The Effect of Word Sampling on Document Clustering

OMAR H. KARAM AHMED M. HAMAD SHERIN M. MOUSSA
Department of Information Systems
Faculty of Computer and Information Sciences
University of Ain Shams, Cairo, 11566
EGYPT
ohkaram@hotmail.com, amhamad13@yahoo.com, sherinmoussa@ieee.org

Abstract: - Many techniques have been used for document clustering that depended on the number of word occurrences in documents. In these techniques, words are considered as dimensions of the clustering space. Since a huge number of words is found in each document, studies were held to reduce this high dimensionality for better performance i.e., words pruning. Sampling was used to choose random documents representatives to which apply clustering techniques instead of using the whole data set, but it was not implemented on words before. In this paper, we study the effect of using word sampling on document clustering as a method of high dimensionality reduction, where a random word sampling technique is presented. The Euclidean and Manhattan distance functions were both used as the similarity measure. A hybrid clustering algorithm is modified to include word sampling. The results are compared with the non-word sampling through the clustering accuracy of the resultant clusters.

Key-Words: - Data mining, Text mining, Document clustering, Agglomerative clustering, k-means clustering, and Word sampling.

1 Introduction

Automated document clustering is an important text mining task since, with the existence of an increasing number of online documents; it is essential to be able to automatically organize such documents into clusters so as to facilitate document retrieval and subsequent analysis [3]. Document clustering is a task that seeks to identify homogeneous groups of documents based on the values of their dimensions (word occurrences).

Given a set of objects and a clustering criterion, partitional clustering obtains a partition of the objects into clusters such that the objects in a cluster are more similar to each other than to objects in different clusters. Examples are the k-means and k-medoid methods that determine k cluster representatives and assign each object to the cluster with its representative closest to the object such that the sum of the distances squared between the objects and their representatives is minimized. On the other hand, a hierarchical clustering is a nested sequence of partitions. An agglomerative hierarchical clustering starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters until all objects are in a single cluster. Divisive hierarchical clustering reverses the process by starting with all objects in cluster and subdividing into smaller pieces [3, 4, 5, 8, 11, 13].

Documents are very high dimensional. Sampling was used to apply clustering techniques on a randomly chosen set of documents instead of the whole data set in order to enhance clustering performance [1, 2, 8, 12]. Sampling has never been discussed on the words level within a single document, where each document is represented by a sample of its words. In [10], multiple subsamples, J, are randomly chosen from the data set and clustered independently producing J estimates of the true cluster locations. To avoid the noise associated with each of the J solutions, a "smoothing" procedure was employed. However, to "best" perform this smoothing, the problem of grouping the K*J points (J solutions, each having K clusters) into K should be solved.

Another way to address high dimensionality is to apply a dimensionality reduction method to the dataset such as the principal component analysis that optimally transforms the original data space into a lower dimensional space by forming dimensions that are linear combinations of given attributes. The new space has the property that distances between points remain approximately the same as before [7].

CLARA is a clustering algorithm that combines a sampling procedure with the classical PAM algorithm [3, 9]. Instead of finding medoids for the entire data set, CLARA draws a random sample from the data set and uses the PAM algorithm to select an optimal set of medoids from the sample.

In this paper, we present an experimental study for using word sampling on document clustering as a method of high dimensionality reduction, where a random word sampling technique is presented. The Euclidean and Manhattan distance functions were both used as the similarity measure. A hybrid clustering algorithm is modified to include word sampling [1, 2]. The results are compared with the non-word sampling through the clustering accuracy of the resultant clusters.

The rest of the paper is organized as follows: In section 2, Clustering Text Documents is discussed, while section 3 presents Document clustering using word sampling and discusses the associated experiments and results. Section 4 is the Conclusion.

2 Documents Preparation

In order to apply any clustering algorithm on text documents, some steps should be followed [4, 5, 6]. These are:

2.1 Preprocessing

Initially, Very common words known as stop words are removed completely from documents. In addition, stemming is applied. In our work, Porter's suffix stripping algorithm is used [14]. Now documents are ready to be mathematically represented.

2.2 Mathematical Representation

In our system, the vector space model is used, where each document is represented by the Term-Frequency (TF) vector, **d**, in the multidimensional space of document words. The (TF) vector is:

$$d_{tf} = (tf_1, tf_2, \cdots, tf_n) \quad (1)$$

where tf_i is the frequency of the ith term in the document. Given a set, S, of documents and their corresponding vector representations, the centroid vector, **c**, is defined to be:

$$\mathbf{c} = \frac{1}{|S|} \sum_{\mathbf{d} \in S} \mathbf{d} \quad (2)$$

which is obtained by averaging the weights of the various terms present in the documents of S, where **d** is the document vector obtained from the set S.

2.3 Dimension Space Construction

Pruning is used to reduce the high dimensionality of document vectors. A threshold P is given, and words having a frequency less than this threshold are removed completely from the document vectors.

At this stage, document vectors lengths are not unified. Therefore, a superset vector is created that contains all words in all the documents after pruning. This superset vector is then compared to each document vector, and each document vector is mapped from its own space to the superset vector space.

2.4 Similarity Measure

Similar document vectors are grouped together in the same cluster. The similarity measure between any two vectors is computed using the Minkowski distance function,

$$d(i,j) = (|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q)^{1/q} \quad (3)$$

where i represents the ith document vector and j represents the jth document vector, x_{ip} is the word number p in the ith document vector; x_{jp} is the word number p in the jth document vector, q is the power of the distance function In this work, the Centroids similarity is used [11], where similarity between two clusters is measured between the centroids of each.

3 Document Clustering Using Word Sampling

3.1 The System

The hybrid clustering system [1, 2] has been modified to apply the concept of word sampling for further reduction of the high dimensionality of documents. It is based on the idea of employing the agglomerative hierarchical clustering algorithm in order to provide the k-means with the initial starting conditions, i.e., k and the initial centroids, thus providing a solution to the initialization problem associated with the k-means. That is, a sample of documents is randomly chosen from the data set to be agglomeratively clustered. Fig. 1 represents the abstract flowchart of the system. The resultant clusters are considered to be the initial centroids and their count is taken as the k value to cluster the whole data set using the k-means clustering algorithm.

The used documents sampling technique groups documents into batches of 10. For each batch except the last, 10 random numbers are generated in the range from 1 to 10 to add documents at the positions of these generated random numbers to the sample while excluding the others. In the last batch, the required sample percentage is multiplied by the number of documents found in this batch to get the number of documents to keep, then random numbers are generated as previously explained.

System Parameters

a) The pruning threshold, P.

b) The sample size used in the agglomerative clustering approach, R.

c) The agglomerative distance threshold, A.

d) The power of the distance function, q.

e) The clustering accuracy measure.

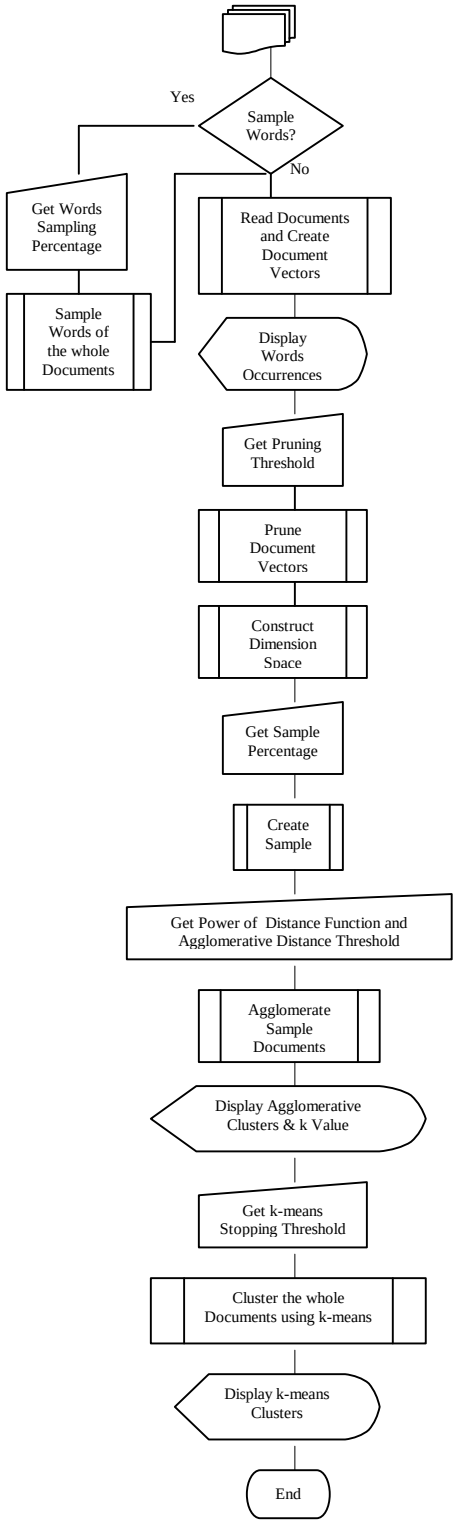

Fig. 1 – General Process.

3.2 Word Sampling Technique

Each document is represented by only a percentage of its words, while the other words are removed. Fig. 2 provides the flowchart of the used word sampling technique. Words percentage is a user-defined parameter. Document vectors are then created using the chosen sampled words and their number of occurrences. To avoid biased sampling, words are processed in batches of 10 words and the words to remove from each batch are selected randomly by generating random numbers from 1 to 10 and select the word at the position of the random number generated. As for the last batch, the required word sample percentage is multiplied by the number of words contained to determine how many words to keep from this batch. Document vectors are then pruned. Hence, High dimensionality is overcome. The specified percentage of word sampling is an important parameter that will be studied in this paper to determine its effect on the accuracy of clustering results. Fig.3 contains the sizes of the created matrices of the constructed non-word sampling dimension spaces, where rows represent the number of contained words (obtained from the superset vector) and columns represent the number of considered document vectors. On the other hand, fig. 4 contains the matrices size of the constructed word sampling dimension spaces at L = 40%. We can notice the great reduction in the matrices sizes of the word sampling dimension spaces compared to those of the non-word sampling.

3.3 Experiments

In order to study the effect of word sampling on document clustering using the hybrid clustering algorithm, the National Science Foundation data set was obtained by downloading 725 abstracts included different subjects. The same experiments were held twice; once without using the word sampling, and the other using the word sampling. As for the non-word sampling experiments, Different pruning thresholds were used with P = 2 to 7. Document samples of R = 10% and 20% were randomly chosen to run the agglomerative clustering algorithm on. In case of the Euclidean distance function, different agglomerative distance thresholds, A = 5, 10, 15, and 20 were applied. In case of the experiments held for the Manhattan distance function, the constructed dimension space has been changed to simplify computations. During the stage of comparing each word in the superset vector with the words of each document vector to construct the dimension space, once a word is found in the considered document vector its word count is

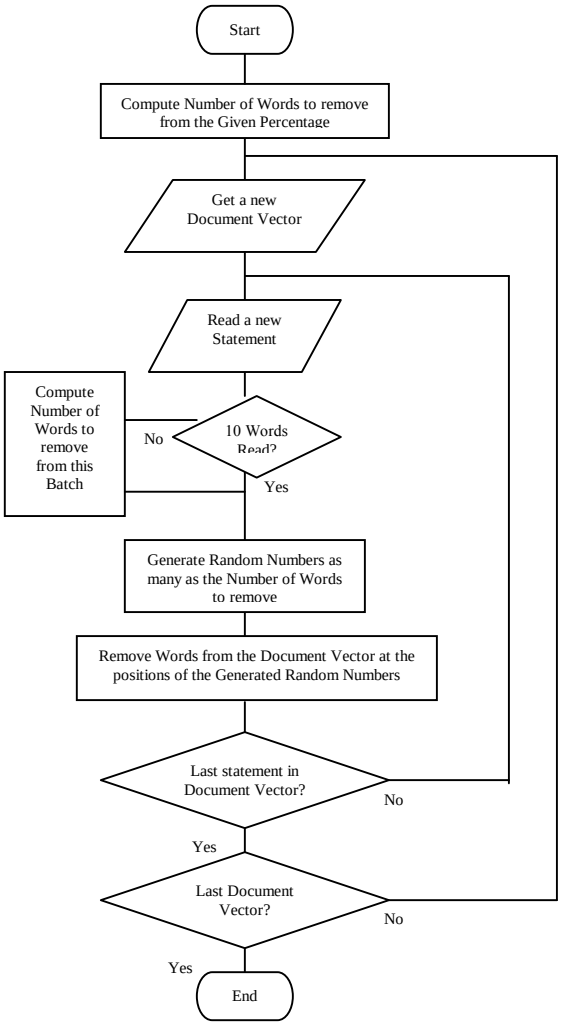

Fig. 2 – Word Sampling.

changed to 1 instead of the actual number of occurrences to indicate that it is found. Thus, the new constructed dimension space contains document vectors of zeros and ones. Another agglomerative thresholds were applied at A = 1, 3, 5. Regarding the word sampling experiments, another pruning thresholds were used P = 2, 3, 4. This change in the P values is due to the great reduction of the words occurrences after word sampling. Whereby, different agglomerative thresholds were applied using the Euclidean distance function at A = 5, 10. The effect of word sampling on the resultant clusters was studied at different word sampling sizes L = 20%, 40%, 60%, and 80%. In order to determine the clustering accuracy, we obtained taxonomy for the used NSF data set. Clustering accuracy represents the percentage of the number of documents that are clustered correctly according to the taxonomy of the documents at similar k value (similar number of clusters).

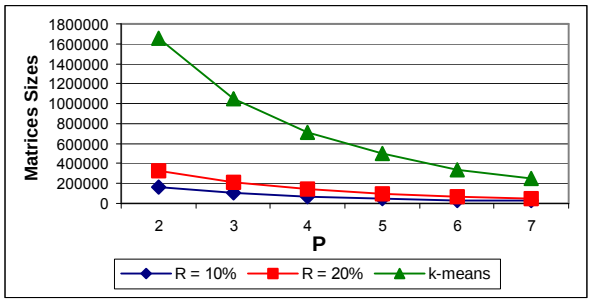

Fig. 3 – Matrices sizes for non-word sampling constructed dimension spaces

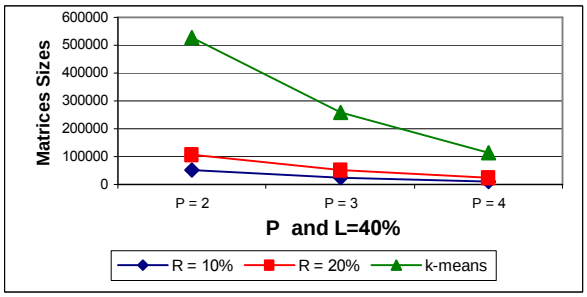

Fig. 4 – Matrices sizes for word sampling constructed dimension spaces with L = 40%

3.4 Results

As shown in Fig. 5 to Fig. 8 for the Euclidean distance function at R = 10%, results stability started from P = 6 to P = 7, while at R = 20% stability started from P = 5 to P = 7 for the different agglomerative thresholds. In case of the Manhattan for both R = 10% and 20%, clustering is not sensitive to the pruning threshold as much as it is sensitive to the smaller variations in the Euclidean. The behavior of the pruning threshold in Manhattan is more consistent, which emphasizes the existing difference in each dimension due to the squaring. In addition, the range of accuracy obtained using the Euclidean distance function at higher pruning thresholds is the same as for the Manhattan especially for the high pruning thresholds. Therefore, it is recommended to use the Manhattan distance function since it is less complex. Regarding the agglomerative threshold effect in the Manhattan distance function for both document sample R = 10% and 20%, the clustering accuracy is consistent for the different pruning thresholds while varying the agglomerative threshold. As for the Euclidean distance function, again the effect of squaring is clear in the wider range in the clustering accuracy percentage across all the agglomerative thresholds. Considering the Euclidean distance function as referred in Fig. 9 up to Fig. 12 for both document sample R = 10% and 20%, clustering accuracy for L = 40% to 80% are virtually all above 75%.

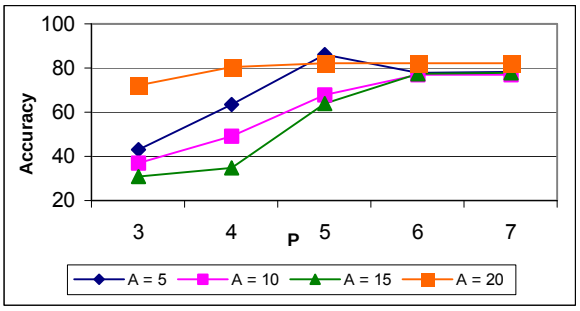

Fig. 5 – Accuracy vs. P at different A with R = 10%, Euclidean distance function

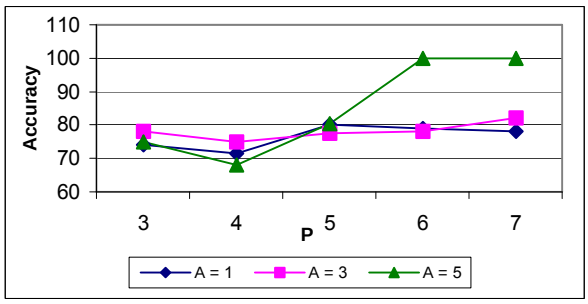

Fig. 6 – Accuracy vs. P at different A with R = 10%, Manhattan distance function

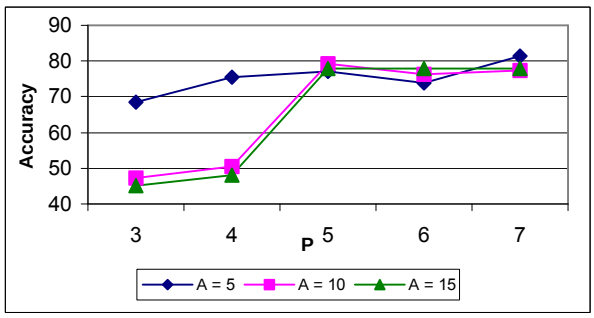

Fig. 7 – Accuracy vs. P at different A with R = 20%, Euclidean distance function

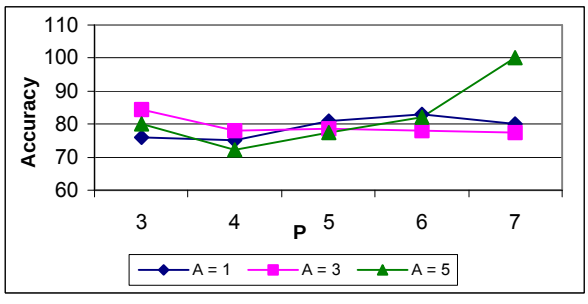

Fig. 8 – Accuracy vs. P at different A with R = 20%, Manhattan distance function

At higher agglomerative threshold A = 10, accuracy percentages are close to each other. In case of the Manhattan, variations in accuracy are higher than those using the Euclidean. Therefore, Euclidean can be used with word sampling while maintaining accuracy. Regarding the comparison between non-word and word sampling in Fig. 13, non-word sampling case is a word sampling with L = 100%. It was found to use word sampling with low pruning thresholds and obtain high accuracy compared to those obtained with non-word sampling (L = 100%).

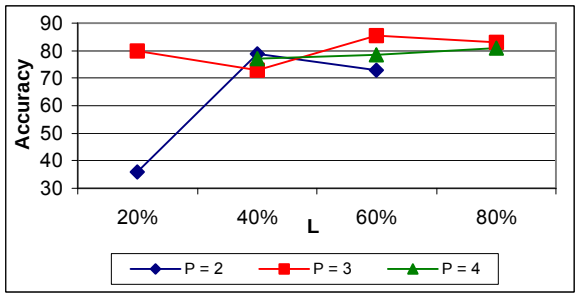

Fig. 9 – Accuracy vs. P at different L with A = 5 and R = 10%, Euclidean distance function

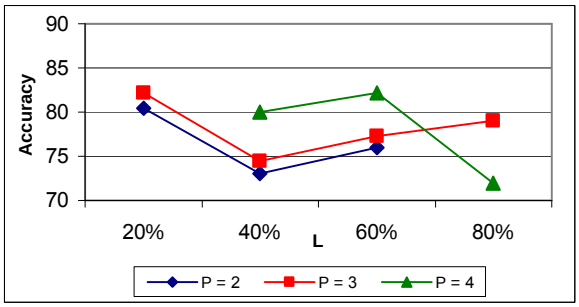

Fig. 10 – Accuracy vs. P at different L with A = 3 and R = 10%, Manhattan distance function

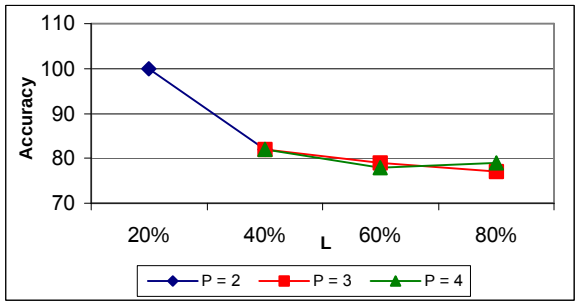

Fig. 11 – Accuracy vs. P at different L with A= 10 and R = 10%, Euclidean distance function

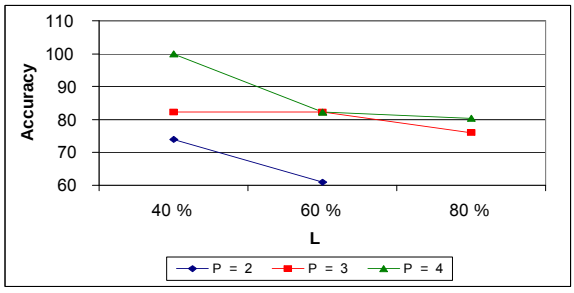

Fig. 12 – Accuracy vs. P at different L with A = 5 and R = 10%, Manhattan distance function

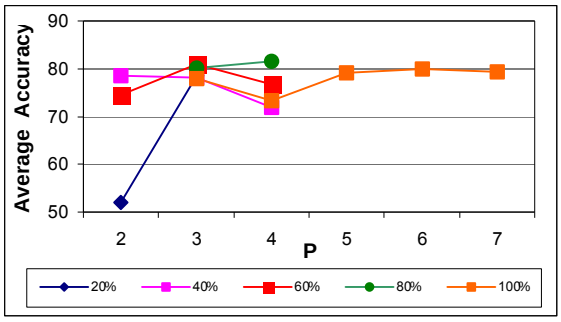

Fig. 13 – Average accuracy vs. P at different word sample size percentages

4 Conclusion

A study is presented for the effect of using word sampling on document clustering as a method of high dimensionality reduction, where a random word sampling technique is used. The results were compared with the non-word sampling through the clustering accuracy of the resultant clusters. It was found that we could use word sampling with low pruning thresholds and obtain a higher accuracy compared to those with non-word sampling. Considering the Euclidean distance function, clustering accuracy is virtually above 75% for higher values of word sampling sizes, while the variations in the clustering accuracy are higher in case of the Manhattan distance function. At higher agglomerative threshold accuracy percentages are close to each other for high values of word sampling sizes. The Euclidean distance function can be used with word sampling while maintaining clustering accuracy for high values of word sampling sizes.

References:

[1] O. H. Karam, A. M. Hamad, and S. M. Moussa. Determining Initial Starting Conditions For Documents Clustering. In Proceedings, 1st International Conference on Intelligent Computing and Information Systems (ICICIS 2002). Cairo, June 2002.

[2] O. H. Karam, A. M. Hamad, and S. M. Moussa. Analysis of Documents Clustering Using Sampled Agglomerative Technique. *In Proceedings, 12th International Conference on Computer Theory and Applications, (ICCTA'2002).* Alexandria, August 2002.

[3] J. Han, M. Kamber. *Data Mining Concepts and Techniques.* Morgan Kaufmann Publishers. Pages 335 – 388 and 428 – 435, 2001.

[4] I. S. Dhillon and D. S. Modha. *Concept Decompositions for Large Sparse Text Data using Clustering*. Technical Report RJ 10147, IBM Almadan Research center, 2000.

[5] M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. *In KDD workshop on Text Mining*, 2000.

[6] J.L. Neto, A.D. Santos, C.A.A. Kaestner, and A.A. Freitas. Document Clustering and Text Summarization. *In Proceedings, 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining (PADD-2000)*, 41-55. London: The Practical Application Company, 2000.

[7] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data*, Seattle, Washington, June 1998.

[8] M. Meila, and D. Hackerman, *An Experimental Comparison of Several Clustering and Initialization Methods*, (Technical Report 98-06). Microsoft Research Redmond, WA, 1998.

[9] C. Wei, Y. Lee and C. Hsu. Empirical Comparison of Fast Clustering Algorithms for Large Data Sets. *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 1998.

[10] P.S. Bradley, and U.M. Fayyad, Refining Initial Points for K-Means Clustering. *Proceedings of the15th Intertional Conference on Machine Learning*, 91-99. Morgan Kaufmann Publishers, Inc., San Francisco, CA.,1998.

[11] M. J. A. Berry, G. Linoff. *Data Mining Techniques for Marketing, Sales, and Customer Support.* John Wiley & Sons. Pages 187 – 215, 1997.

[12] D. Fisher. *Iterative optimization and simplification of hierarchical clusterings*. J. Art. Intell. Res., 4:147--179, 1996.

[13] D. Cutting, D. Karger, J. Pedersen, and J. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. *In 15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92),* pages 318--329, 1992.

[14] M.F. Porter, *An Algorithm for Suffix Stripping*, Program 14 (3), Pages 130-137, July 1980.