International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 7 Issue 1 –MARCH 2014.

A NOVEL APPROACH FOR DOCUMENT RETRIEVAL USING HAC BASED ON MULTI-VIEW POINT SIMILARITY MEASURE

KALAIVENDHAN.K[1], SUMATHI.P[2]
PG Student[1], Assistant Professor[2]
Dept. of Computer Science and Engineering
KSR Institute for Engineering and Technology
Tiruchengode, Namakkal,
TamilNadu

***Abstract* --Clustering is one of the most interesting and important topics in data mining. The aim of clustering is to find relationship among the data objects, and classify them into meaningful subgroups. The effectiveness of clustering algorithms depends on the appropriateness of the similarity measure between the data in which the similarity can be computed. In this paper, a novel method for measuring similarity between data objects particularly text documents is introduced which uses the Multi-View point based Similarity Calculation. With the proposed similarity measure, we then implement Hierarchical Clustering Algorithm which forms the document groups. From the clustered objects, the document retrieval can be done based on the query. The query is preprocessed then it is matched with the documents in the clusters. Ranking is provided for the Clusters with respect to the query matching result. The most relevant Cluster to the query will be retrieved with this approach. From this, more informative assessment of similarity could be achieved between the documents.**

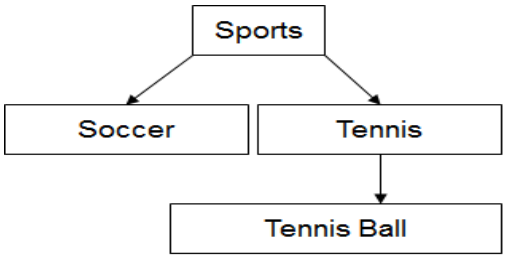
***Index Terms*—Document Clustering, Correlation Similarity, Hierarchical Clustering.**

I. INTRODUCTION

In recent years, an increasing number of usage of data sets have become available. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data. Data mining, the extraction of hidden predictive information from large databases, isa powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyze offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems.

A.Clustering

Clustering is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. Many clustering algorithms published every year.



They can be proposed for very distinct research fields, and developed using totally different techniques and approaches. Nevertheless, according to a recent study, more than half a century after it was introduced, the simple algorithm k-means still remains as one of the best data mining algorithms. It is the most frequently used partitional clustering algorithm in practice. Another recent scientific discussion states that k-means is the favorite algorithm that practitioners in the related fields choose to use.

k-means has more than a few basic drawbacks, such as sensitiveness to initialization and to cluster size, and its performance can be worse than other state-of-the-art algorithms in many domains. In spite of that, its simplicity, understandability, and scalability are the reasons for its tremendous popularity. An algorithm with adequate performance and usability in most of application scenarios could be preferable to one with better performance in some cases but limited usage due to high complexity. While offering reasonable results, k-means is fast and easy to combine with other methods in larger systems.

II. EXISTING SYSTEM

The internal Structure of the data will be find and organize them into a meaningful groups. Existing Systems greedily picks the next frequent item set in the next cluster. The clustering result depends on the order of picking up the item sets. K-means method related fields data's are processed. Used a cosine similarity for find out the dissimilar document object in the cluster. Existing system proposed a multi viewpoint algorithm for move the dissimilar document object from one cluster to another cluster. The second similarity measures similarity between the dissimilar document object and the other cluster groups document objects.

i. Cosine Similarity
ii. Increment Mining

A.Cosine Similarity

It is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgement of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

Note that these bounds apply for any number of dimensions, and Cosine similarity is most commonly used in high-dimensional positive spaces. For example, in Information Retrieval and text mining, each term is notionally assigned a different dimension and a document is characterised by a vector where the value of each dimension corresponds to the number of times that term appears in the document.

Cosine similarity will gives a useful measure of how similar two documents are likely to be in terms of their subject matter. One of the reasons for the popularity of Cosine similarity is that it is very efficient to evaluate, especially for sparse vectors.

B. Incremental Mining

It is Multi View-point Similarity(MVS) algorithm. A Matrix is generated by using MVS. By building matrix the similarity between documents can be identified. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions such as I_R and I_V used for document clustering are proposed based on this new measure. Comparison is made with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the similarity of clusters.

III. PROPOSED SYSTEM

Propose a new method to group the documents into cluster. Cosine similarity is used to find out the dissimilar document object in the cluster. Similarity measures will depend on the text mining. A multi viewpoint algorithm for move the dissimilar document object from one cluster to another cluster. The Correlation similarity will measures similarity between the dissimilar document object and the other cluster groups document objects. Multi-View point based Similarity Calculation is used for measuring similarity between data objects. With the proposed similarity measure Hierarchical Agglomerative Clustering Algorithm (HAC) is implemented in which forms the document groups. From the clustered objects, the document retrieval can be done based on the query. The query is preprocessed then it is matched with the documents in the clusters. Ranking is provided for the clusters with respect to the query matching result. The most relevant cluster for the query will be retrieved with this approach.

A. Correlation Similarity

Correlation similarity is the combination of
(i)Distance Covariance
(ii)Distance Variance

In this, the similarity will be done between the each and every document of the cluster which improves the accuracy similarity of the document clusters.

B. HAC Algorithm

HAC stands for Hierarchical Agglomerative Clustering algorithm which form the structure like tree. This is one of the Hierarchical cluster analysis method which performs “bottom up” approach. The observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

A.Data Preprocessing

Stemming
Stop word removal

Stemming

Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations. Instead, a typically smaller list of "rules" is stored which provides a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include:

- if the word ends in 'ed', remove the 'ed'
- if the word ends in 'ing', remove the 'ing'
- if the word ends in 'ly', remove the 'ly'

Suffix stripping approaches enjoy the benefit of being much simpler to maintain than brute force algorithms, assuming the maintainer is sufficiently knowledgeable in the challenges of linguistics and morphology and encoding suffix stripping rules.

Stop Word Removal

Stop words are words which are filtered out prior to, or after, processing of natural language data (text). There is not one definite list of stop words which all tools use and such a filter is not always used. Some tools specifically avoid removing them to support phrase search.

Any group of words can be chosen as the stop words for a given purpose. For some search machines, these are some of the most common, short function words, such as the, is, at, which, and on.

B.Similarity Matrix

Construct the similarity matrix for document set between two documents. Measure the cosine similarity value for two similarity vectors to compute the similarity value. Multi view point similarity is used to calculate similarity for each two document. Then the most dissimilar object for the cluster group is got. Remove the document from the cluster group and declare that the document as a outlier for the cluster group. Finally, similarity matrix is computed.

C.Document Clustering

Get the similarity matrix. Compute the correlation similarity for each document with this outlier document. Documents are clustered using Hierarchical Agglomerative Clusteringalgorithm by grouping the levels of clusters.

D.Retrieval of Documents

Get the query from the user. Calculate the cluster weight using TF-IDF weighting approach. The weight is calculated by the means content based retrieval. Clusters are Ranked according to cluster weight by using ranking technique then most relevant documents are displayed.

IV.RESULT AND DISCUSSION

In this we use any type of text document for find similarity by using the formula

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

And we get the similarity valu as follows,

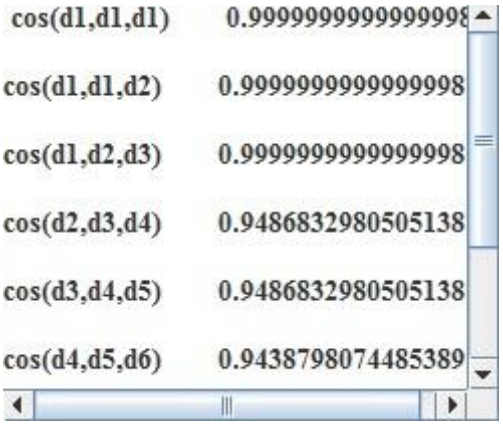


Fig 2.Cosine similarity

After grouping into clusters the HAC algorithm is evaluated and minimized clusters are formed. The dissimilarity document is moved to related cluster from one cluster to another cluster is shown.

Then the document which is misclassified are analysed using HAC and the document is moved corresponding to most similar cluster.

V.CONCLUSION

In this, new techniques called HAC and Correlation similarity is used for any type of text document to display the most relevant document of the clusters. The Correlation similarity and HAC algorithm

will makes similarity and document retrieval more accuracy than the cosine similarity and MVS algorithm. Cluster weight is calculated using weighted approach called TFIDF. Document ranking method is used to rank the documents of the cluster.In this, study is made about domain knowledge and also the literature survey is conducted in the area of clustering techniques and algorithm. The design of proposed system is prepared to solve the problem in the existing system.

REFERENCES

1. DucThangNguyen and CheeKeongChan 'Clustering with Multiviewpoint-Based Similarity Measure', IEEE Trans on Knowledge and Data Eng., Vol. 24, No. 6, 2012.

2. Banerjee, A. and Sra, S. ,'Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions,' J. Machine Learning Research, vol. 6, pp. 1345-1382 , 2005.

3. CharuAggarwal, C. andCheng Xiang, 'A Survey of Text Clustering Algorithms', Proc. SIAM Int'l Conf. Data Mining Workshop Clustering Algorithm's and its Applications, 2005.

4. Dhillon, I.S. '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.

5. C. Ding, and H. Simon, 'A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering', Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 107-114 , 2001.

6. J. Friedman , and J. Meulman, 'Clustering Objects on Subsets of Attributes', J. Royal Statistical Soc. Series B Statistical Methodology, vol. 66, no. 4, pp. 815-839, 2004.

7. J. Ghosh,and S. Zhong, '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.

8. D. Ienco, and R. Meo, 'Context-Based Distance Learning for Categorical Data Clustering', Proc. Eighth Int'l Symp. Intelligent Data Analysis (IDA), pp. 83-94, 2009.

9. P. Lakkaraju, and M. Speretta, 'Document Similarity Based on Concept Tree Distance', Proc. 19th ACM Conf. Hypertext and Hypermedia, pp. 127-132, 2008.

10. V. Leela Prasad, and B. SimmiCintre, 'Analysis of Novel Multi-Viewpoint Similarity Measures', Int'l Journal of Engineering Research and Applications ISSN: 2248-962 Vol. 2, Issue 4,pp.409-420, 2012,

11. S. Merugu, and J. Ghosh, 'Clustering with Bregman Divergences', J. Machine Learning Research, vol. 6, pp. 1705-1749, 2005.

12. D. Modha, and I. Dhillon, 'Concept Decompositions for Large Sparse Text Data Using Clustering', Machine Learning, vol. 42, nos. 1/2, pp. 143-175, 2001.

13. R. D. Nowak, and R. M. Castro, 'Likelihood Based Hierarchical Clustering ',IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229, 2011.

14. M. Pelillo, 'What Is a Cluster? Perspectives from Game Theory', Proc. NIPS Workshop Clustering Theory, 2009.

15. Xu, W. and Y. Gong, 'Document Clustering Based on Non- Negative Matrix Factorization', Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 267-273, 2003.

16. Yi Wang and YipingKe, 'A Model-Based Approach to Attributed Graph Clustering', Proc. Second Int'l Conf. Autonomous Agents (AGENTS '98), pp. 408-415, 2008.

17. F. Zarrinkalam, and M.Kahani , 'A New Metric For Measuring Relatedness Of Scientific Papers Based On Non-Textual Features', Intelligent Information Management,4, 99-107, 2012.

18. Y. Zhao, and G. Karypis, 'Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering', Machine Learning, vol. 55, no. 3, pp. 311-331, 2004.