

AN EFFICIENT TECHNIQUE FOR PERFORMING DOCUMENT CLUSTERING

Miss Anamika Gujriya (Computer Science, Sri Satya sai university sehore Bhopal M.P)

Mr. Manoj Verma (Computer Science,Sri Satya Sai university sehore Bhopal M.P)

ABSTRACT:

Clustering means keeping similar objects together. Document clustering is an extension of clustering, which is related to keeping similar text documents together. Document clustering plays a vital role in development of search engines, where a group of document is required to listed as a result of query in minimum response time. This paper elaborates the concept of document cum text clustering. This paper will provide a survey of recent work done in the field of text clustering. A critical review of modern text clustering techniques will also be provided by this paper. This paper also proposes an efficient technique for document clustering. This proposed techniques uses better measures for cluster centroid selection.

1. INTRODUCTION:

Data mining is a technique that helps to extract important data from a large database [14]. It is the process of sorting through large amounts of data and picking out relevant information through the use of certain sophisticated algorithms. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information.

Data mining extracts novel and useful knowledge from data and has become an effective analysis and decision means in corporation. Results of Data Mining Include:

- Forecasting what may happen in the future
- Classifying people or things into groups by recognizing patterns
- Clustering people or things into groups based on their attributes
- Associating what events are likely to occur together
- Sequencing what events are likely to lead to later events

Clustering is a partition of data into groups of related objects. Each set, called cluster, consists of objects which are similar to each other and dissimilar to the item of other groups. In other language, the principle of a high-quality document clustering approach is to decrease intra-cluster distances between documents. It is shown below in figure 2. In clustering is the allocation and the nature of information that will conclude cluster membership, in conflict to the classification where the classifier learn the association between objects and classes from a so set, i.e. a set of documents properly label by hand, and then replicates the learnt performance on unlabeled data

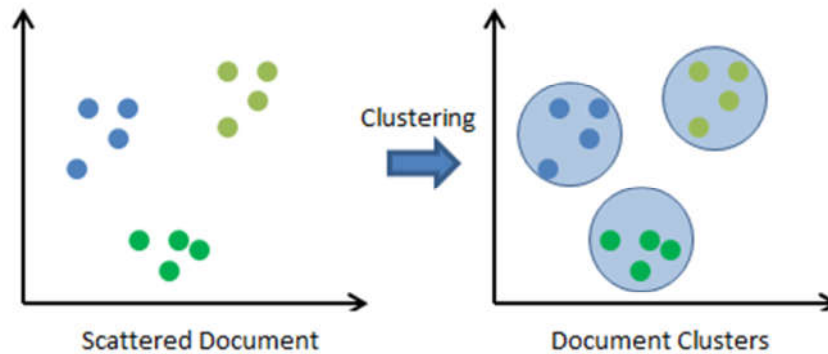


Figure 1: Document Clustering [4]

2. Related work

In 2010, F. Iqbal, et al [1] shows a correlated application domain of mining, e-mails are group by using structural, and domain-specific features. Three clustering methods (K-means, Bisecting K-means and EM) were used. In 2010, Guo-Yan Huang *et al.* [2] posited an way for clustering heterogeneous data streams with uncertainty. A occurrence histogram with H-UCF facilitate to trace characteristic categorical statistic. firstly, creating 'n' clusters by a K-prototype algorithm, the new method proves to be more useful than U Micro in regard to clustering value In 2010, Alam *et al.*, [3] designed a new clustering approach by combination divisional and agglomerative clustering known as HPSO. It developed the cleverness of ants in a decentralized environment. This method proved to be very efficient as it performed clustering in a agglomerative manner

In 2010, Shin-Jye Lee *et al.*, [4] define clustering-based scheme to recognize the fuzzy system. To start the mission, is tried to present a modular method, based on hybrid clustering method. Next, finding the number and position of clusters seemed the prime concerns for evolving such a model. In 2010, Li Taoying *et al.*, [5] lent maintain to an incremental clustering for unqualified data using clustering collection. They initially compact unnecessary attributes if required, and then made use of accurate values of different attributes to form clustering memberships. In 2009, M. Debbabi, et al [6] shows incorporated background for mining mails for forensic study, using classification and clustering method

In 2009, S. Decherchi, et al [7] addressed the difficulty of clustering mails for forensic study where a Kernel-support variation of K-means was apply. The obtained outcome were examine personally, and the creator concluded that they are attractive and valuable from an analysis perspective In 2009, Pallav Roxy and Durga Toshniwal et al [8] The former, capable of maximize middling similarity within clusters and minimize the same among clusters, is a twosome similarity clustering. The latter attempt to generate approach from the manuscript, each technique representing one document set in particular. In 2008, Miha Grcar *et al.* [9] mulled over a method about be short of software extracting method, which is a procedure of extracting information out of resource code. They offered a software extracting task with an integration of manuscript mining and link study technique. This variety of algorithm has also been used by In 2007 N.L. Beebe, et al [10] in organize to cluster the results from keyword searches. The underlying assumption is that the clustered results can increase

the information retrieval efficiency, because it would not be required to review all the documents found by the client anymore

In 2005 B.K.L.Fei, et al [11] shows (self-organize map) SOM-based algorithms used for clustering files with the aim of making the decision-making process achieved by the examiners more efficient. The files were clustered by taking into report their creation dates/times and their extensions In 2005, Agrawal et al [12] a scribed data mining function and their various necessities on clustering procedure. The most important necessities considered are their potential to recognize clusters implanted in subspaces. The subspaces contain elevated value data and scalability. They moreover consist of the understandable ability of outcome by end-users and distribution of unpredictable information transfer.

In 2004, Shehroz Khan and Amir Ahmad, et al [13] predetermined iterative clustering method to evaluate preliminary cluster centers for K-means. This procedure is sufficient for clustering procedure for constant data.

The main negative aspect of K-means approach is that generates empty clusters based on initial center vectors. However, this drawback does not cause any significant problem for static execution of K-means and the problem can be conquer by implementing K-means algorithm for a numeral of times. However, in a small number of applications, the cluster issue poses problems of erratic behavior of the system and affects the overall performance.

3. PROPOSED CLUSTERING METHODOLOGY FOR DOCUMENT CLUSTERING:

The document clustering framework is shown below in figure 3. Input are text documents. Then key words are identified in these documents. Then similarity is measured in these documents. Generally Euclidian distance is used as similarity measure. Then on basis of similarity documents are mapped in the correspondent clusters.

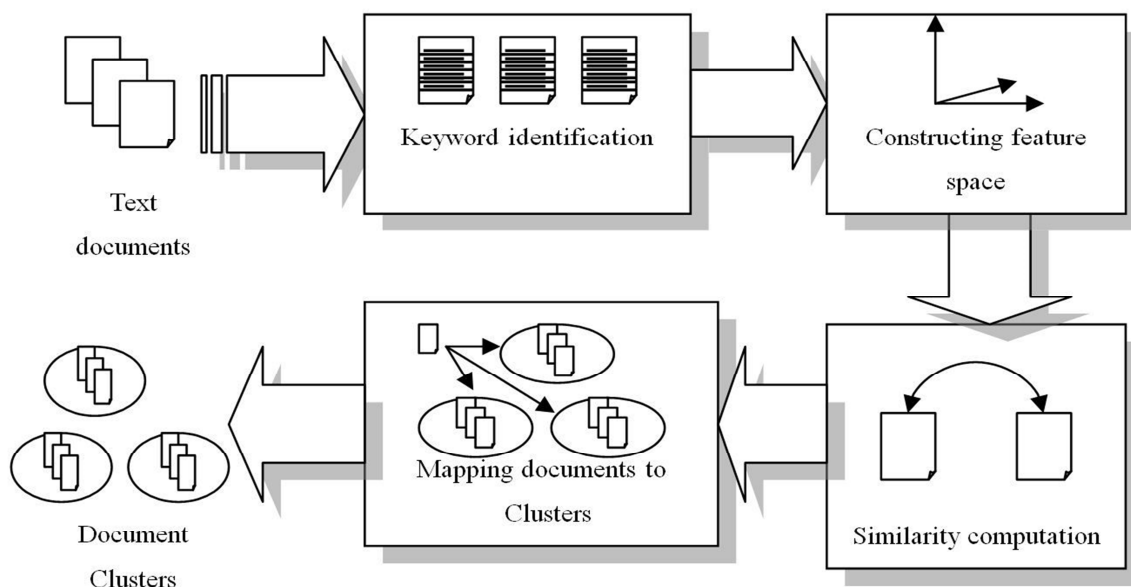


Figure 2: Document Clustering Framework[6]

The outline of the proposed approach is as follows:

1. Pre-Processing element
2. Count the amount of cluster
3. Clustering procedure
4. Query Results

The steps of proposed technique are as follows:

Output: $D = \{d_1, d_2, d_3, \dots, d_i, \dots, d_n\}$

$d_i = \{x_1, x_2, x_3, \dots, x_i, \dots, x_m\}$ k

Input: A value of k clusters.

Step 1. Select $k=2$ original cluster centers C_i randomly from data X_i .

Repeat following steps for every cluster center.

Step 2. Find Euclidean similarity of each data objects X_i from cluster centers and allocate objects to cluster with least distance.

Step 3. Find min and max distance similarity beside with corresponding nearby aim and farthest object

Step 4. Evaluate two sets of items NPT and MPT enclose tightly joined objects to within distance: $avg_dist = (Min_dist + Max_dist) / 3$

Step 5. Choose K i) $NPT_i \cap MPT_i = \Phi$ ii) $NPT_i \cap NPT_j = \Phi$ and $MPT_i \cap MPT_j = \Phi$ If (i) legal then divide C_i and if both (i) and (ii) valid divide both center and assign new center as of corresponding cluster. If either condition is valid then goto step 2.

Step 6. Find mean for each cluster.

Step 7. If cluster value is same then exit.

The above Modified k-means algorithm has additional steps in traditional k-means for better cluster center selection. We use distance for transfer item to appropriate cluster by using these deliberate distance and we find near objects from c For selecting better cluster centers we use sets of densely connected objects. The NPT set enclose objects within avg_dist from min_obj and MPT set enclose objects within avg_dist distance from max_obj

4. Result Analysis:

For this experiment, We have used animal dataset and several other dataset from the news data set, which has been taken from a newspaper where different type of article is being published.

The results obtained from the existing k-means algorithm and the proposed improved incremental algorithms are shown below in following tables and graphs:

Algorithm Name	No of Documents	No of Clusters Created	Accuracy
Existing K-means	15	4	70%
Incremental Proposed	15	4	95%

Table 1: Result Comparison of previous and proposed algorithm

Accuracy is an anonymous for purity higher the purity higher is the accuracy, here shown that correctness of the incremental mining is more than regular K-mean method

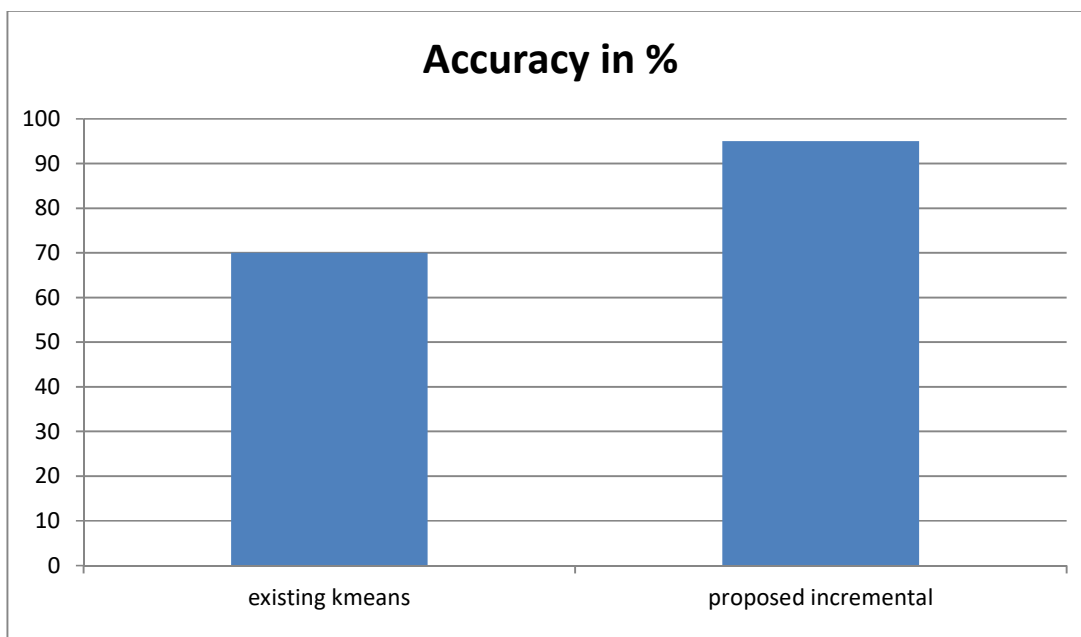


Figure 3: Accuracy Comparison

In clustering purity is an important measure to calculate the overall quality of clustering solution, higher the purity better the superiority of cluster

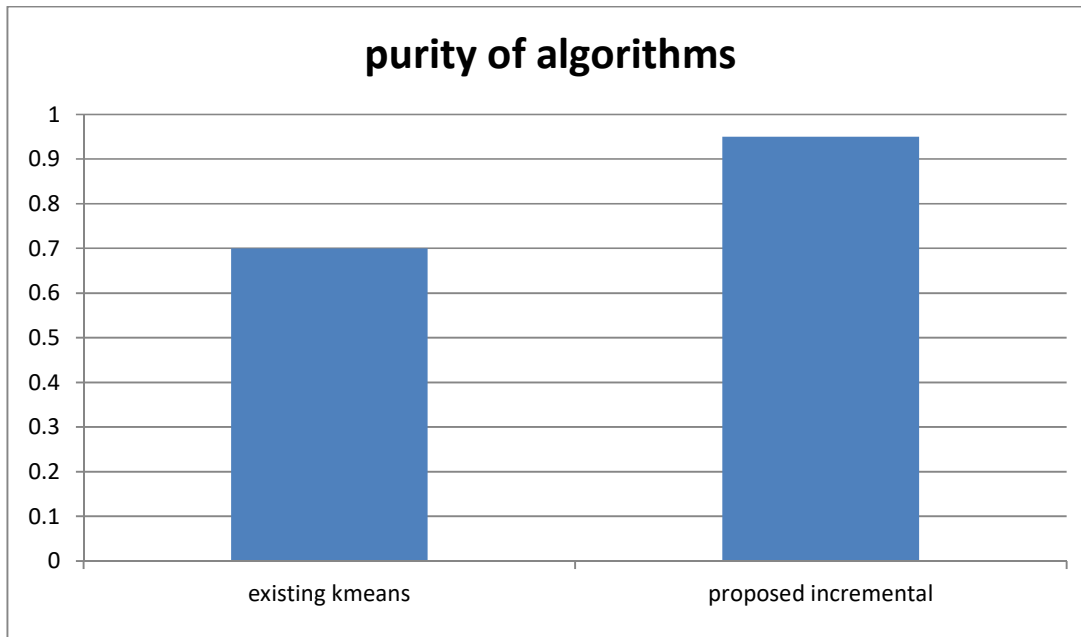


Figure 4: Purity Graph

5. CONCLUSION:

In this paper, the focus is on Document Clustering which is very recent technology, we investigated many existing algorithms. As clustering plays a very vital role in various applications, many researches are still being done. The upcoming innovations are mainly due to the properties and the characteristics of existing methods. This paper presents an introduction to the present document clustering concept along with the methods used for document clustering. A critical review of existing work done by authors on document clustering in recent time is also presented in this paper. This paper also proposed an efficient technique for document clustering. This proposed technique uses better measures for cluster centroid selection. Experimental results have shown that the accuracy of proposed technique is better.

REFERENCES:

- [1] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," *Digital Investigation, Elsevier*, vol. 7, no. 1–2, pp. 56–64, 2010.
- [2] Guo-Yan Huang, Da-Peng Liang, Chang-Zhen Hu and Jia-Dong Ren, "An algorithm for clustering heterogeneous data streams with uncertainty", 2010 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 4, pp. 2059-2064, 2010.
- [3] Alam, S., Dobbie, G., Riddle, P. and Naeem, M.A. "Particle Swarm Optimization Based Hierarchical Agglomerative Clustering", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 2, pp. 64-68, 2010.
- [4] Shin-Jye Lee and Xiao-Jun Zeng, "A three-part input-output clustering-based approach to fuzzy system identification", 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 55-60, 2010.

- [5]. Li Taoying, Chne Yan, Qu Lili and Mu Xiangwei, “Incremental clustering for categorical data using clustering ensemble”, 29th Chinese Control Conference (CCC), pp. 2519-2524, 2010.
- [6] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, “Towards an integrated e-mail forensic analysis framework,” *Digital Investigation, Elsevier*, vol. 5, no. 3–4, pp. 124–137, 2009.
- [7] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, “Manuscript clustering for digital forensics analysis,” *Computat. Intell. Security Inf. Syst.*, vol. 63, pp. 29–36, 2009
- [8] Pallav Roxy and Durga Toshniwal, “Clustering Unstructured Manuscript Documents Using Fading Function”, *International Journal of Information and Mathematical Sciences*, Vol. 5, No. 3, pp. 149-156, 2009
- [9] Miha Grcar, Marko Grobelnik and Dunja Mladenic, “Using Manuscript Mining and Link Analysis for Software Mining”, *Lecture Notes in Computer Science*, Vol. 4944, pp. 1-12, 2008.
- [10] N. L. Beebe and J. G. Clark, “Digital forensic manuscript string searching: Improving information retrieval effectiveness by thematically clustering search results,” *Digital Investigation, Elsevier*, vol. 4, no. 1, pp. 49–54, 2007.
- [11] B.K.L.Fei, J.H.P.Eloff, H.S.Venter, and M.S.Oliver, “Exploring forensic data with self-organizing maps,” in *Proc. IFIP Int. Conf. Digital Forensics*, 2005, pp. 113–123.
- [12]. Aggarwal, C.C. Charu, and C.X. Zhai, Eds. “Chapter 4: A Survey of Manuscript Clustering Algorithms,” in *Mining Manuscript Data*. New York: Springer, 2012.
- [13] Shehroz S. Khan and Amir Ahmad, “Cluster Center Initialization Algorithm for K-means Clustering”, *Pattern Recognition Letters*, Vol. 25, No. 11, pp. 1293-1302, 2004