International Journal of Computer Science and Engineering Communications
Volume.4, Issue.5 (2016): Page.1459-1466
www.scientistlink.com/ijcsec

Co-Clustering the Documents and Words Using K-Means Clustering

M.Indirani[1],P.Revathi[2]
[1,2]Assistant Professor
Department of Information Technology
Hindusthan College of Engineering and Technology, Coimbatore, India

Abstract: *In this paper, we propose a novel constrained coclustering method to achieve two goals. First, we combine information theoretic coclustering and constrained clustering to improve clustering performance. Second, we adopt both supervised and unsupervised constraints to demonstrate the effectiveness of our algorithm. The unsupervised constraints are automatically derived from existing knowledge sources, thus saving the effort and cost of using manually labeled constraints. To achieve our first goal, we develop a two-sided hidden Markov random field (HMRF) model to represent both document and word constraints. We then use an alternating expectation maximization (EM) algorithm to optimize the model. We also propose two novel methods to automatically construct and incorporate document and word constraints to support unsupervised constrained clustering: 1) automatically construct document constraints based on overlapping named entities (NE) extracted by an NE extractor; 2) automatically construct word constraints based on their semantic distance inferred from 20 Newsgroups Data set. The results of our evaluation using 20 Newsgroups demonstrate the superiority of our approaches against the k-Means similarity approaches.*

Keywords : *Datamining, Clustering ,K-means clustering and Pattern set mining*

1 INTRODUCTION

1.1 DATA MINING

Data mining (the analysis step of the "Knowledge Discovery and Data Mining" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

1.2 PROBLEM DEFINITION

When clustering textual data, one of the most important distance measures is document similarity. Since document similarity is often determined by word similarity, the semantic relationships between words may affect document clustering results. Moreover, the relationships among vocabularies such as synonyms, antonyms, hypernyms, and hyponyms, may also affect the computation of document similarity. In existing techniques there is no provision to use the additional knowledge on documents and words to facilitate document clustering.

1.3 REQUIREMENTS

Functional Requirements is the specification of the function that the system must support, whereas non- functional requirements are the constraint on the operation of the system that is not related directly to function of system.

1.3.1 Functional Requirements

Inputs: The major inputs for the Concurrent Clustering Interface are the 20 News Group Document sets and Clusters
Output: The outputs are the accessing the cluster similarity with EM Algorithms.

1.3.2 Performance Requirements

Performance is measured in terms of ease of use of user interface.

2 LITERATURE SURVEY

1. **Annotating Search Results from Web Databases**:

In this paper, a data unit is a piece of text that semantically represents one concept of an entity. It corresponds to the value of a record under an attribute. It is different from a text node which refers to a sequence of text surrounded by a pair of HTML tags. It also perform data unit level annotation. There is a high demand for collecting data of interest from multiple WDBs. For example, once a book comparison shopping system collects multiple result records from different book sites, it needs to determine whether any two SRRs refer to the same book. If ISBNs are not available, their titles and authors could be compared. The system also needs to list the prices offered by each site. Thus, the system needs to know the semantic of each data unit. Unfortunately, the semantic labels of data units are often not provided in result pages. In this paper, we consider how to automatically assign labels to the data units within the SRRs returned from WDBs. In addition, a clustering-shift algorithm is introduced in this paper to explicitly handle the one-to nothing relationship between text nodes and data units while the previous version has a pure clustering algorithm.

2. k-Pattern Set Mining under Constraints:

In this paper, it made a first step toward formally specifying pattern set mining problems and solving them by means of general algorithms. It developed a framework in which a multitude of tasks, including concept-learning,conceptual clustering, redescription mining and tiling, can be formalized. The main idea in this framework is to formalize mining tasks as problems of finding k patterns that together satisfy constraints. In contrast to earlier approaches, where constraints are typically only formalized on the local level, that is, on individual patterns, within this framework we also formalize constraints on the global level, that is on the pattern set as a whole. Both levels of constraints are formalized at the same time, that is, in a single specification; we present a high-level modeling language, independent from underlying frameworks, and show how to use it to formulate many well known tasks. A key feature is hence that we open up the possibility that mining problems are not solved in multiple steps, but also in one single step. Constraint programming is a generic framework for solving combinatorial and optimization problems under constraints. It has been used successfully in numerous applications, including constraint-based mining of individual patterns. The key power of CP lies in its generic approach to problem solving: users model a problem by specifying constraints, and the CP solver will use those constraints to find the solutions. This has the advantage that new problems can be solved by only changing the specification in terms of constraints; a new solver is not needed. A potential issue is that CP solves problems using exhaustive search. A CP system will also attempt to solve the mining problem in one exhaustive search step.

3. Semi-Supervised Clustering via Matrix Factorization

Given the inter-type relationship information and intra-type relationship constraints, we propose a general constrained co-clustering framework to cluster the multiple type data points simultaneously. It show that the traditional semi-supervised clustering methods are special cases of our framework when the data set is of only one single type. Finally the experimental results on several real world data sets are presented to show the effectiveness of our method. The information is considered in the form of pair wise constraints on the same type of data objects.

4. Text Classification from Labeled and Unlabeled Documents using EM

This paper shows that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled documents. This is important because in many text classification problems obtaining training labels is expensive, while large quantities of unlabeled documents are readily available. It introduce an algorithm for learning from labeled and unlabeled documents based on the combination of Expectation-Maximization (EM) and a naive Bayes classifier. It present two extensions to the algorithm that improve classification accuracy under these conditions:

(1) A weighting factor to modulate the contribution of the unlabeled data.

(2) The use of multiple mixture components per class It uses Expectation-Maximization (EM) to learn classifiers that take advantage of both labeled and unlabeled data. EM is a class of iterative algorithms for maximum likelihood or maximum a posteriori estimation in problems with incomplete data (Dempster, Laird, & Rubin, 1977). The unlabeled data are considered incomplete because they come without class labels. The algorithm first trains a classifier with only the available labeled documents, and uses the classifier to assign probabilistically-weighted class labels to each unlabeled document by calculating the expectation of the missing class labels. It then trains a new classifier using all the documents both the originally labeled and the formerly unlabeled and iterates. The paper combine EM with naive Bayes, a classifier based on a mixture of multinomial, that is commonly used in text classification. We also propose two augmentations to the basic EM scheme.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

1. Kullback-Leibler KL divergence:

This is for clustering multivariate normal distributions based upon the symmetric, Kullback-Leibler divergence. Optimal mean vector and covariance matrix of the centroid normal distribution are derived and a set of Riccati matrix equations is used to find the optimal covariance matrix. The solutions are found iteratively by alternating the intermediate mean and covariance solutions. Clustering performance of the new algorithm is shown to be superior to that of non-optimal sample mean and covariance solutions. It achieves a lotheyr overall distortion and flatter distributions of pdf samples across clusters. The resultant optimal clusters theyre further tested on the Wall Street Journal database for adapting HMM parameters in a Structured Maximum A Posterior Linear Regression (SMAPLR) framework.

2. Nonnegative matrix factorization (NMF)

In most IR clustering problems, they directly cluster the documents, working in the document space, using cosine similarity between documents as the similarity measure. In many real-world applications, however, they usually have knowledge on the word side and wish to transform this knowledge to the document (concept) side. They provide a mechanism for this knowledge transformation. To the best of our knowledge, this is model for such type of knowledge transformation. This model uses a nonnegative matrix factorization model X= FSGT , where X is the word document semantic matrix, F is the posterior probability of a word belonging to a word cluster and represents knowledge in the word space, G is the posterior probability of a document belonging to a document cluster and represents knowledge in the document space, and S is a scaled matrix factor which provides a condensed view of X. They show how knowledge on words can improve document clustering, i.e, knowledge in the word space is transformed into the document space.

3.1.1 DISADVANTAGES

1. The existing semi-supervised methods are applicable to 1D clustering only. For text data, these techniques cannot show the relationship between document and word clusters, but also leverage the knowledge transferred between the two sides.
2. They do not include document and word constraints.
3. The results are not optimized.

3.2 PROPOSED SYSTEM

In this work, we utilize Kmeans to initialize the document and word clusters. Before we apply this method, however, we need to initialize Kmeans first. To make the Kmeans algorithm more stable for document and word clustering, we employ a farthest-first traversal method. It aims to find K data points that are maximally separated from each other. In our implementation, at the beginning of initialization, we randomly select a data point as the first cluster center. Then, to identify a new center, we choose a data point that has not been selected previously using the following procedure. We first compare the distances between a candidate data point and all the previously selected centers, and record the minimal distance between this point and the centers. Then the candidate point with the largest minimum distance is selected as the new center. Finally, K centers are selected to initialize the cluster centers of Kmeans. With Kmeans results this system proposes a new constrained coclustering algorithm . CITCC: In addition two novel methods to automatically construct and incorporate constraints into CITCC to help improve document clustering performance. Since both the constraints are automatically constructed by the system, it performs purely unsupervisedocument clustering. Here the system automatically constructs document constraints based on the overlapping named entities extracted by an NE extractor; Then automatically word constraints are formed based on their semantic distance inferred from 20 Newsgroup.

3.2.1 ADVANTAGES

1. It performs better than the existing coclustering algorithms because it allows the system to incorporate additional constraints to guide the clustering towards the ground-truth
2. It performs better than the existing 1D constrained clustering methods since it can take advantage of the co-occurrences of documents and words;
3. It performs better than the existing constrained coclustering approaches on text data since it optimizes a KL-divergence based objective function versus a Euclidean distance-based function that is commonly used by other systems.

4. SYSTEM DESIGN AND IMPLEMENTATION

4.1 SYSTEM DESIGN

System Design is the process of making the newly designed system fully operational and consistent in performance. The following steps have been followed in the implementation of the system.

4.2 SYSTEM IMPLEMENTATION

System implementation is the important stage of project when the theoretical design is tuned into practical system. The main stages in the implementation are as follows:

- Planning
- Training

4.2.1.1.1 MODULE DESCRIPTION

This system consists of six modules.
Dataset Preprocessing
K-means clustering
Document EM
Word EM
Final clustering
Performance Analysis

4.3.1 DATASET PREPROCESSING

4.3.2

We had used 20 News Group Dataset from the Text Retrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. The raw text files are downloaded from http://trec.nist.gov/ . It's then preloaded with C# string builder for Data pre trimming such separation of records from single file entity, removal of fixed tags etc,. We use MS SQL 2008 Data server to load the unprocessed input dataset. The Master Document interface was also designed with over all GUI model.

4.3.3 K- MEANS CLUSTERING

Aims to find K data points that are maximally separated from each other. In our implementation, at the beginning of initialization, we randomly select a data point as the first cluster center. Then, to identify a new center, we choose a data point that has not been selected previously using the following procedure. We first compare the distances between a candidate data point and all the previously selected centers, and record the minimal distance between this point and the centers. Then the candidate point with the largest minimum distance is selected as the new center. Finally, K centers are selected to initialize the cluster centers of Kmeans.

4.3.4 DOCUMENT EM

Document E and M Step is implemented using Dhillon Function and Probabilities.

Algorithm 1 Alternating EM for HMRF^2-ITCC model.

Input: Document and word sets $\mathcal{D}$ and $\mathcal{V}$; cluster numbers K_d and K_v; pairwise constraints $\mathcal{M}$ and $\mathcal{C}$.
Initialize document and word cluster labels using Kmeans.
Initialize $q^{(0)}(\hat{d}_{k_d}, \hat{v}_{k_v})$, $q^{(0)}(d_m|\hat{d}_{k_d})$ and $q^{(0)}(v_i|\hat{v}_{k_v})$.
while $t <$ maxIter and $\delta >$ maxδ **do**
Document E-Step: compute document clusters using the ICM algorithm to minimize

$$\begin{aligned} \mathcal{L}_d^{(t+1)} = & \arg\min \\ & \sum_{k_d}^{K_d} \sum_{d_m : l_{d_m} = k_d} p(d_m) D_{KL}(p(\mathcal{V}|d_m)||p(\mathcal{V}|\hat{d}_{k_d})) \\ + & \sum_{d_{m_1}}^{M} \sum_{d_{m_2} \in \mathcal{M}_{d_{m_1}}} V(d_{m_1}, d_{m_2} \in \mathcal{M}_{d_{m_1}}) \\ + & \sum_{d_{m_1}}^{M} \sum_{d_{m_2} \in \mathcal{C}_{d_{m_1}}} V(d_{m_1}, d_{m_2} \in \mathcal{C}_{d_{m_1}}) \end{aligned}$$

Document M-Step: update parameters

$$q^{(t+1)}(\hat{d}_{k_d}, \hat{v}_{k_v}), \quad q^{(t+1)}(d_m|\hat{d}_{k_d}) \text{ and } q^{(t+1)}(v_i|\hat{v}_{k_v}).$$

and compute $q^{(t+1)}(d_m|\hat{v}_{k_v})$.

4.3.5 WORD EM

Word E and M Step is implemented using iterated conditional mode.

Word E-Step: compute document clusters using the ICM algorithm to minimize

$$\begin{aligned}&\mathcal{L}_v^{(t+2)} = \arg\min \\ &\sum_{k_v}^{K_v} \sum_{v_i : l_{v_i} = k_v} p(v_i) D_{KL}(p(\mathcal{D}|v_i) || p(\mathcal{D}|\hat{v}_{k_v})) \\ + \quad &\sum_{v_{i_1}}^{V} \sum_{v_{i_2} \in \mathcal{M}_{v_{i_1}}} V(v_{i_1}, v_{i_2} \in \mathcal{M}_{v_{i_1}}) \\ + \quad &\sum_{v_{i_1}}^{V} \sum_{v_{i_2} \in \mathcal{C}_{v_{i_1}}} V(v_{i_1}, v_{i_2} \in \mathcal{C}_{v_{i_1}})\end{aligned}$$

Word M-Step: update parameters

$$q^{(t+2)}(\hat{d}_{k_d}, \hat{v}_{k_v}), \quad q^{(t+2)}(d_m|\hat{d}_{k_d}) \quad \text{and} \quad q^{(t+2)}(v_i|\hat{v}_{k_v}).$$

and compute $q^{(t+2)}(v_i|\hat{d}_{k_d})$.

4.3.6 FINAL CLUSTERING

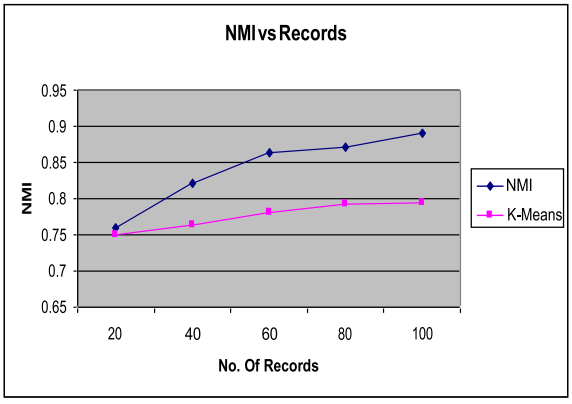
Final Clustering is implemented using (objective function with Document label and word labels). Here the Document labels are compared with varying cluster labels, and with the following cost function the results are updated. The clustered are formed and its cluster labels are also updated.

Compute cost $cost^{(t+2)}$ using Eq. (8) and compute $\delta = (cost^{(t+2)} - cost^{(t)})/cost^{(t)}$.

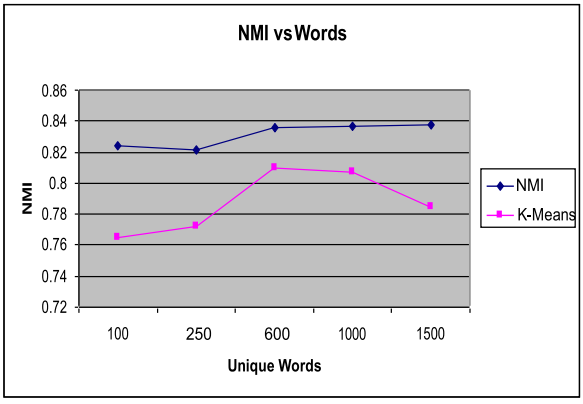
4.3.7 PERFORMANCE ANALYSIS

Performance Analysis is made between Normalized mutual information (NMI) k means vs Proposed).

No. Of Document/Records VS NMI



No. Of Unique Words VS NMI



4.2.1.2 CONCLUSION AND FUTURE ENHANCEMENTS

4.2.1.2.1 CONCLUSION

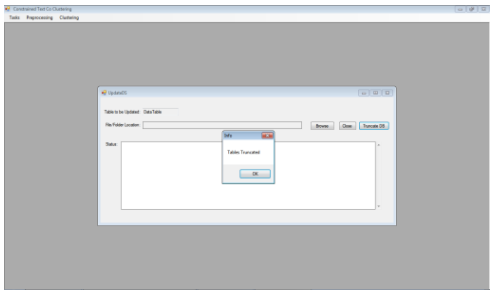
In this project, we have demonstrated how to construct various document and word constraints and apply them to the constrained coclustering process. We proposed a novel constrained coclustering approach that automatically incorporates various word and document constraints into information-theoretic coclustering. Our evaluations on 20newsgroup data sets

demonstrated the effectiveness of the proposed method for clustering textual documents. Furthermore, our algorithm consistently outperformed all the tested constrained clustering and coclustering methods under different conditions.

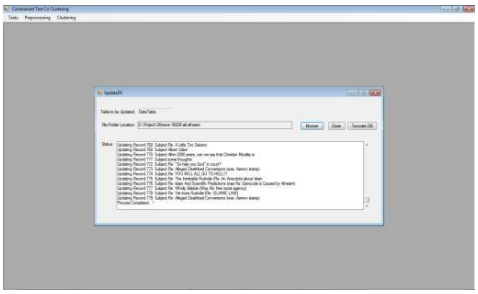
4.2.1.2.2 FUTURE ENHANCEMENTS

There are several directions for future research. Our investigation of unsupervised constraints is still preliminary. We will further investigate whether better text features that can be automatically derived by using natural language processing or information extraction tools. We are also interested in applying CITCC to other text analysis applications such as visual text summarization. Many different classes of machine learning algorithms have been applied to NLP tasks. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. All these scopes are left for future development.

TRUNCATING TABLES

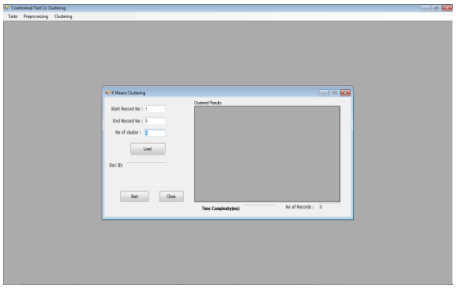


AFTER UPDATING DATABASE

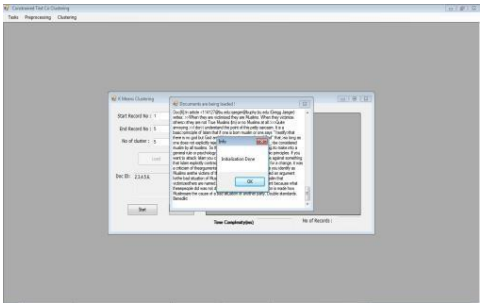


K-MEANS CLUSTER

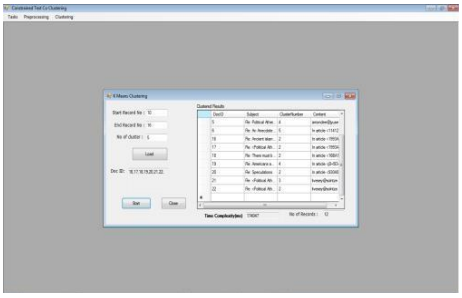
K-MEANS CLUSTER INPUT INITIALIZATION



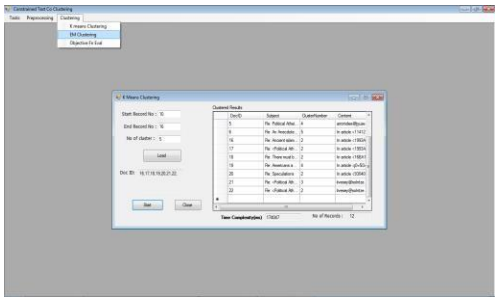
AFTER INITIALIZING



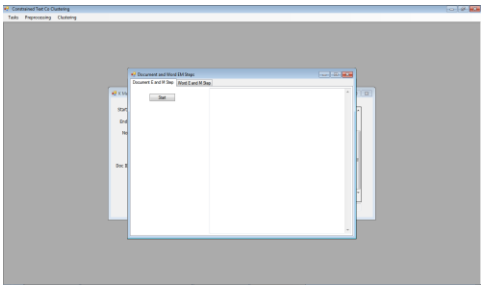
K-MEANS CLUSTERINGRESULT DATAS



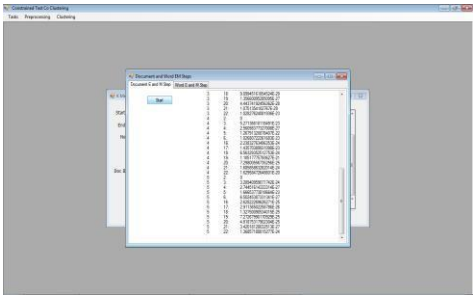
EM CLUSTERING



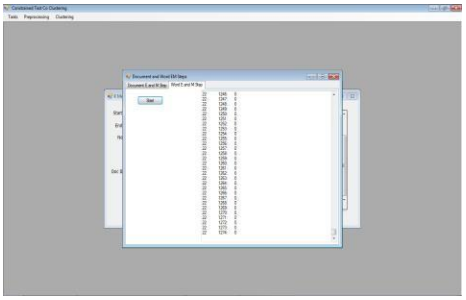
DOCUMENT EM STEP



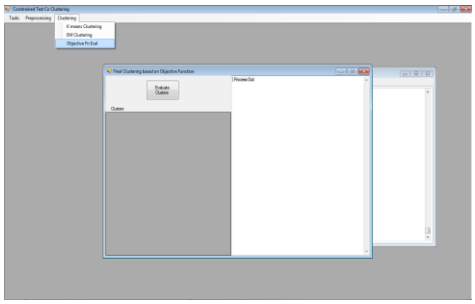
DOCUMENT EM STEP RESULT



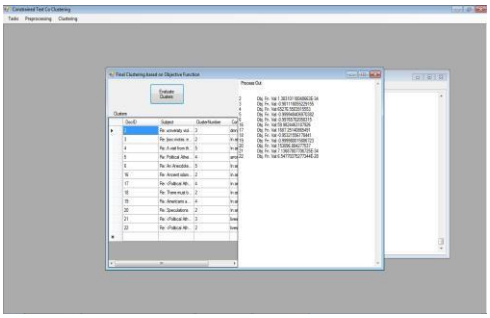
WORD EM STEP RESULTS



OBJECTIVE EVALUATION FUNCTION



FINALCLUSTERING FOR FINAL CLUSTERS



REFERENCES

[1] A. Jain, M. Murty, and P. Flynn, "Data Clustering: A Review,"ACM Computing Surveys,1999.
[2] Y. Cheng and G.M. Church, "Biclustering of Expression Data,"Proc. Int'l System for Molecular Biology Conf. (ISMB)
[3] I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)
[4] I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-Theoretic Co-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)
[5] H. Cho, I.S. Dhillon, Y. Guan, and S. Sra, "Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data,"Conf. Data. Mining (SDM), 2004.
[6] Semi-Supervised Learning, O. Chapelle, B. Scho¨lkopf, and A. Zien,eds. MIT Press, http://www.kyb.tuebingen.mpg.de/ssl-book,2006.

[7] S. Basu, I. Davidson, and K. Wagstaff, Constrained Clustering:Advances in Algorithms, Theory, and Applications. Chapman & Hall / CRC, 2008.
[8] R.G. Pensa and J.-F. Boulicaut, "Constrained Co-Clustering of Gene Expression Data," Proc. SIAM Int'l Conf. Data Mining (SDM),pp. 25-36, 2008.
[9] F. Wang, T. Li, and C. Zhang, "Semi-Supervised Clustering via Matrix Factorization," Proc. SIAM Int'l Conf. Data. Mining (SDM),pp. 1-12, 2008.
[10] Y. Chen, L. Wang, and M. Dong, "Non-Negative Matrix Factorization for Semi-Supervised Heterogeneous Data Co-Clustering," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 10,pp. 1459-1474, Oct. 2010.
[11] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D.S. Modha, "A Generalized Maximum Entropy Approach to Bregman Co-Clustering and Matrix Approximation," J. Machine Learning Research, vol. 8, pp. 1919-1986, 2007.