IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 13, Issue 1 (Jul. - Aug. 2013), PP 27-30
www.iosrjournals.org

Correlation Preserving Indexing Based Text Clustering

Venkata Gopala Rao .S [1], A. Bhanu Prasad[2]
[1](M.Tech, Software Engineering, Vardhaman College of Engineering/ JNTU-Hyderabad, India)
[2](Associate Professor, Department of IT, Vardhaman College of Engineering/ JNTU-Hyderabad, India)

Abstract: *In Document clustering previously they presented new document clustering method based on correlation preserving indexing. It simultaneously maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. Consequently, a low dimensional semantic subspace is derived where the documents corresponding to the same semantics are close to each other with learning level parsing procedure based CPI method. The proposed CPI method with learning level parsing procedure is to find correlation between relational documents to avoid maximum unknown clusters those are not effectual to find exact correlation between documents depend on accuracy of sentences. The proposed CPI method with learning level parsing procedure in document clustering doubles the accuracy of previous correlation coefficient. The proposed hierarchical clustering algorithm behavior is different with CPI in terms of NMI, Accuracy.*

Index Terms*—Document clustering, correlation measure, correlation latent semantic indexing, dimensionality reduction.*

I. Introduction

The aim of document clustering is to automatically group related documents into clusters. Document clustering plays vital role in machine learning and artificial intelligence and has received much attention in recent years. Based on different existing measures number of methods have been proposed to handle document clustering [4],[5],[6],[7],[8],[9]. In existing measures more frequently used measure is Euclidean distance. One method that uses Euclidean distance concept is k-means method, which minimizes the sum of squared Euclidean distance between the data points and corresponding cluster centers.

Through spectral clustering method low computation cost is achieved, in which documents are first projected into low dimensional semantic subspace and then traditional clustering algorithm is applied for document clustering. Latent semantic indexing (LSI) [7] is another spectral clustering method aimed at finding the best subspaces approximation to original document space by reducing the global reconstruction error(Euclidean distance).

Euclidean distance is dissimilarity measure space which describes the dissimilarities rather than similarities between documents. Hence, it is not able to capture non linear manifold structure embedded in similarities between them. Locality preserving indexing (LPI) is different clustering method based on graph partitioning theory. This LPI method applies a weighted function to each pair wise distance to capturing similarity structure rather than dissimilarity structure of the document. It does not overcome the limitation of Euclidean distance. Furthermore, the selection of the weighted functions is often a difficult task.

In this Document clustering previously presented new document clustering method based on correlation preserving indexing. It simultaneously maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. Consequently, a low dimensional semantic subspace is derived where the documents corresponding to the same semantics are close to each other proposed CPI method with learning level parsing procedure. Because previously they proposed algorithm is support to find correlation between documents depend on words. Now we are proposing to find correlation between relational documents to avoid maximum unknown clusters those are not effect able to find exact correlation between documents depend on accuracy of sentences. This is a little double that the previous correlation coefficient is only one way to find the correlation between documents depends on words. In this paper we are proposing to find the correlation between two documents depends on parsers to get accuracy at learning level then we are providing correlation of CPI methods.

II. Related Work

Correlation preserving indexing: Semantic structure usually implicit in high dimensional document space. It is necessary to find a low dimensional semantic subspace in which the semantic structure can become clear. Hence, discovering the intrinsic structure of the document space is often important task of document clustering. Correlation as a similarity measure is suitable for capturing the manifold structure

embedded in the high dimensional document space because the manifold structure is often embedded in the similarities between the documents. The correlation between two vectors (column vectors) u and v is defined as

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

The correlation corresponds to an angle Θ such that

Cos Θ = Corr (u , v).

The association between vectors u and v is stronger when the value of Corr(u,v) is larger.

Online document clustering aims to group documents into clusters, which belongs unsupervised learning and it can be transformed into semi-supervised learning by using the following information:

1. If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster [8].
2. If two documents are far away from each other in the original document space, they tend to be grouped into different clusters.

Document preprocessing: In document preprocessing set of documents are given as inputs to the database. Then randomly choose the one particular document form database. From randomly selected documents identify all unique words and remove stop words for finding similarity between documents.

Stemming is the process for reducing derived words to their stem, base are root form generally a written word form. A stemming algorithm is a process in which the various form of a word are reduced to common form, for example

- suffix Removal to generate word stem
- Grouping words
- Increase relevance

Finally term weighting is to provide the information retrieval and text categorization. In document clustering groups together conceptually related documents thus enabling identification of duplicate words.

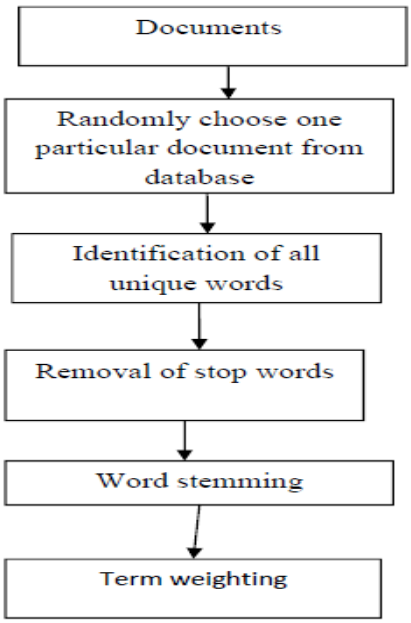

Fig 1. Document preprocessing

III. Actual Work

Preprocessing: Document clustering method based on correlation preserving indexing (CPI) and which explicitly considers manifold structure embedded in the similarities between the documents. Its goal is to find an optimal semantic subspace by simultaneously maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outside these patches. This is different between LSI and LPI, which are based on a dissimilarity measure (Euclidean distance), and which are focused

on detecting the intrinsic structure between widely separated documents rather than on detecting the intrinsic structure between nearby documents. The similarity-measure-based CPI method aims on detecting the intrinsic structure between nearby documents rather than on detecting the intrinsic structure between widely separated documents. As the intrinsic semantic structure of the document space is often embedded in the similarities between the documents and the CPI can effectively detect the intrinsic semantic structure of the high-dimensional document space.

Correlation Preserving Indexing based Documentation clustering: The semantic structure is usually implicit in high-dimensional document space. It is desirable to find a low dimensional semantic subspace in which the semantic structure can become clear. Hence, discovering the intrinsic structure of the document space is often a primary task of document clustering. Since the manifold structure is often embedded in the similarities between the documents, correlation as a similarity measure is suitable for capturing the manifold structure embedded in the high-dimensional document space.

K-means on Document sets: The *k*-means method is the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centers. Since the document space is always of high dimensionality and it is preferable to find a low dimensional representation of the documents to reduce computation complexity.

Documents Classification into clusters: The aim of online document clustering is to group documents into clusters and which belongs unsupervised learning. Further it can be transformed into semi-supervised learning by using the following side information:

1. If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster.
2. If two documents are far away from each other in the original document space, they tend to be grouped into different clusters.

Hierarchical clustering method: Groups the data instances into a tree of clustering in hierarchical clustering methods. There are two major methods in hierarchical clustering methods

1. Agglomerative method
2. Divisive method

Agglomerative method is one which performs the clusters in bottom up fashion. The divisive method is another which splits the data into smaller clusters in a top-down passion. These hierarchical methods can be represented by using dendrograms. These methods are known for their quick termination.

Agglomerative (bottom up) - in agglomerative method data comparison start with first point(singleton) and recursively add two or more appropriate clutters. Finally stops the comparison method when k number of clusters achieved.

Divisive (Top down) - In Divisive method data comparison start with big cluster and recursively divide into smaller clutters. Finally stops the comparison method when k number of clusters achieved.

Semantic-based document mining: The above figure illustrate semantic understand based document mining that satisfies some user needs and these user needs are acquired through mining process such as document clustering, document classification and information retrieval. The semantic understanding based document mining undergoes parsing procedure and parsing step comprises semantic analysis to extract systematic structure descriptions. In this paper proposing CPI based learning level parsing procedure which improves the accuracy compared to CPI clustering algorithm.

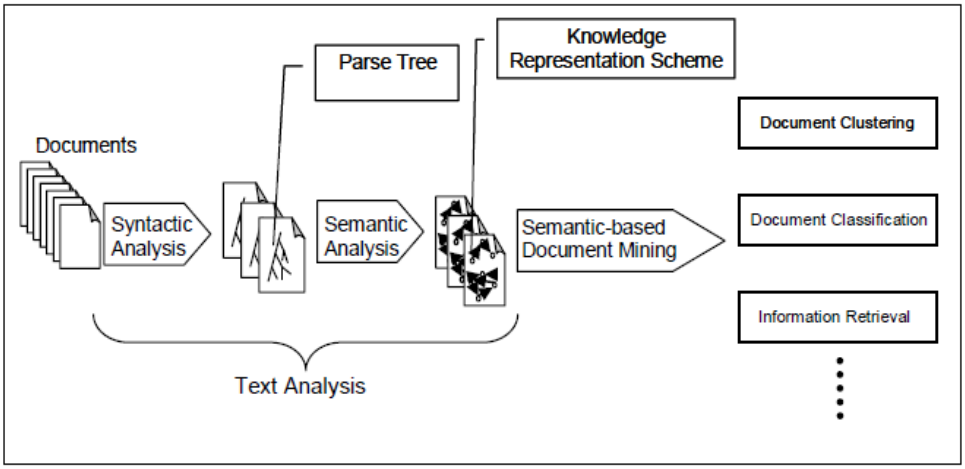

Fig 2. Semantic based document mining

IV. Performance Analysis

This proposed approach of correlation preserving indexing illustrates and evaluates the performance of all the approaches. We analyze our proposed scheme shows or works better than other existing systems (LSI,LPI) in terms of memory, storage, generalization error, performance. Hierarchical clustering algorithm and learning level parsing enhances the accuracy and performance.

V. Conclusion

The proposed system is document clustering method based on correlation preserving indexing and it simultaneously maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. The CPI method with learning level parsing procedure is to find correlation between relational documents to avoid maximum unknown clusters those are not effectual to find exact correlation between documents depend on accuracy of sentences. CPI method has good generalization capability and it can effectively deals with very large size data. The proposed CPI method with learning level parsing procedure in document clustering doubles the accuracy of previous correlation coefficient.

References

[1] Taiping Zhang, yuan yan tan, Bin Fang Young Xiang " document clutering in correlation limilarity measure space" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGNEERING, vol. 24, no.6, june 2012.

[2] R.T. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB), page. 144-155, 1994.

[3] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, page. 264-323, 1999.

[4] P. Pintelas and S. Kotsiantis "Recent Advances in Clustering: A Brief Survey," WSEAS Trans. Information Science and Applications, vol. 1, no. 1, page. 73-81, 2004.

[5] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp. Math. Statistics and Probability, vol. 1, page. 281-297, 1967.

[6] A.K. McCallum and L.D. Baker "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, page. 96-103, 1998.

[7] X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document Clustering with Cluster Refinement and Model Selection Capabilities," Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '02), page. 191-198, 2002.

[8] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, "Indexing by Latent Semantic Analysis," J. Am. Soc. Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[9] j. han and D. Cai, X. He, "Document Clustering Using Locality Preserving Indexing," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 12, page. 1624-1637, Dec. 2005.

[10] y. gong and W. Xu, X. Liu, "Document Clustering Based on Non- Negative Matrix Factorization," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Informaion Retrieval (SIGIR '03), page. 267-273, 2003.

[11] P. Achananuparp, X.-J. Shen and X.-H. Hu "The Evaluation of Sentence Similarity Measures," Proc. 10th International Conference on Data Warehousing and Knowledge Discovery (DaWak), 2008, page. 305-316.

[12] J.-P. Bao, Q.-B. Songand J.-Y. Shen, X.-D. Liu"A New Text Feature Extraction Model and Its Application in Document Copy Detection," Proc. 2nd International Conference on Machine Learning and Cybernetics, 2003, page. 82-87.

[13] S. Manandhar and M. D. Boni "An Analysis of Clarification Dialogue for Question Answering," Proc. HLT-NAACL, 2003, page. 48-55.

[14] R. Mihalcea and C. Corley "Measuring the Semantic Similarity of Texts," Proc. ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, 2005, page. 13-18.

[15] J. Feng, Y.-M. Zhou, and T. Martin, "Sentence Similarity based on Relevance," Proc. IPMU, 2008, page. 832-839.

[16] C. Ho, M. A. A. Murad, S. C. Doraisamy and R. A. Kadir "Word Sense Disambiguation-based Sentence Similarity," 23rd International Conference of Computational Linguistics (COLING), 2010, in press.