ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Visual Text Summarization in Supervised and Unsupervised Constraints Using CITCC

S.Mohan Gandhi[1], T.Suresh Kumar[2]

PG Scholar, Department of Computer Science and Engineering, Nandha College of Technology, Erode, Tamilnadu[1]

Assistant Professor, Department of Information Technology, Nandha College of Technology, Erode, Tamilnadu[2]

Abstract: In this work clustering performance has been increased by proposes an algorithm called constrained information-theoretic co-clustering (CITCC). In this work mainly focus on co-clustering and constrained clustering. Co-clustering method is differing from clustering methods it examine both document and word at a same time. A novel constrained co-clustering approach proposed that automatically incorporates various word and document constraints into information-theoretic co-clustering. The constraints are modeled with two-sided hidden Markov random field (HMRF) regularizations. An alternating Expectation Maximization (EM) algorithm has developed to optimize the model. NE extractor and WordNet methods are proposed to automatically construct and incorporate document and word constraints to support unsupervised constrained clustering. NE extractor is used to construct document automatically based on the overlapping named entities. WordNet is used to construct word constraints automatically based on their semantic distance inferred from WordNet. It can simultaneously cluster two sets of discrete random variables such as words and documents under the constraints extracted from both sides. With this work contains add visual text summarization to increase more clustering performance.

Keywords: coclustering, constrained clustering, document constraints, word constraints, NE extractor, WordNet

I. INTRODUCTION

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and systems. It is the process to extract information from a data set and transform it into an understandable structure. Data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). Text mining, sometimes alternately referred to as text data mining, roughly equivalent to the analytics of text, refers to the process of deriving high-quality information from text. High-quality information is technically derived the devising of patterns and trends through means such as statistical pattern learning. Text mining basically involves the process of deriving the input text (usually parsing, along with the summation of some derived linguistic features and the removal of others, and insertion into a database), deriving patterns in the structured data, evaluation and interpretation of the output. 'High quality' in text mining relatively refers to some related combination of pertinence, novelty, and interestingness. Typical text mining tasks involves text categorization, concept/entity extraction, text clustering, production of granular taxonomies, sentiment analysis, summarization of document, and entity relation modeling.

Document classification or document categorization is a problem in information science, library science, and computer science. Categorization is the process in which ideas and objects are recognized, differentiated, and understood.[1] Categorization implies that objects are grouped into categories, usually for some specific purpose. Ideally, a category

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

illuminates a relationship between the subjects and objects of knowledge. The task is to allocate a document to one or more classes or categories. This may be done "manually" (or "intellectually") or algorithmically. The rational classification of documents has mostly been the province of library science, although the algorithmic classification of documents is used mainly in information science and computer science. The problems are overlapping, and therefore also interdisciplinary research on document classification. The documents to be classified may be texts, images, music, etc. Every kind of document possesses its special classification problems. When not otherwise specifies, text classification is implied.

Documents may be classified according to their subjects or according to other. In the rest of this article only subject classification is considered. There are two main principles of subject classification of documents: The content based approach and the request based approach.

- Content based classification is classification in which the weight given to particular subjects in a document determines the class to which the document is assigned. For example, it is a rule in much library classification that at least 20% of the content of a book should be about the class to which the book is assigned. In automatic classification it could be the number of times given words appears in a document.

- Request oriented classification (or indexing) is classification in which the anticipated request from users is influencing how documents are being classified.

Clustering is a technique commonly used for automatically organizing a large collection of data's. In this work focus on co-clustering and constrained clustering. Clustering can be considered the most important unsupervised learning problem as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. Clustering is to determine the intrinsic grouping in a set of unlabeled data. Co-clustering is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix.

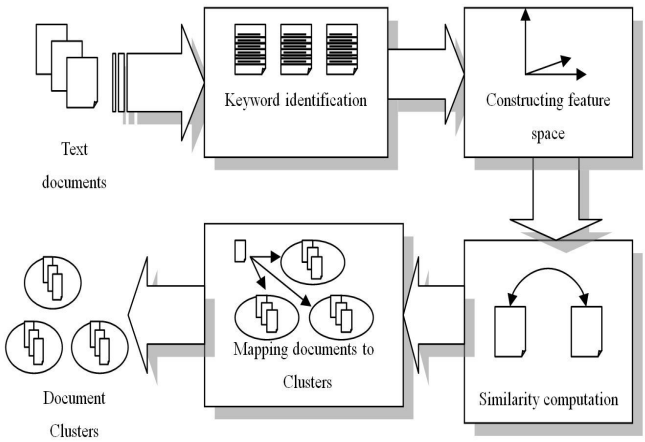

Fig. 1 Word and Document clusters

Co-clustering is differing from clustering methods it examine both document and word at a same time. It is more effective than 1D clustering. Constrained clusters developed for document cluster. Typically, constrained clustering incorporates either a set of must-link constraints, cannot-link constraints, or both, with a Data clustering algorithm. Both a must-link and a cannot-link constraint define a relationship between two data instances.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

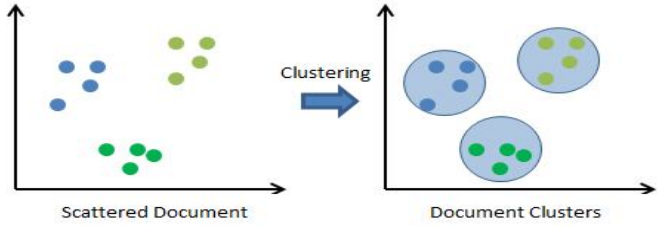

Fig. 2 Document Clusters

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. A HMM (fig.3) can be considered the simplest network. An expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

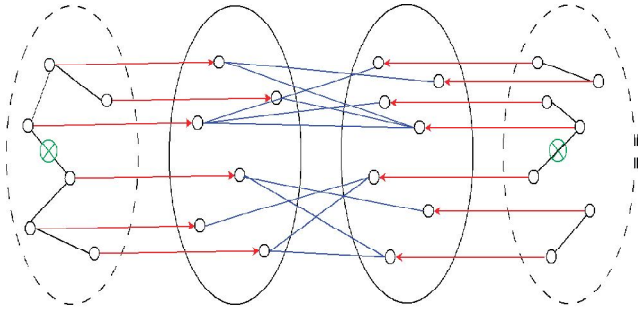

Fig. 3 HMRF-ITCC model

One problem of standard classification approaches is that they do not take into account the predicted labels of the surrounding words. This can be done using probabilistic models of sequences of labels and features. Frequently used is the hidden Markov model (HMM), which is based on the conditional distributions h of current labels L(j)given the previous label L(j-1) and the distribution of the current word t(j) given the current and the previous labels L(j);(j-1).

$$L^{(J)} \sim p\,(L^{(j)} \mid L^{(j-1)}) \qquad t(j) \sim p(t^{(j)} \mid L^{(j-1)}, L^{(j-1)})$$

Pattern mining is a data mining method that involves finding existing patterns in data. An Information filtering system is a system that removes redundant or unwanted information from an information stream using (semi)automated or computerized methods prior to presentation to a human user.

II. PROPOSED SYSTEM

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

In this work clustering performance has been increased by proposes an algorithm called constrained information-theoretic co-clustering (CITCC). It integrates constraints into the information theoretic co-clustering (ITCC) framework where KL-divergence is adopted to better model textual data. The constraints are modeled with two-sided hidden Markov random field (HMRF) regularizations. An alternating Expectation Maximization (EM) algorithm has developed to optimize the model. NE extractor and WordNet methods are proposed to automatically construct and incorporate document and word constraints to support unsupervised constrained clustering. NE extractor is used to construct document automatically based on the overlapping named entities. WordNet is used to automatically construct word constraints based on their semantic distance inferred from WordNet.

III. SYSTEM MODEL

3.1 Initialization

In this work utilize K means to initialize the document and word clusters before apply this method need to initialize K means first. To generate the K means algorithm more stable for document and word clustering, this work utilize a farthest-first traversal method. (Fig.4)It aims to find K data points that are maximally separated from each other.

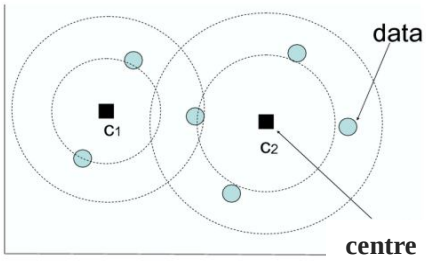

Fig. 4 K means algorithm

In implementation, at the beginning of initialization randomly select a data point as the first cluster center. Then, to identify a new center choose a data point that has not been selected previously using the following procedure. First compare the distances between a candidate data point and all the previously selected centers, and record the minimal distance between this point and centers. Then the candidate point with the largest minimum distance is selected as the new center. Finally, K centers are chooses to initialize the cluster centers of K means.

3.2 Matrix Operation

The original document and word co-occurrences as well as the intermediate parameters are all stored in matrices. The matrices maintain a row-style. Specifically, for a dense matrix use an array to store the row elements; for a sparse matrix, after comparing different hash table in this work choose the COLT1 hash to store the row elements. When implementing K means store the norm of each row beforehand since the computation of the norm of each point is one of the biggest overheads in computing the Euclidean distance or cosine similarity.

$$cos(d_i, d_j) = \frac{d_i \cdot d_j}{\| d_i \| \| d_j \|}$$

Finding similarity between words is a fundamental part of text similarity. Accurate clustering requires a precise definition of the closeness between a pair of objects, in terms of either the pair-wised similarity or distance. When

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

implementing NMF-based clustering approach carefully encode matrix multiplication since it affects the system performance most. When implementing the ITCC-based methods, the computational sequence of KL-divergence is designed to create the computation faster.

3.3 Constraints and ICM inference

The constraints are stored in a set of hash tables choose this data structure to make the constraints symmetric for pairs of data points. It also makes it easier to develop the appointed constraints with neighborhood inference. In this work only use the original constraint set and do not infer new constraints, since the Named Entities-based constraints and the WordNet-based constraints are noisy, which may not fascinate the consistency assumption. For the constrained clustering problem, we use HMRF to specify the prior information for both document and word latent labels. We use a general EM algorithm to find estimation. There are two steps in the EM algorithm: the E-step and the M-step.

- In the E-Step, we update the cluster labels based on the fixed model function from the last iteration.

$$Q(\theta,\theta^t) = \sum_{i=1}^{n}\sum_{y} P(y \mid x_i,\theta^t)\log P(x_i, y \mid \theta)$$

- In the M-Step, we update the model function. Since the latent labels are fixed, the update is not affected by the must-links and cannot-links.

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta,\theta^t)$$

3.4 Document and Word Constraints

We propose new methods to derive "good but imperfect" constraints using information automatically extracted from either the content of a document NE constraints or existing knowledge sources WordNet constraints (Fig.5). Named Entities Extractor is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, locations, organizations, quantities, expressions of times, monetary values, percentages, etc. the document must-link constraints can be constructed from the correlated named entities such as location, person and organization. Explicitly, if there are overlapping NEs in two documents and the number of overlapping NEs is larger than a predefined threshold, we can add a must-link to these documents. WordNet is a lexical database for the English language.

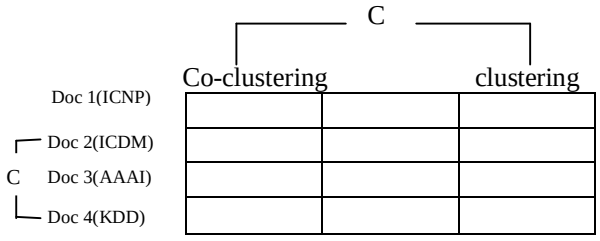

Fig. 5 Document and word co-clustering

It groups English words into sets of synonyms called synsets, general definitions, provides short, and records the multiple semantic relations between these synonym sets. The semantic distance of two words can be computed based on their relationships in WordNet. Since we may construct word must-links based on semantic distances.Named Entity Mining (NEM) is a text mining task in which the information on the named entities of a class is mined from a large amount of data. The classes can be, for example, movie, game, book and music, and the task can be to mine all the titles of movies in a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

textual data collection. WordNet covers some specific terms from every subject related to their terms. WordNet as a lexical database map all the stemmed words from the standard documents into their specifies lexical categories.

IV. CONCLUSION

In this research have demonstrated how to construct various document and word constraints and apply them to the constrained co-clustering process. A novel proposed constrained co-clustering approach that automatically incorporates various word and document constraints into information-theoretic co-clustering. Evaluations on two benchmark data sets demonstrated the effectiveness of the proposed method for clustering textual documents. Furthermore, algorithm consistently outperformed all the tested constrained clustering and co-clustering methods under different conditions. With this work contains add visual text summarization to increase more clustering performance.

REFERENCES

[1] Banerjee A, Dhillon I, Ghosh J, Merugu S, and Modha D S," A Generalized Maximum Entropy Approach to Bregman Co-Clustering and Matrix Approximation", J. Machine Learning Research, volume 8, pp. 1919-1986, 2007.
[2] Basu S, Davidson I, and Wagstaff K, Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall/ CRC, 2008.
[3] Chen Y, Wang L, and Dong M, "Non-Negative Matrix Factorization for Semi-Supervised Heterogeneous Data Co- Clustering", IEEE Transactional Knowledge and Data Engineering., volume 22, no. 10, pp. 1459-1474, Oct. 2010.IEEE Trans. Med. Imag., no. 5, pp. 775–785, May .
[4] Cheng Y and Church G M, " Biclustering of Expression Data", International System for Molecular Biology Conference (ISMB), pp. 93-103, 2000.
[5] Cho H, Dhillon I S, Guan Y and Sra S, "Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data", Proceedings Fourth SIAM International Conference Data Mining (SDM), 2004
[6] Dhillon I S, Mallela S and Modha D, "Information-Theoretic Co-Clustering", Proceedings Ninth ACM SIGKDD International Conference Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003.
[7] Dhillon I S, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning", Proceedings Seventh ACM SIGKDD International Conference Knowledge Discovery and Data Mining (KDD), pp. 269-274, 2001.
[8] Ding C, Li E, Peng W, and Park H, "Orthogonal Nonnegative Matrix T-Factorizations for Clustering," 12th ACM SIGKDD International Conference Knowledge Discovery and Data Mining, pp. 126-135, 2006.
[9] Jain A, Murty M, and Flynn P, "Data Clustering: A Review", ACM Computing Surveys, volume 31, no. 3, pp. 264-323, 1999.
[10] Pensa R, GandBoulicaut J F, "Constrained Co-Clustering of Gene Expression Data," Proceedings SIAM International Conference Data Mining (SDM),
[11] Semi-Supervised Learning, ChapelleO,Scho¨lkopf B, and Zien A, eds. MIT Press, http://www.kyb.tuebingen.mpg.de/ssl-book, 2006.
[12] Shan H and Banerjee A, "Bayesian Co-Clustering", IEEE Eight International Conference Data Mining (ICDM), pp. 530-539, 2008.