iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** XII **Month of publication:** December 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

An Efficient Algorithm for Document Clustering in Information Retrieval

Ms. R. Janani[1], Dr. S. Vijayarani[2]
[1]PhD Research Scholar, [2]Assistant Professor
Department of Computer Science,Bharathiar University,Coimbatore

Abstract - Document clustering is a set of documents can be divided into similar groups called clusters, so that documents within a cluster have high similarity in comparison to other documents in different clusters. It has been considered intensively due to the fact of its extensive applicability in various areas like information retrieval, web mining and search engines like Google. It is determining the similarity between documents and based on the similarity it will group the documents together. It offers efficient representation and visualization of the documents; thus helps in convenient navigation also. The main objective of this research work is to cluster the documents into similar groups based on the content of the documents. In order to perform this task this research work uses two existing documents clustering algorithms, namely K-means and DBSCAN and also this work proposes a new clustering algorithm, E-DBSCAN. From the experimental results it is observed that the E-DBSCAN gives the better clustering accuracy than other algorithms.

Keywords - Document Clustering, preprocessing, K-means, DBSCAN, E-DBSCAN

I. INTRODUCTION

Document clustering is the subset of data clustering, which comprises the concepts from the fields of information retrieval, natural language processing, and machine learning. It organizes the collection of documents into different groups, called as clusters, wherever the documents in each cluster have common properties allowing to distinct similarity measure. The high quality document clustering algorithms play a significant role to effectively navigate, summarize, and organize the documents. Clustering can achieve either disjoint or overlapping partitions. In an overlapping partition, it is probable for a document to appear in multiple cluster [1]. In disjoint clustering, each document seems in exactly only one cluster. Document clustering can be divided into two major subcategories, hard clustering and soft clustering. Hard clustering calculates the hard assignment of the particular document to a cluster. Soft clustering is divided into partitioning, hierarchical, and frequent item set based clustering [2]. The main objective of this research work is to cluster the documents by analyzing the contents of the documents using document clustering algorithms.
This paper organized as follows, section II explains the related work and section III presents the methodology of this research work. Experimental results are given in Section IV and Section V describes the conclusion of this research work.

II. RELATED WORKS

In [3], a simple and efficient variant of K-means, bisecting Kmeans, where centroids are updated incrementally, is introduced. It produces better clusters than those produced by regular K-means. Bisecting K-means has a linear time complexity. The authors have first compared three agglomerative hierarchical techniques, namely Intra-Cluster Similarity Technique (IST), Centroid Similarity Technique (CST), and UPGMA. Results show that UPGMA is the best hierarchical technique, which is then, compared with K-means and bisecting K-means. Bisecting k-means is proving to be superior to UPGMA and regular k-means. The better performance of bisecting K-means is because of production of relatively uniform size clusters.
In [4] is an improved version of the DBSCAN, introducing a sampling technique to address the two issues of DBSCAN and its variations: (1) - to make it effective while dealing with large volume of spatial data objects; (2) - reduces the I/O cost. From the experiments, sampling based IDBSCAN outperforms the DBSCAN in minimizing I/O cost and memory requirement for clustering with no compromise on the quality of cluster. Although, IDBSCAN improved the I/O cost, but still it needs users to specify the value of threshold parameters manually.
In[5] purposed a web fuzzy clustering model. In their paper the experimental result of web fuzzy clustering in web user clustering proves the feasibility of web fuzzy clustering in web usage mining.
In [6] is another important variation of DBSCAN. That tackled the issues associated with most of the density-based clustering algorithms is that they are not effective to perform clustering accurately in the presence of clusters with different densities. Uncu et al. proposed a three level clustering mechanism to provide a solution to this problem. In the first level, it provides appropriate grids

such that density is similar in each grid. In the next level, it merges the cells having same densities. At this level, the appropriate values of MinPts are also identified in each grid. In the final step, the DBSCAN algorithm is applied to these identified parameter values to obtain the required final number of clusters. Although accuracy of GRIDBSCAN is better as compare to DBSCAN but on the other hand GRIDBSCAN may be handy in terms of computational complexity when applied to large spatial data.

III. METHODOLOGY

The main objective of this research work is to cluster the documents by analyzing the contents of the documents using document clustering algorithms. In order to perform this task, this research work uses two existing document clustering algorithms; they are K-means and DBSCAN. This work also proposes a new document clustering algorithm, namely E-DBSCAN. The performance factors are cluster purity, cluster entropy, F-measure, precision and recall. Figure 1 shows the architecture of this research work.

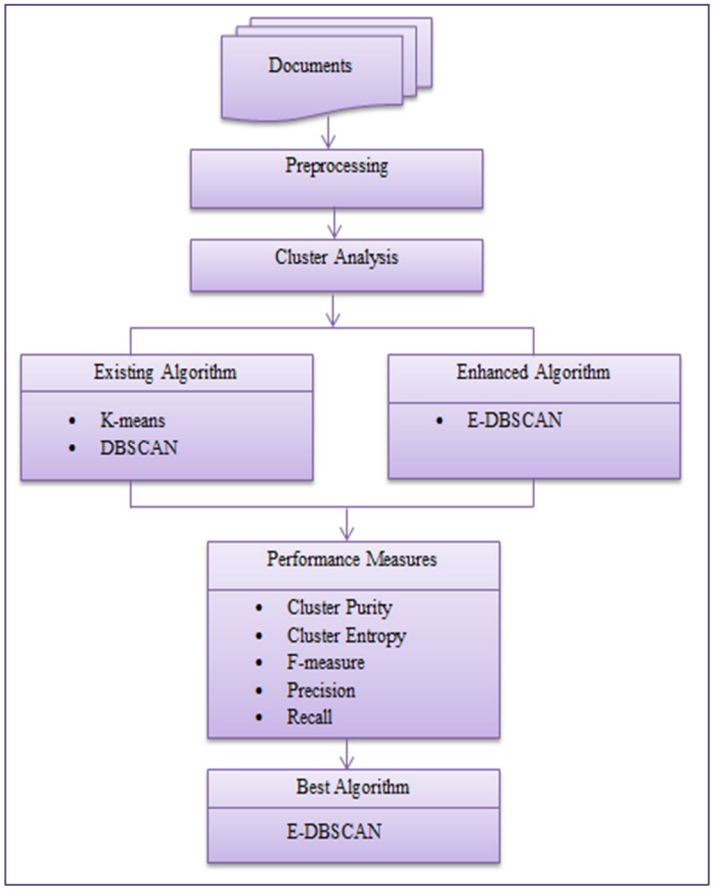


Figure 1. System Architecture

A. *Documents*

The synthetic dataset was created and it has 25 documents. This document contains the text files (.txt) and document files (.docx). The size of the documents varied from 1 kb to 1514 kbs.

B. *Preprocessing*

Document preprocessing is an important task in text mining, information retrieval (IR) and Natural Language Processing (NLP). In text mining, document preprocessing is used for extracting non-trivial knowledge from unstructured text data [7]. Before clustering the documents, the preprocessing techniques are applied on the particular dataset to reduce the size of the dataset. It will increase the efficiency of the document clustering system. In this research work, stemming, stop word removal, numbers and punctuation removal techniques are used.

C. *K-means*

K-means Clustering is a simple and empirical data analysis technique. This is non-hierarchical method of grouping objects together

[8]. This clustering aims to partitioning the number of objects into k clusters in which all observation belongs to the cluster with nearest mean value.

In k-means algorithm, first choose the k initial centroids, where k is a user given parameter, called as the number of clusters. Each and every point is then assigned to the nearest centroid, and each collection of points assigned to a centroid is a cluster [9]. The centroid of each cluster is then updated based on the points assigned to the particular cluster. Repeat these update steps until no point changes clusters or until the centroids remain the same.

Algorithm 1: K-means algorithm

1: Select K points as the initial centroids
2: Repeat
3: from K clusters by assigning each point to its closest centroid
4: Recompute the centroid of the each cluster
5: Until centroids do not change

D. DBSCAN

DBSCAN is the first density based clustering algorithm. It was proposed by Ester et al. in 1996, and it was designed to cluster data of arbitrary shapes in the presence of noise in spatial and non-spatial high dimensional databases [10]. The DBSCAN (Density-based spatial clustering of applications with noise) algorithm can be used to identify the clusters in large spatial datasets by observing at the local density of database elements, using one input parameter. Moreover, the users get an idea on which parameter value that would be appropriate. Consequently, minimal knowledge of the domain is essential [11].

The DBSCAN can also determine what information should be classified as noise or outliers. In spite of this, its working process is quick and scales very well with the size of the database almost linearly. By using the density distribution of nodes in the database, DBSCAN can group these nodes into separate clusters that define the different classes. DBSCAN can find clusters of arbitrary shape [12]. But, clusters that lie close to each other tend to fit into the same class.

Algorithm 2: DBSCAN algorithm

1: Label all points as core, border or noise points
2: Eliminate noise points
3: Put an edge between all core points that are eps of other
4: Make each group of connected core points into a separate cluster.
5: Assign each point to one of the cluster of its associated core points

E. E-DBSCAN

The key idea of E-DBSCAN is that for each object of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of objects (MinPts), which means that the cardinality of the neighborhood has to exceed some threshold [13]. The neighborhood of an arbitrary point 'p' is defined in Equation (1)

$$N_{Eps} = \{q \in D / dist(p, q) < Eps \quad \ldots\ldots\ldots\ldots\ldots (1)$$

Here, D is the database of objects. If the neighborhoods of a point P at least contain a minimal number of points, and then this point is called core point. The core point is defined in Equation 2.

$$N_{Eps}(p) > Minpts \quad \ldots\ldots\ldots\ldots\ldots (2)$$

NEps (P) > MinPts (2) Here Eps and MinPts are the user's specified parameters which mean the radius of the neighborhood and minimum number of points in the neighborhood of a core point respectively. If this condition is not satisfied then this point is considered as non-core point.

Algorithm 2: E-DBSCAN algorithm

```
E-DBSCAN (D, epsilon, min_points):
 1: C = 0
 2: for each unvisited point P in dataset
 3: mark P as visited
 4: spoint = regionQuery (P, epsilon)
 5: if sizeof (spoint) < min_points
 6: ignore P
 7: else
 8: C = next cluster
 9: eclust (P, sphere_points, C, epsilon, min_points)
Eclust (P, sphere_points, C, epsilon, min_points)
 1:  add P to cluster C
 2: for each point P' in spoint
3: if P' is not visited, mark P' as visited
4: spoint = regionQuery (P', epsilon)
5: if sizeof(sphere_points') >= min_points
6: spoint = spoint joined with spoint'
7: if P' is not yet member of any cluster,
   8:  add P' to cluster C
Region Query (P, epsilon)
1: return all points within the n-dimensional sphere centered at P with radius
epsilon
```

IV. EXPERIMENTAL RESULTS

In order to perform this task, the performance factors are cluster purity and cluster entropy. The synthetic dataset was created and it has 25 documents. For this analysis, the existing and enhanced algorithms were implementer in 'R'.

A. Purity

Given the true clustering assignment for the subjects, this function calculates the cluster purity index comparing with clustering assignment determined by integrative NMF algorithms. Higher value of cluster purity indicates better cluster predictive discrimination.

B. Entropy

Given the true clustering assignment for the subjects, this function calculates a cluster entropy index comparing with clustering assignment determined by integrative NMF algorithms. Smaller value of cluster entropy indicates better cluster predictive discrimination.

C. Precision

It is calculated as the fraction of pairs correctly put in the same cluster.

D. Recall

It is calculated as the fraction of actual pairs that were identified.

E. F-measure

It is the harmonic mean of precision and recall.

Table 1 illustrates the cluster analysis of five clusters for K-means, DBSCAN and E-DBSCAN clustering algorithms. Figure 2 Performance analysis of K-means, DBSCAN and E-DBSCAN algorithm. From this figure it is observed that the E-DBSCAN gives high accuracy for all clusters.

Table 1. Cluster accuracy for K-means, DBSCAN and E-DBSCAN algorithm

Number of Clusters	Accuracy (%)		
	K-means	DBSCAN	E-DBSCAN
2	76.5	82.4	89.2
3	93.4	95	97.2
4	97.4	97.9	98.3
5	98.3	98.9	99.1

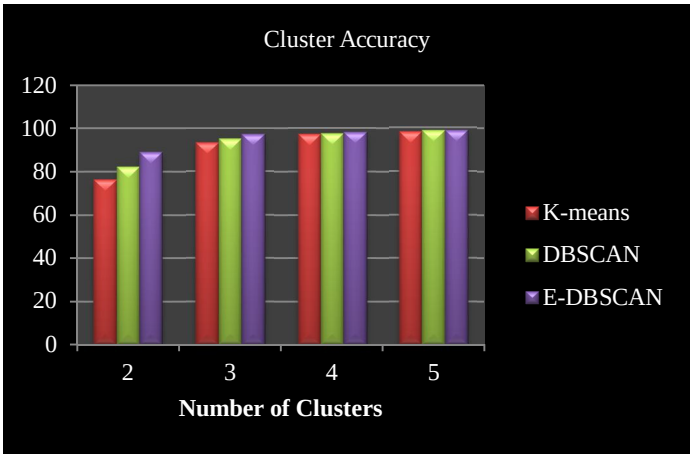


Figure 2. Performance analysis of K-means, DBSCAN and E-DBSCAN algorithm

Table 2 illustrates the performance metrics like cluster purity and cluster entropy for K-means, DBSCAN and E-DBSCAN clustering algorithm. Here cluster purity and entropy is calculated based on the average of all the clusters such as cluster 2 to cluster 5. Figure 3 describes the analysis of cluster purity and entropy methods for K-means, DBSCAN and E-DBSCAN clustering algorithm. From figure 3, it is observed that enhanced DBSCAN performs well when compared to other existing algorithms.

Table 2. Purity and Entropy measures of K-means, DBSCAN and E-DBSCAN algorithm

Performance Measures	K-means	DBSCAN	E-DBSCAN
Cluster Purity	0.9	0.9	1
Cluster Entropy	0.6	0.3	0.2

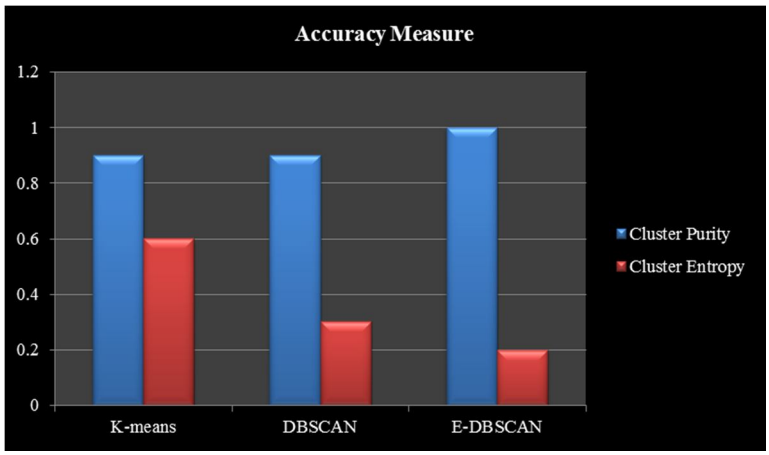


Figure 3. Accuracy Measures of K-means, DBSCAN and E-DBSCAN algorithm

Table 3 illustrates the performance metrics like precision, recall and f-measure for K-means, DBSCAN and E-DBSCAN clustering algorithm. Here precision, recall and f-measure is calculated based on the average of all the clusters such as cluster 2, cluster3, cluster4 and cluster 5.From figure 4, it is observed that enhanced DBSCAN performs well when compared to other existing algorithms.

Table 3. Precision, Recall and F-measure of K-means, DBSCAN and E-DBSCAN algorithm

Performance Measures	K-means	DBSCAN	E-DBSCAN
Precision	39.7	49.1	53.23
Recall	25.7	40.8	50.1
F-Measure	30.46	44.64	51.45

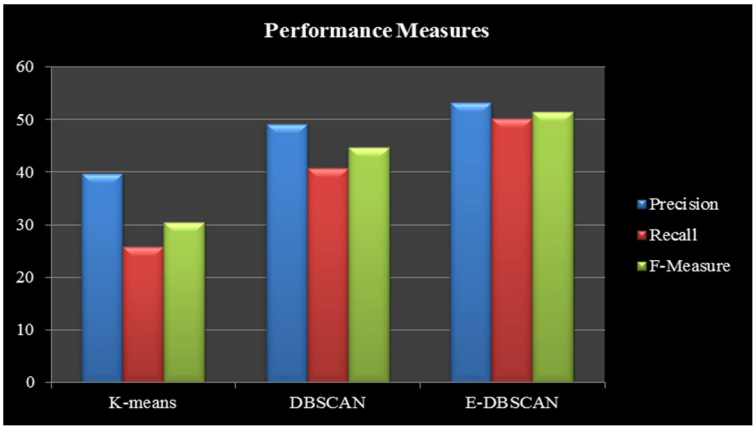


Figure 4. Performance analysis of K-means, DBSCAN and E-DBSCAN algorithm

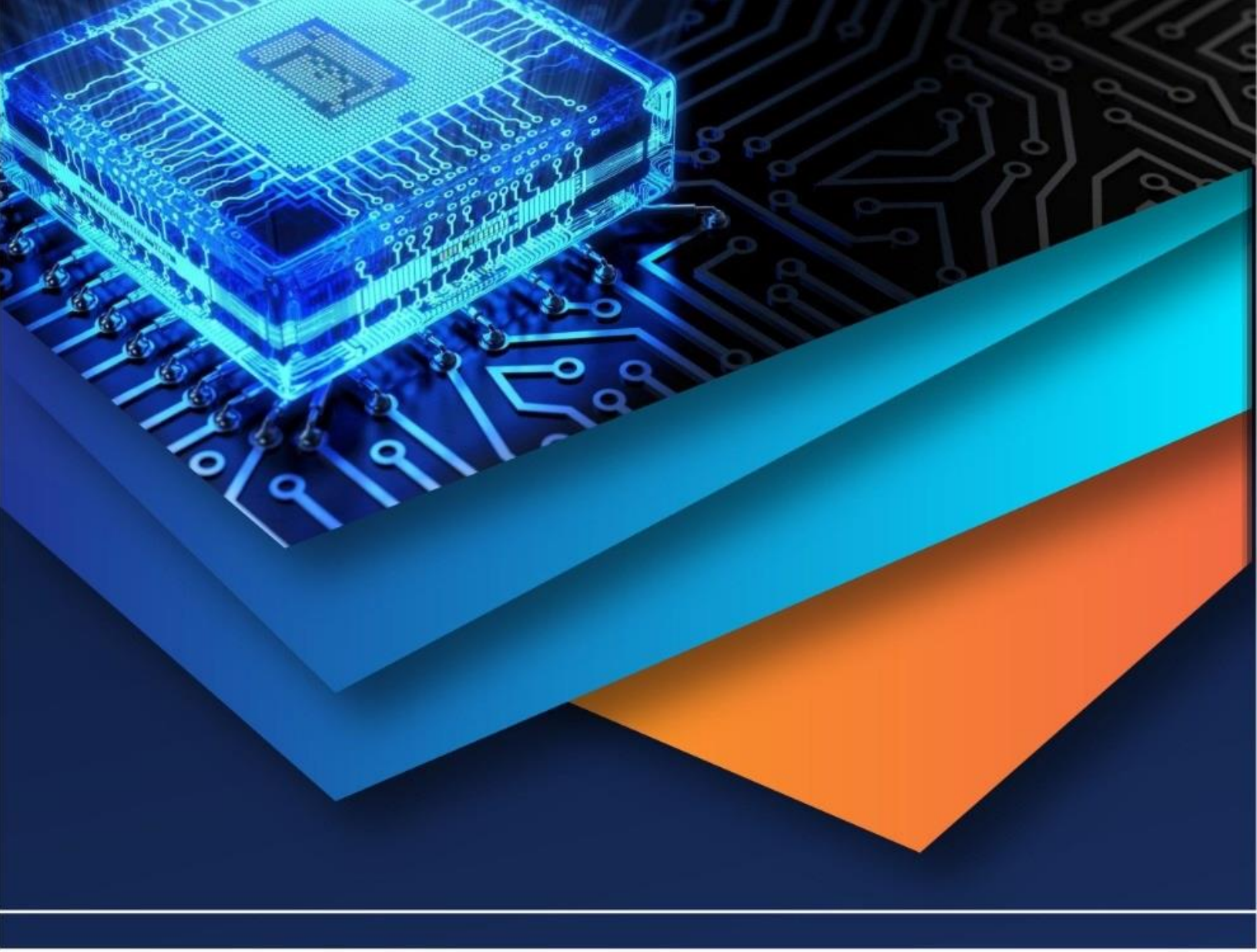
V. CONCLUSION

Document clustering is an essential process used in information retrieval, unsupervised document organization and automatic topic extraction. In this research work, it analyses the performance measures of existing and enhanced clustering algorithm for document clustering. The performance metrics are cluster accuracy, precision, recall, f-measure, cluster purity and cluster entropy. From this analysis, in existing algorithms the DBSCAN clustering algorithm gives the best accuracy for this synthetic data set. From the existing and enhanced algorithms E-DBSCAN gives the best accuracy. In future, the most efficient algorithms have to be developed for clustering all types of documents.

REFERENCES

[1] M. Steinbach, G. Karypis, and V. Kumar, “A Comparison of Document Clustering Techniques,” Proc. Knowledge Discovery and Data Mining (KDD) Workshop Text Mining, Aug. 2000.
[2] Rekha Baghel, Dr. Renu Dhir, “A Frequent Concepts Based Document Clustering Algorithm” , International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010.
[3] Yuqin Li , Xueqiang Lv , Yufang Liu , Shuicai Shi , “Research on Text Clustering Based on Concept Weight”, 2010 Fourth International Conference on Genetic and Evolutionary Computing.
[4] B. Borah and D. K. Bhattacharyya, “An Improved SamplingBased DBSCAN for Large Spatial Databases,” presented in the international Conference on Intelligent Sensing and Information Processing, Chennai, India, January 2004.
[5] Maofu Liu, Yanxiang He and Huijun Hu,” Web Fuzzy Clustering and Its Applications in Web Usage Mining”,
[6] O. Uncu, W. A. Gruver, D. B. Kotak, D. Sabaz, Z. Alibhai, and C. Ng, “GRIDBSCAN: GRId Density-Based Spatial Clustering of Applications with Noise,” 2006 IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan.
[7] Han J, Kambr M. “Data Mining: Concepts and Techniques”. Hand Book. Beijing: Higher Education Press, 2001.
[8] Thangamani.M, Dr. Thangaraj.P ,”Ontology Based Fuzzy Document Clustering Scheme”, Modern Applied Science, Vol. 4, No. 7; July 2010 148 ISSN.
[9] J. Jayabharathy, S. Kanmani and A. Ayeshaa Parveen’ “Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature”, 978-1-61284-486-2/11
[10] Florian Beil, Martin Ester, Xiaowei Xu, “Frequent Term-Based Text Clustering”, 2002 ACM 1-58113-567-X/02/0007.
[11] Jiabin Deng, JuanLi Hu, Hehua Chi, Juebo Wu, “An Improved Fuzzy Clustering Method for Text Mining”, 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing.

[12] Amir Hamzah, Adhi Susanto, F.Soesianto, Jazi Eko Istyanto, "Concept based Text Document Clustering" Proceedings of International Conference on Electrical Engineering and Informatics, Indonesia June 17-19 2007
[13] Jie Ji, Tony Y. T. Chan, Qiangfu Zhao, "Fast Document Clustering Based on Weighted Comparative Advantage", Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA - October 2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY