ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 2, February 2013

Correlation Preserved Indexing Based Approach For Document Clustering

Meena.S.U, P.Parthasarathi

***Abstract*— Document clustering is the act of collecting similar documents into clusters, where similarity is some function on a document. Document clustering method achieves 1) a high accuracy for documents 2) document frequency can be calculated 3) term weight is calculated with the term frequency vector. Document clustering is closely related to the concept of data clustering. Document clustering is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. The correlation preserving indexing method is performed to find the correlation between the documents. The Term Frequency-Inverse Document Frequency (TF-IDF) method is used to find the frequency of occurrence of words in each document. The disadvantage of this method is computation complexity. In this paper Significant Score Calculation method is introduced, where similarity between the words are calculated using word net tool. Here the related words are identified. The 98% accuracy is occurred with significant score calculation for finding correlation preserving indexing.**

***Index Terms*— Correlation Preserving Indexing, Document Clustering, Significant Score Term Frequency-Inverse Document Frequency.**

I. INTRODUCTION

Document clustering [1][7][8][9] is a fundamental operation used in unsupervised document organization, automatic topic extraction and information retrieval. The similarity measure-based CPI method focuses on detecting the intrinsic structure between nearby documents rather than on detecting the intrinsic structure between widely separated documents. Since the intrinsic semantic structure of the document space is often embedded in the similarities between the documents, CPI can effectively detect the intrinsic semantic structure of the high- dimensional document space. CPI can find a low-dimensional semantic subspace in which the documents related to the same semantic are close to each other. Thus, correlation is an appropriate metric for measuring similarity between the documents.

II. METHODOLOGY

Correlation-Based Clustering With TF-IDF

The low-dimensional representation of the ith document $x_i \in X$ in the semantic subspace, where i=1,2,3,…….n.

D1 = If two documents are close to each other in the original document space, then

they tend to be grouped into the same cluster.

D2 = If two documents are far away from each other in the original document space, they tend to be grouped into different clusters.

$$max \sum_{i} \sum_{x_j \in N(x_i)} Corr(y_i, y_j)$$

And

$$min \sum_{i} \sum_{x_j \in N(x_i)} Corr(y_i, y_j)$$

where N(xi) denotes the set of nearest neighbors of xi . The equivalent metric learning

$$d(x, y) = \propto * \cos(x, y)$$

Where $d(x, y)$ denotes the similarity between the documents x and y, $\propto$ corresponds to whether x and y are the nearest neighbors of each other.

Document Representation

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. The TF/IDF weighting scheme assigned to the term t_i in document d_j is given by

$$(tf/idf)_{i,j} = tf_{i,j} \times idf_i$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

is the term frequency of the term t_i in document d_j, where $n_{i,j}$ is the number of occurrences of the considered term t_i in document d_j.

$$idf_i = \log\left(\frac{|D|}{|d_i : t_i \in d|}\right)$$

is the inverse document frequency which is a measure of the general importance of the term t_i, where $|D|$ is the total number of documents in the corpus and $|\{d; ti \in d\}|$ is the number of documents in which the term t_i appears. Let $V = \{t_1, t_2 \ldots t_m\}$ be the list of terms after the stop words removal and words stemming operations. The term frequency vector X_j of document d_j is defined as

$$X_j = [x1j, x2j, \ldots . xmj]$$

$$xij = (tf/idf)_{i,j}$$

Using n documents from the corpus, we construct an m × n term-document matrix X. 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. 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. Mathematically, the correlation between two column vectors u and v is defined as follows

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

Clustering Algorithm Based on CPI

1. Construct the local neighbor patch, and compute the matrices MS and MT .
2. Project the document vectors into the SVD subspace by throwing away the zero singular values. The singular value decomposition of X can be written as $X = U \sum V^T$. Here all zero singular values in _ have been removed. Accordingly, the vectors in U and V that correspond to these zero singular values have
been removed as well.
3. CPI Projection is computed. The matrix value is computed.
4. Documents are clustered in the CPI semantic subspace.

III. RELATED WORKS

1. Document Clustering with Cluster Refinement and Model Selection Capabilities.

Document cluster contains corpus, which employ a richer feature set to represent each document, and use the Gaussian Mixture Model (GMM) together with the Expectation-Maximization (EM) algorithm to conduct an initial document clustering[6]. the model selection capability is achieved by introducing randomness in the cluster initialization stage, and then discovering a value C for the number of clusters N by which running the document clustering process for a fixed number of times yields sufficiently similar results. The problems and limitations associated with traditional IR techniques reside in the following aspects. First, text retrieval results are sensitive to the keywords used by the user to form queries. To retrieve the documents of interest, the user must formulate the query using the keywords that appear in the documents. This is a difficult task, if not impossible, for ordinary people who are not familiar with the vocabulary of the data corpus. Second, as pointed out in, 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.

2. Document Clustering Method Based on Frequent Co-occurring Words

A new document clustering method based on frequent co-occurring words. First employ the Singular Value Decomposition, and then group the words into clusters called word representatives as substitution of the corresponding words in the original documents. Next, we extract the frequent word representative sets by Apriori. Subsequently, each document is designated to a basic unit described by the frequent word representative set, from which we can get the ultimate clusters by hierarchical clustering [6]. The major advantage of our method is that it can produce the cluster description by the frequent word representatives and then by the corresponding words in the clustering process without any extra works. Frequent words are grouped together to form a cluster.

3. Document Clustering Using Locality Preserving Indexing

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 documents can be 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 Nonnegative Matrix Factorization (NMF), our method tries to discover both the geometric and discriminating structures of the document space. Theoretical analysis of our method shows that LPI is an unsupervised approximation of the supervised Linear Discriminant Analysis (LDA) method, which gives the intuitive motivation of our method. The document space is always of very high dimensionality, ranging from several hundreds to thousands. Due to the consideration of the curse of dimensionality, it is desirable to first project the documents into a lower-dimensional subspace in which the
semantic structure of the document space becomes clear. In the low-dimensional semantic space, the traditional clustering algorithms can be then applied. The spectral clustering usually clusters the data points using the top eigenvectors of graph Laplacian which is defined on the affinity matrix of data points. From the graph partitioning perspective, the spectral clustering tries to find the best cut of the graph so that the predefined criterion function can be optimized.

4. Document Clustering Based On Non-negative Matrix Factorization.

A novel document clustering method based on the non-negative factorization of the term document matrix of the given document corpus. In the latent semantic space derived by the non-negative matrix factorization (NMF), each axis captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic (the axis) with which the document has the largest projection value. Topic detection and tracking (TDT) aim to automatically detect salient topics from either a given document corpus or an incoming document stream and to associate each document with one of the detected topics. The TDT problems can be considered as a special case of the document clustering problem and actually most of the TDT systems in the literature were realized by adapting various document clustering techniques. On the other hand, document summarization is intended to create a document abstract by extracting sentences/paragraphs that best present the main content of the original document.

5. A Comparative Study of Generative Models for Document Clustering

The spherical k-means algorithm, which has desirable properties for text clustering, has been shown to be a special case of a generative model based on a mixture of von Mises-Fisher (vMF) distributions. This paper compares these three probabilistic models for text clustering,

both theoretically and empirically, using a general model-based clustering framework. For each model, we investigate three strategies for assigning documents to models: maximum likelihood (k-means) assignment, stochastic assignment, and soft assignment. Our experimental results over a large number of datasets show that, in terms of clustering quality, (a) The Bernoulli model is the worst for text clustering; (b) The vMF model produces better clustering results than both Bernoulli and multinomial models; (c) Soft assignment leads to comparable or slightly better results than hard assignment. We also use deterministic annealing (DA) to improve the vMF-based soft clustering and compare all the model-based algorithms with the state-of-the-art discriminative approach to document clustering based on graph partitioning (CLUTO) and a spectral co-clustering method. Overall, CLUTO and DA perform the best but are also the most computationally expensive; the spectral coclustering algorithm fares worse than the vMF-based methods. The increasing interest in processing larger collections of documents has led to a new emphasis on designing more efficient and effective techniques, leading to an explosion of diverse approaches to the document clustering problem, including the (multilevel) self-organizing map, mixture of Gaussians, spherical k-means, bi-secting k-means, mixture of multinomials, multi-level graph partitioning and co-clustering using bipartite spectral graph partitioning. Most clustering methods proposed for data mining can be divided into two categories: discriminative (or similarity-based) approaches and generative (or model-based) approaches. In similarity-based approaches, one optimizes an objective function involving the pairwise document similarities, aiming to maximize the average similarities within clusters and minimize the average similarities between clusters. Model-based approaches, on the other hand, attempt to learn generative models from the documents, with each model representing one particular document group.

IV. ACTUAL WORK

PREPROCESSING

In this module the preprocessing of database is done. Preprocessing is the phase to remove stop words, stemming and identification of unique words in document. Identification of unique words in the document is necessary for clustering of document with similarity measure. And after that we remove the stop words that is the non informative word for example the, end, have, more etc. We need to eliminate those stop words for finding such similarity between documents.

A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form, for example,

- Removal of suffix to generate word stem
- Groping words
- Increase the relevance

Suffix-stripping algorithm:

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. Finally term weighting is to provide the information retrieval and text categorization. In document clustering groups together conceptually related documents. It also provides metadata characterization the content of given document cluster. Tf–idf, term frequency–inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others.

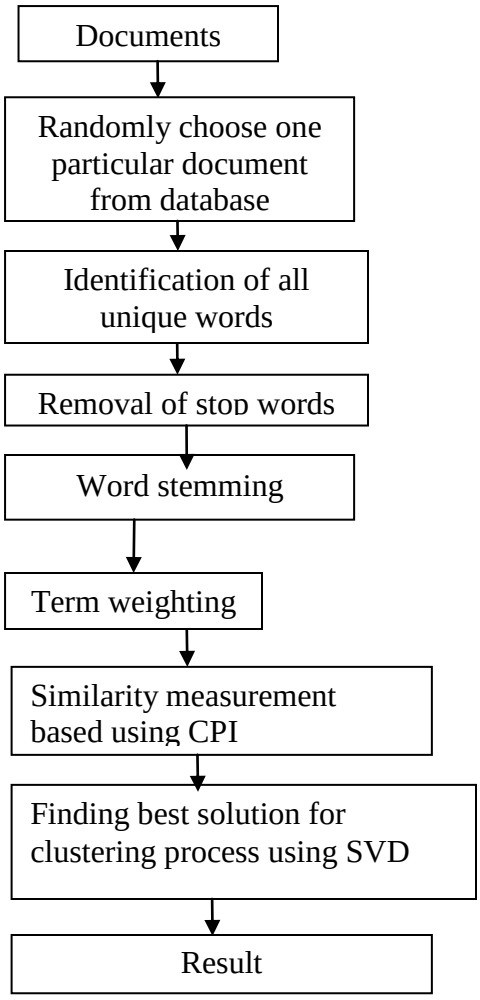

Fig1: Flow of the system

Correlation-Based Clustering

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. 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. Mathematically, the correlation between two vectors u and v is defined as follows

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

Significant Score

The word net tool is used for calculating similarity values between the words. From those words we can identify related words, these words are used when finding document using swarm intelligence techniques. In this technique we used TF-IDF as weighted values of terms and we considered relative words during the searching time only.

But if we use the following scenarios in our approach the weight values of term is calculated with relevance with other terms. This relevance based score calculation is effectively improving our tf-idf based calculations.

Corpus-based frequency profiling works as follows. Assume that we are interested in the significance of word w in the domain document. The domain document contains a total of nd words, and the normative corpus contains nc words. w occurs wd times in the domain document and wc times in the normative corpus. wd and wc are called the observed values of w. Based on the occurrences of w in the domain document and the normative corpus, we can define two expected values for w:

Ed = nd(wd + wc)/ (nd + nc)

Ec = nc(wd + wc)/ (nd + nc)

The log-likelihood value for w is then

$$LL_w = 2(W_d . l_n \frac{W_d}{E_d} + W_c . l_n \frac{W_c}{E_c})$$

Given a log-likelihood value for each term in the domain document, the terms can be ranked, placing the term with the highest LL value, and thus most likely to represent an underlying abstraction, at the top. This corpusbased frequency profiling is the primary technique used successfully by WMatrix . The corpus-based frequency profiling technique described above works well for terms that are single words, in practice it doesn't help with multiword terms. To solve this problem, we synthesize a significance value for all terms using a heuristic based on the number of words of which the term is composed, and the LL value for each word. In its simplest form, the significance value for a term t ={w1, w2;..;wi} is given by the formula:

$$S_t = \frac{\sum_i LL_{wi}}{l}$$

simply calculates the mean of the LL values for all the component words comprising a multiword term.

V. Performance Analysis

In this module the proposed approaches were illustrated and evaluated to compare the performance of all the approaches. We analyze our proposed scheme in terms of memory, storage, computation complexity, generalization error, performance. Based on the comparison and the results from the experiment show the proposed approach works better than the other existing systems.

VI. DATA SET

20 Newsgroups (or NG20)

The 20 newsgroups corpus consists of roughly 20,000 documents that come from 20 specific Usenet newsgroups. We repeated the experiments in[2] and [3] to illustrate the performance of the proposed CPI algorithm and other competing algorithms. The first set of experiments involved binary clustering. In each experiment, we randomly chose 50 documents from the two selected newsgroups and 100 runs were conducted for each algorithm to obtain statistically reliable clustering result. The means and standard deviations of the test results were recorded. The results of statistical significance test show that CPI is more accurate than the other methods with statistical significance for most of the data sets.

VII. CONCLUSION

A new document clustering method based on correlation preserving indexing and the use of natural language processing tool is presented. It simultaneously maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. A low dimensional semantic subspace is derived where the documents corresponding to the same semantics are close to each other. Extensive experiments on NG20 show that the proposed CPI method outperforms other classical clustering methods. CPI method has good generalization capability.

VII. REFERENCES

1. Cai, D. He, X. ; Han, J. "Document clustering using locality preserving indexing "Knowledge and Data Engineering, IEEE Transactions on Dec. 2005.
2. D. Cheng, R. Kannan, S. Vempala, and G. Wang, "*A Divide-and- Merge Methodology for Clustering*" ACM Trans. Database Systems, vol. 31, no. 4, pp. 1499-1525, 2006.
3. G. Lebanon, "Metric Learning for Text Documents," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 497-507, Apr. 2006 .
4. H. Zha, C. Ding, M. Gu, X. He, and H. Simon, "Spectral Relaxation for k-Means," Neural Information Processing Systems, vol. 14 (NIPS 2001), pp. 1057-1064, MIT Press, 2001.
5. S. Zhong and J. Ghosh, "Scalable, Balanced Model-Based Clustering," Proc. Third SIAM Int'l Conf. Data Mining, pp. 71-82, 2003.
6. 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.
7. Taiping Zhang, Yuan Yan Tang, Bin Fang and Yong Xiang "*Document Clustering in Correlation Similarity Measure Space*" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012.
8. 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), pp. 191-198, 2002.

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.