Clustering Forensic Documents to Find Relevant Data Set

Neeraj Gadiya[1] Pankaj Jadhav[2] Vaibhav Gandhi[3] Rahul Thombare[4] Snehal Dongre[5]
[1,2,3,4,5]Department of Information Technology
[1,2,3,4,5]SGR Education's Foundation G. H. Raisoni College of Engineering & Management Chas-Ahmednagar, Maharashtra, India

Abstract— In forensic analysis, large number of data, documents and files are usually examined. Much of the data in those files consists of random data or in unstructured text, whose analysis by computer examiners is difficult to be performed and also a time consuming approach. Automated methods of analysis are very useful in such conditions. Considering this approach we can implement a system which will cluster the document set into different clusters using (K-means) and similarity measures. Algorithms that can cluster documents can prove very useful in such conditions the system will provide the forensic analyst with a clustered set of documents along with relevant set of documents from that particular cluster and that too in less time and with more accuracy.

Key words: Clustering Forensic, Relevant Data Set

I. INTRODUCTION

Clustering can be elaborated, as bounding the similar type of data into one group. Clustering algorithms are typically used for arranging data which is similar in context with their text contents. Forensic data analysis, process in which there is little or no prior ideas about the data .Technically forensic datasets consist of unlabelled objects such as the classes or categories of documents that can be retrieved are unknown .This is the case in which many applications of Computer Forensics and analyst is left with very little in their hands. As per assumption that labeled datasets could be available from previous analyses, there is almost no hope that the same classes would be still valid for the new files- data, obtained from other computers and associated to different investigation processes. It is just that the new data sample will appear from a different collection all together. In this such state, the use of similarity algorithms which can form prior clusters using a sample dataset and then clustering algorithms, finds the relevant set. Hence after this, analysis process is performed by the expert examiner. The reason behind clustering algorithms is that objects within a valid cluster are more similar to each other than they are to objects belonging to a different cluster .Thus,(once a group of similar data is formed) i.e. Cluster is formed, as per concerned with when the clusters are formed the analyst might initially focus on reviewing representative documents from the obtained from more relevant set of clusters. Then, after this primary analysis, he may eventually decide to check other documents from each less relevant cluster. By doing so, one can avoid the difficult task of examining all the documents (individually) and provide the analyst with a very meaningful information within a very short span.

II. PROPOSED SYSTEM

The aim of proposed system is to provide relevant set of cluster to the forensic data analyst. Clustering is performed based on the sample data set not on the basis of documents similarity. So an analyst gets a set of relevant and less relevant cluster as output. Similarity Measures are used to form prior clusters and then the K-Means approach which increases the efficiency in cluster forming approach. Analyst can intelligently select his sample data set depending on type of investigation being performed and thus get a better cluster as output.

Even if system is processing document set this is similar all together still it will end up in formation of different clusters as cluster formation depends on sample data set not on the similarity between documents under process. The case in which data set is appropriate system will accurately cluster the documents in relevant and less relevant cluster.

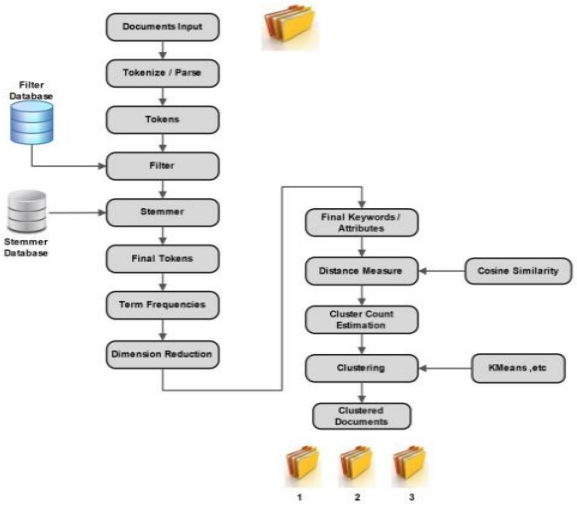

Fig. 1: Proposed System Architecture

III. METHODOLOGIES USED

A. *Porter Stemmer*

Porter stemmer algorithm is used in our proposed system. Which is used for pre-processing of data.ie these algorithms eliminates the stop words and also finds the root word.

B. *Term Frequency*

Term frequency is a weighting scheme that refers to the assignment of weight to each term in the document that depends on the number of occurrences of the term in that document as we know cluster is group of data of similar type of data. And huge data is obtained, from the document and files at time of analysis so in order to calculate occurrences of words appeared in file (document) term freq is used Term frequency is denoted by (tf).formula for term frequency is as follows

>>TF= total occ_words / total words.

C. *Cosine Similarity*

This algorithm calculates the score of similarity of words in the document using these score initial clusters are formed. Formula for the calculating the similarities are as follows.

>>Cosine Similarity =X * Y / square root of X^2 + Y^2.

D. K-Means

K-MEANS is most important algorithm used in proposed system. K-means is popularly used for cluster analysis in data mining. The main objectives of k-means is to partition the initial cluster into sub cluster which are less relevant and more relevant

IV. IMPLEMENTATION

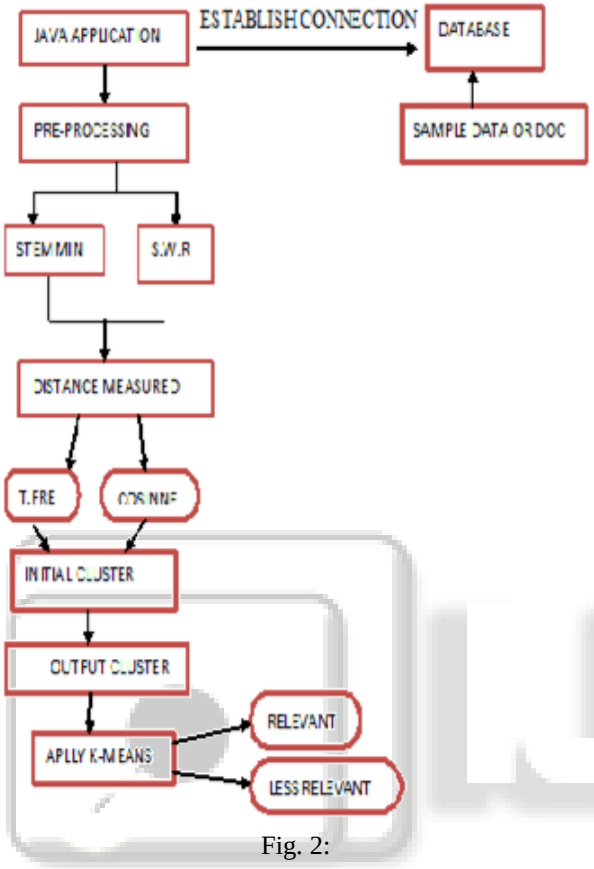

Fig. 2:

- Clustering of set is done on the basic of similarity between the documents.
- When a total distinct set of dataset is under examination system will end up forming a very large number of clusters.
- Even situation in which many un-relevant set of documents have same textual data the system will form a huge cluster of those documents.
- Thus this approach provides a easy examination approach for the analyst team but fails in some major situation which are more found situation at the time of forensic analysis.
- Also the K-means approach requires that the number of clusters to be formed should be fixed in such situation if user selects less value of K then system will form very huge clusters.
- So the analyst requires to take intelligent decisions at the time of formation of clusters.
- Clustering is performed based on the sample dataset not on the basic of documents similarity. So an analyst gets a set of relevant and less relevant cluster as output.
- Similarity Measures are used to form prior clusters and then the K-means approach which increases the efficiency in cluster forming approach.
- Analyst can intelligently select his sample dataset depending on type of investigation being performed and thus get a better cluster as output.
- Even if system is processing document set which is similar all together still it will end up in formation of different clusters as cluster formation depends on sample dataset not on the similarity between documents under process.
- The cases in which dataset is appropriate system will accurately cluster the documents in relevant and less relevant clusters.
- The K-means approach which requires fixed k as input will not be a problem because the clusters as output is always fix that is 2 clusters Relevant and Less Relevant Cluster.
- Less Time consuming approach for analyst team

V. CONCLUSIONS

Forensic analysis on documents is performed by the analyst which helps them to find some hidden facts or knowledge base from which helps them in investigation process. But this approach is carried out manually which is a very time consuming approach and prone to human errors. Thus a system can be developed using a systematic similarity based and clustering approach to cluster the large collection of forensic document set and find relevant set from that collections. This approach will not only save precious time of forensic analyst but also provide with useful information within a very quick time period.

ACKNOWLEDGMENT

We would like to thank all the professors of Information Technology Department of SGR Education's Foundation G. H. Raisoni College of Engineering, Chas-Ahmednagar. We are indebted to Prof. Miss Dongre Snehal our project guide, Prof. Mr. S.Kothari & Prof .Miss Mahajan Jagruti who were very generous in providing us with technical-support, material and otherwise. Her invaluable suggestion and time have helped in making this project possible.

REFERENCES

[1] J.F Gantz. D. Riesel, C. Chute, W. Schlichting, J. McArthur,S.Minton, I. Xheneti,A. Tocheva, and A.Manfrediz, "The expanding digital universe: A forecast f worldwide information growth through 2010, '' inf.data vol 1.pp. 1+21,2007.
[2] B. S. Everett, S. Landau, and M. Lees', Cluster Analysis. London, U. K: Arnold, 2001.
[3] K. Jain and R. C. Dubes, Algorithms for Clustering Data. Engle-wood Cliffs, NJ: Prentice-Hall, 1988.
[4] L. Kaufman and P.Rousseeuw, Finding Groups in Gata: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.
[5] R, Xu and D. C. Wunsch, II, Clustering. Hoboken, NJ: Wiley/IEEE Press, 2009.
[6] Strehl and J. Ghosh,"Cluster ensembles: A knowledge reuse frame-work for combining multiple partition," J. Mach. Learning Res., vol. 3, pp. 583-617, 2002.
[7] E. R. Hruschka, R. J. G. B. Campello, and L. N. de Castro, "Evolving clusters in gene-expression data," Inf. Sci., vol. 176, pp. 1898-1927, 2006.

