ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

A novel approach to the hierarchical co-clustering in text mining with zone contents and metadata

Rucha Bhutada[1], D. A. Borikar[2]

Research Scholar, Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India [1]

Assistant Professor, Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India [2]

Abstract: Nowadays, in many text mining applications, eloquent quantity of information from document is present in the form of text. This text information contains distinct types of data such as metadata and zones where metadata can also be called as side information which includes title, name of author, document provenance information, links in the document, user access behavior from web logs and the content of zone can be abstract, body, conclusion etc. It becomes difficult to cluster both the types of information as this text information contains noise which can either improve the aspect of illustration of mining process or can count up noise to the process. As a result of this, there is a need of upright way to carry through mining process so as to increase the superiority of the text information. Co-Clustering discovers clusters of similar objects with regard to the value as well as clusters of similar features with regard to the object associated by them. Conforming to that, this paper represents review on most clustering and co-clustering techniques containing different kinds of data.

Keywords: clustering, coclustering, metadata, text mining, zone.

I. INTRODUCTION

Data mining is hugely used in normal life, something can be found from it. Many researchers have used manifold techniques such as classification, outlier detection, clustering, regression analysis etc. The clustering is used some special application. Clustering is mechanisms of combining set of physical or abstract objects into classes of similar objects. There are different orders or groups which is called cluster, subsist of objects that are correlated within themselves and unrelated to objects of other order or groups. Clusters of data objects can be taken together as one bunch and so may be examined as structure of data compression. Depicting data by fewer clusters cardinally exhausts certain refined analysis; however acquires interpretation. It exhibits frequent data objects by few clusters, and therefore it figure out data by its cluster. Data modeling brings clustering in a real broad view in mathematics and numerical analysis. Due to enormous amounts of data assembled in databases, cluster analysis has afresh inclined eminently functioning topic in data mining research. Clustering is a demanding field of research in which its plausible utilization poses their peculiar specification. Data mining comprises of many large databases so as to enforce on clustering study appended firm estimating fulfillment. Clustering is also called as data segmentation in some applications because clustering divides large data sets into bunch by considering presence of their resemblance. This cluster analysis has been recognized basis task in data mining. The terminology in the research composes clustering follow the distinct rational mark on the object.

Large number of techniques has recently become known that suitable to requisite and were strongly adapted to substantial problems in data mining for clustering and co-clustering. Clustering deals either with row or column. Han and Kamber have given the classic introduction to contemporary data mining clustering techniques [10]. Hierarchical clustering construct data in hierarchy structure in which the input sets of objects points is recursively disjoined into petite subgroups until these subgroups assumed to be single objects points. The essence of a clear hierarchical clustering deteriorates from its impotence to enforce the arrangement once converge and divide has been implemented. Clustering and co-clustering can be done on different kinds of data. The technique that simultaneously clusters row and column is called as co-clustering. Clustering can also be done on text data, called as text clustering. Text clustering has remodeled progressively significant technique as it is applicable to various areas such as in fast information retrieval, uprooting of automatic document, sentiment analysis, formulation of granular taxonomies. The suggestive clustering algorithms are applied on statistics and probability such as Word Term Frequency [1], Word Meaning Frequency [2], and Frequent Item set [3]. However, in text mining, text clustering is an issue for increasing eloquent quantity of information unregulated data which is present in various fields for instance any network. Data is not present in flawless contents pattern. This text information contained by document possesses different kinds of data such as metadata and zones. Metadata is also called as side information which involves links in the document, user access behavior from web logs, document provenance information, title, name of authors, date of publication etc whereas the contents of zone can be abstract, body, conclusion, subjective free text. Further, it can be difficult to merge this kind of information in the

DOI 10.17148/IJARCCE.2014.31248

clustering process as this information contains noisy which can either cultivate the aspect of representation of mining process or can even count up noise to the process. Co-clustering is a solution to merge two or more than two types of data. It interpolates to simultaneously clustering of more than one data type. Hierarchical co-clustering refers to concurrently clustering of more than two data points or objects points. In plain words, it attempts to acquire the objective of both hierarchical clustering and co-clustering. I. S. Dhillon proposed two algorithms which are information theoretic co-clustering algorithm and bipartite graph partitioning co-clustering algorithm. These are used to exploit the co-clustering of documents and words [7]. This results in developing the upright solution which can increase the excellence of this text information. These concerns induced to the emergence of vigorous widely pertinent data mining clustering and co-clustering approaches studied below.

The remainder of this paper is organized as follows. Section 2 discusses the related work. In section 3, the details of preprocessing, clustering and co-clustering of text data is described. Section 4 draws conclusion from this work.

II. LITERATURE REVIEW

There is a strong association between clustering techniques and several other approaches. Clustering is one of the most typical topics in the field of information retrieval and machine learning. It helps in feature compression and extraction to reduce dimensionality of feature vector by coupling related features into clusters. It has always been applied in statistics. Similarly, text clustering is also an important task for mining text data. The scheme of text clustering is to group the related text documents composed of well organized text data [3]. There are many proposed approaches for clustering and co-clustering. Frequently used words have been found by applying association rule mining. These words then matched with the documents so that clustering has been executed using flat clustering and hierarchical frequent term based clustering. The dimension of vector space model has been reduced in natural way. These methods can also be applicable to transaction data due to the similarity of text data [1]. Context sensitive language, high dimension of the documents do not satisfy the characteristics of text document clustering. Clustering is exploited on topics of documents which are coming frequently term sequences and frequently term meaning sequences. Closeness between words and word meaning contained by documents can be measured using such clustering. Generally, user needs to specify the desired number of clusters as input parameter [2].

Number of clusters is an optional input parameter in the proposed algorithm. Along with these, high clustering accuracy, mere in examining the meaningful cluster description is the advantage of frequent item set based on hierarchical clustering. In this paper also, frequent item sets is calculated from association rule mining [3]. Clustering is applied either on both the types of information, that is, text information and side information or on pure text information by using COATES algorithm and classification methods across many baseline techniques on real and scientific datasets. Further, the results obtained from proposed methods are to reinforce peculiarity of text clustering and classification by using the side information [4]. Cognitive situation has been taken out of the English paragraphs by facilitating semantic analysis. Illustration of features is extracted from elegantly chosen cognitive situation dimension. Formation of cognitive situation matrices yields to build clustering tree. The advantage is number of different cognitive situation can be handled [5].

Simultaneous clustering of documents and words is a problem which can be represented as bipartite graph partitioning, isoperimetric graph partitioning, optimization problem etc. By representing these problems, simultaneous clustering or co-clustering can be performed. Co-clustering of documents and words has been proposed by using bipartite spectral graph partitioning problem which was further resolved as a new spectral co-clustering algorithm that uses singular vector decomposition as the NP-complete objective. The quality of this document and words simultaneous clustering increases by implementing confusion matrix. In addition to this, purity and entropy is also determined by confusion matrix. Text data can be impersonated as contingency table or co-occurrence table [7]. Co-clustering in contingency table has been solved. The two dimensional table is appeared as empirical joint probability distribution and poses as optimization problem. The proposed algorithm gradually increases the processed bilateral information by tangling of rows and columns clustering at all levels. The disadvantage of this method is number of row and column clusters has to be pre-specified [8]. The next goal of this paper can be hard co-clustering in which an algorithm is to be designed for abstract multivariate clustering setting. To solve the problem of bipartite graph, Isoperimetric co-clustering technique can be applied. This algorithm has been implemented using a sparse system of linear equations. This method decreases perimeter and area of the bipartite graph partition under a pertinent definition [9].

Very few algorithms can co-cluster music type of data. Hierarchical co-clustering can also be applied on music data. Co-clustering is applied on artists and tags together, artists and styles together or artists and moods labels together by implementing agglomerative and divisive strategy of hierarchical co-clustering method. This processed is to be performed so as to understand the affiliation among artists/song [11]. Supervised and unsupervised constraints are generally used to achieve many clustering goals which can further be used in deriving document and word constraint. To additional upgrading of clustering fulfillment, there has also been aspiration on bringing co-clustering and constrained clustering together which has deficiency. The existing algorithm uses semi-supervised learning that depends upon human illustrated labels to construct constraints. This

DOI 10.17148/IJARCCE.2014.31248

method is difficult to implement, time consuming and costly. To solve this problem, there is a proposed approach called constrained information theoretic co-clustering algorithm. This algorithm performs better than existing one since it takes the benefit of the co-occurrences of documents and words and adds some constraints to monitor the clustering process. The more work can be done on this system by determining better text features using natural language processing or readily available tools [12].

III. PROPOSED WORK

The thrust of the approach is to induce clustering and co-clustering in which the zone content and metadata provide related hints about the behavior of the underlying clusters and co-clusters and concurrently neglecting those aspects in which clashing intimation are provided. Conventionally, hierarchical co-clustering consists of two major steps:

1. Clustering: the content of zone is clustered into number of clusters, so that similar objects categorize into same clusters.

2. Co-clustering: the content of metadata will be co-clustered so that similar objects categorize into same clusters which are not defined above. Co-clustering will be done inside the clustering so as to make the simultaneously clustering process in better manner.

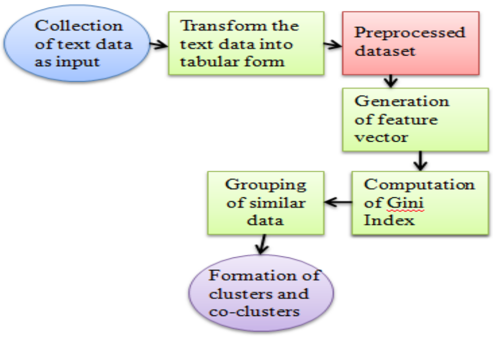
Collection of text data as input
Transform the text data into tabular form
Preprocessed dataset
Generation of feature vector
Computation of Gini Index
Grouping of similar data
Formation of clusters and co-clusters

Fig. 1: steps in generating clusters and co-clusters in text mining

Clustering of text data is one of the text mining tasks. Text mining deals with unstructured data. Text information fails to get the imposed structure of traditional database, though it explicitly has very wide range of information. Thus, it is important to represent thus unstructured data into structured form, so that appropriate patterns and features can be retrieved from this text information. Any text dataset can have noise, missing data or inconsistency. Thus, applying the preprocessing is again a crucial step in text mining. This preprocessing removes punctuations and stop words, stemming and white spaces are required to obtain the well structured data. However, preprocessed data needs to reduce the dimension as it is of large size and also administering further process requires more effort as many algorithm requires numerical representation of objects since such representation facilitate processing and statistical analysis. Feature vector is multi-dimensional vector of numerical features that imitates some objects. Generation of feature vector is to reduce the dimensions of the preprocessed data. Therefore, it is essential to reduce the dimension by using Gini Index. The Gini index measures the inequality among values of a frequency distribution. Mathematically, Gini index is defined as:

$$G = A\,/(A + B) \qquad \text{...Eq (1)}$$

where A is single attribute and B is all the attributes.

In this feature selection process, Gini index will be computed for each attribute with respect to class labels. If the Gini index is γ standard deviation (or more) below the average gini index of all attributes, then these attributes are excluded globally and will never be used in further clustering process. These data are grouped together based on some similarity measure into different clusters. Similarity between all pairs of clusters and co-clusters is computed to form a similarity matrix. These feature zone text data and feature metadata can be clustered and co-clustered into hierarchical structure suitable for browsing which suffers efficiency problems by using k-means and k-medoids algorithm. Retrieving information from such groups turns out extensively manageable task.

IV. CONCLUSION

Co-clustering problems on text data mainly focus on simultaneous clustering which can improve the simple clustering process efficiently. This paper presented a brief review on clustering and co-clustering techniques. The central theme in many of these is providing dimensionality reduction to improve text data co-clustering process. Co-clustering and clustering helps in reducing the number of features which again subsequently help in reducing the number of dimensions.

A very important goal is to achieve high quality co-clustering process. Many forms of text databases contain a large amount of information including zone and metadata which may be used in order to improve the co-clustering technique. Co-clustering on text data usually requires – first, parsing that converts unstructured to structure by exploiting text preprocessing operations. Second, feature extraction from structured text data by computing Gini index. Third, similarity measure has to be calculated on the feature text data by mining knowledge. Fourth, efficient technique for clustering and co-clustering is to be applied to form the clusters and co-clusters with similar text document.
The review presented in this work is being used for further endeavoring distinctive process so as to maximize the co-clustering mechanisms.

ACKNOWLEDGEMENTS

The authors are grateful to the Head, Computer Science and Engineering Department and the Principal, Shri Ramdeobaba College of Engineering and Management for providing adequate facilities to conduct a research. Authors are thankful to the faculty members of Computer Science and Engineering Department for their support and cooperation during this work.

REFERENCES

[1]. F. Beil, M. Ester, X. Xu, Frequent term-based text clustering, in *Proc of ACM SIGKDD International Conf on Knowledge Discovery and Data Mining*, pp. 436-442, 2002.

[2]. Y. Li, S. M. Chung, J. D. Holt, Text document clustering based on frequent word meaning sequences, *Data and Knowledge Engineering*, 64(1), pp. 381-404, 2008.

[3]. B. C. M. Fung, K. Wang, M. Ester, Hierarchical document clustering using frequent itemsets, in *Proc of SIAM International Conference on Data Mining*, 2003.

[4]. Charu C. Aggrawal, Yuchen Zhao, and Philip S. Yu, On the Use of Side Information for Mining Text Data, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 6, June 2014.

[5]. Yi Guo, Zhiqing Shao, Nan Hua, A Hierarchical Text Clustering Algorithm with Cognitive Situation Dimensions, *2nd International Workshop on Knowledge Discovery and Data Mining*.

[6]. Jing Wang, Yang Jing, Yue Teng, Qingling Li, A Novel Clustering Algorithm for Unsupervised Relation Extraction

[7]. I. S. Dhillon, Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning, in *Proc. 7th ACM SIGKDD Int. Conf Knowledge Discovery and Data Mining*, 2001, pp. 269-274.

[8]. I. S. Dhillon, S. Mallela, and D. S. Modha, Information-Theoretic Co-Clustering, in *Proc. 9th ACM SIGKDD International Conf. Knowledge Discovery and Data Mining*, 2003, pp. 89-98.

[9]. Manjeet Rege, Ming Dong and Farshad Fotouhi, Co-clustering documents and words using Bipartite Isoperimetric Graph Partitioning, in *Proc of the Sixth International Conference on Data Mining* (ICDM'06)

[10]. Jiawei Han and Michelie Kamber, *Data Mining Concepts and Techniques* 500 Sansome Street, Suite 400, San Francisco, CA 94111, Morgan Kaufmann

[11]. Jingxuan Li, Bo Shao, Tao Li and Mitsunori Ogihara, Hierarchical Co-Clustering: A New Way to Organize the Music Data, *IEEE Transactions on Multimedia*, VOL. 14, NO. 2, APRIL 2012

[12]. Yangqiu Song, Shimei Pan, Shixia Liu, Furu Wei, Michelle X. Zhou, Weihong Qian, Constrained Text Coclustering with Supervised and Unsupervised Constraints, *IEEE Transactions on Knowledge and Data Engineering*, VOL. 25, NO. 6, JUNE 2013

BIOGRAPHIES

Rucha Bhutada earned her B.E degree in computer science and engineering in 2013 from SGBAU Amravati University, Amravati (India). She is a pursuing Masters in Technology in Computer Science and Engineering from Shri Ramdeobaba College of Engineering and Management, Nagpur-440013. Her areas of interest include Data Mining and Data Hiding

Dilipkumar A. Borikar earned his B. E. (Computer Technology) degree and M.B.A. (Finance and Marketing) in 1998 and 2001 respectively, from RTM Nagpur University, Nagpur (India). He obtained M. Tech. (Information Technology) from the School of Information Technology, Indian Institute of Technology, Kharagpur, West Bengal, India in 2009. He was awarded with Institute Silver Medal of IIT Kharagpur for 2009.

He has been in academics for over 14 years and is currently working as Assistant Professor in Computer Science and Engineering at Shri Ramdeobaba College of Engineering and Management, Nagpur.

He has published 07 technical papers in international conferences and journals. He is a life member of ISTE, New Delhi. His research interest includes information processing, access control and security, soft computing, image processing, and multidimensional databases.

DOI 10.17148/IJARCCE.2014.31248