Hybrid Clustering Framework Using Concurrences and Constraints

[1,*]Dr.L.Jegatha Deborah, [1]S.Dhivya Lakshmi, [1]A.Bharathi.

[1] Department of Computer Science and Engineering, University College of Engineering Tindivanam, Melpakkam – 604 001

E-mail :[*]blessedjeny@gmail.com[*]Corresponding Author

Abstract-Traditionally, the web search engines return thousands of pages in response to a broad query,making it difficult for users to browse or to identify relevant information.For the purpose of quick access to the relevant information, clustering method is a better choice and can be used to automatically group the retrieved and relevant documents of the target domain.In this paper, a new method which combines the techniques of constrained and coclustering methods has been proposed. This combined approach achieves two goals: First, it combines information theoretic coclustering and constrained clustering to improve the clustering performance. Second, additionallythe unsupervised constraints are incorporated into the proposed method to demonstrate the effectiveness of the algorithm. To achieve this goal, a two-sided hidden Markov random field (HMRF) model is developed to represent both document and word constraints.The results of our evaluation over benchmark data sets exhibit that the proposed algorithm is superior compared to other existing approaches.

Index Terms-Coclustering, Constrained clustering, Document Clustering,Gibbs distribution, Pairwise Constraints, Similarity, Unsupervisedconstraints

1 Introduction

Clustering is a well-known technique for the automatic organizationand summarization of a large collection of text [1]. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups. These groups are technically called as clusters.Cluster analysis is an important activity and is defined as the organization of a collection of patterns. These patterns are usually represented as a vector of measurements or a point in a multidimensional space and are grouped into clusters based on some similaritymeasure[2]. Since few decades there have been many approaches to clustering techniques based on data, dimension, shapes and the density parameter.

Coclustering is a specific kind of clustering that examines both document and word relationship at the same time in case of document clustering applications. Coclustering works by finding a pair of maps from row to row cluster and from column to column cluster, with minimum mutual information loss. Since, during the process of clustering mutual information loss is an important component and it has to be greatly reduced for efficient clustering performance. Coclustering is a technique that works by finding minimum cut vertex partitions in a bipartite graph between documents and words. In recent years co-clustering has become an important challenge in market-basket analysis text mining, natural language processing, microarrays and recommendation system analysis.

Constrained clustering are either semi-supervised or unsupervised methodologies. The semi-supervised constrained clustering[3],[4],[5] work by means of providing manually labeled constraints for clustering. In such a situation, the constrained clustering incorporates either a set of must-link constraints, cannot-link constraints, or neither, using a predefined data clustering algorithm or through natural clustering mechanism. These link constraints define a semantic relationship between two data instances. A must-link constraint is used to specify that the two instances in the must link relation should be associated in the same cluster. A cannot link constraint is used to specify that the two instances in the cannot link relation should not be associatedin the same cluster.

To further enhance clustering performance, there has also been some effort on combining coclustering and constrained clustering [6][7][8].To incorporate word and document constraints a newhybrid approach called **Constrained InformativeCoclustering- Kmeans** – CIC-Kmeanshas been proposed in this paper. The proposed technique performed better than the existing 1D constrained clustering method since it can take advantage of the co-occurrences of documents and words. The efficiency of the proposed algorithm is also evident from the fact that the constraints to the algorithm are unsupervised and not semi-supervised. The must-link and cannot-link constraints are given through some knowledge sources automatically.

The remainder of this paper is organized as follows. Section 2 provides a brief summary of existing systems. Section 3 exhibits the system architecture and working of the proposed algorithm. Section 4 illustrates the experimental results of the proposed algorithm with the comparison results shown graphically. Section 5 presents the concluding remarks and some of the future scope of the proposed work.

2. Related Works

Bipartite Spectral algorithm [9] is an algorithm that simultaneously partitions documents and words, and demonstrates that the algorithm offers good global solutions.Spectral coclustering algorithm used in this approach uses the second left and right singular vectors of an appropriately scaled word-document matrix to yield good bipartitionings.This algorithm has good theoretical properties as it provides the optimal solution to real relaxation of the NP-complete coclustering. However, the demerits of this algorithm are factors of inefficiency and instability.

Information Theoretic Coclustering algorithm [10] is an approach that simultaneously clusters both document and word at the same time. The algorithm addresses the concept of optimality wherein an optimal coclustering is one that minimizes the differences called loss in the mutual information between the original random variables and the mutual information between the clustered random variables.Coclustering differs from ordinary one sided clustering in that at all stages the row cluster prototypes incorporate column clustering information, and vice versa. It was theoretically established that the proposed algorithm never increases the loss, and so, gradually improves the quality of coclustering. In their approach the coclustering techniqueis used to annotate the document cluster. The disadvantage of their algorithm lies in the computational effort.

BregmanCoclustering[11] algorithm is used to measure the approximation error to improve the clustering quality.The approximations are based on coclustering concepts,and are expected to exhibit different behavior from the spectral methods typically employed for matrix approximations.They had developed an efficient meta coclustering algorithm based on alternate minimization that is guaranteed to achieve the local optimality for all bregman divergences and theproposed objective function cannot be improved by changing either the row clustering or the column clustering using Lagrange multipliers.Since the methods are iterative in nature and do not involve eigenvaluecomputations, they are significantly faster than the other existing methods and hence, their algorithm are more appropriate for large data matrices.This algorithm uses the sum of the squared residue method,whichis found to be very advantageous.

Semisupervised clustering algorithm [12], [13] is anothertechnique incorporated with limited amounts of supervision in the form of labels on the data or constraints in case of document clustering applications. In contrast, the unsupervised clustering incorporateslabeled data in the following three ways:

- Initial cluster centroids are estimated from the neighborhoods induced from constraints.
- Constraint-sensitive assignment of instances to clusters, where points are assigned to clusters so that the overall distortion of the points from the cluster centroids is minimized, while a minimum number of must-link and cannot-link constraints are violated.
- Iterative distance learning, where the distortion measure is re-estimated during clustering to warp the space to respect user-specified constraints as well as to incorporate data variance.

Non-Negative matrix factorization (NMF) algorithm [14] is useful in the decomposition of multivariate data. The condition of nonnegativity is a useful constraint for matrix factorization and the proposed NMF algorithm is used to minimize the conventional least squares errors and minimizes the generalized KL-divergence.

S. no	Name of the system	Publ, Year	Merits	Demerits
1	Text Classification from Labeled and Unlabeled Documents using EM	2000	Simple and easy to implement	Not Scalable
2	Coclustering of documents and words using Bipartite Spectral Graph Partitioning	2001	Yields good bipartionings.	Inefficient and instable
3	Minimum Sum Squared Residue Coclustering of Gene Expression	2004	Retains all edge weight informa	No semantic document

	Data		tion	similarity.
4	BregmanClustering and Matrix Approximation	2007	Faster and flexible	Highly sensitive to ouliers.
5	Efficient Semi-Supervised Spectral CoClustering with Constraints	2010	Increases efficiency and reduces cost	Speed of execution.
6	Non-Negative Matrix Factorization for Semi supervised Heterogeneous Data Coclustering	2010 October	Easy and reliable during clustering	Inconvenient for large applications

Analysis of related works:

- Deficient techniques for considering both document and word semantic nature.
- Lack of efficient techniques for increased effectiveness of the clustering performance.
- Lack of optimization algorithms for producing optimal clustering results.

Based on the analysis given above, a new hybrid approach has been proposed in this paper which incorporates the techniques of coclustering and constrained clustering methods. This incorporation of hybrid technique increases the accuracy of the clustering results. The coclustering technique incorporated in the hybrid approach uses basically the k-means algorithm which serves as an input to the coclustering algorithm. This is done due to the fact the mutual information loss has to be reduced and to prove that the clustering results obtained are not accidental. Moreover, the constraints given to the algorithm are deemed to be unsupervised since the link constraints given between the documents and the words are not based on human provided labels but from machine based knowledge sources. The document constraints are extracted from the named entity recognizer sources and the word constraints are extracted from the word net similarity. In this wordnet similarity measure, both the nouns and the pronouns are considered form the whole vocabulary of documents. Moreover, for the computation of similarity, two different similarity measures are used and compared namely cosine similarityand Jaccard similarity[15]. Due to the above claims, the proposed algorithm is superior in achieving accurate clustering results compared to other existing algorithms.

3Constrained Informative Coclustering- Proposed Framework:

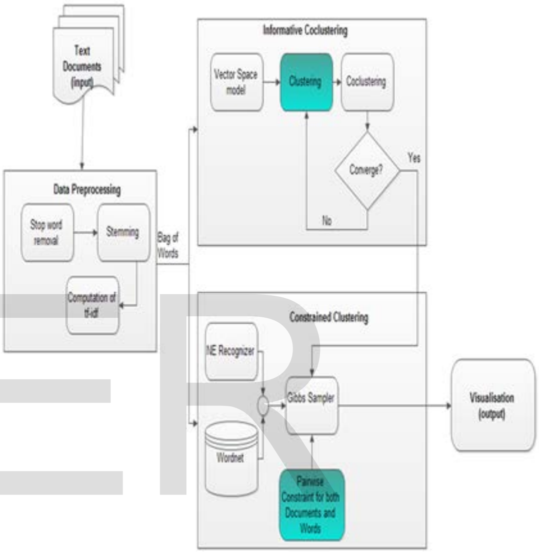

Fig.1 Proposed System Architecture

Document Preprocessing:

Today's real world dataset are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low quality data will lead to low quality clustering result. So pre-processing places a vital role in clustering. Document preprocessing is divided into following stages:

1. Stop-Word Removal:

Stop-words are words that are from non-linguistic view and do not carry meaningful information. Stop-words remove the non-informative behavior words from the text documents and thus reduce noisy data.

2. Stemming:

Suffixes are removed in stemming operation. For example: ing, ion, ment, ement, s, etc...

3. Computation of term frequency- inverse document frequency(tf-idf) :

To compute the *term frequency*, count the number of times each term occurs in each document and sum them all together. For example if we have a set of English text document and wish to determine which document is most relevant to “the white rose”. A simple way to eliminate documents, which does not contains all the three words “a” ,” white” and “rose”. The term "the" is not a good keyword to distinguish relevant and non-relevant documents and terms, unlike the less common words "white" and "rose". Hence an *inverse document frequency* factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

Information Coclustering:

Information Coclustering [8] allows finding a pair of maps from rows to row-clusters and from columns to column-clusters, with minimum mutual information loss. Moreover, the coclustering algorithm does not restrict the number of document and word clusters to be the same. Thus, groups of documents can be extracted that share the same keyword clusters so that different document clusters may share the same keywords.

Vector Space Model:

A starting point for applying clustering algorithms to document collections is to create a vector space model [9]. The basic idea is to extract unique content-bearing words from the set of documents treating these words as features and to then represent each document as a vector in this feature space. Thus the entire document collection may be represented by a word-document matrix A whose rows correspond to words and columns to documents. A non-zero entry in the matrix A_{ij}, indicates the presence of word i in document j, while a zero entry indicates an absence.

The document set and word set are denoted as

D={d_1,d_2,…,d_m} and v={v_1,v_2,…,v_v } respectively. Then the joint probability of $p(d_m,v_i)$ can be computed based on the co-occurrence count of d_m and v_i .The function q(d_m,v_i) is used to approximate the $p(d_m,v_i)$ by minimizing the Kullback-Leibler(KL) divergence [8][16],

q(d_m,v_i) = $p(\hat{d}_{kd},\hat{v}_{kv})$ $p(d_m|\hat{d}_{kd})$ $p(v_i\,|\,\hat{v}_{kv})$ (1)

The Kullback-Leibler divergence of q(d_m,v_i) from $p(d_m,v_i)$ is denoted by $D_{KL}(p(d_m,v_i)$ || q(d_m,v_i)) and is a measure of information lost when q(d_m,v_i) is used to approximate $p(d_m,v_i)$.

D_{KL}(p(D,v) | | q(D,v))=D_{KL}(p(D,v,$\ddot{D}$,$\hat{v}$) | | q(D,v,$\ddot{D}$,$\hat{v}$))(2)

The Kullback-Leibler(KL) divergenceis a fundamental equation of information theory that quantifies the proximity of two probability distribution.

The loss in mutual information[17] can be expressed as a weighted sum of relative entropy between row distribution and row cluster distribution and a weighted sum of relative entropy between column distribution and column cluster distribution respectively given below in (3) and (4).

D_{KL}(p(D,v) | | q(D,v))=
$\sum_{k_d}^{K_d}\sum_{d_m:l_{d_m}=k_d} p(d_m)D_{KL}(p(v|d_m)||p(v|\hat{d}_{k_d}))$ (3)

$D_{KL}(p(\mathrm{D,v})||\mathrm{q}(\mathrm{D,v}))=$
$\sum_{k_v}^{K_v}\sum_{v_i:l_{v_i}=k_v} p(v_i)D_{KL}(p(D|v_i)||p(D|\hat{v}_{k_v}$
)) (4)

Where $\ddot{D}$ and $\hat{v}$ are the cluster sets, p(v | $\hat{d}_{k_d}$) denotes a multinomial distribution based on probabilities.

Constrained Clustering:

The two latent label sets L_d={l_{d1},l_{d2},…,l_{dm}} are introduced for documents and L_v={l_{v1},l_{v2},…,l_{vv}} for words[16]. Then the coclustering technique in the hybrid approach can be mathematically formulated as the log-likelihood[18] of a conditional probability in the exponential family

p(D,v | L_d,L_v)=

exp(-D_{KL}(p(D,v,$\ddot{D}$,$\hat{v}$) | | q(D,v,$\ddot{D}$,$\hat{v}$)))$b_{\phi KL}$(.) (5)

Where $b_{\phi KL}$(.) is normalization constant determined by its divergence type [11]. [18]A likelihood function is a function of the parameters of a statistical model. The likelihood of a set of parameter values D,V given outcome L_d,L_v is equal to the probability of those observed outcomes given those parameter values.

Named Entity Extractor:

Named-Entity Recognition(NER) is also known as entity identification and entity extraction. It is a subtask of information extraction process that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc [19].NE Extractor can be used to find out the document similarity. For example, if two documents share the same people names such as “Deepak Chhabra”, “Glenn Anderson”, and “Selanne”,then both documents are probably about“Hockey sports” both are likely to be in the same document cluster.

For the purpose of finding similar documents, two types of similarity measures are used in the proposed hybrid approach namely cosine similarity and Jaccard similarity measure. These similarity measures are calculated individually and the results are compared for optimality.

WordNet:

WordNet is one of the most widely used and largest Lexical databases [20] of English. It groups nouns, verbs, adjectives and adverbs into sets of synonyms called synsets. The synsets are organized into senses, giving the synonyms of each word, and also into hyponym/hypernym and holonym relationships, providing a hierarchical tree like structure for each term. The semantic distance between two words can be computed based on their relationships in wordnet. The word must-links and cannot-links are based on the computation of the semantic distances in the wordnet. For example, a word must-link is added if the distance between two words is less than a threshold otherwise a word cannot-link is added. These kinds of link constraints are extracted for nouns and pronouns in the hybrid approach from the entire set of the vocabulary of terms in the documents.

Identification of neighborhood Terms:

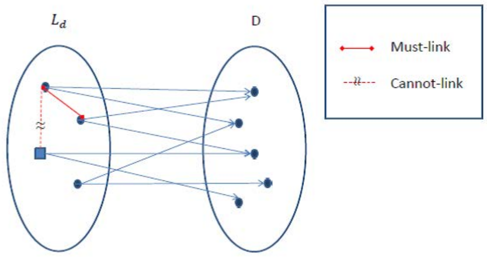

Fig.2 Must-link and Cannot-link Relationships

In order to apply the constraints to the documents and words in terms of must-link and cannot-links, it is necessary to retrieve the neighborhood terms. Such identification facilitates the constrained clustering [21] technique. For the constrained clustering problem, in the proposed hybrid approach Hidden Markov Random Field (HMRF) is used to formulate the prior information for both document and word latent labels[22],[23][24][25]. In this case, the unlabelled documents and words are considered as latent labels[26],[2] which are used for further processing. As shown in Fig.2, we first focus on deriving the constraints forL_d with the latent label document set. The same procedure is followed for word constraints and hence is easy to generalize the derivation to L_v.

Neighborhood Graph Generation:

First, the latent label l_{d_m},the must-link set is denoted asM_{d_m},and cannot-link set as C_{d_m}.The neighbour set of $l_{d_m}$is denoted as N_{d_m}={M_{d_m}, C_{d_m}}. Then the latent labels l_{d_m} (m=1,…,M) is used to construct a neighbourhood graph and the random field defined on this graph is a Markov Random field, following the Markov property:

$$p(l_{d_m} | L_d\text{-}\{l_{d_m}\})= p(l_{d_m} | l_{d_m} \in N_{d_m}.$$

As a result, the configuration of the latent label for documents can be expressed as a Gibbs distribution[12]. Following the generalized potts energy function and its extension[8],

$$p(L_d)=\frac{1}{Z_d}\exp\text{-}\sum_{d_{m1}}^{M}\Big\{ \sum_{dm2\in Nd_{m1}} V(d_{m1}, d_{m2}) \Big\} \quad (6)$$

Gibb’s Sampling:

Gibbs sampling[24] is a particular form of Markov chain Monte Carlo (MCMC) algorithm for approximating joint and marginal distribution by sampling from conditional distributions. Gibbs Sampling algorithm could generate a sequence of samples from conditional individual distributions, which constitutes a Markov chain, to approximate the joint distribution. Gibbs Sampling algorithm has been widely used on a broad class of areas, e.g. Bayesian networks, statistical inference, bioinformatics, econometrics.

The power of Gibbs Sampling is:

1. Approximate joint and marginal distribution

2. Estimate unknown parameters

In each iteration of Gibbs sampling, the embodiment of the invention samples a topic assignment for each word in each document based on topic-word co-occurrences and document-topic co-occurrences.

The posterior probability $p(L_d)$ in Eqn.(6) has 2 components:

the first factor evaluates each label configuration, corresponding to cluster assignments of every point, and gives a higher probability to a configuration that satisfies more of the given must-link and cannot-link constraints. A particular label configuration determines the cluster assignments and hence the cluster representatives. The second factor estimates the probability of generating the observed data points using the conditional distributions. The overall posterior probability of the cluster label configuration of all the points therefore takes into account both the cluster distance measure and the constraints in a principled unified framework.

Incorporation of Pairwise Constraints:

Unsupervised clustering can be significantly improved using supervision in the form of pairwise constraints [12]. For the purpose of providing this kind of supervision, a Must-Link(ML) constraint specifies that the pair of instances that should be assigned to the same cluster, and a Cannot-Link (CL) constraint specifies that the pair of instances should be placed into the different clusters. In some of the application domains, the pairwise constraints can be collected automatically along with the unlabeled data. It augments functions with penalty terms for violating the constraints.

For the constrained clustering framework we have to use the constraint violations [27] to learn the underlying distance measure,the penalty for violating a must-link constraint between distant points should be higher than that between nearby points. This would reflect the fact that if two must-linked points are far apart according to the current distortion measure and are hence put in different clusters, the measure is inadequate and needs to be modified to bring those points closer together. So, the must-link penalty function is chosen to be

V(d_{m1}, d_{m2}□ Md_{m1}) =

$a_{m1,m2}D_{KL}(p(\mathrm{v} \mid d_{m1}) \mid\mid p(\mathrm{v} \mid d_{m2})).I_{ld_m1 \neq ld_m2}$

(7)Analogously, the penalty for violating a cannot-link constraint between two points that are *nearby* according to the current distance measure should be higher than for two *distant* points. This would encourage the distance learning step to put cannot-linked points farther apart. The cannot-link penalty function can be accordinglychosen to be

V(d_{m1},d_{m2}□Cd_{m1})=

$\bar{a}_{m1,m2}(D_{max}\text{-}D_{KL}(p(\mathrm{v} \mid d_{m1}) \mid\mid p(\mathrm{v} \mid d_{m2})).I_{ld_m1 = ld_m2}$ (8)

If a cannot-link is violated the cost is the distance between the cluster centroid both instances are in and nearest cluster centroid to one of the nearest instances.

Where $p(\mathrm{v} \mid d_{m1})$ denotes a multinomial distribution based on the probabilities $p(v_1 \mid d_{m1}),\ldots, p(v_2 \mid d_{m1}))^T$,D_{max} is the maximum value for all the D_{KL} $(p(\mathrm{V} \mid d_{m1}) \mid\mid p(\mathrm{V} \mid d_{m2}))$,$a_{m1,m2}$ and $\bar{a}_{m1,m2}$ are tradeoff parameters to be set emprically, and $I_{true} = 1$,I_{false}=0.

Consequently, the constrained coclustering problem can be formulated as an MAP (Maximum APosterior) estimation for label configurations

$$p(L_{ds},L_v \mid \mathrm{D,v}) \alpha p(\mathrm{D,v} \mid L_d,L_v)p(L_d)p(L_v) \quad (9)$$

As there are two HMRF prior for L_d and L_v, this is called as two-sided HMRF regularization.Mathematically, the objective function can be rewritten as

$$\{L_d,L_v\} = \arg\min D_{KL}(p(\mathrm{D,v},\ddot{D},\hat{v}) \mid\mid q(\mathrm{D,v},\ddot{D},\hat{v}))$$

$$+\sum_{d_{m1}}^{M} \sum_{d_{m2} \epsilon M d_{m1}} V(d_{m1}, d_{m2} \epsilon \, \mathrm{M} d_{m1})$$

$$+\sum_{d_{m1}}^{M} \sum_{d_{m2} \epsilon C d_{m1}} V(d_{m1}, d_{m2} \epsilon \, \mathrm{C} d_{m1})$$

$$+\sum_{vi_1}^{V} \sum_{vi_2 \epsilon M vi_1} V(vi_1, vi_2 \epsilon \, \mathrm{M} vi_1)$$

$$+\sum_{vi_1}^{V} \sum_{vi_2 \epsilon M vi_1} V(vi_1, vi_2 \epsilon \, \mathrm{C} vi_1) (10)$$

By putting the must-link and cannot-link in Eqn (6) for both documents and words then the objective function will be minimized, which will leads to maximizing the posterior probability.

Visualization:

The final stage is to visualize the results after the hybrid algorithm is implemented. The result of the clustering process is enhanced with the help of adding constraints in the form of must-links and cannot-links. In addition, to this the clustering results are further enhanced since the mutual information loss is greatly reduced using Kullback-Leibler divergence function. The final result of the hybrid algorithm is to view the clustering result using a standard visualization SOM visualization tool. The table given in Section 4provides the comparative analysis on the mutual information loss for the various divergence functions.

4 Results and Discussions

To incorporate the word and document constraints, a hybrid approach called constrained informative coclustering combining the techniques of coclustering and constrained clustering is proposed. In the proposed method the coclustering is facilitated in prior by k-means clustering rather than the natural grouping. The claim of the proposed algorithm is that instead of using natural clustering technique, in case if a standard clustering algorithm is used, the mutual information loss can be further reduced. In the hybrid approach, K-means clusteringalgorithm is used for prior clustering the row and column of the matrix which is used for further processing. In such a scenario, the mutual information loss is greatly reduced which is evident from the table given in Table 1.

The experiments were done on several raw input text documents. The result of the CIC-Kmeans method is compared with the other previous algorithms such as Kmeans, Constrained Kmeans(CKmeans),Semi-NMF (SNMF), Constrained SNMF (CSNMF),Tri-factorization of Semi-NMF (STriNMF), Constrained STriNMF (CSTriNMF), ITCC and CITCC.

Table 1 Divergence Functions Vs Mutual Information loss

Divergence Function	Mutual information loss
I-Divergence	0.0432
Bregman Divergence	0.0336
Kullback-Leibler Divergence with Natural clustering	0.0288
Kullback-Leibler Divergence with Kmeans	0.0254

4.1 Result Data Set:

To evaluate the effectiveness of the proposed hybrid approach, the experiment was conducted on the 20-newsgroups dataset. The 20-newsgroups dataset is a collection of approximately 20,000 newsgroups documents, partitioned evenly across 20 different newsgroups. The performance of the proposed hybrid approach has been evaluated against various clustering algorithms by using Normalized Mutual Information (NMI) [16].The NMI between two random variables X and Y is defined as

$$\mathrm{NMI(X,Y)}=\frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

Where I(X;Y) is the mutual information between X and Y. The entropies H(X) and H(Y) are used for normalizing the mutual information to be in the range of[0,1].In practice, we estimate the NMI score [28] using the following formulation:

$$\mathrm{NMI}=\frac{\sum_{s=1}^{k}\sum_{t=1}^{k} n_{s,t}\log\left(\frac{n\,n_{s,t}}{n_s n_t}\right)}{\sqrt{\left(\sum_s n_s\log\frac{n_s}{n}\right)\left(\sum_t n_t\log\frac{n_t}{n}\right)}}$$

Where n is the number of data samples, n_s and n_t denote the amount of the data in class s and cluster t, $n_{s,t}$ denotes the amount of data in both class s and cluster t.In this experiment, the performance of hybrid constrained informative coclustering technique is compared with that of several representative clustering algorithms such as Kmeans, Constrained Kmeans(CKmeans), Semi-NMF (SNMF), Constrained SNMF (CSNMF), Tri-factorization of Semi-NMF (STriNMF), Constrained STriNMF (CSTriNMF), ITCC and CITCC.

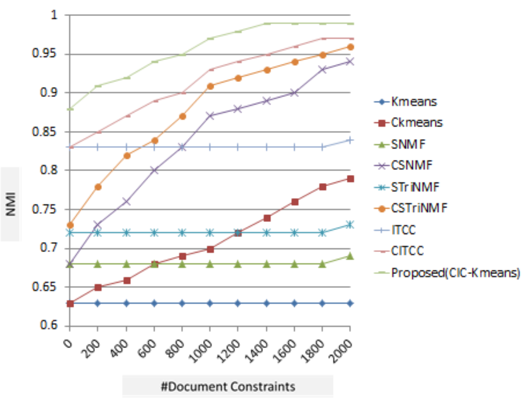

Fig.3a Effects of document constraints

We varied the number of document and word constraints in each experiment by randomly selecting a fixed number of constraints from all possible must-links and cannot-links to investigate their impact on clustering performance.Fig.3a shows the experiment results for document constraints. The x-axis represents the number of document constraints and y-axis denotes the average Normalized Mutual Information [NMI] of five random trials. From the graph it is evident that the proposed CIC-Kmeans consistently performed the well compared to the other existing algorithms. In addition, to evaluate the effect of the number of word constraints on the constrained coclustering performance, we evaluated 1) CITCC (5k) and CSTriNMF (5k) with document constraints plus 5000 word constraints, 2) CITCC (10k) and CSTriNMF (10k) with document constraints plus 10,000 word constraints , 3) CIC-kmeans our proposed algorithm.

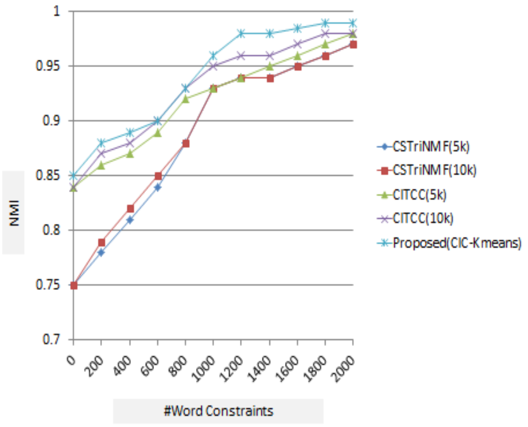

Fig.3bEffects of word constraints

As shown in Fig.3b, ingeneral, more word constraints resulted in better clustering.However, the impact of the word constraintwas not as strong as that of the document constraints. From the graph shown in Fig. 3(b) our proposed Constrained Informative Coclustering – Kmeans algorithm has better performance among the other algorithms for the word constraints.

5 Concluding Remarks

In this paper,we proposed a Constrained Informative Coclustering-Kmeans(CIC-Kmeans) approach that automatically incorporates constraints into information-theoretic co-clustering. Our evaluations on a benchmark data set demonstrated the effectiveness of the proposed method for clustering textual documents.Our algorithm consistently outperformed all the tested constrained clustering and co-clustering methods under different conditions.There are several directions for the future research.For example,we will explore other text features that can be automatically derived by natural language processing (NLP) tools to further improvize unsupervised document clustering performance. We are interested in applying CIC-Kmeans to other image documents in future work.

REFERENCES

[1] A. Jain, M. Murty,and P. Flyn, “Data Clustering: A Review,” ACM Computing Surveys, vol. 31, no. 3,pp. 264-323,1999.
[2]A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In 2000), Workshop on Artificial Intelligence for Web Search (AAAI), pages 58–64, July 2000.
[3]K.Wagstaff, C.Cardie, S.Rogers, and S.Schro d1, “Constrained KMeans Clustering with Background Knowledge,” Proc. 18 th Int'l Con'f. Machine Learning(ICML), pp.577-584,2001.
[4] M.Bilenko, S.Basu, and R.J.Mooney, “Integrating Constraints and Metric Learning in Semi-Supervised Clustering,” Proc.21st Int'l Conf. Machine Learning (ICML), pp.81-88, 2004.
[5] F.G. Cozman, I. Cohen, and M.C. Cirelo, “Semi-Supervised Learning of Mixture Models,” Proc. Int'l Conf. Machine Learning(ICML), pp. 99-106, 2003.
[6]R.G.Pensa and J.-F. Boulicaut, “Constrained Co-clustering of Gene Expression Data,”Proc.SIAM Int'1 Conf.DataMining(SDM), pp.25-36,2008.
[7] F.Wang, T.Li and C.Zhang, “Semi-supervised Clustering via Matrix Factoriztion,”Proc. SIAM Int'1 Conf.DataMining(SDM), pp.1-12,2008.
[8] Y.Chen, L.Wang, and M.Dong, “Non-Negative Matrix Factorization for Semi-Supervised Heterogeneous DataCo-Clustering,” IEEE Trans. Knowledge and DataEng., vol. 22, no. 10, pp. 1459-1474, Oct. 2010.
[9] 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),pp.269-274,2001.
[10]I.S.Dhillon,S.Mallela,andD.S.Modha,”Information-Theoretic Co-clustering,”Proc.Ninth ACM SIGKDD Int'lConf. Knowledge Discovery and Data Mining(KDD),pp.89-98,2003.
[11] A. Banerjee, I. Dhillon, J. Ghosh, S.Merugu, and to D.S.Modha, “A Generalized Maximum Entropy ApproachBregman Co-Clustering and Matrix Approximation,” J. Machine Learning Research, vol. 8, pp. 1919,2007.
[12] S.Basu, M. Bilenko, and R.J. Mooney, “A Probabilistic Framework for Semi-Supervised Clustering,” Proc SIGKDD, pp. 59-68,2004.
[13] M. Bilenko, S. Basu, and R.J. Mooney, “Integrating Constraints and Metric Learning in Semi-Supervised Clustering,” Proc. 21st Int'l Conf. Machine Learning (ICML), pp. 81-88,2004.
[14]P. Wang, C. Domeniconi, and K.B. Laskey, “Latent Dirichlet Bayesian Co-Clustering,” Proc. European Conf.Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), pp. 522-537,2007.
[15] Anne Chao,1 Robin L. Chazdon, 2 Robert K. Colwell 2 and Tsung-Jen Shen 1”A new statistical approachforassessing similarity of species composition with incidence and abundance data”Proc. Ecology letters pp.148-159, 2005.
[16]Yangqiu Song, Shemi Pan, Shixia Liu, Furu Wei, with Michelle X.Zhou, and Weihong“ ConstrainedCoclusteringnSupervised and Unsupervised constraints”Proc., Conf. Knowledge and Data Engineering., pp.1227-1239, 2013.
[17] T. Yang, R. Jin, and A.K. Jain, “Learning from Noisy Side Information by Generalized Maximum Entropy Model,” Proc. Int'l Conf. Machine Learning (ICML), pp. 1199-1206, 2010.
[18] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the RoyalStatistical Society B, 39:1–38, 1977.
[19]Niu,C,Li;W; Ding, J; and Srihari, R.K.2003. A using bootstrapping approach to named entity classification successive learners. In Proceeding of ACL.
[20]G.A.Miller,”Wordnet:A Lexical Database for English,” Comm. ACM , vol.38,pp.39-41,1995.
[21]S.Basu,I.Davidson, and K.Wagstaff, Constrained Clustering: Advances in Algorithms, Theory, and Applications, Chapman & Hall/ CRC, 2008.
[22] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, “Co-Clustering Based Classfication for Out-of-Domain Documents,” Proc 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data mining, pp. 210-219, 2007.
[23]K. Nigam, A.K. McCallum, S. Thrun, and T.M. Mitchell, “Text Classification from Labeled and UnlabeledDocuments using EM,” Machine Learning,vol.39, no. 2/3, pp. 103-134,2000.
[24] M.Bilenko and S.Basu, “A Comparison of Inference Techniques for Semi-Supervised Clustering with HiddenMarkov Random Fields” Proc.ICML WORKSHOPStatistical Relational Learning and it's

Connections toother Fields(SRL '04), 2004

[25] X. Shi, W. Fan, and P.S. Yu, "Efficient Semi-IEEE Supervised Spectral Co-Clustering with Constraints," Proc.10th Int'l Conf. DataMining(ICDM), pp. 1043-1048,2010.

[26] Y. Zhang, M. Brady, and S. Smith. Hidden Markov random field model and segmentation of brain MR images.IEEE Transactions on Medical Imaging, 20(1):45–57, 2001.

[27] Y. Zhang, M. Brady, and S. Smith. Hidden Markov random field model and segmentation of brain MR images.IEEE Transactions on Medical Imaging, 20(1):45–57, 2001.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the RoyalStatistical Society B, 39:1–38, 1977.