Novel Clustering Algorithm for Text Document

[1]Shrejal Bhawsar, [2] Ajay Phulre

Department of Computer Science & Engineering ,
Shri Balaji Institute of Technology and Management, Betul, MP, India
Email – [1] shrejalbhawsar16@gmail.com, [2] aphulre@gmail.com

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. It is used to measure similarity between documents and grouping similar documents together. 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 analyze the information of digital devices. Here we proposed Novel Clustering approach to attain efficient document clustering for digital document analysis on the basis of interested keyword. The accuracy of clustering of documents has been improved by means of this Novel Clustering approach.*

Key Words: *Document Clustering, Digital Document Analysis, Examination, Data Mining.*

1. INTRODUCTION:

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

1.1 Digital Document Analysis

Digital document 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 examine 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. Usually, 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 an evidence.

It also deals with the preservation, identification, extraction as well as documentation of digital evidences .This is task of analyze enormous number of files from computer devices. But in computer document procedure all the essential 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 treated as evidence in the court of law to prove what occurred based on such evidences. Therefore collection of evidence from seized devices is also task of document examiner.

Digital proof is defined as the information and data of investigative value that are stored on, received or transmitted by digital device. Such digital evidences needs to be collected from computer devices in order to confess the case in court of law. 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 document analysis is shown is describe below. The Digital Document examination (DDE) process as defined by DDRWS. After determining items, components, and data related with the unpleasant incident (Identification phase), the next level step is to preserve the crime 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 examine the collection files and perform an in detail systematic search for significant evidence

2. LITERATURE REVIEW:

K.Nagarajan etal. [14] proposed conventional clustering approaches suffer with the scalability of number of attributes based on which the clustering is performed. There are approaches to cluster data points with multiple attributes but suffers with overlapping and multiple iteration needed to perform clustering, also the measure computed for the variation of data points between cluster also will not be effective when doing with multiple attributes. To overcome this problem provided a new graph based approach which represents the relation between the data points and clusters.

H.Chen etal. [15] show an overview of case studies done with relation to their COPLINK project. The project's specific interest was how information overload hindered the effective analysis of criminal and terrorist activities by law enforcement and national security personnel. Their work proposed the use of data mining to aid in solving these issues. In their report they define data mining in the context of crime and intelligence analysis to include entity extraction, clustering techniques, deviation detection, classification, and lastly string comparators.

G.Thilagavathi etal. [20] proposed computer document process is to examine the documents present in suspect's computer. Due to enhance amount of documents and larger size of storage space devices makes very difficult to evaluate the documents on computer.

L.F.C.Nassif etal. [21] proposed an approach that applies document clustering algorithms for the document analysis of computer devices. They illustrated an approach by carrying out wide experimentation with six well known clustering algorithms (K-mean, K-medoids, Single Link, Average Link, complete Link and CSPA) applied to five real world datasets obtained from computer seized.

3. IMPLEMENTED METHODOLOGY:

Our objective will be firstly to collect information i.e. gathering 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 Novel-Clustering algorithm.

3.1 Implemented Novel-Clustering Algorithm

Let's observe the special requirements for good document clustering algorithm: The document model should better conserve 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 achieve this feature in our proposed system we enhance 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 Novel-Clustering algorithm which will gives us the better clustering result .The main idea of Novel-Clustering 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 Novel-Clustering algorithm is very efficient. Due to the modification 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 Novel-Clustering 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 criteria are not meet. Otherwise stop

4. EXPERIMENTAL RESULTS AND DISCUSSION:

For experimental analysis we used sample dataset containing bunch of .txt ,.doc and .pdf files. This dataset is used as an input for clustering purpose. Following fig1dipicts the working of implemented Novel approach where we should specify the keyword on which we want to perform clustering using k-mediods and Novel clustering algorithm after pressing make cluster button both k- mediods and novel algorithm searches the documents that contain specified keyword and its clearly shows that novel algorithm find more documents containing the specified file keyword as compared to k-mediods algorithm because in novel approach we not only used specified keyword for searching but we also find the auxiliary information of specified word and then perform clustering so search operation in novel

algorithm is more wide than that of k-mediods algorithm. Figure 2 indicate the performance bar chart where novel algorithm required more time that of k-mediods .

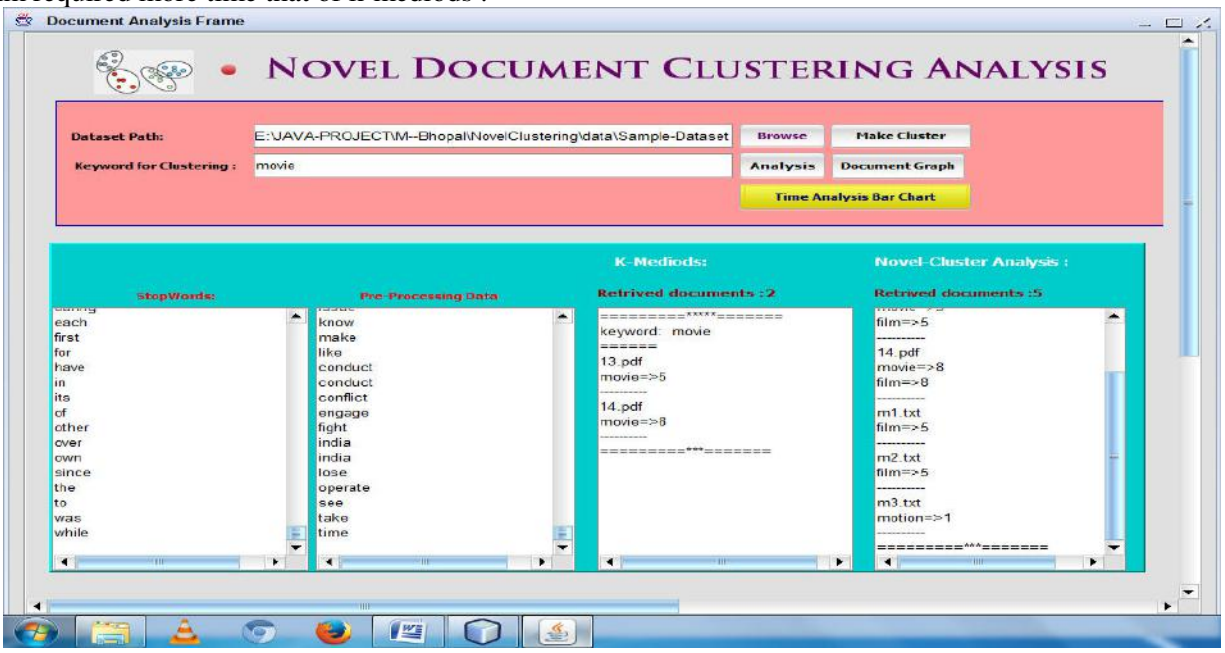

Fig1: Text Document Cluster Analysis Frame

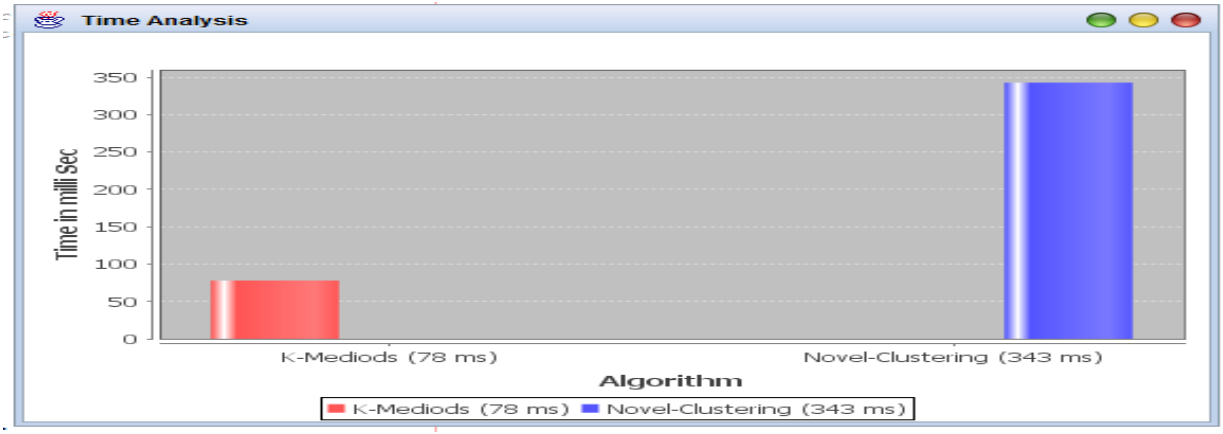

Fig2: K-Mediods and Novel Clustering Algorithm Time Analysis Frame

5. CONCLUSION:

The implemented algorithm has following characteristics

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

This paper conclude that it is hardly 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 Novel-Clustering 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.