Enhancing Digital Forensic Analysis through Document Clustering

Vidhya B, Priya Vaijayanthi R

Abstract— **Digital forensic is the process of uncovering and interpreting process of uncovering and interpreting electronic data for use in a court of law. The goal of the process is to preserve any evidence in its most original form while performing a structured investigation by collecting identifying and validating the digital information for the purpose of reconstructing past events. Digital forensics deals with the analysis of artifacts on all types of digital devices. The role of digital forensics is to facilitate the investigation of criminal activities that involve digital devices, to preserve, gather, analyze and provide scientific and technical evidence, and to prepare the documentation for law enforcement authorities. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. Document clustering involves descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document cluster is generally considered to be a centralized process. Example of document clustering is web document clustering. Application of document clustering can be categorized to two types that are online and offline. Seized digital devices can provide precious information and evidence about facts. Large amount of data analyzed. Digital tools supported. In this paper do the work of extracting document and get a brief knowledge.**

Index Terms— **K-Means, document clustering, Aco**

I. INTRODUCTION

Forensic analysis is the uses of documented and controlled analytical and investigative techniques to identity collect examines and preserve digital information. It is used to variety of data theft incidents. Some forensic analysis services are propriety Information theft reconstruction and analysis deleted data recovery and analysis. Digital forensics in a branch of forensic science encompassing the recovery and investigations of material found in digital devices often in relation to computer crime. Digital forensics investigations have a variety of applications. The most common is to support or refute a hypothesis before criminal or civil, courts; forensics may also feature in the private sector such as during internal corporate investigations or .

Manuscript received Jan, 2014.

***Vidhya B**, Department of Computer Science and Engineering, Sri Vidya College of Engineering and Technology, Virudhunagar, India, +91 9600917355.*

***Priya Vaijayanthi R**, Department of Computer Science and Engineering, Sri Vidya College of Engineering and Technology, Virudhunagar, India.*

Fig.1.Forensic analysis

Intrusion investigation. Issues of this digital forensics are obtaining magnetic residue data, dealing with an intrusion,

From the figure forensic analysis do collect evidence. Looking into the logs, repairing the systems, tracking the hacker, keystroke loggers, finding the spy. Solution of this problem is admissible, authentic, and accurate, complete. Digital evidence has to be: Admissible is must conform to current legal buildings and could depend on legal system. Must prove records. Digital evidence has to be Authentic and reliability. Authenticate is most explicit link data to physical person. Must be self- sustained, strong access controls in place, logs and audit in good shape. Accurate means data process reliability determines content reliability and timings issues might throw you overboard

Clustering algorithm identifies [4] the accurate data from the analysis of little knowledge or no prior knowledge data. Computer forensics have unlabeled [4] objects. In previous analysis have labeled object design or supervised learning setting. Preliminary analysis defines data partition from the data and expert examiner only focus on reviewing representative documents from the obtained set of cluster. In preliminary process avoid the hard work of examiners. After finding relevant document the examiner could pass the analysis of the other document to investigation. Text clustering [1] in digital evidence defines information and data of investigate value. That are stored in digital device or transmitted in digital device. This type of seized device established by digital forensic analysts. It deals with massive amount of data and increasing capacity of data. Investigate activity have two aspects is acquisition and retrieval information extracted from digital device. Forensic acquisition puts most relevant data into the preliminary phase. It is the selective storage. It involves two steps. That is textual information extraction have digital device text files and early analysis (bit-stream acquisition) and textual data analysis via clustering based text mining tool identifying,

ISSN: 2278 – 1323

tracking, extracting and classifying discovering. Text clustering for forensic analysis based on dynamic adaptive clustering model. Digital investigation important for textual [3] evidence. Examples of investigations are e-mails, internet browsing history, instant messaging, word processing documents n/w activity logs. In physical level every byte search at the digital evidence. Second identifies the specific text string. It moves to the next investigation. Text string search have Information Retrieval (IR) overhead, and make noise. Small device have a capacity of 80gb.these problems solved two solution. First one have decrease the number of irrelevant search hits. Second one has present the search hits a manner which enables the investigator to find the relevant hits more quickly. Indexing algorithms and ranking algorithms combines fail in the first solution. At the second solution it works. Main function is improving the (IR) information retrieval. Fuzzy Methods [2] defines crime data analysis and utilization important for intelligence. Intelligence based approach for law enforcement. Must it necessicity. analysis related to type of intelligence. Forensic intelligence defines the accurate, timely and useful products of logically processing forensic case data. Results of forensic intelligence have discipline specific activities. Information technology used to produce the information sets and digital evidence the methods from Artificial intelligence. Artificial intelligence defines the science and engineering of making intelligent machine. Computational intelligence includes a number of computational methods as neural networks, fuzzy systems. Fuzzy methods improve the quality of data analysis phase. Fuzzy tools apply digital investigation. Forensic analysis evidence has computational intelligence methods and techniques and assigning analysis. Evolutionary algorithm and genetic algorithm solve the problem of missing persons. Writer identification solves the problem of hand writing analysis. Fuzzy methods important a role and learning complex data structures and patterns classifying them to make intelligent decisions. Comparing k-means and k-medidos it works best.

II. RELATED WORK

Seized digital devices [1] can provide precious information and evidences about facts. Large amount of data analyzed. Digital text analysis text mining technique used. Difficult to search string. Solve the problem in using forensic acquisition and early analysis and textual information extraction and text clustering. Supervised learning tools to categorize data on already defined categories for investigate purposes. In computer forensic [4] analysis hundreds of thousands of files are usually examined. Much of the data in those files consists of unstructured text, whose analysis by computer examiners is difficult to be performed. To overcome these problems applies clustering algorithms to forensic analysis of computer seized in police investigations. Clustering includes labels. Examiner identifies easy and also content quick search. Difficult to [3] identifies specific text string. To solve this problem using algorithm of ranking and indexing. Automatic approaches for clustering labeling. The assignment of labels to clusters may enable the expert examiner to identify the semantic content of each cluster more quickly. Improve the quality of data analysis. Make a automatic procedure for inferring accurate and [2] easily understandable expert-system-like rules from forensic data. Methodology is based in the fuzzy set theory. TO overcome these problem using fuzzy set theory. It produces the best result comparing k-means and k-mediods. The accuracy of rules inferred was very high and clearly better than the minimum level required to make them usable in a particular string. Complicates reduce communication experts.

In this section, we discuss related work on document clustering and clustering algorithm.

A.DOCUMENT CLUSTERING

Extraction and fast information retrieval or filtering. Related to data clustering. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. Document clustering involves descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document cluster is generally considered to be a centralized process. Example of document clustering is web document clustering. Application of document clustering can be categorized to two types that are online and offline.

B.PRE-PROCESSING STEPS

Stop words doing before clustering algorithm. It defines remove of prepositions, pronouns, articles and irrelevant document, Meta data. It enables snowball steaming. Text mining using traditional satisfies approach. Identifies vector space model. In this model [4] have effectiveness, efficiency, clustering algorithm. Transformation vector selects a number of attributes that have been used namely, cosine-based distance and leven steins-based distance.

C.CLUSTERING ALGORITHM

Machine learning data mining fields using Cluster Ensemble Based Algorithm (CSPA)Mediods have centroids. This property makes it particularly interesting for applications in which1) centroids cannot be computed, and2) distances between pairs of objects are available-MEANS AND k-mediods are sensitive to initialization considering partitioned algorithms. Every partition represented by the dendogram subsequently choosing best results. CSPA algorithm essentially finds a consensus clustering from a cluster ensemble formed by a set of different data partitions. After applying clustering algorithms to the data a similarity matrix computed. Each element of this matrix represents pair-wise similarities between objects. The similarity between two objects is simply the fraction of the clustering solutions in which those two objects lie in the same cluster.

III. IMPLEMENTATIONS K-MEANS ALGORITHM

K-means is one of the simplest unsupervised learning algorithms that partition feature vectors into k clusters so that the within group sum of squares is minimized. K-means clustering is a method of vector quantization originally from signal processing that is popular for cluster analysis in data

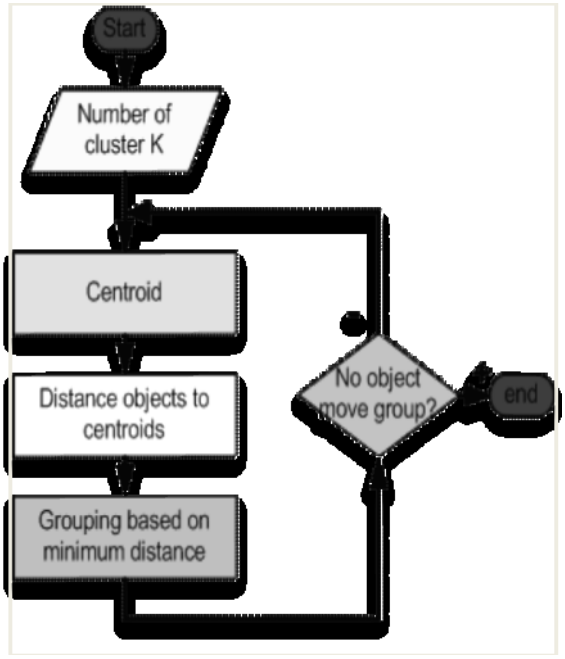

Fig.2. K-Means process

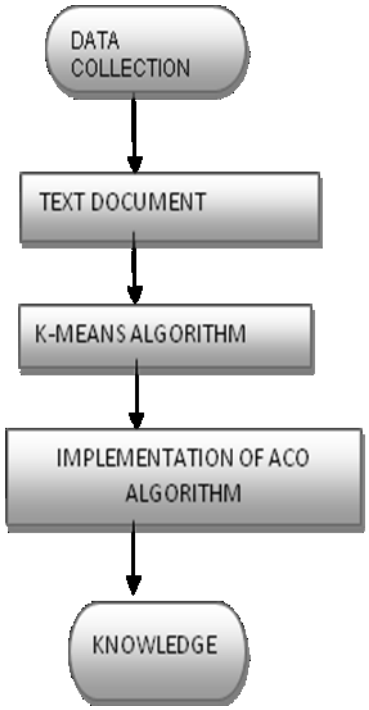

Fig.3 clustering process

Mining. From the figure K-Means follows a simple way to classify a given data set and looks like

STEPS

Step 1: Place randomly initial group centroids into the 2d space.
Step 2: Assign each object to the group that has the closest centroid.
Step 3: Recalculate the positions of the centroids.
Step 4: If the positions of the centroids didn't change go to the next step, else go to Step 2.
Step 5: End.

IV. IMPLEMENTING ACO ALGORITHM

Ant colony optimization algorithm is a very important one among swarm intelligence algorithms. Because Java is an advanced object-oriented and platform-independent computer programming language, in order to use this algorithm in a platform-independent and flexible way, this paper introduces a Java-based implementation package of it. This package includes some sub-packages. There are several classes which are all implemented in Java using object-oriented technology in each sub-package. Users can utilize these classes on computers installed the corresponding Java runtime environment to solve some problems. After the test on two travelling salesman problems, these classes performed properly and efficiently, and the good effect was received.

V. PROBLEM DESCRIPTION

Extraction and fast information retrieval or filtering. Related to data clustering. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. Document clustering involves descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document cluster is generally considered to be a centralized process. From the figure document clustering process is web document clustering. Application of document clustering can be categorized to two types that are online and offline. Seized digital devices can provide precious information and evidence about facts. Large amount of data analyzed. Digital tools supported. Digital text analysis text mining technique used. A Collection of raw text files processed by the text mining tool. Set of documents is $D=\{D_{1,...},D_n\}$.Vectors v' and v'' support the computation of the frequency based distance.$\Delta^{[f]}$ and of the stylistic distance,$\Delta^{[s]}$,respectively. It is not a metric space because does not guarantee the triangular inequality, for this reason equation can be more properly considered a similarity measure. This distance measure has been employed in the well known K-Means clustering algorithm.

VI.CONCLUSION

In this paper collecting text document extract the information in that document in brief formats. It reduces the work of data examiner. It helps to police departments. Because the terrorist missing the evidence of device. It searches and examines gives the knowledge about attacks. So it is very helpful to prevent attacks .and future work is suppose mobile device is caught by police departments. It examines the device give the brief knowledge to the same clustering techniques.

ISSN: 2278 – 1323

VII.REFERENCES

[1]Luis Filipe da Cruz Nassif and Eduardo Raul Hruschka," Document Clustering for forensic Analysis: An Approach for Improving Computer Inspection", IEEE transaction on Information Security, Vol. 8,pp. 46-54,January 2013.

[2]Sergio Decherchi,Simone Tacconi,and Judith Redi,Fabio Sangiacomo,Alessio Leoncini,and Rodolfo Zunino," Text Clustering for Digital Forensics Analysis", Journal of Information Assurance and security", Vol. 5,pp. 384-391,January 2010.

[3]Hammouda k.M,Kamel M.S," Efficient pharase based document indexing for web document clusterimg",IEEE Transactions on knowledge and data engineering, Vol. 16,pp. 1279-1296,2004.

[4]Chim .H,Deng .X," Efficient pharse-based document similarity for clustering",IEEE Transaction on Knowledge and data Engineering,Vol. 20,pp. 1217-1229,2008.

[5]Girolami .M,"Mercer Kernel Based Clustering in featurespace"IEEE Transaction on neural networks,Vol. 13,pp. 2780-2784,2002.

[6]Zadeh A.L,"Outline of new approach to the analysis of complex systems and decision processes",IEEE Transaction on systems,Vol. 1,pp. 28-44,1973.

[7]Mamdani .E,Assilian .S,"An experiment in linguistic synthesis wuth a fuzzy logic controller", Journal of Man-Machine studies,Vol. 7,pp. 1-13,1975.

[8]Fei B.K.L,Eloff J.H.P,and Venter H.s,Oliver M.S,"Exploring forensic data with self-organizing maps", proceedings of the IFIP International Conference on Digital forensics,pp.-113-123,2005.

[9]Beebe N.L,Clark J.G,"Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results", Proceedings of the Conference Digital Investigation,pp. 49-54,2007.

[10]Stoffel .K,COtofrei .P,and Han.D,"Fuzzy methods for forensic analysis", Proceedings of the International Conference soft computing and pattern Recognition,pp. 23-28,2010.