IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 30-32

Digital Data Clustering

[1]Prof. Ashish Mohod, [2]Guddi Ramtekkar, [3]Nidhi Girde, [4]Manish Yadav, [5]Pranay Shinde

[1]Faculty of CSE Dept. of Priyadarshini J L College of engineering, Nagpur
[2-5]Students of CSE Dept. of Priyadarshini J L College of engineering, Nagpur

Abstract: *Today, we use all new technologies and technical methods in digital world to create huge digital documents so, the examination of such huge set of document is difficult and more important task. Document clustering is automatic organization of documents into clusters so that documents within a cluster have high similarity in comparison to documents in other clusters. Itis a widely studied problem in Text Categorization. It is the process of partitioning or grouping a given set of documents into disjoint clusters where documents in the same cluster are similar.The study of similarity measure for document clustering is not based on keywords generally domain based clustering is done. Our main objective is to improve the accessibility and usability of text mining for various applications. So, to do test document analysis time limit is an also major factor. So it's a not easy task for examiner to do such analysis in quick period of time. That's why to do the digital document analysis within short period of time, requires particular techniques to make such complex task in a simpler way. Such special technique called document clustering. So, clustering algorithms are of great interest. This document clustering analysis is very helpful for any document investigation to analyse the information of digital devices. Here we proposed a modified novel-k-representative algorithm which uses Jaccard distance measure for computing the most dissimilar k documents as centroids for k clusters. Our experimental results demonstrate that our proposed K-representative algorithm with Jaccard distance measure for computing the centroid improves the clustering performance of the simple K-means algorithm.The accuracy of clustering of documents has been improved by means of this modified novel-k-representative approach.*

I. Introduction:

In recent times digital technology especially in the computer world there is enormous increase in digital documents. So, extraction of relevant data from such vast set of digital document is much more important task for that we need to do digital data clustering and analysis.

1.1 Digital Data Clustering And Analysis

Digital data clustering and analysis is the branch of systematic document analysis process for investigation of matter found in digital devices interrelated to computer. Digital evidence equivalent to particular incident is any digital data that provides suggestion about incident. The important part of digital document process is to analyze the documents that present on suspect's computer. Due to increasing count of documents and larger size of storage devices makes very difficult to analyze the documents on computer. Normally, digital documents is the use of investigation and analysis technique to collect and protect evidence from exacting computing device in a way that is proper for presentation in as a evidence.

It also deals with the preservation, designation, extraction as well as certification of digital evidences .This is task of analyze tremendous number of files from computer devices. But in computer document process all the necessary information and files are stored in digital form. This digital information stored in computer devices has a key factor from an investigative point of view which handled as evidence in the court of law to prove what come to pass based on such evidences. Therefore collection of evidence from seized devices is also task of document examiner.

Digital evidence is defined as the information and data of fact-finding value that are stored on, received or transmitted by digital device. Such digital evidences needs to be accumulated from computer devices in order to confess the case in court of justice. So such digital proof have a great asset for the document examiner .So the key factor to improve such document analysis process requires document clustering technique The process of digital data clustering and analysis is shown is describe below.The Digital Document examination (DDE) process as defined by DDRWS. After determining items, constituents, and data related with the unpleasant incident (Identification phase), the next level step is to preserve the criminal scene by stop or prevent several actions that can harm digital information being collected (Preservation phase). Follow that, the next level step is collect digital information that might be related to the incident, for example copying files or recording network

traffic (Collection phase). Next step, the investigator conducts an in detail efficient search of evidences related to the incident being analysis such as filter, validation and pattern matching techniques (Examination phase) [1].

The examiner can put the evidence together and tries to develop theories concerning events that occurred on the suspect's computer (Analysis phase). In the examination phases investigators often utilize certain document tools to help analyzethe collection files and performs in detail systematic search for important evidence

II. Literature Review:

K.Nagarajan etal. [14] proposed ordinary bunching approaches endure with the adaptability of number of qualities dependent on which the grouping is performed. There are ways to deal with group information focuses with various characteristics yet endures with covering and different emphasis expected to perform bunching, likewise the measure figured for the variety of information focuses between group additionally won't be viable while doing with numerous properties. To beat this issue gave another chart based methodology which speaks to the connection between the information focuses and groups.

H.Chen etal. [15] demonstrate an outline of contextual analyses finished with connection to their COPLINK venture. The task's particular intrigue was the way data over-burden prevented the viable investigation of criminal and fear based oppressor exercises by law requirement and national security faculty. Their work proposed the utilization of information mining to help in settling these issues. In their report they characterize information mining with regards to wrongdoing and insight examination to incorporate substance extraction, bunching systems, deviation location, order, and ultimately string comparators.

G.Thilagavathi etal. [20] proposed PC report process is to analyze the records present in speculate's PC. Because of improve measure of reports and bigger size of storage room gadgets makes extremely hard to assess the archives on PC.

L.F.C.Nassif etal. [21] proposed a methodology that applies report grouping calculations for the archive investigation of PC gadgets. They showed a methodology via doing wide experimentation with six understood grouping calculations (K-mean, K-medoids, Single Link, Average Link, total Link and CSPA) connected to five genuine world datasets got from PC seized.

III. Proposed Methodology

Our objective will be firstly to collect information i.e. assembling the dataset. After this remove stop words and the unique words along with count from those data sets will be our next objective. Once the search keywords are input we will then perform the clustering using the modified k-means algorithm

3.1 Implemented Clustering Algorithm

Let's observe the special requirements for good document clustering algorithm:The document model should better preserve the relationship between words like synonyms in the documents since there are different words of same meaning.Relate a meaningful label to each final cluster is necessary.The high dimensionality of text documents must be reducing.So to accomplish this feature in our proposed system we heighten approach to improve document clustering in document analysis. For that we were implementing hybrid approach to accomplish this proposed approach. We implementing new text clustering algorithm such as modified k-means algorithm which will gives us the better clustering result .The main idea of modified k-means algorithm is to use the relative attribute frequencies of the clusters mode in the dissimilarity measures in the K-mode objective function [21].

It has been shown that modified k-means algorithm is very efficient. Due to the alteration proposed in forming representatives for clusters of categorical objects, the dissimilarity between a categorical object and the representative of a cluster is defined based on simple matching as follows.

3.1.1 Steps of modified k-means algorithm

i. Initialization and partition of Dataset randomly using file extension.
ii. Calculate centroid value Ci, one for each cluster.
iii. For each ci, calculate the dissimilarities d (Ci, Ql), l = 1, Reassign Ci to cluster C l (from cluster C6, say) such that the dissimilarity between ci and Ql is less. Update both Ql and Q6.
iv. Repeat Step (iii) if convergence standard are not meet. Otherwise stop

IV. Conclusion:

The implemented algorithm has following characteristics

- It performed clustering on .txt,.doc and .pdf files
- On large dataset it will take more time

- It will match with relevant data

This paper conclude that it is barely possible to get a more general algorithm, which can work the best in clustering all types of datasets like web ,txt etc. Thus we tried to implement novel text clustering algorithms which can work well in categorical or numerical datasets. The working of algorithm is described in implemented methodology, the modified k-means algorithm, suits the set of documents in which the required classes are related to each other and we require a strong basis for each cluster. Thus, this algorithm can be very effective in applications like a search engine for a particular keyword.

References:

[1]. M. R. Clint, M. Reith, C. Carr, and G. Gunsch, an Examination of Digital Forensic Models, 2003.
[2]. https://en.m.wikipedia.org/wiki/information_retrieval.
[3]. A. Kao and S. R. Poteet, "Natural Language processing and Text mining", Springer Verlag London Limited, 2007.
[4]. https://en.m.wikipedia.org/wiki/informtion_extraction.
[5]. Y. Zhao, G. Karypis, and U. M. Fayyad, "Hierarchical clustering algorithms for document datasets", Data Mining Knowledge Discovery, vol.10, 2005.
[6]. Aggarwal, C. C. Charu, and C. X. Zhai, Eds., "Chapter 4: A Survey of Text Clustering Algorithms", Mining Text Data, New Springer, York, 2012.
[7]. D.Napoleon and P.Ganga Lakshmi, "An Enhanced K-means Algorithm to Improve the Efficiency Using Normal Distribution Data Points", International Journal on Computer Science and Engineering (IJCSE), vol. 02, issue 07, 2010.
[8]. G.Gandhi and R. Srivastava, "Analysis and implementation of modified K-medoids algorithm to increase scalability and efficiency for large dataset", International Journal of Research in Engineering and Technology (IJRET), Vol.03 Issue-06, Jun-2014.
[9]. K. Murugesan and J. Zhang, "Hybrid Bisect K-Means Clustering Algorithm", Department of Computer Science, University of Kentucky Lexington, USA.
[10]. R. Mundhe, A.Maind and R.Talmale, " Information Retrieval Using Document Clustering for Forensic Analysis" International Journal of Recent Advances in Engineering & Technology (IJRAET), Vol.2, Issue -5, 2014.
[11]. B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver, "Exploring forensic data with self-organizing maps", Proceedings IFIP Int. Conf. Digital Forensics, 2005.
[12]. W.Liao, Y.Liu and A. Choudhary, "A Grid-based Clustering Algorithm using Adaptive Mesh Refinement", Appears in the 7th Workshop on Mining Scientific and Engineering Datasets 2004.
[13]. K. Stoffel, P. Cotofrei, and D. Han, "Fuzzy methods for forensic data analysis", IEEE International Conference Soft Computing and Pattern Recognition, 2010.
[14]. K. Nagarajan and Dr. M. Prabakaran ,"A Relational Graph Based Approach using MultiAttribute Closure Measure for Categorical Data Clustering", The International Journal Of Engineering And Science (IJES) ,Vol. 3, 2014.
[15]. H. Chen, W. Chung, Y. Qin, M.Chau, J.Xu, G.Wang, R. Zheng, and H. Atabakhsh. "Crime data mining: an overview and case studies", Proceedings of the 2003 annual national conference on Digital government research ,Digital Government Research Center, 2003 pages 1–5.
[16]. G. Forman, K. Eshghi, and S.Chiocchetti, "Finding similar files in large document repositories", KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM, New York, NY, USA, 2005.
[17]. A.B.Schatz and G.Mohay, "A correlation method for establishing provenance of timestamp in digital evidence", Digital Investigation, volume 3, supplement1, 6th Annual Digital Forensic Research Workshop, 2006, pp. 98–107.
[18]. T. Abraham, "Event sequence mining to develop profiles for computer forensic investigation purposes", ACSW Frontiers '06: Proceedings of the 2006 Australasian workshops on Grid computing and e-research, Australian Computer Society, Australia, 2006, pp. 145–153.
[19]. J.G.Clark and N.L.Beebe, "Digital forensics text string searching: Improving information retrieval effectiveness by thematically clustering search results", In Digital Investigation, vol.4, 6th Annual Digital Forensic Research Workshop, 2007, pp. 49–54.
[20]. G. Thilagavathi and J. Anitha, "Document Clustering in Forensic Investigation by Hybrid Approach", International Journal of Computer Applications, vol. 91, April 2014.
[21]. L.F.D.C Nassif and E.R. Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection", IEEE Transactions on Information Forensics and Security, vol.8, issue 1, January 2013.