International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 3, Issue 11, ISSN No. 2455-2143, Pages 30-34
Published Online March 2019 in IJEAST (http://www.ijeast.com)

PRAGMATIC APPROACH FOR DIGITAL DATA CLUSTERING USING I-DELEGATE ALGORITHM

Mr. Ashish P Mohod
Department of CSE
PJLCE, Nagpur, Maharashtra, India

Mr. Sagar Tete
Department of CSE
PJLCE ,Nagpur, Maharashtra, India

***Abstract*— Now a day, we use all new scientific and technical process in digital world to create huge digital documents so, the analysis of such huge set of document is really difficult and more important work. Document clustering should be an automatic process where documents are partition into clusters having high likeness based on input term. It is a popularly studied problem in text classification but generally the study of analogy measure for document clustering is not based on keywords normally domain based clustering is done. Our main aim is to improve the accessibility, scalability and usability of text mining for various applications. So, to do text document analysis within a stipulated time is a key factor. So it's not an easy work for examiner to do such analysis in quick period of time. That's why to do the digital document analysis within less period of time, requires particular techniques to make such difficult task in a simpler way. Such special technique called document clustering. So, clustering algorithms are of great advantage. Here we proposed a I-Delegate algorithm which uses Jaccard distance measure for computing the most dissimilar k documents as centroids for k clusters. Our pragmatic experimental results display that our implemented I-delegate 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 I-delegate approach due to synonym identification and delegates that synonym to clustering process along with near index approach for better result.**

***Keywords*— Documents clustering; I-delegate Algorithm; Jaccard similarity coefficient; K-Means Algorithm**

I. INTRODUCTION

In recent times digital science especially in the computer era has tremendous increase in digital documents. So, extraction of desired related data from such enormous set of digital document with accuracy and within less time is much more crucial work and for that we need to do digital data clustering and analysis.

A. Digital Data Clustering and Analysis

Digital data clustering and analysis play an important role in examination of matter found in digital devices associated to computer. The important part of digital document process is to analyze the documents that present on suspect's computer in case of digital forensic analysis. Due to increasing count of documents and larger size of storage devices makes very difficult to analyze the documents on computer. Normally, digital documents are use for investigation and inspection technique to collect and guard evidence from exacting computing device in such a way that it is suitable for displaying as evidence.

It also deals with the conservation, classification, extraction as well as certification of digital evidences. This is task of analyze enormous number of files from computer devices. But in computer document process all the essential data and files are stored in digital form. This digital information stored in computer devices has a key factor from an examination point of view which directed as evidence in the court of law to prove what come to pass based on such evidences. Therefore set of evidence from absorbed devices is also task of document examiner.

Digital evidence is defined as the information and data of fact-finding assess that are gathered on, received or transmitted by digital device. Such digital evidences needs to be collect from computer devices in instruct to admit the case in court of equity. So such digital attestation has a great advantage for the document auditor. So the key factor to improve such document analysis process requires document clustering technique. The method of digital data clustering and analysis is shown is characterize below. The Digital Document examination (DDE) process as defined by DDRWS. After determining elements, contents, and data related with the offensive incident (Identification phase), the next level step is to defend the criminal scene by stop or defend several act that can damage digital information being gathered. Follow that,

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 3, Issue 11, ISSN No. 2455-2143, Pages 30-34
Published Online March 2019 in IJEAST (http://www.ijeast.com)

the next level step is collect digital information that might be associated to the incident, for example copying files or recording network traffic. Next step, the investigator conducts an in detail effective search of evidences related to the event being analysis such as filter, validation and pattern matching techniques [1].

The examiner can put the evidence concurrently and tries to develop ideas concerning events that appear on the suspect's computer. In the examination phases investigators generally apply certain document device to help analyze the set files and act in detail efficient search for important evidence

II. LITERATURE REVIEW

Chen H. et. al. [1] demonstrate an outline of contextual analyses finished with connection to their COPLINK proposed method. The task's particular intrigue was the means by which data over-burden obstructed the compelling investigation of criminal and fear based oppressor exercises by law implementation and national security work force. Their work proposed the utilization of information mining to help in understanding these issues. In their report they characterize information mining with regards to wrongdoing and insight examination to incorporate substance extraction, bunching methods, deviation identification, order, and in conclusion string comparators.

Nassif L.F.C et. al. [6] proposed a methodology that applies archive grouping calculations for the report investigation of PC gadgets. They showed a methodology via doing wide experimentation with six surely understood bunching calculations (K-mean, K-medoids, Single Link, Average Link, total Link and CSPA) connected to five genuine world datasets acquired from PC seized.

Mascarnes S. et al [9] proposed a novel document clustering model that allows a surveyor to cluster semantically related documents stored on a suspect's digital devices with the help of subject suggestions initially provided to him. Surveyor provided subject suggestion word or bunch of words improves the accuracy and speeds up the process of searching the evidences in forensic analysis.

Vidhya B. et. al [10], studied a variety of text clustering and document clustering technique for forensic digital analysis. To improve digital forensic analysis they proposed K-mean algorithm and ant colony optimization algorithm. This was very important among swarm intelligent algorithm. K-mean was one of the simplest algorithms for document clustering which was efficient to giving better clusters form huge amount of datasets.

Nagarajan K. et. al. [12] focus on conventional clustering approaches that suffer with huge number of attributes base on which the clustering was performed and thus overlap and numerous iteration required to perform clustering,. To defeat this issue they provided a graph based approach which represents the relation between the data points and clusters. They were also used threshold values to select the data point which are closure to each other here the threshold value is selected based on number of attributes the data point has. There proposed method produces better results compare to other approaches discussed in that period of time and they have been use there method with various data sets.

Thilagavathi G. et. al. [13] proposed PC archive process which is used to look at the records present in suspects computer. In this approach because of improve measure of reports and bigger size of storage makes extremely hard to assess the records on PC. Here the challenge is multithreaded approach to search entire HDD using thread based approach.

III. IMPLEMENTED METHODOLOGY

Our main objective is to find required dataset i.e. identifying the desire dataset. The next step is preprocessing which help to reduce the noise, computational complexity and dimensionality. Preprocessing involves tokenization, removal of stop words and Stemming process. After this identify the most significant words using TF-TDF calculation and then apply K-Means and I- Delegate algorithm to get resultant clusters which are forms according to the input query of investigator. It will also show the performance analysis of K-means and I-delegate algorithm. It can be used for application such as forensic data analysis if the used dataset is related to crime.

A. Implemented Clustering Algorithm

Let's follow the special demands for good document clustering algorithm: The document model should better maintain the relationship between words like synonyms in the data set since there are different words of same meaning. Concern a meaningful label to each final cluster is essential. The high dimensionality of text documents must be decrease. So to achieve this feature in our implemented system we were implementing hybrid I-delegate approach to accomplish this. This new I-delegate algorithm gives us the better clustering result. The principal idea of hybrid I-Delegate algorithm is to use the relative assign frequencies and Jccard function to create cluster of similar type along with ranking of clusters. Ranking can give investigator an overall view of dataset files, which is beneficial for examiner to grasp the most important information in a short time.

It has been shown that new I-delegate algorithm is very productive. Due to the sequence proposed in forming representatives for clusters of absolute objects, the difference between a categorical object and the representative of a cluster is defined based on simple matching as follows.

B. Steps of I-delegate algorithm

i. Initialization and partition of dataset randomly based on dataset file extension.
ii. Perform preprocessing to reduce computational complexity and noise

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 3, Issue 11, ISSN No. 2455-2143, Pages 30-34
Published Online March 2019 in IJEAST (http://www.ijeast.com)

iii. Find term frequency using TF-TDF calculation. (TF-IDF is mainly used to determine the weight of each term in information retrieval and text mining.)

iv. for every input query, generate expansion vector by analyzing related words for each input query in Extended Synonym List (ESL). (Extended Synonym List is an event specific list which contains dataset related synonyms terms which can be used by specific domain and it will not present in dictionaries).

v. Calculate centroid value Ci, one for each cluster using similarity measure of the search words with dataset file using jaccard similarity coefficient. Following formula is used to find mean distance as follows:

$$J(X,Y)= |X \cap Y| / |X \cup Y|$$

vi. For each ci, calculate the dissimilarities d(Ci,Rl), l = 1, Reassign Ci to cluster C l (from cluster C5, say) such that the dissimilarity between ci and Rl is less. Update both Rl and R5

vii. Repeat Step (iii) if convergence standard are not meet. Otherwise stop

IV. Experiment Results and Discussion

This section covers the results and the discussion of the implemented work. Here we include the output and analysis of implemented work. We have produced result for crime related dataset using two algorithms. The first algorithm is existing algorithm K-mean and second one hybrid algorithm that is I-delegate proposed by us to be better. The results for the I-delegate algorithm have been generated to compare them with the existing algorithm.

A. Experimental Evaluation

We were using MYSQL 5.6 database for experimental setup to maintain and update datasets. Here we also discuss dataset on which we practically apply text mining I-delegate algorithm and its parameter. On the basis of that we decide efficiency of our proposed system. We have used crime related dataset for performing clustering result .In which we collected data related to crime investigation in which we have taken 250 documents which contain some .txt files,. pdf files , .doc and .docx files in that dataset these files are related to laws, corruption, crime , thief etc. Above Figure 1 shows bar chart of no of retrieved documents after applying two clustering algorithms and from that analysis we depicted that which clustering algorithm is more efficient.

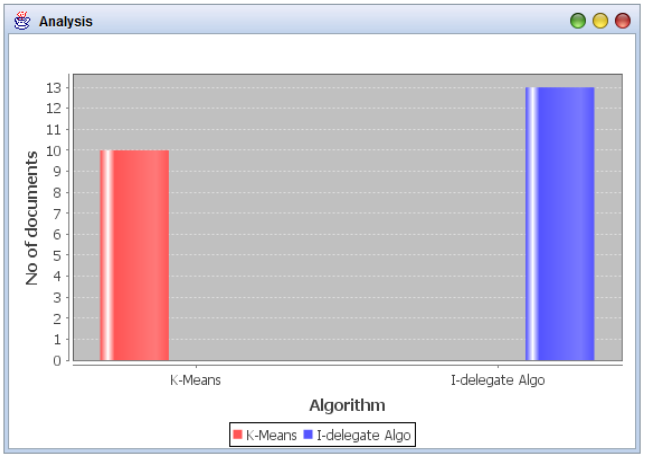

Fig. 1. Number of Retrieved Documents

In this analysis red bar shows K-mean algorithm result it retrieved 10 documents, blue bar shows I-delegate algorithm result it retrieved 13 documents which are related to crime query word in whole dataset which contain 200 files if we overlook result the I-delegate will gives best result than K-mean algorithm.

B. Result for K-Means Algorithm:

Table1 shows results of K-mean algorithm in this we calculate the result of precision and recall for finding the accuracy of K-mean algorithm using formulas where as Local, corruption, law, legal entered as search queries for finding the evidence.

Table -1: Precision Recall Result of K-Mean Algorithm For Sample Keywords

K- Mean Result					
Keywords/Parameter	**Crime**	**Local**	**Corruption**	**Law**	**Legal**
Total no of relevant result in system	45	8	10	38	15
No of retrieved records	28	4	3	30	6
No of relevant records	26	3	2	28	4
No of relevant record not retrieved	17	4	7	8	9
No of irrelevant record retrieved	2	1	1	2	2
Precision	0.92	0.75	0.66	0.93	0.66
Recall	0.60	0.42	0.22	0.77	0.30

C. Result for I-delegate Algorithm

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 3, Issue 11, ISSN No. 2455-2143, Pages 30-34
Published Online March 2019 in IJEAST (http://www.ijeast.com)

Table-2 Depicts Results of I-Delegate Algorithm

I-Delegate Algorithm Result					
Keywords/ Parameter	**Crime**	**Local**	**Corruption**	**Law**	**Legal**
Total no of relevant result in system	45	8	10	38	15
No of retrieved records	40	6	8	35	10
No of relevant records	38	5	5	32	8
No of relevant record not retrieved	5	2	2	6	5
No of irrelevant record retrieved	2	1	3	3	2
Precision	0.95	0.83	0.71	0.91	0.80
Recall	0.84	0.71	0.62	0.84	0.61

This table shows the result of precision and recall for finding the accuracy of I-delegate algorithm using precision and recall formula. Table 2 shows the results obtained by taking five sample keywords i.e. crime, local, corruption, law, legal entered as search queries for finding the evidence. To test the scalability of the proposed system, experiments have been conducted on demo dataset consisting of 250 documents related to crime. The result is shown in figure 2.

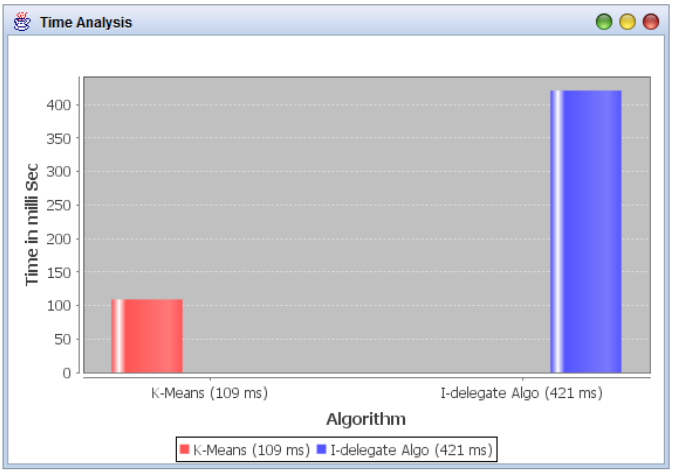

Fig. 2 Shows time analysis of K-Means and I- delegate algorithm

Further, the files in the dataset are duplicated so that the scalability can be measured starting from 10 documents going up to 250 documents. The implemented system consists of two main phases: Preprocessing and Clustering. The objective of this scalability experiment is to measure the runtime of each phase to ensure it does not grow proportionally as the data set sizes increases.

Table -3 Shows Scalability of Implemented Algorithm

Number of documents (Samples)	**Time(ms)**	
	K-Means	**I-delegate**
10	109	421
25	140	588
50	202	687
75	244	705
100	282	795

IV.CONCLUSION

This paper conclude that it is almost possible to get a more common algorithm, which can act the best in clustering all types of datasets containing files type like , .txt ,.doc and .pdf files etc. Thus we tried to implement I-delegate text clustering algorithms which can act well in absolute or digital datasets. The overall working of algorithm is described in implemented methodology, the hybrid I-delegate algorithm, suits the dataset in which the required family are connected to each other. The main advantage is it retrieves information more effectively but it takes more searching time due to synonym search sot it provides efficient way of representing and visualizing the documents. Thus, this algorithm can be very effective in applications like a search engine for a particular query term. Finally we would conclude that though many algorithms have been proposed for clustering but it is still an open problem and looking at the rate at which to perform clustering on large documents, it will become an essential part of the application

V. REFERENCE

[1] Chen H., Chung W., Qin Y. et. al (2003) "Crime data mining: an overview and case studies", Proceedings of the annual national conference on Digital government research ,Digital Government Research Center, 2003 pages 1–5.

[2] Zhao Ying, Karypis George, and Fayyad U. et.al (2005), "Hierarchical clustering algorithms for document datasets", Data Mining Knowledge Discovery, *Volume 10*, Issue 2, pp. 141–168

[3] Schatz A. and Mohay G. (2006), "A correlation

method for establishing provenance of timestamp in digital evidence", Digital Investigation, volume 3, supplement1, 6th Annual Digital Forensic Research Workshop, pp. 98–107.

[4] Clark J. and Beebe N. (2007), "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, pp. 49–54.

[5] Napoleon D. and P. Ganga Lakshmi (2010), "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 pp. 2409-2413

[6] Nassif L.F.D.C and Hruschka E. (2013), "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection", IEEE Transactions on Information Forensics and Security, vol.8, issue 1, pp 46-54.

[7] Niwattanakul S, Singthongchai J. , Naenudorn E. and Wanapu S, (2013) "Using of Jaccard coefficient for keywords similarity", International Multi Conference of Engineers and Computer Scientists (IMECS), vol.1, Hong Kong, pp. 13-15.

[8] Gandhi Gopi and Srivastava Rohit (2014), "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, pp. 2515-2525

[9] Mascarnes S. and Gomes J. (2014), "Subject based Clustering for Digital Forensic Investigation with Subject Suggestion", International Journal of Computer Applications (0975 – 8887) Vol. 102 – No.11

[10] Vidhya B. and Priya Vaijayanth R. (2014), "Enhancing Digital Forensic Analysis through Document Clustering", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue 1, pp.48-57

[11] Umale B. and Nilav M, "Survey on Document Clustering Approach for Forensics Analysis", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 pp. 3335-3338

[12] Nagarajan K. and Prabakaran M.(2015), "A Relational Graph Based Approach using Multi-Attribute Closure Measure for Categorical Data Clustering", The International Journal Of Engineering And Science (IJES) ,Vol .*7*, no. 2, pp. *3*-12

[13] Thilagavathi G. and Anitha J. (2016), "Document Clustering in Forensic Investigation by Hybrid Approach", International Journal of Computer Applications, vol. 91, Issue 4, pp. 141–168