ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Achievement on Document Clustering using Association Preserve Indexing

Prof. Sagar R. Mali

Assistant Professor, Dept. of CSE & IT, AITRC, Vita., Shivaji University, Kolhapur, Maharashtra India

ABSTRACT: Document clustering is a technique for unconfirmed document organization, repeated topic removal and fast information recovery. In association preserving indexing, the documents are first assign into a low dimensional semantic space. The documents with in the cluster are highly related to each other while the documents outside the cluster are dissimilar. The document space is always of high dimensional and it is preferable to find a low dimensional representation of the documents to reduce calculation difficulty. The essential geometrical structure of the document space is often set in the similarities between the documents. Consider association as a similarity measure for detecting the fundamental geometrical structure of the document space than Euclidean space.

KEYWORDS: Document Clustering, Association Preserving Indexing, Invalid learning, Basic Semantic Structure, Various Structures.

I. INTRODUCTION

Document Clustering is a regular group of text documents into clusters. Cluster is a subset of objects which are "similar". A subset of objects such that the space between any two objects in the cluster is less than the space between any object in the cluster and any object not located inside it. A connected area of a multidimensional space containing a relatively high density of objects. Based on various distance measures, a number of methods have been planned to handle document clustering. A typical and widely used distance measure is the Euclidean space. The k-means method is one of the methods that use the Euclidean space, which minimizes the sum of the squared Euclidean space between the data points and their equivalent cluster center. Since the document space is always of high dimensional, it is preferable to find a low-dimensional representation of the documents to reduce calculation difficulty than k- means. Low calculation cost is achieved in spectral clustering methods, in which the documents are first projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. However, because of the high dimensional of the document space, a certain representation of documents usually resides on a nonlinear various embedded in the similarities between the data points. Unfortunately, the Euclidean distance is a variation measure which describes the dissimilarities rather than similarities between the documents. Thus, it is not able to effectively capture the nonlinear various structures embedded in the similarities between them.

Document clustering involves the use of descriptors and their extraction. Descriptor is the sets of words that describe the content in the document within the cluster. It is generally considered to be a centralized process. Examples includes web document clustering for search users. Document clustering can be categorized to two types, offline and online. Online applications are usually constrained by efficiency problems when compared offline applications.

II. RELATED WORK

Association Preserving Indexing:

An association Preserving Indexing The usage of correlation as a similarity measure can be found in the canonical association analysis (CAA) method. The CAA method is to find projections for paired data sets such that the correlations between their low-dimensional representatives in the projected spaces are equally maximized. Specifically, given a paired data set consisting of matrices X= {x1; x2; . . .; xn ;} and Y {y1; y2; . . . ; yn;}

We would like to find directions wx for X and wy for Y that maximize the association between the projections of X on wx and the projections of Y on wy. This can be expressed as

$$\text{Max}_{wx,\,wy} \frac{\langle Xwx, Ywy \rangle}{\|Xwx\|.\|Ywy\|}$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Where <. > And ||.|| denote the operators of inner product and norm, respectively. As a powerful numerical technique, the CAA method has been applied in the field of pattern detection and machine learning. Rather than finding a projection of one set of data, CAA finds projections for two sets of matching data X and Y into a single hidden space that projects the matching points in the two data sets to be as nearby as possible. In the application of document clustering, while the document matrix X is available. So the CAA method cannot be directly used for clustering. In this paper, we propose a new document clustering method based on association preserving indexing (API), which clearly considers the various structures embedded in the similarities between the documents. It aims to find an optimal semantic subspace by at the same time maximizing the association between the documents in the local patches and minimizing the correlations between the documents outside these patches.

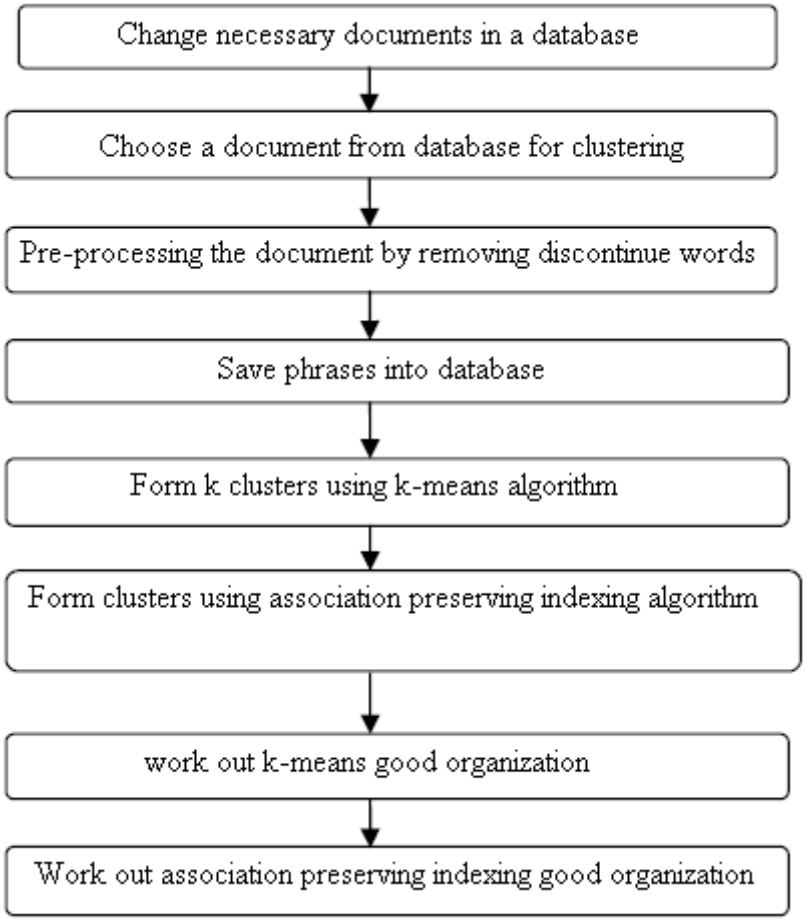

Figure 1

Above the figure 1:-

1. The change necessary documents in a database file.
2. The database file in choose document file for clustering.
3. Pre-processing the document by removing discontinue words.
4. Save the phrases into database file.
5. From k clusters using k-means algorithm.
6. The clusters using association preserving indexing algorithm.
7. Output of the good organization.
8. Output of the association preserving indexing algorithm.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Clustering Algorithm Based on API:

Given set of documents x1, x2, x3...xn ϵ IRn. Let X denotes the document matrix. The algorithm for document clustering based on API can be summarized as follows:

1. Create the local near patches, and work out the matrices MsW and λMW.
2. Task the document vectors into the subspace by throwing away the zero particular values. The particular value breakdown of X can be write as X=UΣVT.
3. Here all zero particular values in Σ have been removed. So, the vectors in U and V that match up to these zero particular values have been removed as well. Thus the document vectors in the subspace can be obtained by $\tilde{X}=U^{T}X$.
4. Calculate API Projection. Based on the multipliers λ1 , λ2,.... λn one can calculate the matrix
   $M = \lambda 0\ ^{*}MT + \lambda 1^{*}x1x1^{T}+....+\lambda n^{*}xnxn^{T}$
5. Let WCPI be the solution of the general Eigen value problem MSW = λMW. Then, the low dimensional representation of the document can be compute by Y=WTCPI $\tilde{X}$ = WTX.

III. K- MEANS ALGORITHM

The k-means clustering algorithm is known to be efficient in clustering great data sets. This clustering algorithm was developed by Macqueen, and is one of the simplest and the best known unverified learning algorithms that solve the well-known clustering problem. The K-Means algorithm aims to divider a set of objects, based on their attributes, into *k* clusters, where *k* is a predefined or user-defined stable. The main idea is to define *k* centroids, one for each cluster. The centroids of a cluster is formed in such a way that it is closely related to all objects in that cluster.

Basic K-Means Algorithm:-

1. Choose k number of clusters to be determined fig a.
2. Choose k objects at random as the first cluster center fig b.
3. Do again
4. Give each object to their closest cluster fig b.
5. Work out new clusters, i.e. work out mean points.
6. Pending
7. No changes on cluster centers fig c.
8. No object changes its cluster fig c.

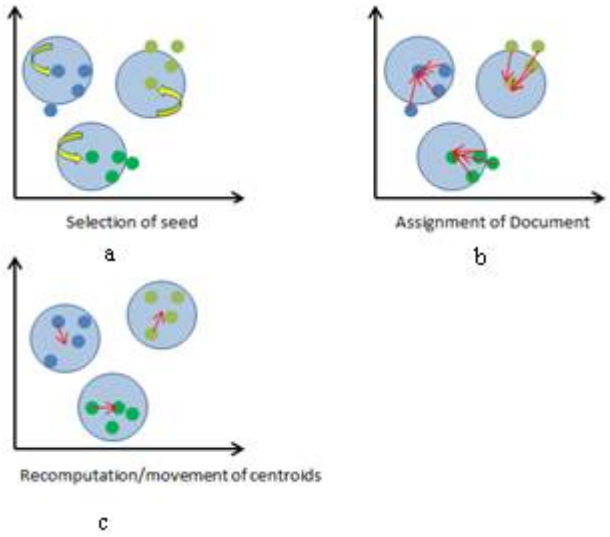

Fig (2) Process of Document clustering

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

IV. METHODOLOGY

Correlation-Based Clustering With TF-IDF:-

The low-dimensional illustration of the 'i'th document xi ϵ X in the semantic subspace, where i=1, 2, 3 ….n.

1. D1 = If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster.
2. D2 = If two documents are far away from each other in the original document space, they tend to be grouped into different clusters.

$$\text{Max } \Sigma i\ \Sigma xj\epsilon N(xi)\ corr(yi,yj)$$
$$\& \min \Sigma i\ \Sigma xj\epsilon N(xi)\ corr(yi,yj)$$

Where N (xi) denotes the set of nearest neighbors of xi. The matching metric learning $d(x,y)=\alpha*\cos(x,y)$
Where d(x, y) denotes the similarity between the documents x and y, α correspond 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 rate vector can be computed as follows:

1. Change 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. Work out the term frequency vector using the TF/IDF weighting scheme. The TF/IDF weighting scheme assign to $(tf/idf)i,j = tfi,j*idfi$

$$tfi,j = \frac{ni,j}{\Sigma knk,j}$$

is the term frequency of the term ti in document dj, where ni,j is the number of occurrences of the measured term ti in document dj.

$$idf_i=\log\left(\frac{|D|}{|djti\epsilon d|}\right)$$

is the inverse document frequency which is a measure of the general significance of the term ti, where $|D|$ is the total number of documents in the corpus and $|\{djti\epsilon d\}|$ is the number of documents in which the term ti appears. Let
V = {t1, t2. . . tm } be the list of terms after the stop words elimination and words stemming operations. The term frequency vector xj of document dj is defined as

$$X_j= [x1j, x2j,.....,xm_j]$$
$$xij= (tf/idf)_{i,j}$$

Module Description:

1. Pre-processing:

A new document clustering method based on association preserving indexing (API), which openly considers the various structures embedded in the similarities between the documents. It aims to find an most select semantic subspace by at the same time as maximizing the association between the documents in the local patches and minimizing the association between the documents outside these patches This is different from K-means, which are based on a difference measure (Euclidean space), and are focused on detecting the fundamental structure between widely separated documents rather than on detecting the fundamental structure between nearby documents. The similarity-measure-based API method focuses on detecting the fundamental structure between nearby documents rather than on detecting the fundamental structure between widely separated documents. Since the fundamental semantic structure of the document space is often embedded in the similarities between the documents, CPI can successfully detect the fundamental semantic structure of the high-dimensional document space.

2. Documentation clustering based on Association Preserving Indexing:

In high-dimensional document space, the semantic structure is usually hidden. It is attractive to find a low dimensional semantic subspace in which the semantic structure can become clear. Hence, discovering the fundamental structure of the document space is often a primary concern of document clustering. Since the various structures are often embedded in the similarities between the documents, association as a similarity measure is suitable for capturing the various structures embedded in the high-dimensional document space.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

3. K-means on Document sets:

The *k*-means method is one of the methods that use the Euclidean space, which minimizes the sum of the squared Euclidean space between the data points and their matching cluster centers. Since the document space is always of high dimensional, it is preferable to find a low dimensional representation of the documents to reduce calculation difficulty.

4. Arrangement of Documents into clusters:

Document clustering aims to group documents into clusters, which belongs unverified learning. However, it can be changed into semi-supervised learning as:

A. If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster.

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

Pre-processing is the phase to remove stop words, stemming and classification of single words. Classification of single 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 useful word for example the, end, have, more etc. The stop words which should be removed are given directly. We need to eliminate those stop words for finding such similarity between documents. Stemming is the process for dropping derived words to their stem; base or root forms generally a written word form. The stem need not be identical to the root of the word it is usually enough that related words map to the same stem, even if this stem is not in itself a valid root. A stemming algorithm is a process in which the variation forms of a word are reduced to a common form. For example, Removal of suffix to generate word stem grouping words Increase the relevance .Finally term weighting is to provide the information recovery and text classification. In document clustering groups together abstractly related documents. Thus enable classification of duplicate words. 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 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 recovery and text classification. In document clustering groups together theoretically 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.

V. RESULTS AND DISCUSSIONS

Performance comparison of k-means and API:

Data set	*k-means*	*CPI*
DOC1	74.14±16.10	97.08±4.41
DOC2	63.92±11.49	83.30±11.40
DOC3	67.78±14.18	86.18±12.07
DOC4	64.50±10.87	76.38±12.36
DOC5	73.56±13.51	95.18±7.07

Initially, select a file from set of documents in a folder by providing path of the file. After pre-processing the content, it removes stop words and form phrases. Later these phrases are separated as keywords which have semantic meaning. Based on these keywords clusters are formed Using k-means algorithm and association preserving indexing algorithms. Calculate and compare efficiencies of both k-means and correlation preserving indexing algorithms. As a result association preserving indexing performs better results than k-means. Association preserving indexing gives results with low calculation cost.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

VI. CONCLUSION

In this paper, we present a new document clustering method based on association preserving indexing. It at the same time maximizes the association between the documents inside the clusters and minimizes the association between the documents outside the clusters. Accordingly, a low dimensional semantic subspace is derived where the documents matching to the same semantics are close to each other. It reduces the computational cost.

REFERENCES.

1. http://en.wikipedia.org/wiki/Document_clustering
2. http://www.milanmirkovic.com/wpcontent/uploads/2012/10/pg 0-49_"Similarity_Measures_for_Text_Document_Clustering.pdf"
3. Claudio Carpineto, Stanislaw Osiński, Giovanni Romano, Dawid Weiss. "A survey of Web clustering engines". ACM Computing Surveys (CSUR), Volume 41, Issue 3, Article No. 17, (July 2009)ISSN:0360-0300
4. D. Cheng, R. Kannan, S. Vempala, and G. Wang, "A Divide-and- Merge Methodology for Clustering" ACM Trans. Database Systems, vol. 31, issue no. 4, pp. 1499-1525, 2006.
5. 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, issue NO. 6, JUNE 2012.
6. Jain AK, Murty MN, Flynn PJ. "Data clustering: a review". ACM Computing Surveys. VOl 31 Issue No.3Page no.264–323. 1999.
7. Wei Xu, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization". In Proc. 2003 Int. Conf. on Research and Development in Information Retrieval (SIGIR'03), pages 267–273, Aug. 2003.
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. D.K. Agrafiotis, H. Xu,"A Self-Organizing Principle for Learning Nonlinear Manifolds", Proc. Nat'l Academy of Sciences USA, Vol. 99, No. 25, pp. 15869-15872, 2002.
10. D.M. Blei, A.Y. Ng, M.I. Jordan,"Latent Dirichlet Allocation," J. Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
11. S. Zhong, J. Ghosh, "Generative Model-Based Document Clustering: A Comparative Study", Knowledge of Information System, Vol. 8, No. 3, pp. 374-384, 2005.
12. K. Funkunaga and P. Navendra, "A Branch and Bound Algorithm for Computing k-Nearest Neighbors," IEEE Trans. Computers, vol. 24, no. 7, pp. 750-753, 1975.

.

BIOGRAPHY

Mr. S. R. Mali is an Assistant Professor in the Computer Science and Engineering/Information Technology Department of Adarsh Institute of Technology and Research Center, Vita. Shivaji University, Kolhapur (Maharashtra), India. He received Master of Technology (M .Tech) degree in the year 2015 from JNTU, Hyderabad, India.