INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

A REVIEW ON FORENSIC DOCUMENT ANALYSIS USUNG APRIORIALGORITHM

AMIT N. BORKAR[1], PROF. PRAVIN S. KULKARNI[2]

1. Department of C. Tech, Rajiv Gandhi College of Engineering, Research& Tech., Chandrapur.
2. Gondwana University, Gadchiroli, Maharashtra, India.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: In recent times the global of Digital technologies especially in computers world, we find a tremendous increase in crimes like ethical hacking, rackets on different domain packets, unauthorized entrance. Hence, there is a need to track such unauthorized access. Our forensic document analysis using apriori algorithm provides an approach which is used to find the evidence by analyzing such massive set of documents. In forensic analysis frequently we examine huge amount of files it may be hundreds or thousands in number. First, using k-representative algorithm, we group the retrieved documents into the meaningful categories list which is the most central process. Hence we are specifying an approach for forensic analysis using document clustering algorithm helpful in police investigation.

Keywords: Document clustering, forensic science, k-means, k-representative.

Corresponding Author: MR. AMIT N. BORKAR

Access Online On:

www.ijpret.com

How to Cite This Article:

Amit N. Borkar, IJPRET, 2015; Volume 3 (9): 71-75

PAPER-QR CODE

INTRODUCTION

Today computer forensic become emergent need due to highly increase in crime linked with the internet and computer. In clustering of documents, computer forensic play an important role of identifying the evidence in police investigation for crime, using computer. This activity exceeds the capability of expert analysis and understanding of data.

For computer forensic analysis generally we requires some computer forensic tools that can subsist in the form of computer software. Introducing such tools are helpful for forensic investigation while dealing with computer investigation. On the other hand as because storage media expanding in size of large volume ,day by day, it becomes difficult task for investigator to locate their points of interest from large pool of data. In addition, it may be difficult for investigators because of the way in which data is present and may result in misinforming. As a result, the method will acquire a very large amount of time for analyzing big volumes of data. Sometimes it is also possible, that tools computer forensic tools may be meaningless for their generated data due to the fact that current tools of computer forensic is not capable of presenting a clear virtual outline of all the stuff (files) originated on the storage medium, as storage medium can store data in huge amount.

In more realistic and practical situation field expert (e.g. forensic examiners) are sparse and have partial time for performing examinations. Thus after finding an appropriate documents, it become sensible to suppose that the examiner might prioritized the analysis of supplementary documents belonging to the significant cluster, because it is likely that these can also be appropriate and applicable to the investigation. Such approaches based o clustering of documents, can certainly advance the progress of analysis of seized computer. Here the number of clusters is a grave bound of various algorithms and it is frequently unknown a priori. Till now the automatic assessment of the cluster number is not evolved in the investigation of the computer forensic literature. Truly, we couldn’t specify one work which is reasonably close in its domain application and that intelligence the competent of the algorithm estimating the cluster in specific number, back to the sixties it was suprising that lack of studies over hierarchical clustering algorithm.

Literature Survey:

The biographer in this article[1] demonstrate the proposed approach by targeting extensive test by conducting different experimental test of six well-known clustering algorithms(Single link, Complete link, Average link, k-means, k-medoids, and CSPA) that where applied to five different dataset of real-world obtains from investigation carried over computer seized. By considering different combination of parameter for the experiment, it resultant in more than 15 instantiation of algorithm. In accumulation, to get the involuntarily approximate number of cluster two validity indexes which are comparative were used. The studies related to the literature are usually limited than what we study. In our experiment, we obtained best result for our application by applying Average and Complete link. If duly initialized, k-means and k-medoids (partitional algorithm) also defer extremely excellent result.

Two process of k-representative instance memberships to cluster and cluster re-estimating are describe where representative is used by replacing centroids as centroids are present only in numerical domain. Representative show the occurring ratio among the possible value of features of the clusters member. Value difference matrix is inserted to compute the distance between the instances, specified in [2].

Exploratory of data analysis where done by clustering algorithms, when there is no or little prior knowledge about the data [3].

For mining e-mails for forensic analysis in integrated surrounding via classification and clustering algorithm, was present in [5].In an application domain which is related to email were grouped by using structural, domain-specific, syntactic, and lexical features[7]. The problem of e-mails clustering for forensic analysis was also introduced, using three clustering algorithm (k-means, Bisecting k-means and EM), where K-means of kernel-based variant was applied [8]. The result obtained was individually analyzed and then it concluded that the result were fascinating and useful for investigation point of view. More recently, a mining association rules from forensic data using a FCM-based method was described [4].

Forensic data analysis using Fuzzy method once again specifies an involuntary process and a methodology for inferring exact and effortlessly comprehensible expert-system-like rules for forensic data. For the most part of data analysis environment the algorithm and methodology used were proven to be easily implemented. By interacting the applicability of different types of fuzzy methods to improve the quality and efficiency of the data analysis phase for investigation in crime, the fuzzy set hypothesis would get implemented [4].

For forensic investigation in mining prints write from e-mails anonymous, basically they were written by multiple anonymous author throw collecting e-mails and focusing on the problem of mining the styles of writing those e-mails. The basic way for anonymous e-mail is to be first cluster by the Stylometric (the application of the study of linguistic style, mainly written language .i.e. Stylometry, but it actually applied for music and to fine-art paintings successfully) features and then pull out the write print, i.e., the inimitable writing style, from each cluster.[5]

This paper speak about the various computer forensic tool that are available on the market, For instance forensic Toolkit, Encase and Pro Discover are the list of available tools.The difference about these tools that some are built for single purpose only while other are designed to provide a whole range of functionalities. Looking over the examples of these functionalities are hashing verification, report generation, advanced searching capabilities and etc. Some computer forensic tools have the common functionalities by difference only in there GUI [6].

PROPOSED SYSTEM:

In existing system applied the k-means method to categorical objects, two main problems are encountered, namely, the formation of cluster centers and the calculation of dissimilarity between objects and cluster centers. To overcome this problem and for obtaining our objective we are proposed k-Representatives Algorithm.

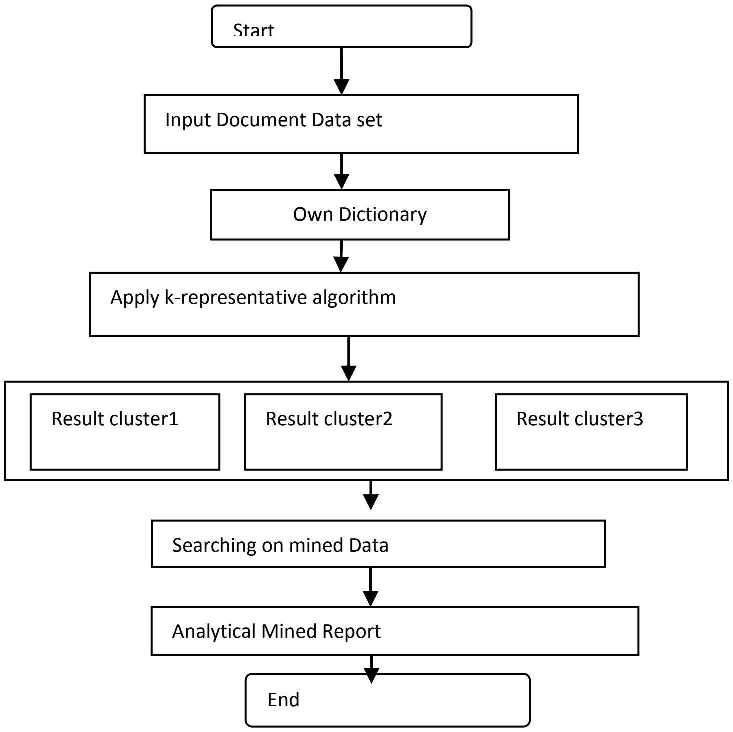

Fig.1.Flow of proposed system

CONCLUSION:

By doing the survey on computer forensic analysis it can be concluded that document clustering is not an easy step at all. There is enormous data to be cluster in compute forensic so to overcome this problem, this paper presented an approach that applies document clustering methods to forensic analysis of computers seized in police investigations. Again by using multithreading technique there will be document clustering for forensic data which will be useful for police investigations.

REFERENCE:

1. Luis Filipe da Cruz Nassif and Eduardo Raul Hruschka "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection" IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 8, NO. 1, 1556-6013 IEEE-2012.

2. Jae Heon Park and Sang Chan Parkk-representatives Algorithm: a Clustering Algorithm with Learning Distance Measure for Categorical Values" KAIST (Korea Advanced Institute of Science and Technology), Department of Industrial Engineering.

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

4. K. Stoffel, P. Cotofrei, and D. Han, "Fuzzy methods for forensic data analysis," in Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition, 2010, pp. 23–28.

5. R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis framework," Digital Investigation, Elsevier, vol. 5, no. 3–4, pp. 124–137, 2009.

6. B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver, "Exploring forensic data with self-organizing maps," in Proc. IFIP Int. Conf. Digital Forensics, pp. 113–123,2005.

7. F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," Digital Investigation, Elsevier, vol. 7, no. 1–2, pp. 56–64, 2010.

8. S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, "Text clustering for digital forensics analysis," Computat. Intell. Security Inf. Syst., vol. 63, pp. 29–36, 2009.