Mining and Investigation of Criminal Records from Digital Devices

Mohammad Azmat Javed[1] Siddhant Jaiswal[2]

[1,2]Department of Computer Science & Engineering

[1,2]G. H. Raisoni College of Engineering, Nagpur (M.S), India

Abstract— This paper presents the methods of document clustering for forensic investigation purpose. Lots of the data present in those files consists of irregular text and there is a necessity to make those texts structured. The analysis of irregular texts is difficult to be performed by computer examiners. For that there is a need of automated methods of analysis. Document clustering can be used for forensic analysis of digital devices in police investigations. Document clustering is very useful in digital investigation. The clusters of either relevant or irrelevant documents can facilitate computer examiners to efficiently focus on the most relevant documents instead of inspecting all of them. Adaptive bisecting k-means algorithm is proposed to cluster the document.

Key words: Document Clustering, Text Mining, Forensic Analysis, Digital Devices

I. INTRODUCTION

It is the world of technology and new innovations. The use of digital devices is increasing rapidly. The amount of information within the digital world enlarged and it continues to grow exponentially. This massive amount of data has direct impact in Computer Forensics. Computer forensics is a method that encompasses the elements of law and computer science to collect and examine data from computers so that they can be used as evidence in court. In order to investigate it is required to examine hundreds and thousands of files per computer. It is hard task to analyze and interpret the data. For that there is a need of an automated method those are used in machine learning and data mining.

The clustering algorithms prove to be good in order to find particular type of patterns from the textual documents that are obtained from the systems which are seized during police investigation. It also improves the analysis process done by experts.

Clustering algorithms are used to cluster the data into similar type of documents. In the case of computer forensics document clustering algorithms would help the examiners to focus on the specific type of the records which are being examined.

In computer forensic analysis document clustering can be very useful because it groups the information which is related to each other. So in case of investigation of criminal records the clustering algorithm will group the information/data into particular type of pattern so that the examiner will only analyze those parts of document which are related to their investigation purpose [1].

Clustering algorithms [2] also find out the relevant data from the little analysis or without any prior knowledge.

To cluster the data Adaptive Bisecting k-means algorithm is proposed. Adaptive Bisecting k-means algorithm gives better clusters, faster output and no blank clusters if data is available. Adaptive Bisecting k-means algorithms shows best result because it decides the number of clusters at run time.

II. BACKGROUND

Commercial forensic toolkits are very expensive. And not every investigation team has much fund to buy these toolkits. So there is an option to use open source toolkits but this has less functionality and required more technical skills. Computer devices which are seized in the digital forensic investigation process consist of random text. The analysis of those texts is difficult to be performed and time factor is also very important during the investigation process. So there is a need of a method which distinguishes the information either relevant or irrelevant for the investigation point of view. So in that case clustering technique can be used to gain the desired purpose.

Researchers have used different methods and algorithms in order to investigate textual information from digital devices. In [3], they show the method to find out the textual evidence from the digital devices. Emails, internet browsing history, instant messaging are described as an example of evidence.

In [4], they give several tools for computer forensic analysis and demonstrated how unsupervised learning neural network model i.e. the self-organizing map (SOM) can help computer forensic investigators in decision making. With the help of SOM the process of conducting the analysis is easier during a computer investigation. They mainly focus on self-organising map (SOM) – it will be helpful for examiner to identify all the data present in the digital devices and locate the specific type of record.

III. PROPOSED METHOD

Proposed method consists of different modules. First module is data set collection. In the second module preprocessing is done on the collected data set. Adaptive Bisecting k-means algorithm is implemented in the third module. In the fourth module forensic analysis has been performed on the clustered data obtained from the third module.

A. Data Set:

Data set is collected from Denver Open Data Catalog. It is a crime data set. This data set contains the criminal offenses in the city and country of Denver.

B. Extract Process:

In this process the field which data set is containing is extracted. As in the data set which is used in this paper consist of different number of fields like Offense ID,

Offense Code, Offense Type, Location, Date of Incident, Date of Report. Because of this it becomes easy for examiners to know which types of records are present in the data set which is being examined.

C. Implementation of Adaptive Bisecting K-Means Algorithm:

The main variation in bisecting k-means algorithm is that in the Adaptive Bisecting k-means algorithm the value of cluster is chosen at run time.

D. Forensic Analysis:

After the cluster has been performed analysis has been done on the clustered data obtained from the previous step. For identifying or grouping particular type of crime Apriori algorithm is used [5]. In this step area wise analysis is done on the clustered data set. For example in the particular are which type of crime is occurred and how many times it is happened.

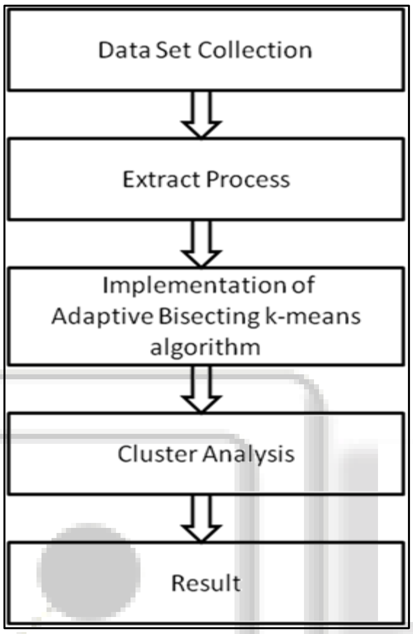

Fig. 1: Process of Investigation of Criminal Records

IV. Evaluation Measure

To evaluate data clustering algorithm reference partition is used. In order to investigate criminal records in our case reference partitions can be construction be the expert examiners and what type of clusters are expected in the data set.

V. Results and Discussion

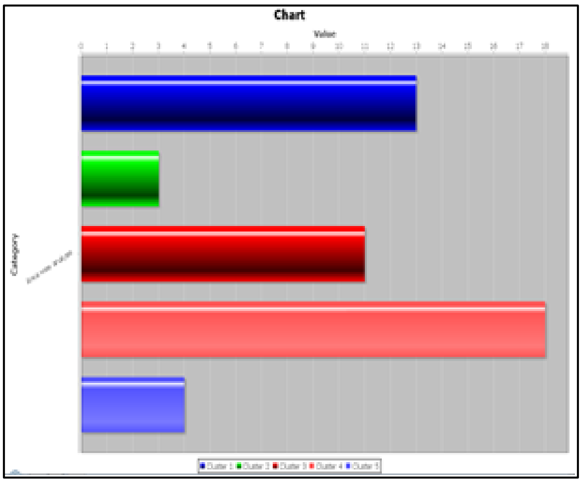

Fig. 2: Divide Into 5 Clusters

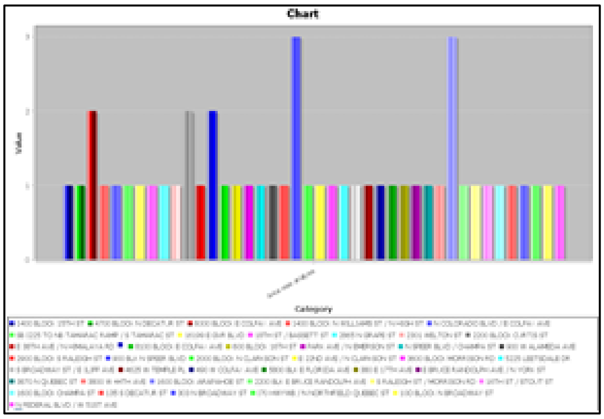

Fig. 3: Number of Occurrence

VI. Conclusion

This paper presents the review on various techniques for document clustering for investigation of criminal records from the data sets. As in the previous methods there were many drawbacks such as there is no method for cluster labeling. Cluster labeling will allow to examiner to identify the content of the cluster within the cluster. To cluster the data Adaptive Bisecting k-means algorithm is proposed. Adaptive Bisecting k-means algorithm gives better clusters, faster output and no blank clusters if data is available. Adaptive Bisecting k-means algorithms shows best result because it decides the number of clusters at run time.

References

[1] M. Azmat Javed, S. Jaiwsal, “Review on Mining and Investigation of Criminal Records from Digital Devices”, International Conference on Innovations in Information, Embedded and Communication Systems, pp. 781-783, March 2015.

[2] L. Liu, J. Kang, J. Yu, and Z. Wang, “A comparative study on unsupervised feature selection methods for text clustering,” in Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Engineering, 2005, pp. 597–601.

[3] B. S. Everitt, S. Landau, and M. Leese, Cluster Analysis. London, U.K.: Arnold, 2001.

[4] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, andM. S. Oliver, “Exploring forensic data with self-organizing maps,” in Proc. IFIP Int. Conf. Digital Forensics, 2005, pp. 113–123.

[5] Priyanka, Er. Vinod Kumar Sharma,” APRIORI ALGORITHM FOR MINING FREQUENT ITEMSETS –A REVIEW” International Journal of Computer Application and Engineering Technology Volume 3-Issue 3, July 2014. Pp. 232-236.