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. 3, Issue. 5, May 2014, pg.713 – 720

RESEARCH ARTICLE

Implementation of Hierarchical Clustering with Multiviewpoint-Based Similarity Measure

Ashish Moon[1], Vinod Nayyar[2]

[1]Department of M.Tech CSE,
R.T.M.N.U. Nagpur, India

[2]Asst. Prof. Department of M.Tech CSE,
R.T.M.N.U. Nagpur, India

[1] ashishmoon2k@gmail.com; [2] vinodnayyar5@gmail.com

Abstract— Clustering is one of the most important data mining or text mining algorithm that is used to group similar objects together. The aim of clustering is to find the relationship among the data objects, and classify them into meaningful subgroups. The effectiveness of clustering algorithms depends on the correctness of the similarity measure between the data in which the similarity can be computed. This paper focus on implementation of Agglomerative hierarchical clustering with Multiviewpoint based similarity measure for effective document clustering. The experiment is conducted over sixteen text documents and performance of the proposed model is analysed and compared to existing standard clustering method with MVS. The experiment results clearly shows that the proposed model Hierarchical Agglomerative Clustering with Multiview Point based Similarity Measure perform quite well.

Keywords— Hierarchical Agglomerative Clustering, Document Clustering, Similarity Measure, Text Mining, K-Mean Clustering Algorithm, Multiviewpoint-Based Similarity Measure

1 INTRODUCTION

One of the important researches in data mining is text mining which refers to the process of automatically extracting information from a usually large amount of different unstructured textual sources. In text mining, the goal is to discover unknown information, something that no one yet knows are to be extracted from large database. Clustering is extensively used for getting text with highest accuracy. It is used for grouping a set of objects into classes of similar objects and is the most interesting concept of data mining. Purpose of Clustering to group essential structures in data and organize them into meaningful subgroup for further analysis. It also makes search mechanism too easy and reduces the bulk of operations and computational cost. There have been many clustering algorithms in the data mining. The most favourite is K-means and top 10 among all data mining algorithms [1]. Even though it is a top most algorithm, it has a few basic drawbacks such as sensitiveness to initialization and to cluster size [2]. It means one should need to specify the number of clusters in advance. In spite of that, it is still popular due to its simplicity, understandability, and scalability. While offering best outcome, K-means is quick and simple to combine with other methods in larger systems. To meet various requirements k-means has many variants. For instance spherical k-means (uses cosine similarity) is used to cluster text documents while original k-means can be used to

clustering using Euclidean distance [3], [4].

A hierarchical clustering algorithm [8] creates a hierarchical decomposition of the given set of data objects. Depending on the decomposition approach, hierarchical clustering algorithms are classified into agglomerative and divisive. An agglomerative clustering is bottom-up approach in such way that each object is assign to a separate cluster and merges the object with the shortest distance to form a large cluster. Generally the problem of clustering can be thought as optimization process, by optimizing similarity measures, the optimal clusters can be formed and its performance is improved. The efficiency of clustering algorithms depends on the accuracy of the similarity measure to the data. Hence the similarity measure plays a very important role in the success or failure of a clustering method. A variety of similarity measures have been proposed so far and widely used measures are cosine similarity, Jaccard coefficient and Pearson correlation coefficient. To improve the accuracy of document clustering, Correlation similarity measure is integrated to Hierarchical Agglomerative Clustering with Multiviewpoint Similarity Measure. The proposed work is motivated by research of similarity measures in document clustering. Similarity measures play a vital role in clustering the documents.

Remaining of this paper is organized as follows: Related work for similarity measures and clustering are reviewed in section 2. Pre-processing and its steps are explained in section 3. In Section 4, Multiviewpoint clustering based similarity measures is explained. In that the partitional K-means clustering and Hierarchical Agglomerative clustering are discussed. The experiments on sixteen text documents are presented in section 5. Finally, conclusion is discussed in section 6. The overview of proposed model is shown in Fig 1.

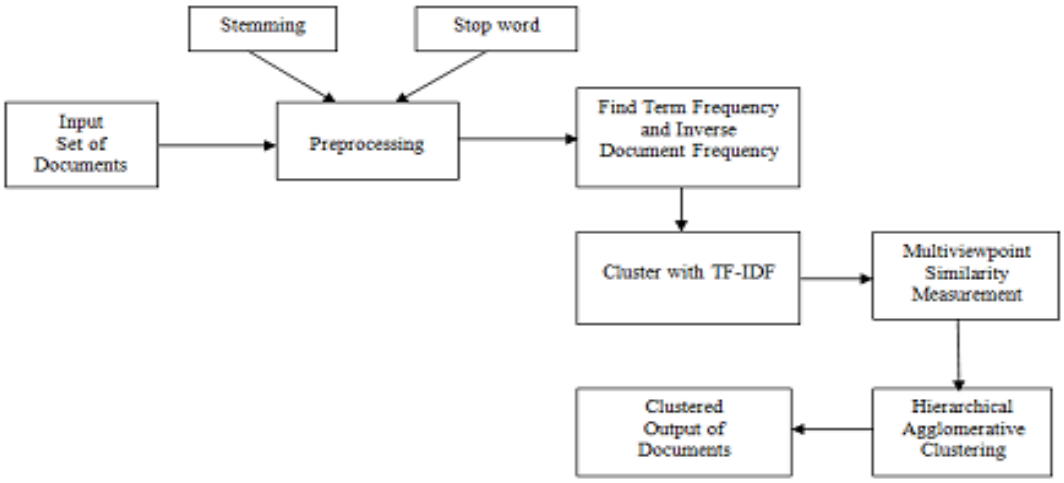


Fig. 1: Overview of Proposed Model

2 RELATED WORK

Table 1 summarizes basic notations that are used to represent documents and related concepts. Each document in a corpus corresponds to an m-dimensional vector d, where m is the total number of terms. Document vectors are often subjected to some weighting schemes, such as the standard Term Frequency-Inverse Document Frequency (TF- IDF), and normalized to have unit length.

Notation	Description
N	number of documents
M	number of terms
C	number of classes
k	number of clusters
D	Document vector, $\|d\|=1$
$S = \{ d_1, \ldots, d_n \}$	Set of all the documents
$D = \sum_{d_i \in S_r}^{S_r} d_i$	composite vector of all the documents

$D_r = \sum_{d_i \in S_r} d_i$	composite vector of cluster r
C =D / n	centroid vector of all the documents
Cr=Dr/ n_r	centroid vector of cluster r, n_r = |Sr|

Table 1: Notations

The standard definition of clustering is to arrange data objects into separate clusters such that the intra cluster (documents within a cluster) similarity with the inter cluster (documents from different cluster) dissimilarity is maximized. Clustering requires a precise definition of the closeness between a pair of documents, in terms of either the pair wise similarity or distance. There are varieties of similarity measure have been proposed and widely applied, such as cosine similarity, Jaccard coefficient, Euclidean distance and Pearson Correlation Coefficient [7]. They are compared by Strehl et al. [6] and concluded that cosine and extended Jaccard are the best ones on web documents.

Cosine Similarity

When documents are represented in terms vectors, the similarity of two documents corresponds to the correlation between the vectors. In a sparse and high dimensional space, cosine similarity is extensively used. It is also a popular similarity score in text mining and information retrieval [5].

$$\max \sum_{r=1}^{k} \sum_{di \in Sr} \frac{di^t Cr}{||Cr||}$$

Jaccard Coefficient

It sometimes referred to as the Tanimoto coefficient, measures similarity as the intersection divided by the union of the objects. For text documents, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms.

$$\mathrm{Sim}_{\mathrm{eJacc}}[Ui, Uj] = \frac{Ui^t\ Uj}{\|Ui\|^2 + \|Uj\|^2 - Ui^t\ Uj}$$

Euclidean Distance

It is common distance between two points and can be without difficulty measured with a ruler in two or three dimensional space. Euclidean distance is one of the most popular measures:

$$\text{Dist (di, dj)} = ||di - dj||$$

It is used in the traditional k-means algorithm. The goal of k-means is to minimize the Euclidean distance between objects of a cluster and that cluster's centroid:

$$\min \sum_{r=1}^{K} \sum_{di \in Sr} ||di - cr||^2$$

Particularly, similarity of two documents vector di and dj, Sim (di, dj), is defined as the cosine of angle between them. For unit vectors, this equals to their inner product:

$$\mathrm{sim}(\mathrm{di}, \mathrm{dj}) = \cos(\mathrm{di}, \mathrm{dj}) = \mathrm{di}^{\mathrm{t}}\,\mathrm{dj}$$

Pearson Correlation Measure
It provides a method for clustering a set of objects into the set of objects into the best possible number of clusters, without specifying that number in proceed. The normalized Pearson correlation defined as:

$$S(x_i, x_j) = \frac{(X_i - \overline{x_I})^T\ (x_j - \overline{X_J})}{\|X_i - \overline{x_I}\| x_j - X_J\|}$$

Where $\overline{x_I}$ denotes the average feature value of x over all dimensions.

3 PREPROCESSING

A database consists of massive volume of data which is collected from heterogeneous sources of data. Due to this data tends to be inconsistent and noisy. If data is inconsistent, then there is a possibility that mining process can lead to confusion which may give inaccurate results. In order to get consistent and accurate data, pre-processing [10] is applied on the data. It is done in two steps i.e. removal of stop word and stemming.

3.1 Stop word removal

The most common words in any text document do not provide meaning of the documents. Those are prepositions, articles, and pronouns etc. These words are treated as stop words. These words do not provide any useful information to us such as "and", "the", "which", "is" etc. It is often useful to get rid of these words otherwise they might mislead the clustering process by including frequent terms that are not informative to us. This process also shrinks the text data and improves the system performance.

3.2 Stemming

Word Stemming is a technique for reducing inflected (or sometimes derived) words to their stem, base or root form. Many words in the English language can be reduced to their base form or stem e.g. connection, connections, connective, connected and connecting belong to connect. Case sensitive systems could have another problems when making a comparison between a word in capital letters and another with the same meaning in lower case. This is essential to keep away from treating different variations of a word distinctly. Word stemming was done using the popular Porter stemming algorithm.

4 MULTIVIEWPOINT-BASED CLUSTERING WITH SIMILARITY MEASURE

The cosine similarity can be expressed in the following form without changing its meaning:

$$\mathrm{Sim}\ (d_i, d_j) = \cos\ (d_i\text{-}0, d_j\text{-}0) = (d_i - 0)^t (d_j - 0)$$

Where 0 is vector 0 that represents the origin point. According to this formula, the measure takes 0 as one and only reference point. In single viewpoint based similarity measure consider only one reference point which is at origin while in multiviewpoint based similarity measure considers more than one point of reference for clustering the documents.

4.1 PARTITIONAL K-MEANS CLUSTERING WITH MVS

K-means clustering [9] is a partitioning method which splits the data into k partitions, where each partition represents a cluster. Due to its straightforwardness and easy to use it is top 10 in data mining domain. Euclidean distance measure is used in K-means algorithm. The main reason of the K-means algorithm is to reduce the distance, as per Euclidian measurement, between objects in clusters. The K-Means clustering algorithm is shown in Fig. 2.

K-Means Clustering Algorithm
Input: Set of data points and set of centres.
Output: Desired Cluster

Step 1: Choose the number of clusters, *k*.
Step 2: Randomly generate k clusters and determine the cluster centres, or directly generate *k* random points as cluster centres.
Step 3: Assign each point to the nearest cluster centre.
Step 4: Recompute the new cluster centres.
Step 5: Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).

Fig. 2: K-means Clustering Algorithm

4.2 HIERARCHICAL AGGLOMERATIVE CLUSTERING WITH MVS

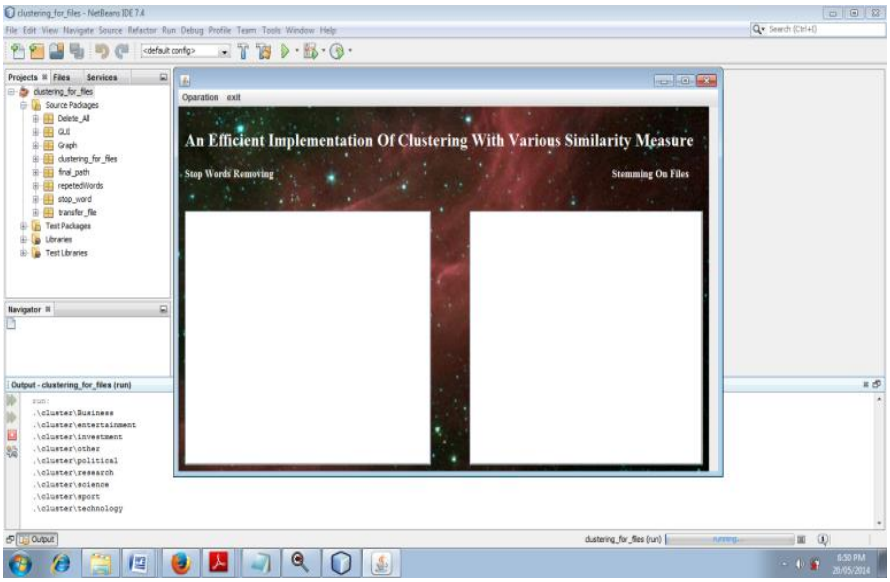
Hierarchical document clustering [8] categorizes clusters into a tree or a hierarchy that facilitates browsing. Hierarchical clustering methods are classified into agglomerative (bottom-up) and divisive (top-down). An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most suitable clusters to form large cluster. The Hierarchical Agglomerative clustering algorithm is shown in Fig. 3.

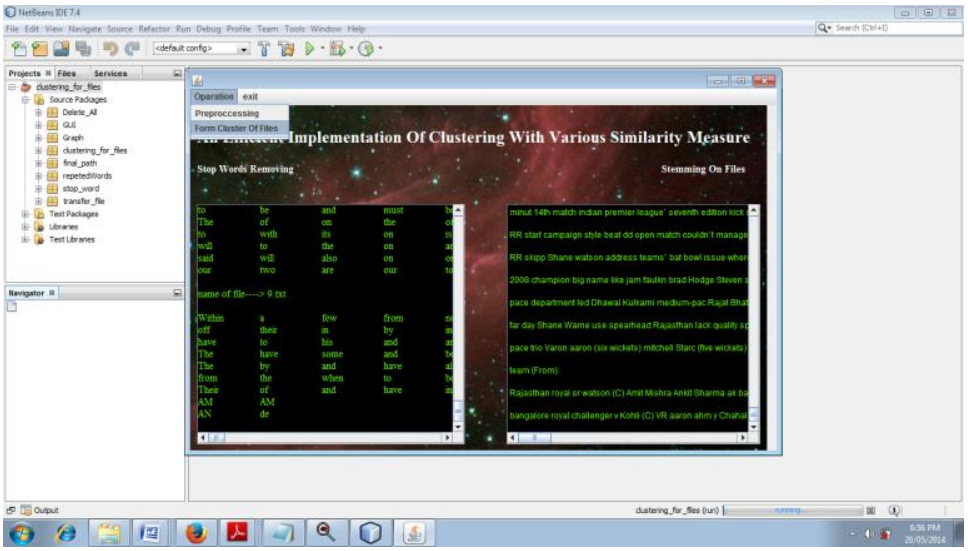
Hierarchical Agglomerative Clustering Algorithm
Input : Number of clusters
Output : Desired cluster

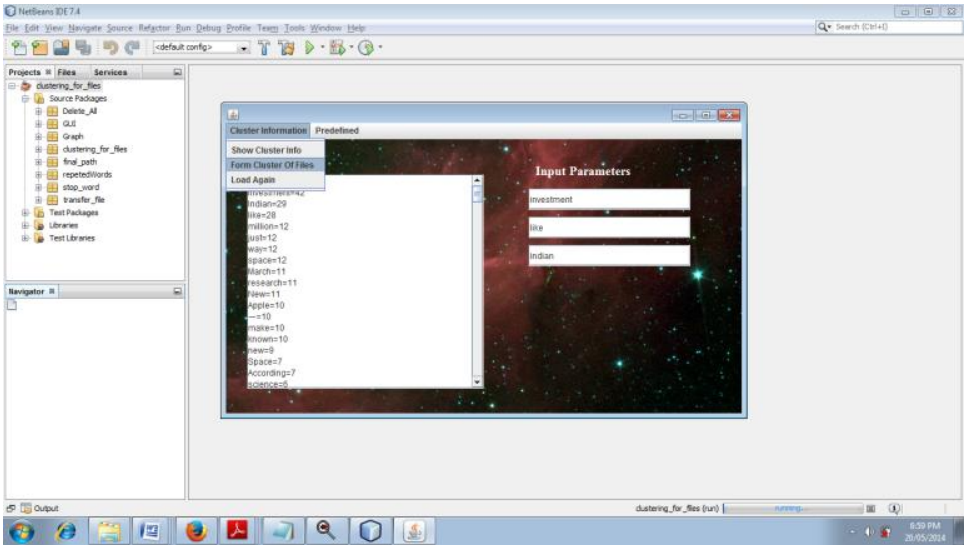
Step 1: Start with N clusters, each containing a single entity, and an N × N symmetric matrix of distances (or similarities).
Let dij = distance between item i and item j.
Step 2: Search the distance matrix for the nearest pair clusters (i.e., the two clusters that are separated by the smallest distance).
Denote the distance between these most similar clusters A and B by dAB.
Step 3 : Merge clusters A and B into a new cluster, labelled S. Update the entries in the distance matrix by
a. Deleting the rows and columns corresponding to clusters A and B, and
b. Adding a row and column giving the distances between the new cluster S and all the remaining clusters.
Step 4: Repeat steps (2.) and (3.) a total of N-1 times.

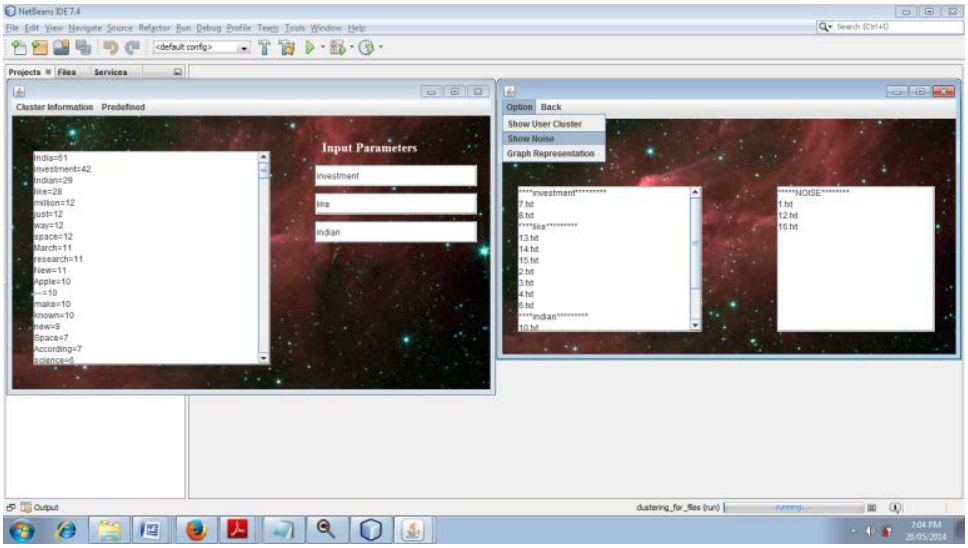
Fig. 3: Hierarchical Agglomerative Clustering Algorithm

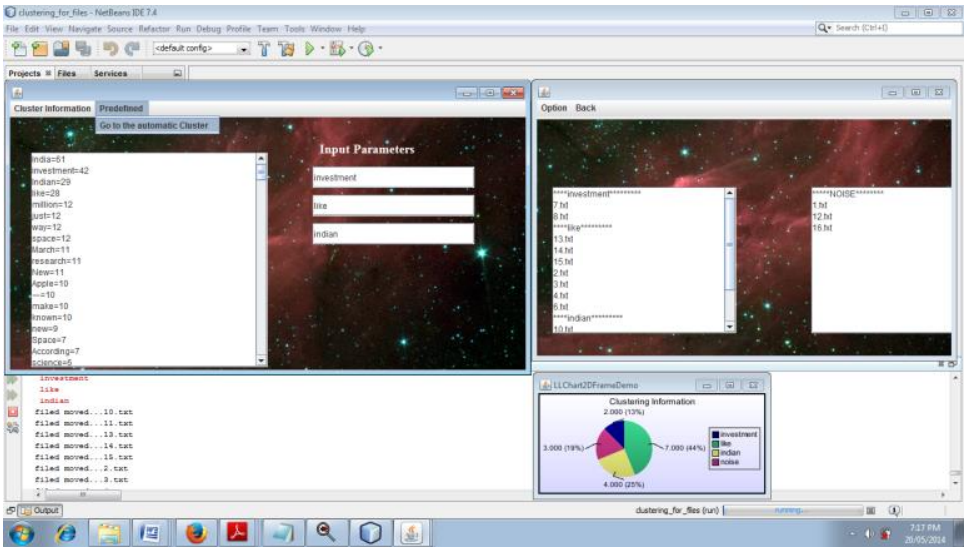
5 RESULTS

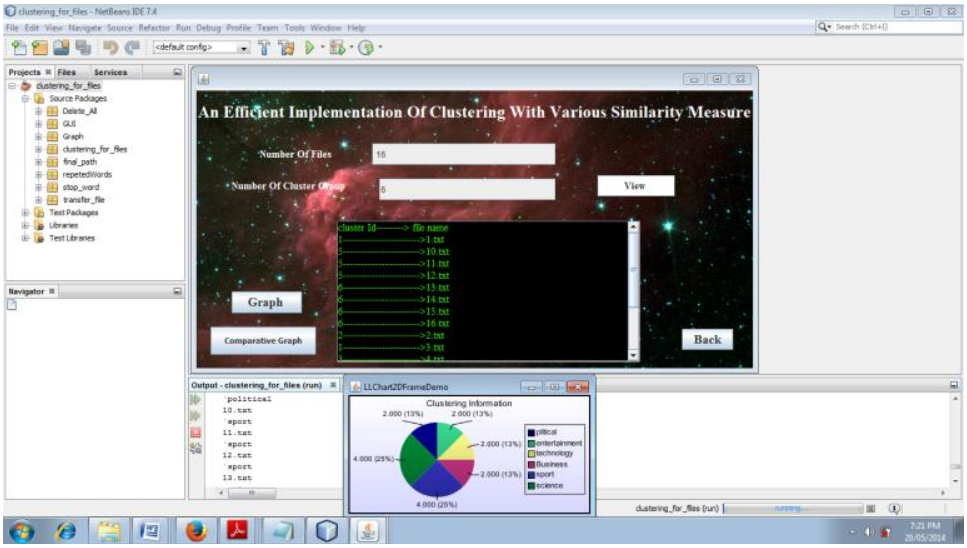












6 CONCLUSION

In this paper, Hierarchical Agglomerative clustering with multiviewpoint-based similarity method is implemented. The experiment is conducted over sixteen text documents. The experimental results show that hierarchical agglomerative clustering with multiviewpoint-based similarity measure is potentially more suitable for text documents clustering and retrieval. It is also suitable for sparse and high dimensional data compared with partitional MVS clusters. The new method can decrease the time cost, reduce the space complexity and improve the accuracy of clustering. Specially, in the document clustering, the newly proposed algorithm measuring result show great advantages.

REFERENCES

[1] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.

[2] I. Guyon, U.V. Luxburg, and R.C. Williamson, "Clustering: Science or Art?" Proc. NIPS Workshop Clustering Theory, 2009.

[3] I. Dhillon and D. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, vol. 42, nos. 1/2, pp. 143-175, Jan. 2001.
[4] S. Zhong, "Efficient Online Spherical K-means Clustering," *Proc. IEEE Int'l Joint Conf. Neural Networks (IJCNN),* pp. 3180-3185, 2005.
[5] C.D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval.* Cambridge Univ. Press, 2009.
[6] A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-Page Clustering," *Proc. 17th Nat'l Conf. Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI),* pp. 58-64, July 2000.
[7] Anna Huang, "Similarity Measures for Text Document Clustering", NZCSRSC 2008, April 2008, Christchurch, New Zealand.
[8] E. Mooi and M. Sarstedt, "Cluster Analysis", DOI10.1007/978-3-642-12541-6_9, ©Springer-Verlag Berlin Heidelberg 2011.
[9] A.K. JAIN, M.N. MURTY and P.J. FLYNN, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999.
[10] A. Anil Kumar, S.Chandrasekhar, "Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering".

