Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Extension Techniques for Text Document Clustering

P. Chiranjeevi*, T. Supraja, P. Srinivasa Rao
Asst. Professor, ACE Engineering College
Hyderabad, India

Abstract: In today's era of World Wide Web, the on-line information is increasing exponentially day to day. Text Document clustering is the one of the fastest growing research technique for organizing documents in an unsupervised manner. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups .Various document clustering techniques clusters the documents with high intra-similarity and low inter-similarity space. Many clustering techniques are not effectively clustering the documents locally but globally the optimal solution can be obtained high-speed and high quality clustering algorithms. In this paper, a brief survey on optimization method to text document clustering.

Keywords: Document clustering, correlation measure, correlation latent semantic indexing, dimensionality reduction Text Clustering, Vector Space Model, Genetic Algorithm, Particle Swarm Optimization

I. INTRODUCTION

Text Document clustering is the application of cluster analysis to textual documents. It has applications in automatic document organization, topic extraction and fast information retrieval or filtering. The use of descriptors and descriptor extraction involves in document clustering.Descriptors are defined as sets of words that explain the contents within the cluster. Document clustering is considered as a centralized process. Examples of document clustering involves web document clustering for search users. A number of methods have been proposed to handle document clustering based on various distance measures.

Euclidean distance is a widely used distance measure. The *k*-means is partition clustering method that uses the Euclidean distance method to minimize the sum of the squared Euclidean. The document space is of high dimensionality, so it is required to find a low-dimensional representation of the documents to reduce its computation complexity.

With web search engines, a user can navigate browse and locate the documents. Naturally search engines give many documents, particularly a lot of information which we are relevant to the topic and some may contain irrelevant documents with poor quality. Cluster analysis or clustering plays an important role in converting such large amount of documents to get back by search engines into meaningful clusters. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to objects in other clusters. Document clustering is similar to data clustering. Cluster analysis has its roots in many data mining research areas, including data mining, information retrieval, pattern recognition, web search, statistics, biology and machine learning. Even if a lot of significant research effort has been done in [3,15,20,22,26,33], the more challenges in clustering is to improve the quality of clustering process.C**lustering**[3] is an example of **unsupervised learning**. In the context of machine learning, clustering contracts with **supervised learning.** Classification assigns data objects in a collection to aim categories or classes. The important job of classification is to concise predict the target class for each instance in the data.

So classification algorithm requires training data. Unlike classification, clustering does not require training data. Clustering does not assign any pre-defined label to each and every group. Clustering groups a set of objects and finds whether there is *some* relationship between the objects

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods
- Frequent pattern-based clustering
- Constraint-based clustering

II. DOCUMENTCLUSTERING

Document clustering is one of the most critical techniques for organizing documents in unsupervised manner. Clustering of documents is used to group documents into heterogeneous topics. The major complexity in document clustering is its high dimension. It requires efficient algorithms which can solve this high dimensional clustering. A document clustering is a major topic in information retrieval area Example includes search engines. The basic steps used in document clustering process are shown in figure 1.

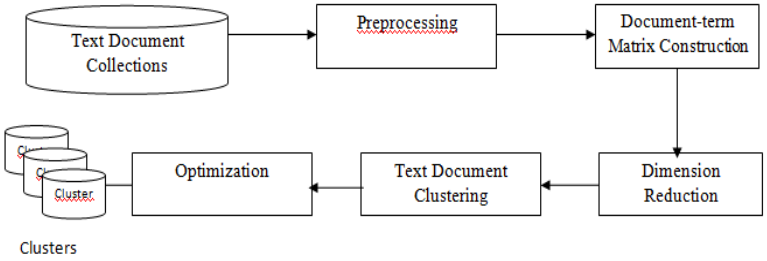

Figure 1.Flow diagram for representing basic Steps in text clustering

2.1 Preprocessing

The text document preprocessing basically consists of a process to strip all formatting from the article, including capitalization, punctuation, and extraneous markup. Then the stop words are removed. Stop words term (i.e., pronouns, prepositions, conjunctions etc) are the words that don't carry semantic meaning. Stop words can be eliminated using a list of stop words. Stop words elimination using a list of stop word list will greatly reduce the amount of noise in text collection, as well as make the computation easier. The benefit of removing stop words leaves us with condensed version of the documents containing content words only.

The next process is to stem a word. Stemming is the process for reducing some derived words into their root form. For English documents, a popularly known algorithm called the Porter stemmer [7] is used. The performance of text clustering can be improved by using Porter stemmer.

2.2 Text Document Encoding

The next process is to encode the text document. In general, documents are transformed into document term matrix (DTM) which is a mathematical matrix whose dimensions are the terms and rows are documents. A simple DTM is the vector space model [1] model which is widely used in IR and text mining [23], to represent the text documents. It is used in indexing, information retrieval and relevancy rankings and can be successfully used in evaluation of search results from web search engines.

Let **D** = (D1, D2, . . . ,DN) be a collection of documents and **T**=(T1, T2, . . . , TM) be the collection of terms of the document collection **D**, where N is the total number of documents in the collection and M is the number of distinct terms. In this model each document Di is represented by a point in an m dimensional vector space, Di = (wi1, wi2, . . . ,wim), i = 1, . . . , N, where the dimension is the total number of unique terms in the document collection. Many schemes have been proposed for measuring wij values, also known as (term) weights. One of the more advanced term weighting schemes is the tf-idf (term frequency-inverse document frequency) [23]. The tf-idf scheme aims at balancing the local and the global weighting of the term in the document and it is calculated by

$$w_{ij} = tf_{ij} * \log\left(\frac{n}{df_j}\right) \quad (1)$$

Wheretf_{ij} is the frequency of term i in document j, and df_{ij} denotes the number of documents in which term j appears. The component log (n/df_{ij}), which is often called the idf factor, defines the global weight of the term j.

2.3 Dimension reduction techniques

Dimension reduction can be divided into feature selection and feature extraction. Feature selection is the process of selecting smaller subsets from larger set of inputs and Feature extraction transforms the high dimensional data space to a space of low dimension. The goal of dimension reduction methods is to allow smaller amountofdimensions for broader comparisons of the concepts contained in a text collection. In this paper, one dimension reduction technique is discussed.

2.3.1 Latent Semantic Indexing (LSI)

A well known text retrieval technique is Latent Semantic Indexing, which uses Singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in a collection of text. The singular value decomposition reduces the dimensions by selecting dimensions with highest singular values. For text processing LSI can be effectively used because it conserves the polysemy and synonymy in the text. A key feature of LSI is its ability to retain latent structure in word and hence it improves the clustering efficiency. Once a term-document matrix X (M×N) is constructed, assuming there are m distinct terms and n documents. The Singular Value Decomposition computes the term and document vectors by transforming TDM matrix X into three matrices P, S and Q, which is given by

$$X=PSQT \quad (2)$$

where

P : left singular vector matrix

Q : right singular vector matrix

S : diagonal matrix of singular values.

LSI approximates X with a rank *k* matrix.

$$Xk=PkSkQTk \quad (3)$$

Where*Pk*is defined to be the first *k* columns of the matrix P and QT*k*is included the first *k* rows of matrix QT.

S*k*=diag(s1,…..sk) is the first *k* largest singular values.

When LSI is used for text document clustering, a document Di is represented by [11]

$$Di=DTiPk \quad (4)$$

Then the text corpus can be organized by another representation of document-term matrix D(N×M) and the corpus matrix is organized by

$$C=DPk \quad (5)$$

2.4 Similarity Measurement

The similarity (or dissimilarity) between the objects is typically computed based on the distance between document pairs. The most popular distance measure as stated by [3] are:

Table 1. Formulas for Similarity Measurement

Name	Formula
Euclidean Distance	$\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+...(x_{in}-x_{jn})^2}$ Where i=$(x_{i1}, x_{j2}...x_{in} and)$ j=$(x_{j1}, x_{j2}...x_{jn})$
Manhattan Distance	$\|(x_{i1}-x_{j1})\|+\|x_{i2}-x_{j2})\|+...=\|((x_{in}-x_{jn})\|$
Minkowski Distance	$(\|(x_{i1}-x_{j1})^p\|+\|(x_{i2}-x_{j2})^p\|+\ldots+\|(x_{in}-x_{jn})^p\|)^{1/p}$ Where p is a positive integer
Jaccard Coefficient	J(A,B)=(A∩B)/(AUB) Where A and B are Documents
Cosine Similarity	$S(x,y)=\frac{x^t.y}{\|\| x \|\|\| y \|\|}$ Where x^t is a transposition of vector x , ‖x‖ is the Euclidean norm of vector x, ‖y‖ is the Euclidean norm of vector y,

2.5 Evaluation of Text Clustering

The quality of text clustering can be evaluated by using the popularly used external indexes [22]: F-measure, Purity and Entropy. These measures are called external quality measures because the results of clustering techniques are compared with known classes (i.e) it requires that the documents be given class labels in advance. A widely used quality measures for the purpose of text document clustering [22] are F-measure, Purity and Entropy which can be defined as follows:

2.5.1. F-measure

The F-measure combines the precision and recall values used in information retrieval. The *precision P(i,j)* and *recall R(i,j)* of each cluster j for each class i are calculated as

$$\mathbf{P(i,j)} = \frac{\beta_{ij}}{\beta_j} \quad (6)$$

$$\mathbf{P(i,j)} = \frac{\beta_{ij}}{\beta_i} \quad (7)$$

where

βi : is the number of members of class i

βj : is the number of members of cluster j

βij: is the number of members of class i in cluster j

The corresponding *F-measure F(i,j)* is given by the following equation:

$$\mathbf{F(i,j)} = \frac{2*P(i,j)*R(i,j)}{P(i,j)+R(i,j)} \quad (8)$$

Then the F-measure of a class i can be defined as

$$\mathbf{F}= \sum_i \frac{\beta_i}{n} \max_j (F(i,j)) \quad (9)$$

where n is the total number of documents in the collection. In general, the larger the F-measure gives the better clustering result.

2.5.2. Purity

The purity measure of a cluster represents the percentage of correctly clustered documents and thus the purity of a cluster j is defined as

$$\textbf{Purity(j)}= \frac{1}{\beta_j} \max_i (\beta_{ij}) \quad (10)$$

The overall purity of a clustering is a weighted sum of the cluster purities and is defined as

$$\textbf{Purity} = \sum_n \frac{\beta_j}{n} Purity(j) \quad (11)$$

In general, the better clustering result is given by the larger the purity value.

2.5.3. Entropy

The Entropy of a cluster can be defined as the degree to which each cluster consists of objects of a single class. The entropy of a cluster j is calculated using the standard formula,

$$e_j = -\sum_{i=1}^{L} p_{ij} \log p_{ij} \quad (12)$$

where

L : Number of classes

Pij : Probability that a member of cluster j belongs to class i.

The total entropy of the overall clustering result is defined to be the weighted sum of the individual entropy value of each cluster. The total entropy e is defined as

$$\textbf{e} = \sum_{j=1}^{k} \frac{\beta_j}{n} e_j \quad (13)$$

where

k : Number of clusters

n : Total number of documents in the corpus.

In general, the better clustering result is given by the smaller entropy value.

2.6 Datasets

For evaluating the effectiveness of the document clustering algorithms, different text collections are available. These collections are useful for research in information retrieval, natural language processing, computational linguistics and other corpus-based research. The following real text datasets have been selected for clustering purpose. The datasets are:

Reuters-21578: Reuters-21578 test collection contains 21578 text documents. The documents in the Reuters-21578 collection are originally taken from Reuters newswire in 1987. The Reuters-21578 contains 22 files. Each of the first 21 files (reut2-000.sgm through reut2-020.sgm) contains 1000 documents, while the last (reut2-021.sgm) contains 578 documents. The documents are broadly divided into five broad categories (Exchanges, People, Topics, Organizations and Places). These categories are further divided into subcategories. The Reuters-21578 test collection is available at [9].

20NewsGroup: 20 Newsgroups data set contained 20,000 newsgroup articles from 20 newsgroups on a variety of topics. The dataset is available in http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/news20.html. It is also available in the UCI machine learning dataset repository available at http://mlg.ucd.ie/datasets/20ng.html. This dataset was assembled by Ken Lang.

Hamshahri:Hamshahri collection was used for evaluation of Persian information retrieval systems. Hamshahri collection consists of about 160,000 text documents and 100 queries in Persian and English languages. There are totally 50350 query document pairs upon which relevance judgements are made. The judgement result is binary judges, either ―1‖ (relevant) or ―0‖ (irrelevant).

III. RELATED WORKS

Document clustering had been widely studied in computer science literature. Several soft computing techniques have been used for text document clustering and some of these are discussed here. Wei Song, Soon Cheol Park [11] proposed a variable string length genetic algorithm for automatically evolving the proper number of clusters as well as providing near optimal data set clustering. GA was used in conjunction with the reduced latent semantic structure to improve clustering efficiency and accuracy. The objective function used was Davis-Bouldin index. It produces much lower cost of computational time due to the reduction of dimensions.

K. Premalatha, A.M. Natarajan [17] presented a new hybrid model of clustering based on GA and PSO which were used to solve document clustering problem. The use of Genetic Algorithm in this model is to attain an optimal solution. Due to the simplicity of PSO and efficiency in navigating large search spaces for optimal solution, it is combined with GA. This hybrid model avoided the premature convergence and improved the diversity.

EisaHasanzadeh [27], [32] developed a text clustering algorithm based on PSO and applied latent semantic indexing (PSO+LSI) for reducing dimension. Latent Semantic Indexing (LSI) was used to reduce the high dimension of textual data. Because of the main problem of text clustering algorithm is very high dimension; it is avoided by using LSI. PSO family of bio-inspired algorithms had successfully been merged with LSI. [27] used an adaptive inertia weight (AIW) that does proper exploration and exploitation in search space. This model produced better clustering results over PSO+Kmeans using vector space model. The proposed work PSO+LSI are faster than PSO+Kmeans algorithms using the vector space model for all numbers of dimensions.

StutiKarol ,VeenuMangat [43] introduced hybrid PSO based algorithm. The two partitioning clustering algorithms Fuzzy C-Means (FCM) and K- Means each hybridized with Particle Swarm Optimization (KPSO and FCPSO). The performance of hybrid algorithms provided better document clusters against traditional partitioning clustering techniques (K-Means and Fuzzy C Means) without hybridization. It is concluded that FCPSO deals better with overlapping nature of dataset than KPSO as it deals well with the overlapping nature of documents.

Nihal M. AbdelHamid, M. B. Abdel Halim, M. WaleedFakhr[36] introduced the Bees Algorithm in optimizing the document clustering problem. The Bees algorithm avoids local minima convergence by performing global and local search simultaneously. This proposed algorithm has been tested on a data set containing 818 documents and the results have revealed that the algorithm achieved its robustness. This model was compared with the Genetic Algorithm and K-means and it was concluded that Bees algorithm outperforms GA by 15% and the K-means by 50%. And also the results revealed that the Bees Algorithm takes more time than the Genetic Algorithm by 20% and the K-means by 55%.

KayvanAzaryuon,BabakFakhar [45] proposed an upgraded the standard ant's clustering algorithm by changing the Ants' movement completely random for clustering. This model provided increasing quality and minimizing run time when compared with the standard ant clustering algorithm and the K-means algorithm. His proposed algorithm has been implemented to cluster documents in the Reuters-21578. The Results have shown that the proposed algorithm presents a better average performance than the standard ants clustering algorithm, and the K-means algorithm.

IV. CONCLUSION

This paper has presented a survey on the research work done on text document clustering based on Extension Techniques. This survey starts with in-depth introduction about clustering in data mining and explored various research papers related to text document clustering. Here mainly focused on types of data clustering techniques and Moreoverresearch works have to be carried out based on semantic to make the quality of text document clustering.

REFERENCES

[1] Gerard M. Salton, Wong. A & Chung-ShuYang , (1975), ―A vector space Model for automatic indexing‖, Communications of The ACM, Vol.18,No.11,pp. 613-620.

[2] Henry Anaya-Sánchez, Aurora Pons-Porrata& Rafael Berlanga- Llavori, (2010), ―A document clustering algorithm for discovering and describing topics‖, Elsevier, Pattern Recognition Letters,Vol. 31 ,No.15,pp. 502–510.

[3] Jiaweihan& Michelin Kamber ,(2010),‖Data mining concepts and techniques‖,Elsevier.

[4] Berry, M. (ed.), (2003), "Survey of Text Mining: Clustering, Classification, and Retrieval", Springer, New York.

[5] Xu, R.&Wunsch, D, (2005),‖Survey of Clustering Algorithms‖, IEEE Trans. on Neural Networks Vol.16, No.3, pp. 645-678

[6] Xiaohui Cui & Thomas E. Potok,(2005), ―Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm‖, Journal of Computer Sciences (Special Issue): 27-33, Science Publications

[7] Porter, M.F., (1980), ―An algorithm for suffix stripping. Program‖, 14: 130-137.

[8] T.E. Potok& Cui X, (2005), ―Document clustering using particle swarm optimization‖, in proceedings of the IEEE Swarm Intelligence Symposium, .pp.185-191.

[9] http://www.daviddlewis.com/resources/testcollections/reuters21578. International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.6, December 2013 43

[10] Lotfi A &Zadeh, (1994),"Fuzzy Logic, Neural Networks, and Soft Computing," Communication of the ACM, Vol. 37 No. 3, pp. 77-84.

[11] Wei Song & Soon Cheol Park,(2009), ―Genetic Algorithm for text clustering based on latent semantic indexing, Computers and Mathematics with applications‖, 57, 1901-1907.

[12] Goldberg D.E.,(1989), ―Genetic Algorithms-in Search, Optimization and Machine Learning‖, Addison- Wesley Publishing Company Inc., London.

[13] Wei Song & Soon Cheol Park, (2006), ―Genetic algorithm-based text clustering technique‖, in Lecture Notes in Computer Science, pp.779-782.

[14] R. Eberhart& J. Kennedy, (1995) " Particle swarm optimization‖, in proceedings of the IEEE International conference on Neural Networks, Vol.4,pp. 1942–1948.

[15] Van der Merwe.D.W&Engelbrecht A.P. , (2003), ―Data clustering using particle swarm optimization‖, in Proc. IEEE Congress on Evolutionary Computation,Vol.1, pp. 215-220.

[16] Nihal M. AbdelHamid, M.B. AbdelHalim& M.W. Fakhr,(2003), ―Document clustering using Bees Algorithm‖, International Conference of Information Technology, IEEE, Indonesia .

[17] K. Premalatha& A.M. Natarajan, (2010), ―Hybrid PSO and GA Models for Document Clustering‖, International Journal of Advanced Soft Computing Applications, Vol.2, No.3, pp. 2074-8523.

[18] Shi, X.H., Liang Y.C., Lee H.P., Lu C. & Wang L.M.,(2005) ―An Improved GA and a Novel PSO-GA-Based Hybrid Algorithm‖, Elsevier , Information Processing Letters, Vol. 93,No. 5, pp.255-261.

[19] Kennedy, J. &Eberhart, R.C, (2001), ―Swarm Intelligence‖, Morgan Kaufmann 1-55860-595-9

[20] Jain, A.K., M.N. Murty& P.J. Flynn, (1999), ―Data clustering: A review‖, ACM Computing Survey, 31:264-323.

[21] PriyaVaijayanthi, Natarajan A M & Raja Murugados, (2012), ― Ants for Document Clustering‖, 1694-0814 .

[22] Michael Steinbach , George Karypis&Vipin Kumar, (2000), ―A comparison of document clustering techniques‖, in KDD Workshop on TextMining.

[23] RicardoBaeza-Yates &BerthierRibeiro-Neto, (1999), ―Modern Information Retrieval‖, ACM Press, New York, Addison Wesley.

[24] Shafiei M., Singer Wang, Zhang.R., Milios.E, Bin Tang, Tougas.J&Spiteri.R, (2007), ―Document Representation and Dimension Reduction for Text Clustering‖ , in IEEE Int. Conf. on Data Engg. Workshop, pp. 770-779.

[25] Wei Song , Cheng Hua Li & Soon Cheol Park,(2009), ―Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures‖, Elsevier, Expert Systems with Applications, Vol.36,No.5,pp. 9095–9104.

[26] J. Bezdek , R.Ehrlich& W. Full ,(1984), ―FCM: The Fuzzy C-Means Clustering Algorithm‖, Computers & Geosciences, Vol. 10, No. 2-3, pp. 191-203.

[27] EisaHasanzadeh& Hamid Hasanpour, (2010),―PSO Algorithm for Text Clustering Based on Latent Semantic Indexing‖, The Fourth Iran Data Mining Conference , Sharif University of Technology, Tehran, Iran

[28] MagnusRosell , (2006), ―Introduction to Information Retrieval and Text Clustering‖, KTH CSC.

[29] Tan,Steinbach&kumar, (2009), ―Introduction to data mining‖, Pearson education .

[30] Trappey.A.J.C.,Trappey.C.V.,Fu-Chiang Hsu &Hsiao.D.W.,(2009), ―A Fuzzy Ontological Knowledge Document Clustering Methodology‖, IEEE transactions on systems, man, and cybernetics—part B: Cybernetics, Vol. 39, No. 3,pp. 806-814.

[31] Jiayin KANG &Wenjun ZHANG, (2011), ―Combination of Fuzzy C-means and Harmony Search Algorithms for Clustering of Text Document‖, Journal of Computational Information Systems,Vol. 7,No. 16,pp. 5980-5986.

[32] EisaHasanzadeh, MortezaPoyan rad and Hamid AlinejadRokny,(2012),‖Text clustering on latent semantic indexing with particle swarm optimization (PSO) algorithm‖, International Journal of the Physical Sciences Vol. 7(1), pp. 116 – 120.

[33] Nock.R&Nielsen.F.,(2006), ―On Weighting Clustering‖,IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 28, No. 8,pp.1223-1235.

[34] Shady, FakhriKarray&Kamel.M.S.,(2010), ‖An Efficient Concept-Based Mining Model for Enhancing Text Clustering‖ IEEE Trans. on Knowledge & Data Engg., Vol. 22, No. 10,pp. 1360-1371.

[35] Pham D.T., Ghanbarzadeh A., Koç E., Otri S., Rahim S., &M.Zaidi "The Bees Algorithm – A Novel Tool for Complex OptimisationProblems",in proceedings of IPROMS Conference, pp.454–461. International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.6, December 2013 44

[36] Nihal M. AbdelHamid, M. B. Abdel Halim, M. WaleedFakhr,(2013), ―BEES ALGORITHM-BASED DOCUMENT CLUSTERING‖, ICIT 2013 The 6th International Conference on Information Technology.

[37] Karypis.G, Eui-Hong Han &Kumar.V., (1999), ―Chameleon: A Hierarchical Clustering Algorithm using Dynamic Modelling‖, IEEE Computer, Vol. 32, No.8, pp. 68-75.

[38] Zhang.T., Raghu Ramakrishnan&Livny.M.,(1996), ―Birch: An Efficient Data Clustering Method for very Large Databases‖, In Proceedings of the ACM SIGMOD international Conference on Management of Data, pp. 103-114.

[39] S. Guha, R. Rastogi, and K. Shim,(1999), ―ROCK: A robust clustering algorithm for categorical attributes‖, International Conference on Data Engineering (ICDE―99), pp. 512-521.

[40] S.N. Sivanandam and S.N.Deepa,(2008), ―Principles of Soft Computing‖, Wiley-India.

[41] Ivan BrezinaJr.ZuzanaČičková ,(2011), ―Solving the Travelling Salesman Problem Using the Ant Colony OptimizationManagement Information Systems‖,Vol. 6, No. 4,pp. 010-014.

[42] Ming-Chuan Hung & Don-Lin Yang,(2001), ―An Efficient Fuzzy C-Means Clustering Algorithm‖,in proceedings of the IEEE conference on Data mining,pp.225-232.

[43] Stuti Karol , VeenuMangat, (2012), ― Evaluation of a Text Document Clustering Approach based on Particle Swarm Optimization‖, CSI Journal of Computing , Vol. 1 , No.3.

[44] Pham D.T., Ghanbarzadeh A, Koc E, Otri S, Rahim S and Zaidi M,(2005), ―The Bees Algorithm‖, Notes by Manufacturing Engineering Centre, Cardiff University, UK .

[45] KayvanAzaryuon , BabakFakhar,(2013), ―A Novel Document Clustering Algorithm Based on Ant Colony Optimization Algorithm‖, Journal of mathematics and computer Science Vol.7 , pp. 171-180.

[46] "Document Clustering in Correlation SimilarityMeasure Space" Taiping Zhang, *Member, IEEE,* Yuan Yan Tang, *Fellow, IEEE,* Bin Fang, *Senior Member, IEEE,* and Yong Xiang

