A Literature Survey on Coclustering and Constraint Clustering for Text Documents

MITESH SHARMA
M.E.Scholar
Computer Science and Engineering
M.B.M. Engineering College
Jai Narayan Vyas University, Jodhpur

AMIT MISHRA
Associate Professor
Computer Science and Engineering
Jodhpur Institute of Engineering and Technology, Jodhpur

DEVEAN PUROHIT
Assistant Professor
Computer Science and Engineering
Jodhpur Institute of Engineering and Technology, Jodhpur

***Abstract*—Nowadays Information Retrieval is one of the fastest growing technologies to retrieve information from text documents. There are lots of methods and techniques are used for Information Retrieval from documents. In this paper we explained two technique co clustering and constrained clustering, and how it used to retrieve information from documents. And how both techniques adopts supervised and unsupervised constraints in the document. Most Existing clustering algorithm clusters documents or words independently. But coclustering clusters document and word constraints simultaneously. First coclustering method simultaneously extracts both document and word constraint in the text document. It automatically extracts unsupervised constraints based on the existing knowledge sources. NE extractor automatically extract document constraint based on overlapping named entities in a set of documents and simultaneously WordNet automatically extracts word constraints which are constructed based on semantic distance between words in a document. Additionally incorporate Constraint clustering, which provide must-link and cannot-link constraints between the documents and word constraints.**

Keywords—coclustering, constraint clustering, word constraint,document constraint.

I. INTRODUCTION

1. *Clustering*

Clustering is a technique or process of grouping data into similar classes and differentiates data from dissimilar data. It is an unsupervised learning [1][2][4]. Machine automatically organizing and grouping the large set of data. It segments a large collection of dataset into group of classes according to the similarity measures and means value of data sets.

Real World there are lots of clustering techniques are available. Such as K means algorithm, agglomerative algorithm, connectivity based algorithm, density based algorithm etc. In this paper we explain coclustering and constrained clustering, and how the two techniques are interested to cluster the text documents.

2. *Document Clustering*

Document clustering is one of the crucial techniques for organizing or grouping documents in an unsupervised manner. Unlike relational database documents are stored in text databases, which consist of huge amount of text documents such as news articles, books, digital libraries etc. Documents contain both structured and unstructured data. So it is hard to grouping the documents into cluster classes.

Information retrieval (IR) is a field to mining information from large set of documents. Text databases are stored semistructured data, which either unstructured or structured data. It retrieve relevant document in a large collection of document datasets.

Document clustering falls into two categories. First one Similarity based document clustering and Model based document clustering. Objective function of similarity based approach is finding the pairwise similarity between documents. So similar documents are grouped into one class. Most document cluster uses the similarity based document clustering. In Model based clustering

Existing document clustering based on co-occurrence of word constraints distributed in the documents. Depends on the word constraints documents will be grouped into clusters. In this paper we explain to construct both document and word constraints grouped to cluster the documents. Coclustering algorithm deals with both document clustering and word clustering.

II. ALGORITHM USED FOR DOCUMENT CLUSTERING

1. *Coclustering*

Coclustering is an emerging data mining technique that deals with two sets of data (dyadic data). Most clustering technique is one dimensional clustering. It clusters either document constraints clustering or word constraints clustering. But coclustering clusters both document and word constraints. In text document this algorithm simultaneously clustering the both word constraints and document constraints. Both constraints will be modeled by using bipartite spectral graph algorithm [5].Words in the documents are clustered basis on the co-occurrence of words in the documents. As well as document constraints will be extracted by overlapped names in the documents.

2. *Constraint clustering*

Constraint clustering is mainly used for semi-supervised clustering. Constraint clustering enforce must link and cannot link between document and word constraints.

Must-link constraints between two documents must be in same group of clusters
Cannot-link constraints between two documents should be in different group of clusters.

Both must-link and cannot-link constraint defines a relationship between two documents. Some of the constraints clustering algorithms are COP Kmeans and PCKmean algorithms.

3. *Semisupervised clustering*

Semisupervised contain small amount of labeled data used to cluster the unlabeled data. It falls between without any labeled data and with labeled data. It is applicable to both classification and clustering. In Semi-supervised clustering labeled data is combined with large set of unlabeled data to provide the better clustering. In text document contain large group of unlabeled data and small amount of labeled data [1][6]. In some semi-supervised algorithm need human added labeled are bias to construct unlabeled constraints. Semi-supervised clustering falls into two categories. One is semi-supervised with labeled seeding points [1][3][6]. Moreover, in this method labeled data are used to generate the initial seed point that explores the clustering algorithm. Right seed labels lead the clustering towards the best way. Second one is semi-supervised clustering with labeled constraints [1][9].

III. TRANSFORM WORD SPACE TO DOCUMENT SPACE

Clustering segmenting large set of data into group of related clusters or classes. So the data in the group are related to the same class and data belonging to the different classes are dissimilar. However, document clustering cluster the related document in similar groups. Most clustering techniques works in a document space. Clustering the documents constraints also works in document space only. By using cosine similarity or similarity measures between the documents are helps to use document clustering that all works in a document space. But in real world application knowledge on the word side are used to cluster the documents. So, the knowledge in the word side is transform to the document space.

Word Space contains two forms. One is categorization of words and second one is pairwise relationship between words. Transform categorization of word is easy way. But the form of pairwise relation is difficult one. There are two types of pairwise relation. (1) Must-link word pair and (2) Cannot-link word pair. Moreover, knowledge on the word side will transform to the document side [7]. Additional information or human added information in Word side can help the clustering the large group of

document. Hence, it requires transformation of word space to document space. By using Non Negative matrix factorization function is used to transfer the knowledge from word space to document space. Co occurrence of words in the documents will be modeled in matrix formation.

IV. FROM ITCC TO CITCC MODEL

Integrated Theoretic Coclustering (ITCC) method only used co clustering algorithm to retrieve document and word constraints. To enhance the performance of document clustering we incorporate the additional clustering algorithm as constrained clustering to provide a pairwise similarity between document and word constraints. Constrained clustering is mainly used for semi-supervised clustering. So the novel enhanced method called constrained information-theoretic coclustering (CITCC).

CITCC Model handles both supervised and unsupervised constraints. To handle unsupervised constraint in document is difficult. Unsupervised constraints such as document constraint and word constraint will be retrieving by using two tools such as Named-Entity (NE) extractor and WordNet [8]. NE Extractor automatically extracts document constraint based on overlapping named entities in the set of documents. It extracts document constraints based on the existing knowledge source. The document constraints are constructed based on three class entity such as *person, location, organization*. For example, if two or more documents talks about the same people names such as “Mahatma Gandhi”, “Jawaharlal Nehru”, “Subash Chandra Bosh” then the documents related about the Independence of India, and all the documents are grouped into same document cluster.

To improve the document clustering performance to incorporate additional lexical word constraint developed by the WordNet based similarity [8]. WordNet is an online lexical database that provides similarity between words in the document. Relationship among word such as synonyms, antonyms, hypernyms and hyponyms are provided the similarity between words in the documents.

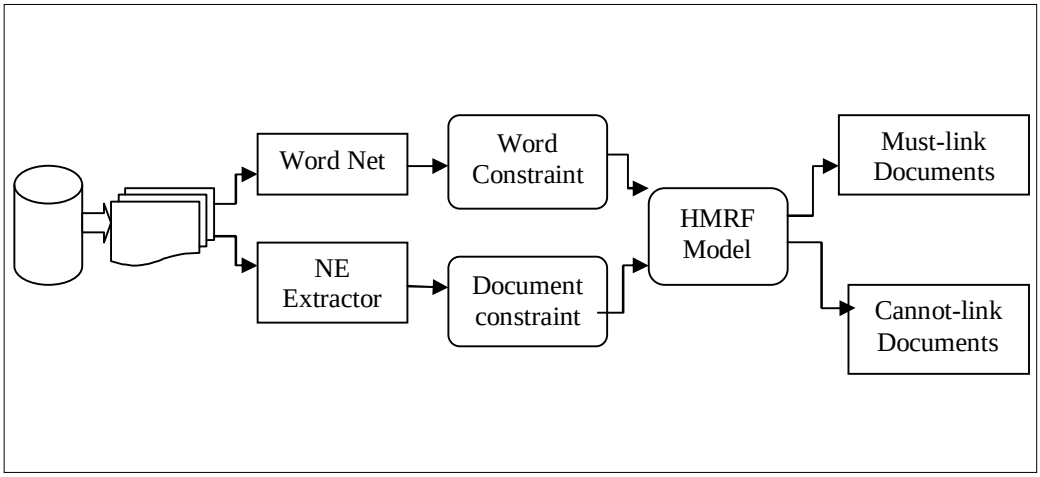

Fig. 1. Example Model of retreiving constraints from document

Figure.1. illustrate the model to retrieving the document constraints and word constraints from the document by using WordNet and Named Entity Extractor tool. Once, document constraints and word constraints are retrieving by using NE Extractor and WordNet, and then two sided HMRF (Hidden Markov Random Field) Model is used to model the document constraints and word constraints. Then after Expectation-Maximization Model is used to optimize the HMRF Model. HMRF Model also useful to incorporate the semi-supervised constraints. In this Framework it takes the data constraints as pairwise must-link constraints and cannot-link constraints. So must-link constraints are grouped into one cluster of documents and cannot-link constraints are grouped into another cluster of documents.

V. CONCLUSION

This Survey paper exploits almost all technical publication related to the novel coclustering and constraint clustering and how this algorithms are used to cluster the text documents and how it handle unsupervised constraints in the documents. Coclustering is most effective algorithm than the one dimensional algorithm. Because it performs automatically retrieve word constraints and document constraints in the text document. Additionally, HMRF (Hidden Markov Random Field) model is used to find the joint probability between word constraint and document constraints. Then after, EM Method is used to optimize the model. In this paper we consider the future work is to retrieve the text images, because text documents also contain the text images. And also we suggest incorporating word space constraints automatically by using natural language processor.

REFERENCES

[1] Yangqui Song, Shimei Pan, Shixia Liu, Furu Wei, Michelle X. Zhou, Weihong Qian," Constraint Text Co-Clustering with supervised and unsupervised constraints, "IEEE VOL. 25, NO. 6, JUNE 2013.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Second Edition.

[3] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, Tom Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM", Machine Learning,1-34.

[4] Inderjit S.Dhillon, Subramanyam Mallela, Dharmendra S.Modha ,"Information-Theoretic Co-clustering", Proc. Ninth ACM SIGKDD Int'l Conf Knowledge Discovery and Data Mining(KDD), pp.89-98,2003.

[5] Inderjit S.Dhillon,"Coclustering Documents and words using Bipartite Spectral Graph Partitioning",Proc.Seventh ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining.pp.269-274,2001.

[6] Sugato Basu, Arindam Banerjee, Raymond Mooney,"Semi-supervised Clustering by Seeding", Proc 19th Int'l Conf on Machine Learning(ICML-2002),pp.19-26,Sydney, Australia, July 2002.

[7] Tao Li, Chris Ding Yi Zhang, Bo Shao," Knowledge Transformation from Word Space to Document Space", Proc.31st Ann.Int'l ACM SIGIR Conf Research and Development in Information Retrieval(SIGIR),pp.187-194,2008.

[8] G.A. Miller, "Wordnet: A Lexical Database for English," Comm.ACM, vol.38, pp. 39-41,1995.

[9] M.Bilenko,S.Basu,and R.J.Moooney,"Integrating Constraints and Metric Learning in Semi-Supervised Clustering" Proc.21st Int'l Conf.Machine Learning(ICML),pp.81-88,2004.