International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur,Tamilnadu, India.

ENHANCED DOCUMENT CLUSTERING APPROACH USING MULTI VIEW POINT APPROACH

D.Hemavathi,
Research Scholar,
Department of Computer Science,
Kongunadu Arts and Science College,
Coimbatore,Tamilnadu,India.

Abstract: Document clustering intends to automatically group associated documents into clusters. This proposed work presents a new spectral clustering technique called Correlation Preserving Indexing (CPI), which is performed in the correlation similarity measure space. In this framework, the documents are projected into a low dimensional semantic space in which the correlations between the documents in the local patches are maximized while the correlations between the documents outside these patches are minimized simultaneously. Since the intrinsic geometrical structure of the document space is often embedded in the likeness among the documents, correlation as a similarity determine is more appropriate for detecting the intrinsic geometrical structure of the document space than Euclidean distance. Accordingly, the proposed CPI technique can effectively find out the intrinsic structures embedded in high-dimensional document space. In addition, the proposed work considers the major variation among a traditional dissimilarity/similarity measures and is that the former uses only a single viewpoint, which is the origin, while the latter make use of many different viewpoints, which are objects assumed to not be in the same cluster with the two objects being measured. More informative assessment of similarity could be achieved by using multiple viewpoints.

Keywords: Correlation Preserving Indexing (CPI), Viewpoints, Cluster, Correlations.

I.INTRODUCTION

Based on a variety of distance measures, a number of techniques have been planned to handle document clustering. A usual and extensively used distance measure is the Euclidean distance. The K-means technique is one of the way that use the Euclidean distance, which reduce the sum of the squared Euclidean distance among the data points and their related cluster centers. Since the document space is forever of high dimensionality, it is preferable to find a low-dimensional representation of the documents to minimize calculation complexity. Low calculation cost is reached in spectral clustering techniques, in which the documents are first predictable into a low-dimensional semantic space and then a usual clustering algorithm is useful to finding document clusters. Latent Semantic Indexing (LSI) is one of the efficient spectral clustering techniques, intended at finding the best subspace estimation to the original document space by reduce the global reconstruction error (Euclidean distance).

Though, because of the high dimensionality of the document space, a certain demonstration of documents regularly resides on a nonlinear manifold embedded in the similarities among the data points. Unluckily, the Euclidean distance is a dissimilarity compute which describes the dissimilarities rather than similarities between the documents. Thus, it is not able to effectively capture the nonlinear manifold structure embedded in the similarities among them. An effective document clustering technique must be able to locate a low-dimensional depiction of the documents that can best conserve the similarities among the data points.

Locality Preserving Indexing (LPI) technique is a dissimilar spectral clustering technique based on graph separation theory. The LPI technique applies a weighted function to each pair wise distance attempting to focus on capturing the resemblance structure, rather than the difference structure, of the documents. Though, it does not overcome the essential restriction of Euclidean distance. In addition, the selection of the weighted functions is often a difficult task. Concurrently, clustering still needs more robust dissimilarity or similarity procedures. The project is motivated by investigations from the above and similar research findings. It appears that the nature of similarity measure plays a very important role in the success or failure of a clustering method.

Clustering is a group of data into a subsets in which the manner of identical substances are grouply collected together that gives an instances within the different group. The instance is also arranged into an effectively depiction that gives a very characterizes into a document is sampled. Clustering of objects is as ancient as the human need for recitation the salient individuality of men and objects and recognize them with a type. Therefore, it squeeze a choice of scientific authority: from mathematics and statistics to biology and genetics, the entire of which uses different terms to describe the topologies formed using this analysis. From biological "taxonomies", to medical "syndromes" and genetic "genotypes" to developed "group technology" the trouble is same: forming groups of unit and transfer individuals to the proper groups within it. Since Clustering is the grouping of similar instances/entities, a number of compute that can choose whether two objects are related or dissimilar is necessary.

A common move towards the clustering trouble is to treat it as an optimization process. A best partition is found by optimizing a exacting purpose of similarity (or distance) among data. Basically, there is an concealed supposition that the true inherent arrangement of data could be properly describe by the similarity formula defined and fixed in the clustering important factor. Hence, efficiency of clustering algorithms under this advance depends on the accurateness of the similarity measure to the data at hand. For example, the original k-means has sum-of-squared-error intent function that use Euclidean distance. In a very sparse and high dimensional realm like text documents, spherical k-means, which uses cosine similarity (CS). Euclidean distance as the calculate, is deemed to be more appropriate. Correlation coefficient is consistent angular division through centering the co-ordinates to its mean value. The value between -1and +1. Correlation measure similarity rather than distance or dissimilarity. Similarity is fairly tricky to calculate. Similarity is compute that reflects the strong point of connection between two objects or two features. This quantity is usually having series of either -1 to +1 or normalize into 0 to 1. If the similarity between attribute i and attribute j is denoted by Sij, we can measure this quantity in numerous ways depending on the level of measurement that we have. Dissimilarity measure the discrepancy between the two objects based on several features. Variation may also be viewed as measure of chaos between two objects.

II.RELATED WORK

In the proposed scheme "Efficient and Effective Clustering Methods for Spatial Data Mining" the authors Raymond T. Ng and Jiawei Han stated that spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. They explored whether clustering technique have a position to play in spatial data mining. To this last part, they developed a new clustering technique called CLAHANS which is based on randomized search. They also developed two spatial data mining algorithms that use CLAHANS. Their examination and experiments show that with the support of CLAHANS, these two algorithms are very efficient and can guide to finding that are hard to locate with current spatial data mining algorithms. In addition, experiments carry out to evaluate the performance of CLAHANS with that of existing clustering techniques show that CLAHANS is the most efficient.

Data mining in common is the search for unseen patterns that may exist in large databases. Spatial data mining in particular is the discovery of attractive relationships and characteristics that may be present implicitly in spatial databases. Because of the massive amounts (regularly, terabytes) of spatial data that may be find from satellite images, medical equipments, video cameras, etc., it is costly and often impractical for users to examine spatial data in detail. Spatial data mining plans to automate such a knowledge discovery procedure. Thus, it plays an significant role in a) extracting attractive spatial patterns and features; b) capturing intrinsic relationships among spatial and non-spatial data; c)presenting data regularity concisely at higher levels. Many brilliant studies on data mining have been carry out, such as those reported in considers the trouble of inferring categorization functions from samples; studies the trouble of mining association rules among sets of data items; proposes an attribute oriented approach to knowledge discovery; develops a visual feedback querying system to carry data mining; and includes many attractive studies on various issues in knowledge discovery such as finding functional dependencies among attributes.

Though, most of these studies are concerned with knowledge finding on non-spatial data, and the learn most relevant to our focus here is which studies spatial data mining. More specifically, recommend a spatial data dominant knowledge removal algorithm and a non spatial data dominant one, both of which plan to get out high level associations among spatial and non spatial data.

On the other hand, clustering is a complex trouble combinatorial, and difference in assumptions and contexts in various communities had made the relocate of useful generic thought and methodologies slow to happen. That paper presents an overview of pattern clustering techniques from a statistical pattern recognition point of view, with an aim of providing useful opinion and references to fundamental concepts accessible to the broad community of clustering practitioners. The clustering troubles have been deal with in many contexts and by researchers in many regulation; this replicate its broad appeal and usefulness as one of the steps in exploratory data analysis.

III.BACKGROUND STUDY

PROBLEMS IN PRESENT SYSTEM

Document clustering intends to automatically group associated documents into clusters. This project presents a new spectral clustering technique called Correlation Preserving Indexing (CPI), which is execute in the correlation similarity measure space.

In this framework, the documents are projected into a low-dimensional semantic space in which the correlations between the documents in the local patches are maximized while the correlations between the documents outside these patches are minimized simultaneously. Since the intrinsic geometrical arrangement of the document space is often embedded in the similarities among the documents, correlation as a similarity compute is more suitable for detecting the intrinsic geometrical construction of the document space than Euclidean distance.

Consequently, the proposed CPI method can effectively discover the intrinsic structures embedded in high-dimensional document space. In addition, the project considers the major dissimilarity among a established dissimilarity/similarity measures and is that the former uses only a particular viewpoint, which is the origin, while the latter operate many different viewpoints, which are objects assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved and so it is implemented in this proposed work.

IV.PROPOSED METHODOLOGY

Proposed System

The existing system is implemented in the proposed system also. In addition, for better clustering, data cleaning process such as stemming and stop word removal is applied. Then synonym word replacement is made which increases the relativity among the documents. Two dimensions are averaged into one dimension and so three dimension data is converted into two dimension data and then clustering is applied. For example, a document's location in a 2D plane is fitted with X axis means the time in which the document is prepared. The count of main words present in the document is plotted as Y-axis. For three dimensional data, e.g., the synonym words count in the document is also to be considered, then the average of main words and synonyms words is prepared and plotted as Y axis. The main reason for the development of the proposed system is the two objects to be measured must be in the same cluster, while the points from where to establish this measurement must be outside of the cluster, so that similarity can be measured well. The existing system is implemented in the proposed system also. In addition, for better clustering, data cleaning process such as stemming and stop word removal is applied. Then synonym word replacement is made which increases the relativity among the documents. Two dimensions are averaged into one dimension and so three dimension data is converted into two dimension data and then clustering is applied. For example, a document's location in a 2D plane is fitted with X axis means the time in which the document is prepared. The count of main words present in the document is plotted as Y-axis. For three dimensional data, e.g., the synonym words count in the document is also to be considered, then the average of main words and synonyms words is prepared and plotted as Y axis. The main reason for the development of the proposed system is the two objects to be measured must be in the same cluster, while the points from where to establish this measurement must be outside of the cluster, so that similarity can be measured well.

1) Data Cleaning Process

In this method, word and stem word is added in to 'stem word' table. Likewise, stop word is added into 'stop word' table. The word and its synonym word is added into synonym table. In stem word process, removal of prefixes and suffixes has been performed. In stop word process, removal of articles and lexical words has been added. In synonym word process, the word is replaced by a unique word where it has same meaning with different words in various document.

2) Correlation

A correlation co-efficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other. However, correlation does not imply causation. There may be, for example, an unknown factor that influences both variables similarly.

In this method, two documents are selected. Then the vector standards for two documents are found out. The cosine comparison measure is applied. Then the correlation among two documents is found out using the following formula.

$$Corr(u,v)=\frac{u^T v}{\sqrt{u^T v}\,\sqrt{u^T v}}=\left\langle \frac{u}{\|u\|},\frac{v}{\|v\|}\right\rangle$$

Correlation Formula

For example, the string "I have to go to school" is present in one document the string "I have to go to temple" is present in other document.

Then the data is prepared such that

[i , have , to , go, school , temple] = [1,1,2,1,1,0]

[i , have , to , go , school , temple] = [1,1,2,1,0,1]

[i , have , to , go , school , temple] = [1,1,2,1,0,1]

Formula: cos = 1*1 + 1*1 + 2*2 + 1*1 + 1*0 + 0*1 / sqrt((1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 0^2) * 1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 0^2)

If the value derived by above formula is 1, then the documents are exactly matched, otherwise if it is 0, then it doesn't similar.

3) K-Means Clustering For 3-D Data

K-means clustering algorithm is used to cluster a collection of 2D data points. The data points are taken from the database with any two columns such as entry time of document versus average of main words count and synonym words count. In K-means clustering, the target is to cluster a set of data points to a predefined number of clusters. An iteration of the algorithm produces a set of cluster centers where it is compared with the set of cluster centers produced during the previous iteration. The total error is the difference between the cluster centers produced at nth iteration and the cluster centers produced at (n-1) th iteration.

The iterations continue until the error reduces to a predefined threshold value. In K-means clustering, each map function gets a portion of the data, and it needs to access this data split in each iteration. These data items do not modify over the iterations, and it is loaded once for the complete set of iterations. The variable data is the present cluster centers calculated during the previous iteration and hence used as the input value for the map function. This method is used to reduce the three dimensions data into two dimensions and then apply clustering.

4) Multi View Point Based Similarity Measure

In this method, instead of viewing two documents similarity from the same cluster (the documents belongs to), viewing in different viewpoints is considered. This will assists analyze both inter cluster similarity and intra cluster similarity. Given a large enough number of viewpoints and their variety, it is reasonable to assume that the majority of them will be useful. Hence the result of misleading viewpoints is constrained and compacted by averaging step it can be seen that this technique

offers more informative assessment of similarity than the single origin point based likeness measure.

The data mining and their method are well known technique for automated information analysis. According to the nature of information, the data mining algorithms are functional for dissimilar kinds of pattern recovery from raw data. In this accessible work the clustering practice is main area of study. The clustering is an unsupervised approach of data examination, where the data objects are evaluated on the basis of their internal comparison and the user distinct groups of data is prepared. But the clustering approaches are not much correct that is needed to be improving for truthful pattern classification. With this inspiration the proposed work is focused on document based cluster analysis method. In order to organize the document based clustering the k-mean algorithm is one of the most admired approaches, therefore in this obtainable work an improved document clustering advance using the multi-view approach is proposed and executed. The presented approach includes two phase of clustering first information with the pre-defined patterns or groups, and in next phase utilizing the domain information for performing the cluster for incoming documents. During the training process the proposed system implement the noise reduction technique using the stop word removal and the special character removal technique. In next process the feature removal technique is used where the two technique are apply first is implement on the basis of word frequency in a particular domain and secondly the significance of a word in a given domain.

V.CONCLUSION

A new document clustering method based on correlation preserving indexing is used which results in better correlation identification is achieved. Maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches Considers both single view point and multiple viewpoints so that inter and intra cluster similarity can be analyzed effectively. In existing system, text retrieval results are sensitive and also traditional text search engines cover only one end of the whole spectrum of information retrieval needs, which is a narrowly specified search for documents matching the user's query. They are not capable of meeting the information retrieval needs from the rest part of the spectrum. The above problems can be solved to certain degrees by clustering documents according to their topics and main contents. In the future, it would also be possible to apply the clustering using hierarchical clustering algorithms so that intra and inter cluster similarity can be analyzed well. It would be interesting to explore how they work on other types of sparse and high-dimensional data. Document clustering methods have been receiving more and more attentions as a basic and enabling tool for efficient organization, navigation, retrieval, and summarization of vast volumes of text documents. With a good document clustering technique, computers can automatically categorize a document corpus into a meaningful cluster hierarchy, which allow an efficient browsing and navigation of the corpus.

VI.REFERENCES

[1]. I. Dhillon and D. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, vol. 42, nos. 1/2, pp. 143-175, Jan. 2001.
[2]. S. Zhong, "Efficient Online Spherical K-means Clustering," Proc. IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.
[3]. D. Lee, J. Lee,—Dynamic dissimilarity measure for support based clustering‖, IEEE Trans. on Knowl. and Data Eng., Vol. 22, No. 6, pp. 900–905, 2010.
[4]. R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. John Wiley & Sons, 2001.
[5]. A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions," J. Machine Learning Research, vol. 6, pp. 1345-1382, Sept. 2005.
[6]. W. Xu, X. Liu, and Y. Gong, "Document Clustering Based on Non-Negative Matrix Factorization," Proc. 26th Ann. Int'l ACM SIGIR
[7]. Conf. Research and Development in Informaion Retrieval, pp. 267- 273,2003.
[8]. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J.Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.
[9]. S. Zhong and J. Ghosh, "A Comparative Study of Generative Models for Document Clustering," Proc. SIAM Int'l Conf. Data Mining Workshop Clustering High Dimensional Data and Its Applications,2003.
[10]. 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.
[11]. Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. In Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science, pages 238250, Vancouver, Canada, November 2002
[12]. Taiping Zhang, Member, IEEE, Yuan Yan Tang, Fellow, IEEE, Bin Fang, Senior Member, IEEE, and Yong Xiang "Document Clustering in Correlation Similarity Measure Space" IEEE Trans. KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012
[13]. Duc Thang Nguyen, Lihui Chen, Senior Member, IEEE, and Chee Keong Chan Clustering with Multiviewpoint-Based Similarity Measure IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 6, June .
[14]. Hamidreza Mirzaei School of Computing Science, Simon Fraser University, Burnaby, Canada hmirzaei@cs.sfu.ca "A Novel Multi-View

Agglomerative Clustering Algorithm Based on Ensemble of Partitions on Different Views" 2010 International Conference on Pattern Recognition.

[15]. Anwiti Jain, Anand Rajavat, Rupali Bhartiya, "An Efficient Modified K-Means Algorithm To Cluster Large Data-set In Data Mining", International Journal of Advanced Research in Computer Science and Electronics Engineering Volume 1, Issue 3, May 2012

[16]. Gurjit Kaur, Lolita Singh, "Data Mining: An Overview", IJCST Vol. 2, Issue 2, June 2011, ISSN: 2229-4333(Print) | ISSN: 0976-8491(Online)

[17]. "An Introduction to Data Mining: Discovering hidden value in your data warehouse", http://www.thearling.com/text/dmwhite/dmwhite.htm

[18]. Manoj and Jatinder Singh, "Applications of Data Mining for Intrusion Detection", International Journal of Educational Planning & Administration. Volume 1, Number 1 (2011), pp. 37-42

[19]. M. Rajalakshmi, M. Sakthi, "Max-Miner Algorithm Using Knowledge Discovery Process in Data Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11, November 2015

[20]. SMRITI SRIVASTAVA & ANCHAL GARG, "DATA MINING FOR CREDIT CARD RISK ANALYSIS: A REVIEW", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Vol. 3, Issue 2, Jun 2013, 193-200 [21] Dipti Verma and Rakesh Nashine, "Data Mining: Next Generation Challenges and Future Directions", International Journal of Modeling and Optimization, Vol. 2, No. 5, October 2012

[21]. Tanu Verma, Renu, Deepti Gaur, "Tokenization and Filtering Process in RapidMiner", International Journal of pplied Information Systems (IJAIS) – Foundation of Computer Science FCS, New York USA Volume 7– No. 2, April 2014

[22]. Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009

[23]. Rene Witte and Qiangqiang Li, Yonggang Zhang and Juergen Rilling, "Text Mining and Software Engineering: An Integrated Source Code and Document Analysis Approach", e IET Software Journal, Vol. 2, No. 1, 2008

[24]. Nanasaheb Mahadev Halgare, Dharmaraj V. Biradar, "MPROVED ALGORITHM ON DYNAMIC CLUSTERING USING METAHEURISTICS IN ADVANCE DATA MINING", International Journal of Enterprise Computing and Business Systems ISSN (Online): 2230-8849 Volume 6 Issue 1 January - June 2016

[25]. E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, "Webace: a web agent for document categorization and exploration," in AGENTS '98: Proc. of the 2nd ICAA , 1998, pp. 408–415.

[26]. J. Friedman and J. Meulman, "Clustering objects on subsets of attributes," J. R. Stat. Soc. Series B Stat. Methodol., vol. 66, no. 4, pp. 815–839, 2004.

[27]. L. Hubert, P. Arabie, and J. Meulman, Combinatorial data analysis: optimization by dynamic programming . Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2001.

[28]. R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. New York: John Wiley & Sons, 2001.

[29]. S. Zhong and J. Ghosh, "A comparative study of generative models for document clustering," in SIAM Int. Conf. Data Mining Workshop on Clustering High Dimensional Data and Its Applications, 2003.

[30]. Y. Zhao and G. Karypis, "Criterion functions for document clus-tering: Experiments and analysis," Dept. of Computer Science, Uni. of Minnesota, Tech. Rep., 2002.

