ISSN No. 0976-5697

Volume 2, No. 1, Jan-Feb 2011

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Survey on Text Mining in Clustering

S.Logeswari*
Department of CSE
Bannari Amman Institute of Technology
Tamil Nadu, India
slogesh@rediffmail.com

K.Premalatha
Department of CSE
Bannari Amman Institute of Technology
Tamil Nadu, India
kpl_barath@yahoo.co.in

D.Sasikala
Department of CSE
Bannari Amman Institute of Technology
Tamil Nadu, India
Sasiramesh04@gmail.com

Abstract: Text mining has important applications in the area of data mining and information retrieval. One of the important tasks in text mining is document clustering. Many existing document clustering techniques use the bag-of-words model to represent the content of a document. It is only effective for grouping related documents when these documents share a large proportion of lexically equivalent terms. The synonymy between related documents is ignored. It reduces the effectiveness of applications using a standard full-text document representation. This paper emphasis on the various techniques that are used to cluster the text documents based on keywords, phrases and concepts. It also includes the different performance measures that are used to evaluate the quality of clusters.

Keywords: Document Clustering, Latent Semantic Indexing, Vector Space Model, tf-idf, precision, recall, F-measure.

I. INTRODUCTION

Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. Clustering analyzes data objects without consulting a known class label. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity). Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

Text data is ubiquitous in nature (Huang et al, 2006) because text data are inherently unstructured and fuzzy. As the volume of text data gets increased, the management and analysis of it becomes unprecedentedly important. Text mining is an emerging technology for handling the increasing text data. The phrase text mining is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful information.

II. DOCUMENT CLUSTERING

Document clustering is one of the essential functions in text mining. Document clustering is to divide a collection of text documents into different categories so that the documents within the category express the same issue. Text mining becomes more challenging because of its characteristics like volume, dimensionality, sparsity and complex semantics involved in it. These characteristics require clustering techniques to be scalable to large and high dimensional data, and able to handle sparsity and semantics. Typical text clustering activity involves with the document representation, document similarity measure and clustering techniques.

A. *Document Preprocessing*

It is necessary to convert the document collection into the form of document vectors. First to determine the terms used to describe the documents, the following procedure was used in earlier experiments. The general references about preprocessing included in Bote et al (2002), Hammouda and Kamel (2004), Karoui et al (2006) , Huang (2008) , Muflikhah and Baharudin (2009) and Premalatha and Natarajan (2009).

- Extraction of all the words from each document.
- Elimination of the stopwords from a stopword list generated with the frequency dictionary of Kucera.
- Stemming the remaining words using the Porter stemmer, which is the most commonly used stemmer for English (Porter, 1980).
- Formalizing the document as a dot in the multidimensional space and represented by a vector d, such as d = {w_1, w_2, , w_n}, where w_i (i = 1, 2, . n) is the term weight of the term t_i in one document. The most widely used weighting scheme combines the term frequency with inverse document frequency (TF–IDF). The weight of term i in document j is given by:

$$w_{ji} = tf_{ji} \times idf_{ji} = tf_{ji} \times \log_2 \left(n / df_{ji} \right)$$

where tf_{ji} is the number of occurrences of term i in document j; df_{ji} indicates the term frequency in the collection of documents, and n is the total number of documents in the collection.

B. *Latent Semantic Indexing*

Latent Semantic Indexing (LSI) method is one of the methods used in association with domain semantics to extend the vector space model to overcome some of the retrieval problems such as the dependence problem or the vocabulary problem. In LSI the associations among terms and documents are calculated and exploited in the retrieval process. An advantage of this approach is that queries can retrieve documents even if they have no words in common.

Foltz and Dumais (1992) compared the performance of LSI and keyword vector matching for filtering of Netnews articles. They found that the LSI filtering improved prediction performance over the keyword matching methods.

C. *Similarity and Performance Measures*

Perfect clustering requires a precise significance of the closeness between a pair of objects, in terms of either the pairwised similarity or distance. A variety of similarity or distance measures have been proposed (Huang, 2008) and widely applied, such as cosine similarity and the jaccard correlation coefficient. Measures such as euclidean distance and relative entropy have been applied in clustering to calculate the pair-wise distances.

Table I. Similarity Measures

Measures	***Function***
Euclidean Distance	d(x,y)= $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Manhattan Distance	d(x,y) = $\sum_{i=1}^{n}\lvert x_i - y_i \rvert$
Minkowski Distance	d(x,y)= $\left(\sum_{i=1}^{n}\lvert x_i - y_i \rvert^p\right)^{1/p}$
Cosine similarity	Cos (ɵ)= $\frac{a.b}{\lVert a\rVert\times\lVert b\rVert}$
Jaccard Coeffieient	d(a,b)= $\frac{a.b}{\lVert a\rVert^2+\lVert b\rVert^2-a.b}$

The Information Retrieval community uses a variety of performance measures to evaluate the effectiveness of scoring functions. There are three types of quality measures: external, when there is a priori knowledge about the clusters, internal, which assumes no knowledge about the clusters, and relative, which evaluates the differences between different cluster solutions. The external measures are applied to both categorization and clustering, while internal and relative measures are applied only to clustering. The six popular external measures — precision, recall, F measure, average precision, reciprocal rank, and normalized discounted cumulative gain are discussed by McSherry and Najork (2008).

Table II. Performance Measures

Measures	***Functions***
Recall	$\frac{\lvert\{relevantdocuments\}\cap\{retrieveddocuments\}\rvert}{\lvert\{relevantdocuments\}\rvert}$
Precision	$\frac{\lvert\{relevantdocuments\}\cap\{retrieveddocuments\}\rvert}{\lvert\{retrieveddocuments\}\rvert}$
F measure	F= $2\cdot precision\cdot recall/(precision+recall)$

Zhao et al (2004) introduced two measures to evaluate the overall quality of clustering solutions purity and entropy. Entropy is a commonly used measure in information theory. Originally it is used to characterize the (im)purity of an arbitrary collection of examples.

D. *Representation of Textual Documents*

It is essential to translate the text data into an efficient and meaningful way so that they can be analyzed. In information retrieval and text mining, text data of different formats is represented in a common representation model. The vector space and probabilistic models are the two major examples of the statistical retrieval approach. To determine the relevance of documents with respect to a query, both models use statistical information in the form of term frequencies. Zhu et al (2007) proposed a probabilistic model called Field Independent Clustering Model (FICM). FICM takes advantage of both the discrimination ability of the field and the power of selecting the best model for each field. Along with these two models, Boolean models, Suffix Tree Document Clustering models (Eiben et al 2005), Document Index Graph (DIG) models (Hammouda and Kamel 2004) and Text Streaming models are also used to represent textual documents.

In Information retrieval and text mining, the vector space model represents the documents and queries as vectors in a multidimensional space. The terms in the documents are used as dimensions to build an index to represent the documents. The formation of an index involves lexical scanning to identify the significant terms, where morphological analysis reduces different word forms to common stems, and the occurrence of those stems is computed. Query and document surrogates are compared by comparing their vectors, using, for example, the cosine similarity measure. The vector space model can assign a high ranking score to a document that contains only a few of the query terms if these terms occur infrequently in the collection but frequently in the document.

III.SURVEY OF RELATED WORKS

Shen et al (2004) introduced a domain-specific concept based information retrieval system for the ease of retrieving organizational internal documents, i.e., technical reports, professional references and even emails. In order to improve the quality of the final search result, the domain knowledge is used to resolve ambiguity resulted from keyword comparison. By comparing the precision and recall, it is proved that the accuracy of concept-based search is better than that of keyword-based search.

Cui and Potok (2005) proposed a hybrid document clustering algorithm based on Particle Swarm Intelligence and K-means algorithms. The main drawback of K-means algorithm is that it is sensitive to the selection of the initial partition and may converge to a local optima. Hybrid Particle Swarm Optimization (PSO)+K-means document clustering algorithm that performs fast document clustering and can avoid being trapped in a local optimal solution as well.

Gao and Wang (2005) presented an efficient text clustering algorithm based on Concept Indexing (CI). It is used to reduce the dimensionality of text feature matrix in the Rival Penalized Competitive Learning (RPCL) text clustering. RPCL is a kind of competitive neural network. The number of clusters is determined more effectively by using this method and also it improves the efficiency of clustering.

Myat and Hla (2005) proposed a Formal Concept Analysis (FCA) method for clustering documents according to their formal contexts. At the corpus level, the Concept hierarchy is built using the formal concepts of the documents. The term weighting model is used to reduce less useful concepts from these formal concepts and the association and correlation mining techniques are used to analyze the relationship between the terms in the document corpus.

Csorba and Vajk (2006) proposed a topic based document clustering technique using confidence as a measure for rejecting the classification of documents with ambiguous topic. This is achieved by applying a confidence measurement for every classification result and by discarding documents with a confidence value less than a predefined lower limit. This measure returns the classification for a document only if it feels sure about it otherwise the document is marked as unsure. Beside this ability the confidence measurement allows the use of a much stronger term filtering, performed by a novel, supervised term cluster creation and term filtering algorithm, which is presented in this paper as well.

Karoui et al (2006) developed an integrated framework for extracting the concepts from web pages based on contextual clustering technique. This method takes advantage from structural HTML document features and the word location to identify the appropriate term context. This method improves word weighting, the selection of the semantically closer and the relevant extracted ontological concepts.

Malik and Kender (2006) proposed a sub-linearly, scalable, hierarchical document clustering method that performs better than the existing agglomerative, partitioning and frequent-itemset based methods both in terms of clustering quality and runtime performance.

Hamzah et al (2007) proposed a method to transform term-document matrix into concept-document matrix in order to reduce the dimension which will be significantly improve the performance of concept-based clustering. A concept-document matrix was constructed in this method by utilizing cluster centre. Results show that by using concept-based clustering the performance of clustering can significantly be improved compare to word-based clustering.

Hoeglinger and Pears (2007) presented a concept-based mining method for data stream which is integrated with a high-speed decision tree learner. This approach uses the content of the data stream itself in order to decide which information is to be erased. Several methodologies, all based around minimizing the overall information loss when pruning the decision tree, are reviewed. Hoeffding trees are used to overcome the storage limitations of classic decision tree.

Looks et al (2007) analyzed the general problems of concept mining, discovering useful associations, relationships, and groupings in large collections of data. Statistical transformation algorithms are used to reduce the content of multilingual, unstructured data into a vector that describes the content. Concept based mining is introduced to improve the performance and scalability of clustering algorithms. A streaming hierarchical partitioning method is developed for extracting semantic content from voluminous data streams.

Yuan and Yuan (2007) presented a concept based mining method based on Core words of a class. The existing popular text categorization methods mainly focus on keywords, which can not deal with synonyms and polysemes scenario properly. In this method, the core words are identified and extracted. How-net maps are used to map keyword space to concept space based on these core words, and finally complete the text categorization process in the concept space. Both Naïve Bayes and k-Nearest Neighbor text categorization methods are used to evaluate the performance of this method. The results indicate that the new core words oriented concept mapping can effectively improve text categorization precision compared with other keywords oriented text categorization methods.

Amine et al (2008) presented an unsupervised classification of texts based on Kohonen Self-Organizing Maps (SOM), which gather a certain number of similar objects without prior information. The majority of classification approaches use supervised learning methods are not feasible for significant flows of data. Unsupervised classification methods are used to discover latent (hidden) classes automatically while minimizing human intervention. The performance of the proposed method is verified in two approaches, one based on the use of WordNet and the other on the use of n-grams. The results obtained show that in spite of the good results obtained by the n-grams method, adding lexical knowledge in the representation makes it possible to build a better classification.

Barresi et al (2008) proposed an indexing technique which goes beyond the syntax of terms; trying to capture their explicit meaning from their context and to derive a set of concepts that are used to represent the documents. This approach overcomes some of the major drawbacks deriving from the use of bag of words and term frequency based indexing techniques. The performance of this method is measured by using unsupervised performance measures called cluster internal cohesion and external isolation. WordNet's morphological processing is used instead of traditional stemming algorithms. The experimental results showed that the proposed technique leads to more cohesive and separated clusters and allows for a significant reduction of the vector dimension.

Cao et al (2008) introduced a document clustering method which is based on named entities as objectives into fuzzy document clustering. The named entities are used to define document semantics. Traditional keyword-based document clustering techniques have limitations due to simple treatment of words and hard separation of clusters. In this proposed model, the traditional keyword-based vector space model is adapted with vectors defined over spaces of entity names, types, name-type pairs, and identifiers, instead of keywords. Then, hierarchical fuzzy document clustering can be performed using a similarity measure of the vectors representing documents.

Chim and Deng (2008) presented a phrase-based document similarity model. Suffix Tree Document (STD) model is used to represent the document data. The STD model preserves all sequential characteristics of the sentences in each document. By mapping all nodes or (or phrases) in the suffix tree into a m-dimensional term space of the Vector Space Document model, the new phrase-based document similarity successfully connects the two document models and inherits their advantages. The significant improvement of the clustering quality indicates that the word order preservation is critical document clustering. The group-average hierarchical clustering algorithm with the phrase-based document similarity is a highly accurate and efficient practical document clustering solution.

Yang et al (2008) proposed a fuzzy concept mining model based on FCA. FCA is used to extract concepts from the text documents and to build a conceptual hierarchy from the given data. Fuzzy FCA incorporates fuzzy set theory into FCA in order to analyze vague data set of uncertainty information.

Luo et al (2009) presented a technique for document clustering based on neighbors. Usually, the cosine function is used to measure the similarity between two documents in the criterion function, but it may not work well when the clusters are not well separated. In order to solve this problem, the information about neighbors and link for the family of k-means algorithms are used. A new similarity measure is used which is a combination of the cosine and link functions. A new heuristic function is used for selecting a cluster to split based on the neighbors of the cluster centroids. The experimental results on real-life data sets demonstrated that the proposed methods can significantly improve the performance of document clustering in terms of accuracy without increasing the execution time much.

Mahadavi and Abolhassani (2009) proposed a Harmony *K*-means Algorithm (HKA) that deals with document clustering based on Harmony Search (HS) optimization method. In this method the problem of finding a globally optimal partition of a given set of documents into a specified number of clusters is considered and it is proved by means of finite Markov chain theory that the HKA converges to the global optimum. Experimental results reveal that the HKA algorithm converges to the best known optimum faster than other methods and the quality of clusters are comparable.

Muflikhah and Baharudin (2009) proposed a concept space approach for document clustering which is the integration of both information retrieval method and document clustering. The method is known as Latent Semantic Index (LSI) approach which used Singular Vector Decomposition (SVD) or Principle Component Analysis (PCA). It is proposed with the aim to reduce the matrix dimension by finding the pattern in document collection with refers to concurrent of the terms. Each method is implemented to weight of term-document in VSM for document clustering using fuzzy c-means algorithm.

Premalatha and Natarajan (2009) presented a procreant PSO algorithm for document clustering. It combines PSO with GA, a population-based heuristic search technique, which can be used to solve combinatorial optimization problems, modeled on the concepts of cultural and social rules derived from the analysis of the swarm intelligence and also based on crossover and evolution. In order to avoid the premature convergence, they proposed a modification in simple PSO and it is applied to the document at the corpus level. It includes reproduction using crossover when stagnation in the movement of the particle is identified and carries out local search to improve the goodness of fit. Reproduction provides faster convergence and better solution.

Sui and Liu (2009) presented a method of automatic construction of Ontology hierarchies based on FCA. Thus the feature extraction can be turned into the extraction of attribute values. The distance between two concepts is computed based on both the weight of the attributes and the weight of the attributes values. F-measure is used for cluster formation.

Verma et al (2010) proposed a document clustering algorithm using squared distance optimization through Genetic Algorithm (GA). Genetic algorithm is often used for document clustering because of its global search and optimization ability over heuristic problems. In this algorithm, search ability of genetic algorithm has exploited with a modification from the general genetic algorithm by not using the random initial population. They found that algorithm is working better than k-means algorithm as it has less probability to be trapped in local optimal solutions. Genetic algorithm has substantially decreased calculation complexity and it also increases the effectiveness of result as initial population was not random. Furthermore less number of iterations is required during execution to converge to a global optimal solution. With the increase in the number of clusters the performance of algorithm gets improved.

Kiran et al (2010) presented a hierarchical clustering algorithm using closed frequent itemsets that use Wikipedia as an external knowledge to enhance the document representation. High dimensionality is a major challenge in document clustering. Some of the recent algorithms address this problem by using frequent itemsets for clustering. But, most of these algorithms neglect the semantic relationship between the words. On the other hand there are algorithms that take care of the semantic relations between the words by making use of external knowledge contained in WordNet, Mesh, Wikipedia, etc but do not handle the high dimensionality. But this algorithm addresses both these problems. F-Score is used to evaluate the performance of this method and the experiments showed that this method is better than existing approaches.

Ponmuthuramalingam and Devi (2010) proposed a frequent term based approach of clustering. It provides a natural way of reducing a large dimensionality of the document vector space. This approach is based on clustering the low dimensionality frequent term sets. Four algorithms are proposed for this approach based on the attributes namely minimum support threshold and matching threshold. The results are compared with the standard Frequent Term base Clustering (FTC) to show their competency. It is proved that the proposed algorithms have higher F-measure value and better cluster quality than FTC algorithm.

Ranjan et al (2010) presented an algorithm for document clustering based on internal criterion function. Most commonly used partitioning clustering algorithms (e.g. k-means) have some drawbacks as they suffer from local optimum solutions and creation of empty clusters as a clustering solution. The proposed algorithm usually does not suffer from these problems and converge to a global optimum, its performance enhances with the increase in number of clusters. The performance of the algorithm enhances with the increase in the number of clusters.

Shehata et al (2010) developed a concept based mining model for text mining. This work bridges the gap between natural language processing and text mining disciplines. The new proposed model composed of four components, is to improve the text clustering quality. By exploiting the semantic structure of the sentences in documents, a better text clustering result is achieved. The Primary component of the proposed method is the sentence-based concept analysis, document-based concept analysis, corpus level analysis and the concept-based similarity measure. The quality of text clustering achieved by this model significantly surpasses the traditional single term-based approaches.

IV.SUMMARY

Clustering is the operation of grouping documents automatically on the basis of some similarity measures, without specifying categories. Document clustering has many important applications in the area of data mining and information retrieval. This survey provides the document preprocessing, document representations, similarity measures and performance measures. It also includes the survey on the various clustering techniques based on terms, phrases and concepts.

V. REFERENCES

[1] A. Amine, Z. Elberrichi, L. Bellatreche, M. Simonet and M. Malki, "Concept-Based Clustering of Textual Documents Using SOM," Proceedings of Sixth IEEE International Conference on Computer System and Applications, pp. 156-163, 2008.

[2] S. Barresi, S. Nefti and Y. Rezgui, "A Concept Based Indexing Approach for Document Clustering," Proceedings of IEEE International Conference on Semantic Computing, pp. 26-33, 2008.

[3] T. H. Cao, H. T. Do, Hong and T. T. Quan, "Fuzzy Named Entity-Based Document Clustering," Proceedings of IEEE International Conference on Fuzzy Systems, pp.2028-2034, 2008.

[4] H. Chim and X. Deng, "Efficient Pharse-Based Document Similarity for Clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 9, pp. 1217-1229, 2008.

[5] Csorba and Vajk, "Term Clustering and Confidence Measurement in Document Clustering," Proceedings of IEEE International Conference on Computational Cybernetics, pp. 1-6, 2006.

[6] S. M. Eiben, B. Stein and M. Potthast , " The Suffix Tree Document Model Revisited," Proceedings of the 5th international conference on knowledge management (I-KNOW 05), Journal of Universal Computer Science, pp.596-603, 2005.

[7] P. W. Foltz and S. T. Dumais, "Personalized information delivery: An analysis of information filtering methods," Communications of the IJCAI, vol.35, no.12, pp.51-60, 1992.

[8] M. Gao and Z. Wang, "RPCL Text Clustering on Concept Indexing," Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, pp. 18-21, 2005.

[9] V. Gupta and G. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1, pp. 60-76, 2009.

[10] K. M. Hammouda and M. S. Kamel, "Efficient Phrase-Based Document Indexing for Web Document Clustering," IEEE Transactions on knowledge and data engineering, vol.16, no.10, pp.1279-1296, 2004.

[11] S. Hoeglinger and R. Pears, "Use of Hoeffding Trees in Concept Based Data Stream Mining," Proceedings of the Third International Conference on Information and Automation for Sustainability, pp. 57-62, 2007.

[12] A. Huang, "Similarity Measures for Text Document Clustering," Proceedings of the New Zealand Computer Science Research Student Conference, pp.49-56, 2008.

[13] J. Z. Huang, Michael and L. Jing, "Text Clustering: Algorithms, Semantics and Systems," PAKDD06 Tutorial., 2006.

[14] I. Kao, C. Tsai and Y. Wang, "An Effective Particle Swarm Optimization Method for Data Clustering," Proceedings of the 2007 IEEE International Conference on Engineering and Engineering Management, pp. 548-552, 2007.

[15] L. Karoui, M. Aufaure and N. Bennacer, "A New Extraction Concept Based on Contextual Clustering," Proceedings of IEEE International Conference on Intelligent Agents, Web Technologies and Internet Commerce, pp. 91-96, 2006.

[16] G. Kiran, R. Shankar and V. Pudi, "Frequent Itemset based Hierarchical Document Clustering using Wikipedia as External Knowledge," Proceedings of International Conference on Knowledge-Based and Intelligent Information Engineering Systems, pp. 1-11, 2010.

[17] S. Loh, L. Wives and M. J. Palazzo, "Concept-Based Knowledge Discovery in Text Extracted from the Web," Journal of ACM SIGKDD, vol. 2, no. 1, pp. 29-39, 2000.

[18] M. Looks, A. Levine, A. Covington, R. Loui, J. Lockwood and Y. Cho, "Streaming Hierarchical Clustering for Concept Mining," Proceedings of IEEE Aerospace Conference, pp. 1-12, 2007.

[19] C. Luo , Y. Li and S. Chung, " Text document clustering based on neighbors, " Proceedings of Data & Knowledge Engineering , vol.68 , pp. 1271–1288, 2009.

[20] M. Mahdavi and H. Abolhassani, " Harmony *K*-means algorithm for document clustering," Proceedings of Data Mining and Knowledge Discovery, Springer, vol.18 , pp. 370–391, 2009.

[21] H. Malik and J. R. Kender, "High Quality, Efficient Hierarchical Document Clustering Using Closed Interesting Itemsets, "Proceedings of Sixth International Conference on Data Mining Issues, pp. 991-996, 2006.

[22] F. McSherry and M. Najork, "Computing Information retrieval Performance Measures Efficiently in the Presence of Tied Scores, " ECIR 2008, LNCS 4956, pp. 414-421, 2008.

[23] L. Muflikhah and B. Baharudin, " Document Clustering using Concept Space and Cosine Similarity Measurement," Proceedings of International Conference on Computer Technology and Development, pp. 58-62, 2009.

[24] N. Myat and K. Hla, "A Combined Approach of Formal Concept Analysis and Text Mining for Concept Based Document Clustering," Proceedings of IEEE International Conference on Web Intelligence, pp. 1-4, 2005.

[25] K. Premalatha and A. M. Natarajan A M, "Procreant PSO for fastening the convergence to optimal solution in the application of document clustering," Current Science, vol. 96, no. 1, pp. 137-143, 2009.

[26] P. Ponmuthuramalingam and T. Devi, "Effective Term Based Text Clustering Algorithms," International Journal on Computer Science and Engineering, vol. 2, no. 5, pp. 1665-1673, 2010.

[27] M. F. Porter, "An algorithm for suffix stripping," Program, pp. 130–137, 1980.

[28] A. Ranjan , E. Kandpal, H. Verma and J. Dhar, " An Analytical Approach to Document Clustering Based on Internal Criterion Function **,"** International Journal of Computer Science and Information Security, vol. 7, no. 2, pp. 257-261, 2010.

[29]Rowena Chau and Chung-Hsing Yeh, "Fuzzy Conceptual Indexing for Concept-Based Cross-Lingual Text Retrieval," IEEE Transactions on Internet Computing, vol. 8, no. 5, pp. 14-21, 2004.

[30] S. Shehata , F. Karray and M. Kamel, "Enhancing Text Clustering using Concept-based Mining Model," Proceedings of the IEEE Sixth International Conference on Data Mining, pp. 1043-1048, 2006.

[31] S. Shehata, Fakhri and M. S. Kamel, "An Efficient Concept Based Mining Model for Enhancing Text Clustering," IEEE Transactions on Knowledge and Data Engineering, vol.22, no.10, pp. 1360-1371, 2010.

[32] L. Shen, K. Lim and H. Lob, "Domain-specific Concept-based Information Retrieval System," Proceedings of IEEE International Conference on Engineering Management , vol. 2, pp. 525-529, 2004.

[33] Z. Sui and Y. Liu, “Inducting Concept Hierarchies from Text based on FCA,” Fourth International Conference on Innovative Computing, Information and Control, pp. 1080-1083, 2009.

[34] H. Turtle and W. B. Croft, “Evaluation of an inference network-based retrieval model**,”** ACM Transactions on Information Systems – Special issue on research and development in information retrieval, vol.9, no.3, pp. 187-222, 1991.

[35] H. Verma , E. Kandpal, B. Pandey and J. Dhar , ” A Novel Document Clustering Algorithm Using Squared Distance Optimization Through Genetic Algorithms, “ International Journal on Computer Science and Engineering , vol. 2, no. 5, pp. 1875-1879, 2010.

[36] M. Wang, X. Wang and C. Xu C, "An Approach to Concept Oriented Text Summarization,” Proceedings of IEEE international Symposium on Communications and Information Technology, vol. 2, pp. 1290-1293, 2005.

[37]R.Xu and D. Wunsch, “Survey of Clustering Algorithms,” IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, 2005.

[38] K. Yang , E. Kim, S. Hwang and S. Choi, “Fuzzy Concept Mining based on Formal Concept Analysis,” International Journal of Computers, vol. 2, no. 3, pp. 279-290, 2008.

[39] F. Yuan and J. Yuan, “A Concept Mapping Method Based On Core Words Of Class,” Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, pp. 19-22, 2007.

[40] J. Zakos and B. Verma, “Concept-based Term Weighting for Web Information Retrieval,**”** Proceedings of the IEEE Sixth International Conference on Computational Intelligence and Multimedia Applications, pp. 173-178, 2005.

[41] Y.Zhao and G. Karypis, “Empirical and theoretical comparisons of selected criterion functions for document clustering,” Machine Learning, vol.55, no.3, pp.311-331, 2004.

[42] S. Zhu, I. Takigawa , S. Zhang and H. Mamitsuka,, “ A Probabilistic Model for Clustering Text Documents with Multiple Fields,” ECIR 2007, Springer-Verlag , LNCS 4425, pp.331-342, 2007.