A Literature Review of Various Techniques for Performing Document Clustering

Zenab Qureshi[1], Prof. Priyanka Dubey[2]
[1,2]*Alpine Institute of Technology, Ujjain*

***Abstract*- The measure of information recorded and the content data accessible on the web has been surprisingly expanding, gathering and enlarging with every day. Such information and data which is accessible in high voluminous structure is really not accessible in a structure which is reasonable for content handling as the information accessible is generally unclear, amorphous or unstructured. Content mining is a sub field of information mining which goes for investigating the valuable data from the recorded assets. Content mining has three significant difficulties. They are high dimensionality, embraced remove measures, accomplishing quality bunches and improved classifier exactnesses.**

Grouping of archive is significant with the end goal of report association, rundown, subject extraction and data recovery in a proficient manner. At first, grouping is connected for upgrading the data recovery procedures. Recently, bunching strategies have been connected in the territories which include perusing the assembled information or in ordering the result given by the web indexes to the answer to the question raised by the clients. In this paper, we are giving an exhaustive review over the archive bunching.

Index Terms- Document Clustering, Term Frequency, Preprocessing, Stemming, Clustering Algorithms.

1. INTRODUCTION

Document clustering is an automatic clustering of text documents in to clusters so that documents in within a cluster have higher similarity and dissimilarity with documents in another cluster.

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, while maximizing inter-cluster distances A similarity calculation lies at the heart of document clustering approach.

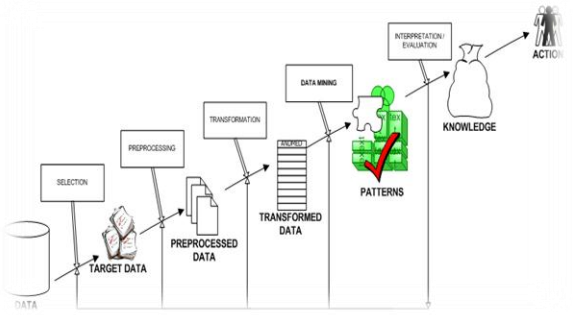

Figure 1: structure of data mining

The difference between clustering and classification is that the clustering is used in unsupervised learning but classification is not. In clustering, it is the division and framework of the information that will decide cluster membership.

Text clustering is the process of grouping similar documents into clusters. Text clustering is. accomplished by representing the documents as a set of terms of indexes associated with numerical weights. The goal is always to cluster the given text documents, such that they get clustered based on the similarity measures with a reasonable accuracy. During text clustering, the documents need to be preprocessed before analyzing the data. The dimensions of the vector that represent the documents need to be reduced. The following aspects can be considered as the most important issues that need to be looked into and decided upon for the purpose of text clustering.

- Document representation
- Suffix tree representation
- Analysis of similarity measures/distance criteria (Clustering)
- Clustering algorithms

2. 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

The Computer Forensics study only reports the utilization of algorithms that suppose that the quantity of clusters is famous and fixed a-priori by the client. Aimed at calming this assumption, which is often impractical in practical applications, a common method in other domains involves estimating the quantity of clusters from documents. Essentially, one stimulate dissimilar records partitions and then evaluate them with a comparative authority index in order to guess the best value for the quantity of clusters [2], [3], [14]. This job makes use of such approach, thus potentially facilitating the work of the specialist examiner—who in perform would hardly know the quantity of clusters a-priori.

Document clustering is the procedure of categorize manuscript document into a systematic cluster or collection, such that the documents in the similar cluster are similar whereas the documents in the other clusters are dissimilar. It is one of the very important courses of action in manuscript mining.

In 2009, Malay Pakhira shows a customized edition of the K-means algorithm that effectively eradicates this empty cluster problem. In fact, in the experiment done in this observe, this algorithm showed better presentation than that of traditional approach

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

In 2004, Crescenzi et al. [14] cited an method that robotically take out data from large data web sites. The "data grabber" interrogates a huge web site and infers a plan for it describing it as a directed graph with nodes. It elaborates classes of structurally similar pages and arcs representing links between these pages. After locating the classes of curiosity, a library of wrappers can be created, one per class with the assist of an external wrapper generator and in this way suitable data can be extracted

In 2003, Likas et al. [15] shows the universal K-means clustering procedure that build preliminary centers by recursively separating data space into rambling subspaces using the K-dimensional tree method. The cutting hyper plane used in this method is the plane that is vertical to the max variance axis resultant by (PCA). Division was accepted out as far as each of the leaf nodes possess less than a previous amount of data illustration or the predefined number of buckets has been produce. The preliminary midpoint for K-means is the center of statistics that are present in the concluding documents

3. CONCLUSION

In this paper, the emphasis is on Document Clustering which is ongoing innovation, we researched many existing calculations. As grouping assumes an extremely crucial job in different applications, numerous inquires about are as yet being finished. The up and coming developments are for the most part because of the properties and the qualities of existing strategies. This paper displays a prologue to the present report bunching idea alongside the strategies utilized for archive grouping. A basic audit of existing work done by creators on archive bunching in ongoing time is additionally exhibited 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, and C.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

[14] Crescenzi valter, Giansalvatore Mecca, Paolo Merialdo and Paolo Missier, "An Automatic Data Grabber for Large Web Sites", VLDB , pp. 1321-1324, 2004

[15] Likas, A., Vlassis, N. and Verbeek, J.J. "The Global k-means Clustering algorithm", Pattern Recognition , Vol. 36, No. 2, pp. 451-461, 2003.