A COMPARATIVE STUDY OF FUZZY MODELS IN DOCUMENT CLUSTERING

G.MANIMEKALAI
M.Phil Scholar
Department of Computer Application
PSGR Krishnammal College for Women
Coimbatore, India.
mekalawin.mphil@gmail.com

K.SATHIYAKUMARI, V.PREAMSUDHA
Assistant Professors, Department of Computer Application
PSGR Krishnammal College for Women
Coimbatore, India.

Abstract— **The availability of large quantity of text documents from the World Wide Web and business document management systems has made the dynamic separation of texts into new categories as a very important task for every business intelligence systems. Text document clustering is one of the emerging and most needed clustering techniques used to cluster documents with regard to similarity among documents. It is used widely in digital library management system in the modern context. Document clustering is widely applicable in areas such as search engines, web mining, information retrieval, and topological analysis. There are several clustering approaches available in the literature to cluster the document. But most of the existing clustering techniques suffer from a wide range of limitations. The existing clustering approaches face the issues like practical applicability, very less accuracy, more classification time etc. Thus a novel approach is needed for providing significant accuracy with less classification time. In recent times, inclusion of fuzzy logic in clustering provides better clustering results. One of the widely used fuzzy logic based clustering is Fuzzy C-Means (FCM) Clustering. In order to further improve the performance of clustering, this thesis uses Modified Fuzzy C-Means (MFCM) Clustering. The documents are ranked using Term Frequency–Inverse Document Frequency (TF–IDF) technique. From the experimental results, it can be observed that the proposed technique results in better clustering when compared to the FCM clustering technique.**

Keywords- Purity, Entropy, TF-IDF, Fuzzy C Mean, Modified Fuzzy C Mean

I. INTRODUCTION

The victory in digital revolution and the advances in the Internet have ensured that enormous volumes of high-dimensional data are available to all users. The World Wide Web (WWW) has played a significant role in making the data, even from geographically distant locations, easily accessible to users all over the world. The availability of huge quantity of text documents from the World Wide Web and business document management systems has made the dynamic separation of texts into new categories as a very important task. The text database, which consists of documents, is usually very large. The Word Wide Web is such a database, where the major question in the areas of information retrieval and text mining is how to investigate and utilize this kind of text database.

Data mining is an evolving and emerging area of research and development both in academia as well as in industry. It involves interdisciplinary research and development encompassing diverse domains. New techniques and directions are being proposed in the literature every day (Karin, 2004; Kriegel *et al*., 2003a; Nierman and Jagadish, 2002). By and large, there are two types of data mining tasks, namely, descriptive data mining tasks and predictive data mining tasks (Banerjee and Ghosh, 2002). Descriptive data mining tasks portray the general properties of the existing data. They discover human-interpretable patterns that depict the data. Examples include association rule discovery, sequential pattern discovery, clustering, characterization, etc. Predictive data mining attempts to do predictions based on inference on available data. They use variables to predict unknown or future values of other variables. Some predictive data mining techniques are classification, regression, anomaly detection, change/evolution analysis, etc. The research work proposed here is focused on clustering of text

documents. Clustering of text data is a realm that comes under text data mining. A brief introduction to text clustering is discussed in the next section.

The main aim of this research is to develop a document clustering technique for very high accuracy. The time utilized for active clustering of documents is more if large databases are taken up for clustering. In the case of determining the initial clusters also, varying clusters would result for the same dataset. The proposed clustering algorithms involve the grouping of electronic documents, extracting important contents from the document collection and supports effective management of digital library documents. Contents of digital documents are analyzed and grouped into various categories.

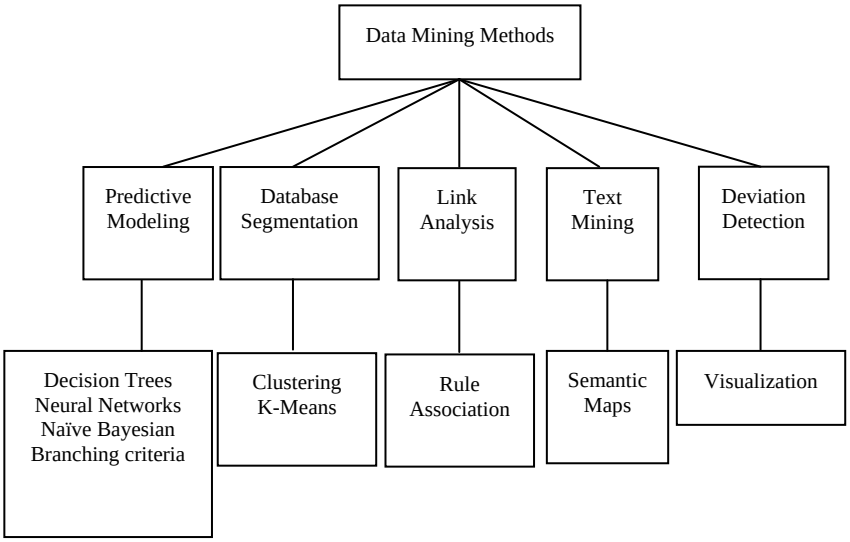

Figure 1: Data mining methods

Clustering Technique

Categorizing data into sensible groups is an indispensable factor for understanding and learning. For instance, a common scheme of scientific classification puts organisms into a system of ranked taxonomy like domain, kingdom, phylum, class, etc. Cluster analysis is defined as the study of methods for grouping or clustering objects depending on the measured or perceived inherent characteristics or similarity. A category label is not used in cluster analysis.

Cluster analysis is widely used in disciplines that involve analysis of multivariate data. It is complicated to list the various scientific fields and applications that have utilized clustering techniques as well as the thousands of published algorithms. Image segmentation is an important problem in computer vision, which can be formulated as a clustering problem. It is also applied to group customers into different clusters for efficient marketing. Text documents can be clustered to generate topical hierarchies for efficient information access or retrieval.

Retrieval result can be gradually increased by several methods; one of the approaches is clustering the retrieval result before showing it to the user. The idea behind it is that the retrieval result usually covers several topics and the user may be interested in just one of them. Document clustering is basically grouping up of similar document into individual groups. It is an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters. Unlike document classification, no labeled documents are provided in clustering; hence, clustering is also known as unsupervised learning. Browsing can be improved by organizing the hierarchical documents into clusters in the form of tree or a hierarchy.

II. LITERATURE SURVEY

Document clustering is the process of categorizing text document into a systematic cluster or group, such that the documents in the same cluster are similar whereas the documents in the other clusters are dissimilar. It is one of the vital processes in text mining. Due to growth and development in the filed of internet and computational technologies, various clustering techniques have been proposed in the literature. Especially, text mining has gained lot of importance and it is demanding various tasks such as production of granular taxonomies, document summarization etc., for the scope of developing higher quality information from text.

Text mining is a knowledge concentrated technique where the user communicates with a document collection by using analysis tools. This is equivalent to data mining approach. It extracts the useful information from large volume of unstructured text. Text document used to identify simplified subset of document features that can be used to represent the particular document as the whole. This feature is said to be a representational model. Each document in a collection is made up of large number of features, so that it affects the system approach, performance and design.

The most widely used fuzzy clustering algorithm is Fuzzy c-means (Bezdek, 1984), a variation of the partitional k-means algorithm. In fuzzy c-means each cluster is represented by a cluster prototype and the membership degree of a document to each cluster depends on the distance between the document and each cluster prototype. The closest the document is to a cluster prototype, the greater is the membership degree of the document in the cluster. Fuzzy Clustering and Fuzzy Merging algorithm (FCFM) (Looney, 1999) is another fuzzy approach, that tries to overcome the fact that fuzzy c means doesn't take into account about the distribution of the document vectors in each cluster. The FCFM uses Gaussian weighted feature vectors to represent the cluster prototypes. If a document vector is equally close to two prototypes, then it belongs more to the widely distributed cluster than to the narrowly distributed cluster.

In the year 1973 Dunn developed the Fuzzy C Means algorithm and later in 1981 Bezdek enhanced it. Fuzzy C Means algorithm is extensively used in pattern recognition. Fuzzy C Means algorithm uses the iterative process, which rejuvenates cluster centers for individual data point. Fuzzy C Means algorithm repetitively iterates the cluster centers to the exact location with in data set elements. The performance of Fuzzy C Mean algorithm is based on the initial centroids selected. The mean of all data points in the Fuzzy C Means algorithm is calculated as the centroid of a cluster and is weighted by their degree corresponding to the cluster.

III. METHODOLOGY

The main aim of this research is to develop a document clustering technique for very high accuracy. The time utilized for active clustering of documents is more if large databases are taken up for clustering. In the case of determining the initial clusters also, varying clusters would result for the same dataset. The proposed clustering algorithms involve the grouping of electronic documents, extracting important contents from the document collection and supports effective management of digital library documents. Contents of digital documents are analyzed and grouped into various categories.

The convergence time for the usage for existing clustering techniques is more and also the number iterations required is also very high. So, new and efficient techniques are needed to reduce convergence time and number of iterations. Moreover, in distributed data mining, the adoption of a flat node distribution technique would have an impact upon its scalability. To address these issues of modularity, flexibility and scalability, the proposed techniques are formulated.

A. Fuzzy Clustering

Conventional clustering techniques create partitions in which each pattern belongs to one and only one cluster. Therefore, the clusters in a hard clustering are disjoint. Fuzzy Clustering approach associates each pattern with every cluster using a membership function. The output of such techniques is a cluster, but not a partition.

B. Fuzzy C Mean Clustering (FCM)

Documents to be clustered are denoted by $X = \{ X_K \,|k = 1, 2....n \}$, and is represented as an n x m matrix.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

Each data sample, x_k, is defined by m features, i.e., $X_k = \{X_{k1}, X_{k2}, X_{k3}, ..., X_{km}\}$ where each x_k in the universe X is an m-dimensional vector of m elements or m features.

A large family of fuzzy clustering algorithms is based on minimization of the fuzzy c-means function formulated as

$$J_m(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \|x_k - v_i\|_A^2$$

Where, $U = [u_{ik}] \in M_{fc}$

is a fuzzy partition matrix of X, and u_{ik} is the membership of the kth data point in the ith class.

$$V = [v_1, v_2, v_3, ..., v_c]\, v_i \in R^n$$

is a vector of cluster prototypes (centers), which have to be determined,

$$D_{ikA}^2 = \|x_k - v_i\|_A^2 = (x_k - v_i)^T A (x_k - v_i)$$

is the Euclidean distance between the ith cluster center and the kth data set and $m' \in [1, \infty)$ is a new parameter called a weighting parameter controls which determines the amount of fuzziness in the classification process. The function J_m can have a large number of values, the smallest one associated with the best clustering**.**

Algorithm

Step 1: Given n data points $X = \{X_1, X_2, ..., X_n\}$

Fix the number of centers to $c, 2 \le c < n$

Initialize the random partition matrix and m' > 1.

Step 2*:* Set 1=0, 1, 2, ... , compute the c-mean vector for i=l, 2, ... , c,

$$V_i^{(l)} = \frac{\sum_{k=1}^{n}\left(u_{ik}^{(l-1)}\right)^{m'} x_k}{\sum_{k=1}^{n}\left(u_{ik}^{(l-1)}\right) m'}$$

Step3*:* Compute the distance: $i = 1, 2, ..., c, k = 1, 2, ..., n.$

$$D_{ikA} = \left\| x_k - v_i^{(l)} \right\|_A$$

Step 4: Update $U^{(1)}$ to $U^{(1+1)}, i = 1, 2, ..., c, k = 1, 2, ..., n.$

$$u_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c}\left(D_{ikA} / D_{jkA}\right)^{2/(m'-1)}}$$

Step 5*:* Compare $U^{(1)}$ and $U^{(1-1)}$ $U^{(l-1)}$: if $\left\| U^{(1)} \text{ and } U^{(1-1)} \right\| < \varepsilon$, stop: otherwise $i = 1+1$ and go to Step 2.

In the FCM algorithm, the following parameters must be specified. Moreover, the fuzzy partition matrix, U, must be initialized.

C. Fuzziness Parameter

The weighting exponent m' significantly influences the fuzziness of the resulting partition. As m' approaches one from above, the partition becomes hard $\left(U_{ik} \in \{0,1\}\right)$ and v_i are ordinary means of the clusters. As $m' \to \infty$, the partition becomes completely fuzzy $\left(\mu_{ik} = 1/c\right)$ and the cluster means are all equal to the mean of X. If all other algorithmic parameters are fixed, then increasing m' will result in decreasing Jm. No theoretical optimum choice of m' has emerged in the literature. Convergence of the algorithm tends to be slower as the value of m' increase.

D. Termination Criterion

The FCM algorithm stops iterating when the norm of the difference between U in two successive iterations is smaller than the termination parameter ε. The termination tolerance ε value is between 0 and 1.

E. Modified Fuzzy C-Means (MFCM) Clustering

One of the important premises in document clustering is that similar documents have related feature values, and the probability that they belong to the same cluster is great. The spatial information is important in clustering, but the standard FCM algorithm does not fully utilize it. To exploit the spatial information, a modified membership function is defined as follows:

$$u_{ij} = \frac{u_{ij}^{m} S_{ij}^{n}}{\sum_{k=1}^{c} u_{kj}^{m} S_{kj}^{n}}$$

Where $S_{ij} = \sum_{k \in N(x_j)} u_{ik}$ is called spatial function, and $N\left(x_j\right)$ represents a square window centered on particulate document in the spatial domain. The spatial function S_{ij} represents the probability that the document x_j belongs to i^{th} cluster.

The spatial function of a document for a cluster is large if the majority of its neighborhood belongs to the same cluster. In a homogenous region, the spatial functions enhance the original membership, and the clustering result remains unchanged. However, for misclassified documents, it will reduce the weighting of a noisy cluster by the labels of its neighbors. As a result, misclassified documents can be easily corrected.

There are two steps at each clustering iteration. The first step is to calculate the membership function in the spectral domain and the second step is to map the membership information to the spatial domain and then compute the spatial function from that.

The iteration proceeds with the new membership that is incorporated with the spatial function. The iteration is stopped when the maximum difference between two cluster centroids at two successive iterations is less than a threshold. After the convergence, defuzzification is applied to assign each document to a specific cluster for which the membership is maximal. The Modified FCM algorithm (MFCM) can be described as follows:

Step 1: Set the number of clusters c and the parameter m. Initialize the fuzzy cluster centroid vector $V = [v_1, v_2, ..., v_c]$ randomly and set $\varepsilon = 0.01$.

Step 2*:* Compute u_{ij} by

$$u_{ij} = \left(\sum_{k=1}^{c} \left(\frac{d\left(x_j, v_i\right)}{d\left(x_j, v_k\right)} \right)^{2/(m-1)} \right)^{-1}$$

Step 3*:* Compute v_i by

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

Step 4*:* Update u_{ij} by

$$u_{ij} = \frac{u_{ij}^m S_{ij}^n}{\sum_{k=1}^{c} u_{kj}^m S_{kj}^n}$$

Step 5*:* Update v_i by

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

Repeat Steps 4 and 5 until the following termination criterion is satisfied:

$$\left| v_{new} - v_{old} \right| < \varepsilon$$

The main aim of this research is to develop a document clustering technique with very high accuracy. The time utilized for active clustering of documents is more if large databases are taken up for clustering. In the case of determining the initial clusters also, varying clusters would result for the same dataset. The proposed clustering algorithm involves the grouping of electronic documents. Contents of digital documents are analyzed and grouped into various categories.

F. Ranking using Term Frequency–Inverse Document Frequency (TF–IDF)

The term (word) frequency/inverse document frequency (TF-IDF) approach is commonly used to weigh each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance 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. Variations of the tf–idf-weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance to a user query.

Formula to compute term frequency is given by

$$TF = \frac{Term\ Occurances}{Occurances\ of\ all\ the\ terms}$$

Formula to compute inverse document frequency is

$$IDF = \log\left(\frac{Total\ Documents}{No.of\ documents\ containing\ term}\right)$$

We then compute the TFIDF weights by multiplying each cell in the “TF normalized” table by the corresponding IDF value.

Formula to compute inverse document frequency is

$$TFIDF\ rank = TF\ X\ IDF$$

Finally, the documents are clustered using Modified Fuzzy C-Means (MFCM) clustering algorithm and the ranking is performed using Term Frequency–Inverse Document Frequency (TF–IDF).

G. Entropy

Entropy is the degree to which each cluster consists of objects of a single class. For each cluster, the class distribution of the data is calculated initially, i.e., for cluster j we compute p_{ij}, the probability that a member of cluster i belongs to class j as $p_{ij} = m_{ij} / m_i$ where m_i is the number of objects in cluster i and m_{ij} is the number of objects of class j in cluster i. Using this class distribution, the entropy of each cluster i is calculated using the standard formula, $e_i = -\sum_{i=1}^{L} p_{ij} \log_2 p_{ij,}$, where L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each clusters weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{k} \frac{m_i}{m} e_i,$, where K is the number of clusters and m is the total number of data points.

H. Purity

Purity is another method of the extent to which a cluster objects of a single class. Purity of cluster i is $p_i = \max_j p_{ij},$, the overall purity of as clustering is

$$purity = \sum_{i=1}^{k} \frac{m_i}{m} p_i$$

IV. EXPERIMENTAL RESULTS

A. General Experimental Setting Bow Toolkit

Given a corpus, were stemmed all the documents, stop words removed and tf.idf vector was constructed for each document by using the *bow toolkit* (McCallum, 1996).

The idf of each term was computed using the following formula:

Where *n* is the total number of documents and df *(w*) is the number of documents that the term *w* appears.

$$idf(w) = \log_2 \left(\frac{n}{df(w)} \right)$$

The rainbow library for preprocessing the dataset. Stop words which do not contribute to the semantics of the document (like a, the, he, she and similar pronouns) and html tags are removed from each file. Stemming is done to combine words with same semantics but different forms/tense. For example (sleeping will be truncated to sleep.)

B.Twenty Newsgroups Dataset

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. It was originally collected by Ken Lang the 20 newsgroups collection has become a popular data set for experiments in text applications and machine learning techniques, such as text classification and text clustering. The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). Here the three-newsgroup dataset selected are talk.politics, alt.atheism, and comp.windows.x.

This is a very standard and popular dataset used for evaluation of many text applications, data mining methods, machine learning methods, etc. Its details are as follows:

- Number of unique documents = 5,000

- Number of categories = 20
- Number of unique words after removing the stop words = 3,000

The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian).

Table 1: Purity of Clustering

Clustering Method	***Talks.politics***	***Alt.atheism***	***Comp.windows***
FCM	0.899	0.898	0.987
Modified FCM	0.975	0.930	0.992

Figure 1: Purity Comparison for the Proposed Technique and Existing Technique

Table 1 and figure 1 shows comparison of the purity of classification for the proposed method with the existing methods. From the table 4.1, it can be observed that for talk.politics category, the purity of classification using FCM algorithm is 0.899 and for the proposed method the purity is higher i.e. 0.975. When the alt.atheism category is considered, the higher purity i.e. 0.930 is achieved by the proposed technique, whereas, the accuracy using FCM is 0.898.When comp.windows.x is considered, the better purity is achieved using the proposed technique i.e. 0.992 and 0.987 respectively.

The resulted entropy is provided in table 2 and figure 2. It can be seen that the resulted entropy is minimum for using MFCM, whereas it is higher for the other existing method. This clearly shows the improvement of the modified fuzzy c means clustering when compared to the existing clustering techniques.

Table 2: Entropy for Different Clustering Methods

Clustering Method	***Talks. Politics***	***Alt.atheism***	***Comp.windows***
FCM	0.355	0.312	0.105
Modified FCM	0.113	0.252	0.041

Figure 2: Entropy Comparison for the Proposed Technique and Existing Technique

C. Dense Dataset

All rows in a dense dataset file are comma-separated lists of real values. The first line of a dense dataset file should be a list of the input and output attribute names, and the second line should be blank. All remaining lines are treated as records. Such dataset files are referred to as csv files. Any attribute of a csv file can serve as the output, and the output attribute is selected at dataset load time. Some flexibility is gained if the output is the last attribute.

Table 3: Purity of Classification

Clustering Method	***Category 1***	***Category 2***	***Category 3***
FCM	0.675	0.899	1.000
Modified FCM	0.898	0.973	1.000

Figure 3: Purity for the Proposed Technique and Existing Technique

Table 3 and figure 3 shows comparison of the purity of classification for the proposed method with the existing methods. From the result, it can be observed that the proposed technique results in higher purity value when compared to the existing approach.

Table 4: Entropy for Different Clustering Methods

Clustering Method	***Category 1***	***Category 2***	***Category 3***
FCM	**0.718**	**0.339**	**0.600**
Modified FCM	**0.524**	**0.254**	**0.411**

The resulted entropy is provided in table 4 and figure 4. It can be seen that the resulted entropy is minimum for using MFCM, whereas it is higher for the other existing method. This clearly shows the improvement of the clustering when compared to the existing clustering techniques.

Figure 4: Entropy Comparison for the Proposed Technique and Existing Technique

V. CONCLUSION

The study uses two clustering algorithms to group documents .The algorithms were evaluated based on their purity and entropy. After the study, it is clear that on the whole SMO performs comparatively better than the other algorithms taken for study when applied to postoperative dataset. Clustering of document is important for the purpose of document organization, summarization, topic extraction and information retrieval in an efficient way. Initially, clustering is applied for enhancing the precision or recall in the information retrieval techniques. In recent times, clustering technique is applicable in the areas, which involves browsing a gathered data, or in categorizing the outcome provided by the search engine for the reply to the query provided by the user. In the future, this work can be extended by applying different classification techniques like Gaussian mixture models and RBF classifiers. Document clustering can also be applicable in producing the hierarchical grouping of document. In order to search and retrieve the information efficiently in Document Management Systems (DMSs), the metadata set should be created for the documents with enough details. But, the only one metadata set is not enough for the while document management systems. This is because the various document types need various attributes for distinguishing appropriately.

VI. FUTURE WORK

This thesis presents a new technique for clustering the documents, which used Modified Fuzzy C-Means algorithm for clustering and Term Frequency–Inverse Document Frequency (TF–IDF) technique for ranking. The problem that still occur in this clustering and also in the real world is how to determine exactly how many concepts actually presences in document collection. The problems included are

a) For realistic instances hundreds of unique keywords are resulted, so each individual is a vector of several hundreds real numbers. And it is known that the size of an individual needed for clustering algorithm to evolve satisfactory solutions grows exponentially with the length of the representation. So, it searches a way to reduce the dimension of clustering space so that the algorithm can be applied to large dataset.

b) A constraint is given that the longest continuous common subsequence shorter than 4, maybe it is better fit for Chinese characters rather than for other languages. Future study to solve this problem by using statistical method applied to find the optimal cluster number might be the answer of this problem. Furthermore, comparative studies with other reduction dimension methods need to be done.

REFERENCES

[1] Malay K. Pakhira, “A Modified k-means Algorithm to Avoid Empty”, International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009.

[2] Likas, N. Vlassis and J.J. Verbeek, “The Global k-means Clustering algorithm”, Pattern Recognition , Volume 36, Issue 2, 2003, pp. 451- 461.

[3] K. Jain, M. N. Murty and P. J. Flynn, “Data clustering: a review”, ACM Computing Surveys (CSUR), Volume 31 Issue 3, Sept. 1999.

[4] Khaled S. Al-Sultana and M. Maroof Khan, “Computational experience on four algorithms for the hard clustering problem”, Pattern Recognition Letters, Volume 17 Issue 3, 1996.
[5] S Z Selim and M A Ismail, “K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality”, EEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 6, Issue: 1, pp 81-87, (1984).
[6] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan, “Automatic Subspace Clustering of High Dimensional Data”, Data Mining and Knowledge Discovery, 2005.
[7] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey “Scatter/Gather: a cluster-based approach to browsing large document collections”, Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, 1992.
[8] M. Meila, D. Heckerman, “An experimental comparison of model-based clustering methods”, Machine Learning, vol. 42, 2001, pp. 9–29.
[9] Barry de Ville, "Text Mining with “Holographic” Decision Tree Ensembles", Data Mining and Predictive Modeling.
[10] Qiaozhu Mei and ChengXiang Zhai, "A Mixture Model for Contextual Text Mining", Research Track Poster.
[11] Miha Grcar, Marko Grobelnik and Dunja Mladenic, "Using text mining and link analysis for software mining", Proceedings of the 3rd ECML/PKDD international conference on Mining complex data, 2007.
[12] Valter Crescenzi , Giansalvatore Mecca , Paolo Merialdo , Paolo Missier and Università Roma Tre “An Automatic Data Grabber for Large Web Sites”, 2004.
[13] Pallav Roxy, and Durga Toshniwal, “Clustering Unstructured Text Documents Using Fading Function”, International Journal of Information and Mathematical Sciences, Vol 5, NO. 3 2009.
[14] Shady Shehata, Fakhri Karray and Mohamed S. Kamel, "An Efficient Model For Enhancing Text Categorization Using Sentence Semantics", International Journal of Computational Intelligence, 2010.
[15] Jun Zhai, Yan Chen, Qinglian Wang and Miao Lv “Fuzzy ontology models using intuitionistic fuzzy set for knowledge sharing on the semantic web”, 12th International Conference on Computer Supported Cooperative Work in Design, 2008.
[16] A. Hinneburg and D.A. Keim. Optimal gridclustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In Proc. of VLDB-1999, Edinburgh, Scotland, September 2000. Morgan Kaufmann, 1999.
[17] H. Schuetze and C. Silverstein. Projections for efficient document clustering. In Proc. of SIGIR-1997, Philadelphia, PA, July 1997, pages 74–81. Morgan Kaufmann, 1997.
[18] Liping Jing,” Survey of Text Clustering”, Department of Mathematics, The University of Hong Kong, HongKong, China, , ISBN: 7695-1754-4/02.
[19] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier and L. Lakhan, “Computing iceberg concept lattice with Titanic”, Journal on Knowledge and Data Engineering, Vol. 42, No. 2, 2002, pp. 189-222.
[20] S. Pollandt, Fuzzy-Begriffe: Formale Begriffsanalyse unscharfer Daten, Springer Verlag, Berlin- Heidelberg, 1996.
[21] S. Vaithyanathan, B. Dom, “Model-based hierarchical clustering”, Proc. 16th Conf. Uncertainty in Artificial Intelligence, 2000, pp. 599–608.
[22] Eman Abdu, and Douglas Salane, ―A spectral-based clustering algorithm for categorical data using data summaries, International Conference on Knowledge Discovery and Data Mining, ACM, Article no. 2, 2009.
[23] I. S. Dhillon, D. S. Modha, “Concept decompositions for large sparse text data using clustering”, Machine Learning, vol. 42, 2001, pp. 143-175.
[24] Zamir, O.Etzioni, “Web Document Clustering: A Feasibility Demonstration,” in Proceedings of the 21st International ACM SIGIR Conference on Research and Development.
[25] S. Pierre and Blondel V., Automatic Discovery of Similar Words , Survey of Text Mining: Clustering, Classification, and Retrieval Springer (Ed.) (2008).
[26] M. Steinbach, G. Karypis, V. Kumar, “A comparison of document clustering techniques”, KDD Workshop on Text Mining, 2000, pp. 109-110.