Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 3, March 2013, pg.89 – 94

RESEARCH ARTICLE

Grouping and Categorization of Documents in Relativity Measure

V.Asaithambi[1], D.John Aravindhar[2], V.Dheepa[3]

[1]PG student of Computer Science and Engineering, Hindustan University, Padur, Chennai, India
[2]Assistant Professor, Department of Computer Science and Engineering, Hindustan University, Padur, Chennai, India
[3]Assistant Professor, Department of Computer Science and Engineering, Hindustan University, Padur, Chennai, India

[1]asaivenkat@gmail.com; [2]jaravindhar@hindustanuniv.ac.in; [3]vdeepa@hindustanuniv.ac.in

Abstract— This paper presents a spectral clustering method called correlation through preserving indexing (CPI), which is to perform in the correlation similarity measure space. The documents are considered into a low dimensional semantic space, the correlations between the documents in the local patches are maximized and correlations between the documents outside these patches are minimized. The intrinsic structure of the document space is included in the similarities between the documents. Correlation is the similarity measure for finding the intrinsic structure of the document space than Euclidean distance. Simultaneously, the proposed CPI methods can effectively finding the intrinsic structures included in high-dimensional document space. The effectiveness of the new method is implemented by extensive experiments conducted on various data sets and by comparison with existing document clustering methods.

Key Terms: - Document Clustering, Correlation Latent Semantic Indexing, Dimensionality Reduction, Correlation Measure.

I. INTRODUCTION

Document clustering is grouping the related document into particular clusters. It is the important tasks in machine learning. A typical and widely used distance measure is the Euclidean distance. Spectral clustering methods are to implement the low computation cost, which the documents are converted into a low dimensional semantic space and traditional clustering algorithm is applied to finding document clusters. Latent semantic indexing (LSI) is one of the effective spectral clustering methods, aimed to find the subspace approximation to the original document space by minimizing the global reconstruction error.

The Euclidean distance means the dissimilarity measure. it specifies dissimilarities than similarities between the documents. Thus, it is not able to effectively capture the nonlinear manifold structure available in the similarities between the documents. An effective document clustering method must be able to find a low dimensional representation of the documents. Locality preserving indexing (LPI) method is a different spectral clustering method based on graph partitioning theory. The LPI method applies a weighted function to focus on capturing the similarity structure, rather than the dissimilarity structure, of the documents.

A new document clustering method based on correlation through preserving indexing (CPI), considers the manifold structure embedded in the similarities between the documents. It aims 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.

II. RELATED WORK

Raymond Tang et al. [1] Propose a Spatial data mining is to find the interesting relationships and characteristics that may exist implicitly in spatial databases. The clustering methods have a role to play in data mining. Data mining is the search for hidden patterns that may exist in large databases. Because of huge amounts of spatial data that may be obtained from satellite images and from various sources it is often costly and unrealistic for users to examine spatial data in detail. Spatial data mining aims to automate such a knowledge discovery process. Thus it plays an important role in

- Extracting interesting spatial patterns and features
- Capturing intrinsic relationships between data
- Presenting data concisely and
- Helping to reorganize spatial database to accommodate data semantics and better performance

Macqueen et.al [2] proposes a process for partitioning an N-dimensional population into k sets on the basis of a sample. The process, which is called k-means, appears to give partitions which are reasonably efficient in the sense of within class variance. if p is the probability mass function for the population, S = {S1, S2, , Sk} is a partition of EN, and the conditional mean of p over the set Si, then W2(S generated by the method, tends to be low, primarily because of intuitive considerations, corroborated to some extent by mathematical analysis and practical computational experience.

Macqueen et.al [3] Unsupervised learning deals with instances, which have not been pre classified in any way and do not have a class attribute associated with them. The scope of applying clustering algorithms is to discover useful but unknown classes of items. Unsupervised learning is an approach of learning where instances are automatically placed into meaningful groups based on their similarity. This paper introduces the fundamental concepts of unsupervised learning while it surveys the recent clustering algorithms. Moreover, recent advances in unsupervised learning, such as ensembles of clustering algorithms and distributed clustering, are described.

S Deng Cai et.al.[6] propose a novel document clustering method, which aims to cluster the documents into different semantic classes. The document space is generally of high dimensionality, and clustering in such a high dimensional space is often infeasible due to the curse of dimensionality. By using Locality Preserving Indexing (LPI), the document scan is projected into a lower dimensional semantic space in which the documents related to the same semantics are close to each other. Different from previous document clustering methods based on Latent Semantic Indexing (LSI) or Non-negative Matrix Factorization (NMF), our method tries to discover both the geometric and discriminating structures of the document space.

Joy deep Ghosh et.al.[8] propose a detailed empirical study of generative approaches to text clustering obtained by applying four types of document-to-model assignment strategies (hard, stochastic, soft and deterministic annealing (DA) based assignments) to each of three base models, namely mixtures of multivariate Bernoulli, multinomial, and von Misses Fisher (vMF) distributions.

Shi David M.Blei et.al.[10] propose a latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document and efficient approximate inference techniques based on variation methods and an EM algorithm for empirical Bayes parameter estimation. The report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

III. PROPOSED METHOD

A. SYSTEM ARCHITECTURE

System description shows the relationship between different components of the system, to understand the overall concept of system, in which the principle parts or functions are represented by blocks connected by lines that show the relationships of the blocks. The documents are categorized and grouped based on the relativity measure of the documents.

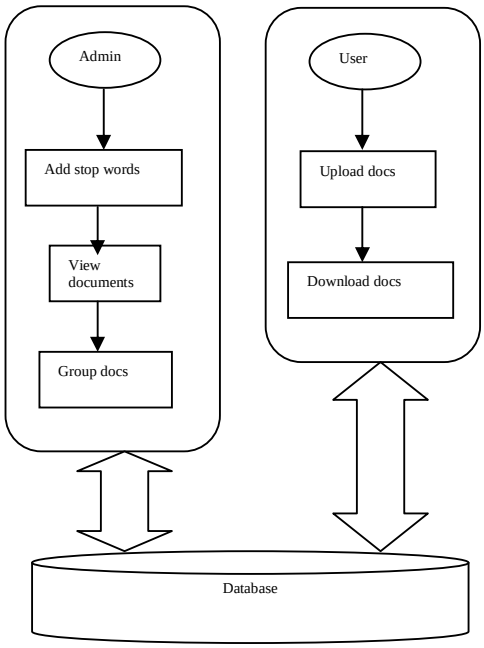

Figure 3.1 System Description

The document clustering is to be implemented through the Euclidean distance. The Euclidean distance is the dissimilarity measure and it is based on dissimilarity measure of the documents. The non-Euclidean distance is based on the similarity of the measures. The documents are analyzed and grouped based on the relative of the entire words in the documents. The relative categorization of the documents are analyzed and grouped into the one particular cluster, the non-relative documents are placed into the new cluster.

B. MODULE DESCRIPTION

a) Admin

The admin module is the major controller of all the processes of the systems.

- *Add Stop Words*

Stop words are common words that have no semantic content. Removal of these words from user query does not affect the extraction of exact results but reduces the computational time. The admin has the rights to add stop words so they can be safely removed from the document content to extract a list of terms.

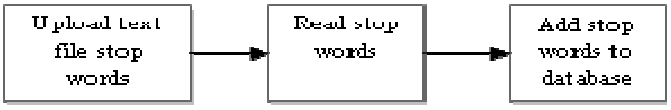

Figure 3.2 Adding Stop Words

- *View Documents*

Admin view the total documents upload by users. And admin can also view user details and can also view the feedback posted by the user. Admin can do documents processing like removing the document, and adding new document.

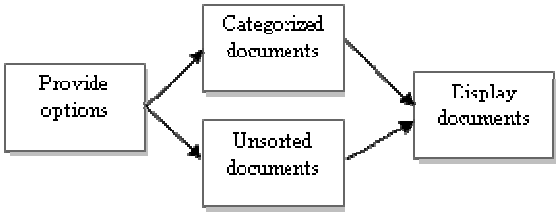

Figure 3.3 Viewing Documents

- *Analyze Documents*

Admin who has the rights to control the website can do operations like make new clusters based on the documents uploaded by the users. The admin reads the documents and extracts the terms of importance by removing stop words. By doing so the correlation between the documents is identified. If a close relation between them is found they are placed under same cluster otherwise the documents are placed in a new cluster.

b) User

This module is used to upload the documents of the user to cluster.

- *Authentication*

Legal access means when a user enters the username and password. The username and password will be matched with the database. If the user name and password matches with the existing username and password in the database, then the user will be allowed to access the site.

- *Upload Documents*

Users after logging in can upload documents containing the information and the contents of text documents can be of any type. By doing so other users can benefit from the resource and knowledge sharing is achieved that is a part of semi supervised learning.

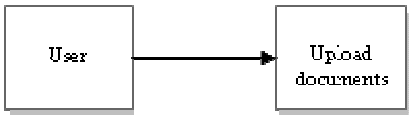

Figure 3.4 Uploading Documents

- *Search*

The users can search the website for specific resources by entering their queries. The users can download the documents if a match is found against the user entered query. If no match is found then the user will be intimated with an indication message.

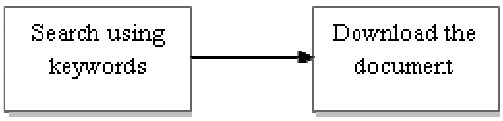

Figure 3.5 Searching Documents

C. Documentation Clustering Based On Correlation through Preserving Indexing

In high-dimensional document space, the semantic structure Is usually implicit. 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 concern Of document clustering.

Online document clustering aims to group documents into clusters, which pertains to unsupervised learning. However, it can be transformed into semi supervised learning by using the following side information.

- If two documents are close to each other in the Original document space, then they tends to be Grouped into the same cluster [8].
- If two documents are far away from each other in the original document space, they tend to be grouped into different clusters. Based on these assumptions, proposes a spectral clustering in the correlation similarity measure space through the nearest neighbors graph learning.

IV. DOCUMENT CLUSTERING BASED ON CPI

- *Document Representation*

In all experiments, each document is represented as a term frequency vector. The term frequency vector can be computed as follows:

1. Transform the documents to a list of terms after words stemming operations.
2. Remove stop words. Stop words are common words that contain no semantic content.
3. Compute the term frequency vector using the TF/IDF weighting scheme.

- *Generalization Capability*

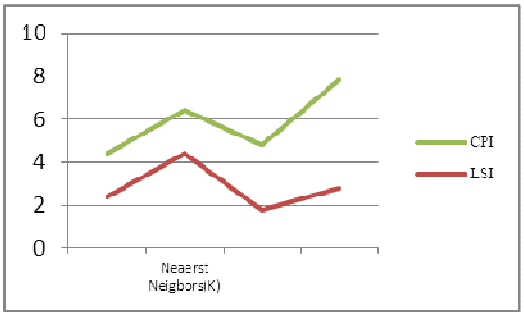

Figure 4.1 The accuracy of CPI

The LSI, LPI, and CPI methods try to find a low-dimensional semantic subspace by preserving the relational structure among the documents, where the mapping between the original document space and the low-dimensional semantic subspace is explicit. In practical applications, the part of documents to learn such mapping, then transform the documents into the low-dimensional semantic subspace which can reduce computing time. The performance on the new samples reflects the generalization capability of the methods.

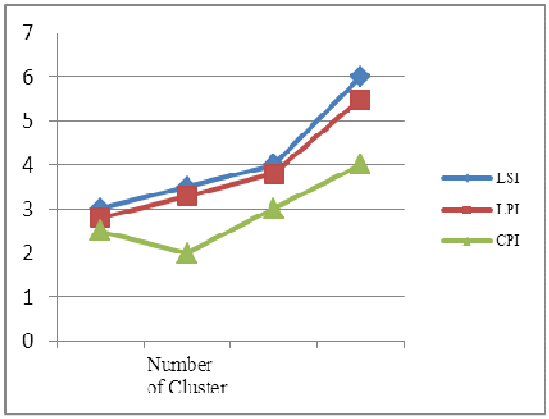

Figure 4.2 The generalization capability of the LSI, LPI, and CPI methods using correlation distance.

In order to test the generalization capability of LSI, LPI, and CPI, the CPI method has smaller generalization error than LSI and LPI methods. This means that the CPI method has better generalization capability. Another interesting result lies in the generalization error of LSI and LPI with Euclidean distances and correlation distance performed much better than the LSI and LPI.CPI can find a low-dimensional semantic subspace in which documents related to the same semantic are close to each other.

V. CONCLUSION AND FUTURE ENHANCEMENT

In this project included a new document clustering method based on correlation through preserving indexing. It simultaneously maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. Furthermore, the CPI method has good generalization capability and it can effectively deal with data with very large size. In future work, the proposed CPI method has good generalization capability and it effectively to deal with data of very large size and all word file types.

The proposed concept might suffer from limitations on the size of the documents being uploaded. In this module users can upload documents of large size. So the barrier on the size of the document to be uploaded is defeated.

REFERENCES

[1] 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), pages: 144-155, 1994.

[2] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth

Berkeley Symp. Math. Statistics and Probability, vol. 1, pp. 281-297, 1967.

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

[4] D. Cai, X. He, and J. Han, "Document Clustering Using Locality Preserving Indexing," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 12, pp. 1624-1637, Dec. 2005.

[5] S. Zhong and J. Ghosh, "Generative Model-Based Document Clustering: A Comparative Study," Knowledge of Information System, vol. 8, no. 3, pp. 374-384, 2005.

[6] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation" J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[7] X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document Clustering with Cluster Refinement and Model Selection Capabilities," Proc. 25thAnn. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '02), pp: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] W. Xu, X. Liu, 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 (SIGIR '03),pp. 267-273, 2003.

[10] S. Zhong and J. Ghosh, "Scalable, Balanced Model-Based Clustering," Proc. Third SIAM Int'l Conf. Data Mining, p p. 71-82, 2003.

