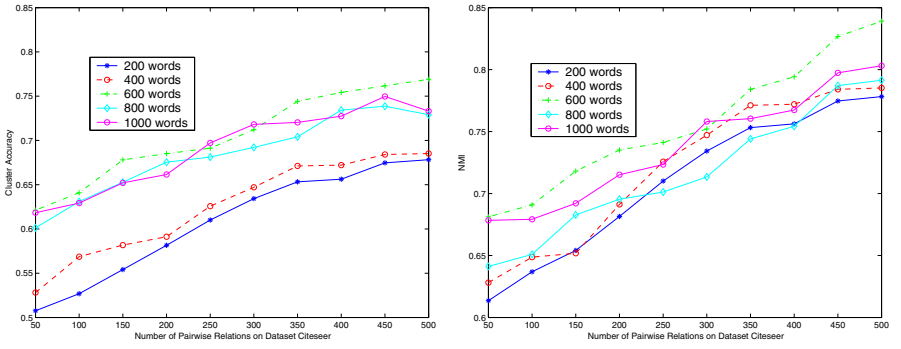
(b) NMI with different numbers of words on 3 dataset.

Fig. 1. Accuracy and NMI results with different numbers of words on 3 dataset

Both Accuracy and NMI show clear benefits of having more words: the performance increases as the amount of words grows, as shown in Fig.1. This indicates the addition of word semantic information can greatly help the clustering performance. It also shows a great variation with the increase of words. When the size of words increases beyond a certain value, the quality of clustering fluctuates and suddenly drops and then becomes stable.

Experiments on Pairwise Constraint of Documents. We conducted experiments for our framework by varying the number of pairwise constraints and size of words. Results from all these document collections indicate that generally as more and more constraints are added to the dataset being clustered, the performance of the clustering method becomes better, confirming previous discussion on the effect of increase of more labeled data. Due to the limitation of this paper, we only present NMI and Cluster Accuracy on Citeseer in Fig.2.

Our finding can be summarized as follows: (1) As long as the constraints are provided, our model always outperforms the traditional constrained methods. (2) The model performs much better with the increase of constraints.



(a) Cluster Accuracy on Citeseer (b) NMI results on Citeseer

Fig. 2. Accuracy and NMI results with different numbers of words and pairwise documents on Citeseer

5 Conclusions and Future Work

In this paper, we consider the problem of semi-supervised document co-clustering. We have presented a novel orthogonal semi-supervised nonnegative matrix tri-factorization model. We also have provided theoretical analysis of the correctness and convergence of the algorithm. The ability of our proposed algorithm to integrate double constraints makes it efficient for document co-clustering.

Our work leads to several questions. We incorporated the word prior knowledge as a specification of the initial word cluster. It would also be interesting to make use of pairwise constraints on the word side. In particular, a further interesting direction is to actively select informative document pairs for obtaining user judgments so that the clustering performance can be improved with as few supervised data as possible.

Acknowledgments. This work is supported by the National Science Foundation of China (No. 60933004, 60903141, 60775035), the National Basic Research Priorities Programme (No. 2007CB311004), 863 National High-Tech Program (No.2007AA-01Z132), and National Science and Technology Support Plan (No.2006BAC08B06).

References

1. Xu, W., Liu, X., Gong, Y.: Document Clustering Based on Non-negative Matrix Factorization. In: Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, pp. 267–273 (2003)
2. Dhillon, I.S.: Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, pp. 269–274 (2001)

3. Lee, D., Seung, H.S.: Algorithms for Non-negative Matrix Factorization. In: Proceedings of 15th Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, vol. 13, pp. 556–562 (2001)
4. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal Nonnegative Matrix Trifactorizations for Clustering. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, pp. 126–135 (2006)
5. Long, B., Zhang, Z., Wu, X., Yu, P.S.: Spectral Clustering for Multi-type Relational Data. In: Proceedings of 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, USA, pp. 585–592 (2006)
6. Chen, Y.H., Wang, L.J., Dong, M.: Semi-supervised Document Clustering with Simultaneous Text Representation and Categorization. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS (LNAI), vol. 5781, pp. 211–226. Springer, Heidelberg (2009)
7. Wang, F., Li, T., Zhang, C.S.: Semi-Supervised Clustering via Matrix Factorization. In: Proceedings of The 8th SIAM Conference on Data Mining, Atlanta, Geogia, pp. 1–12 (2008)
8. Li, T., Ding, C., Zhang, Y., Shao, B.: Knowledge Transformation from Word Space to Document Space. In: Proceedings of the 31st Annual International ACM SIGIR conference on research and development in information retrieval, Singapore, pp. 187–194 (2008)
9. Li, T., Zhang, Y., Sindhwani, W.: A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, pp. 244–252 (2009)
10. Ding, C.H., Li, T., Jordan, M.I.: Convex and Semi-nonnegative Matrix Factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence 99(1), 195–197 (2008)
11. Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S.: Constrained K-means Clustering with Background Knowledge. In: Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA, USA, pp. 577–584 (2001)
12. Dhillon, I., Mallela, S., Modha, D.S.: Information-Theoretic Co-clustering. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, pp. 89–98 (2003)