International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

AN INVESTIGATION OF VARIOUS DATA MINING BASED CLUSTERING TECHNIQUES FOR PERFORMING CLUSTERING OF TEXT DOCUMENTS

Miss Anamika Gujriya[1], Mr. Kailash Patidar[2], Mr. Manoj Varma[3]

[1,2,3]Computer Science, Sri Satya Sai University Of Technology And Medical Science ,sehore 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.

I.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

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

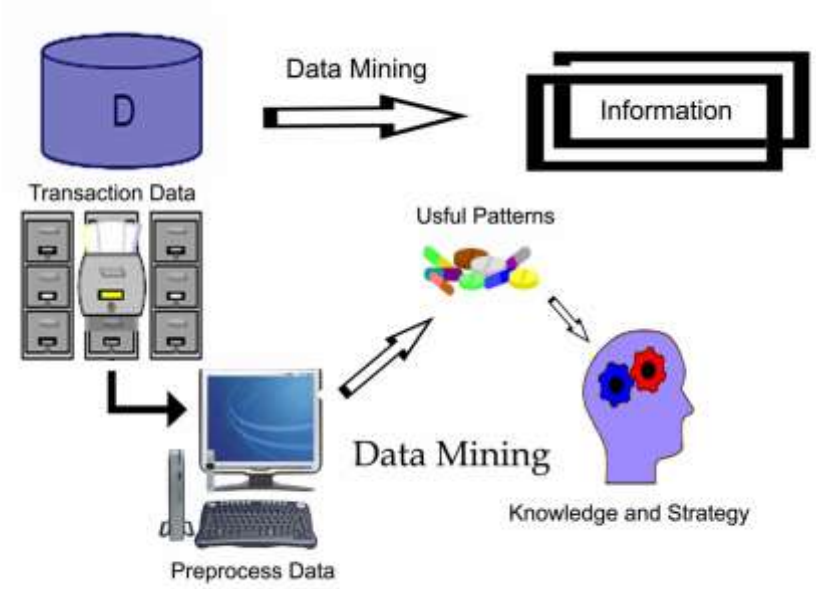

Figure 1 Data Mining [1]

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

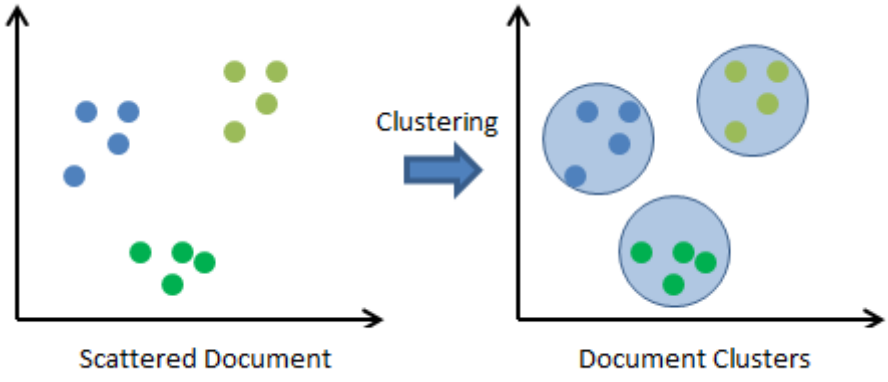

Figure 2: Document Clustering [4]

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.

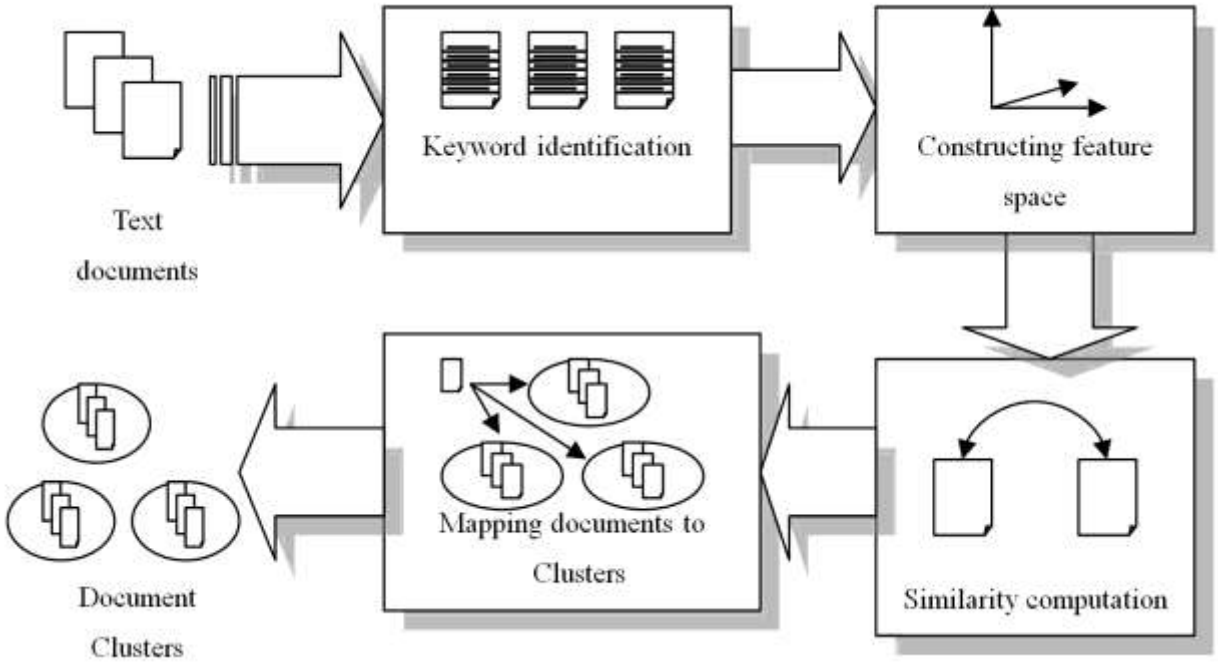

Figure 3: Document Clustering Framework[6]

II.LITERATURE REVIEW

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 Liping, emphasized that the expansion of internet and computational processes haspaved the mode for various clustering methods. document mining mainly has gained a bunch of importance and it strain a range of tasks such as construction of granular taxonomies, document summarization etc., for developing a advanced quality data from documents.

In 2010,Guo-Yan Huang *et al.* [2] posited an way for clustering heterogeneous data streamswith 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. So, taking input, output, generalization and specialization, a HCA has been designed. This three-part enter production clustering method accept lot of clustering characteristics all together to recognize the problem Only a small number of researchers have focused awareness on partition unconditional data in an incremental mode. Designing an incremental clustering for categorical data is a critical problem.

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 technique is concerned with the inter links between instances. Retrieval and knowledge based approaches are the two main tasks used in constructing a tool for software component .An learning frame work named LATINO was urbanized by Grcar *et al.* (2006). LATINO, an open spring principle data mining platform, offers document mining, link analysis, machine learning, etc. Similarity-based approach and model-based approaches

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.

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.

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.

III.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.

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,andC.X.Zhai, Eds. "Chapter4: ASurvey of Manuscript Clustering Algorithms," in *Mining Manuscript Data*.NewYork: 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