International Journal Of Engineering Research & Management Technology

ISSN: 2348-4039

July- 2014 Volume 1, Issue-4

Email: editor@ijermt.org www.ijermt.org

Document Clustering For Digital Devices: An Approach to Improve Forensic Analysis

M. Murali
M Tech II year
Dept of CSE,SVCET
Chittoor

K.Thyagarajan
Associate Professor
Dept of CSE, SVCET
Chittoor

K. Dasaradharami Reddy
Assistant Professor
Dept of CSE, SVCET
Chittoor

ABSTRACT-
In today's digital world, information in computers has great importance and this information is very essential in context for future references and studies irrespective of various fields. So surveying of such information is critical and important task in computer forensic analysis, a lot of information present in the digital devices is examined to extract information so Automated method is in great interest. Digital investigation for forensic is a very specialized field and to conduct an investigation accurately, and without losing or misinterpreting data. Forensic analysis consumes less time .Computer consists of hundreds of thousands of files which contain unstructured text or information, so Document clustering algorithms is of great interest for automated analyze. Analyzing the shape of the Pareto front helps decision makers understand the solution space and possible tradeoffs among the conflicting objectives. Documents Clustering helps to improve analysis of documents under consideration. Document clustering analysis is very useful for crime investigations to analyze the information from seized digital devices like computers, laptops, hard disks, tablets. There are total seven algorithms used for clustering of documents like K-means, K-medoids, single link, complete link, Average Link, and Cluster-based Similarity Partition Algorithm (CSPA). In addition to that improved Stemming Algorithm is used. These seven algorithms are very efficiently used to cluster the digital documents. These algorithms make the analysis very fast by clustering very close documents in one cluster. Also two validity index are used to find out how many clusters are formed from the huge unstructured data. The aim of clustering is to find structure in data is therefore exploratory in nature.
Key words: Forensic Analysis, Clustering, Digital investigation.

I. INTRODUCTION

Estimates that are proposed by IT IS are digital data density increased 18 times in latest 5 to 6 years. Digital world contains very important, complex an unstructured data, clustering algorithms play important role in forensic analysis [1] of such digital documents. In particular domain related to paper, hundreds of thousands of documents are examined and this surveillance which exceeds capabilities of expertise, which monitors or analyze such documents. So it is very prime requirement to make data simple and to use some techniques which boosts the analysis of complex, unstructured documents. In [2] digital forensics as a discipline is not particularly new however, in the past it was usually associated with law enforcement investigations of computer-related crimes which is used for experts. More recently, it is becoming increasingly common for high profile corporations, especially financial services companies, to have fulltime resources dedicated to battling the onslaught of cybercrime and malware keying in on these profitable institutions.. Clustering algorithms [3] have been studied for decades Therefore, we decided to choose a set of seven algorithms in order to show the potential of the proposed approach, namely: the partitional K-mean [4], K-medoids [5] the hierarchical Single/Complete/Average Link [6] the cluster ensemble algorithm Known as CSPA and improved stemming algorithm are used for clustering the documents in digital devices.

TABLE I
SUMMARY OF ALGORITHMS AND THEIR PARAMETERS

Acronym	Algorithm	Attributes	Distance	Initialization	K-estimate
Kms	K-means	All	Cosine	Random	Simp.Sil
Kms100	K-means	100>TV	Cosine	Random	Simp.sil
Kmd100	K-medoid	100>TV	Cosine	Random	Simp.sil
Kmdlevs	K-medoid	Name	Lev.	Random	Silhouette
AL100	Average Link	100>TV	Cosine	Random	Simp.sil
CL100	Complete Link	100>TV	Cosine	Random	Silhouette
SL100	Single Link	100>TV	Cosine	Random	Silhouette

Sil: SilhouetteAL: Average LinkCl: Complete Link

In more practical and realistic scenario, domain experts are scare and have limited time available for performing examinations. Our goal is to help decision makers identify groups of strongly related solutions from unstructured data and make them understand data easily, so that they can understand more easily the range of design choices, identify areas where strongly different solutions achieve similar levels of objectives, and decide first between the major groups of solutions before deciding for a particular variant within the chosen group.

II. LITERATURE SURVEY

The literature on computer forensic only reports the use of algorithms for clustering the document in digital devices for examining officially by police department where cluster is known and fixed a priori by the user. Essentially, one includes different data partitions and assesses them with relative validity index in order to estimate the best value for the number of clusters. Literature on computer forensic is huge we Discuss the important concepts regarding this paper.

a) Computer Forensic

Several criminal activities are being committed nowadays such as cyber terrorism, internet fraud, viruses, illegal downloads, falsification of document, child pornography, counterfeiting, economic espionage, benefit fraud, financial fraud, internal security policy violations, system vulnerability exploits, e-discovery, and sensitive data leakage investigations If the computer and its contents are examined by anyone other than a trained and experienced computer forensics specialist [7], the usefulness and credibility of that evidence will be tainted. A highly skilled computer forensic analyst is someone who understands the discipline as well as understands the use of computer forensic tools. Evidences can be collected from various sources hard drives, memory, system logs, network traffic, IDS (both NIDS, HIDS), and physical security

When a team of computer forensic [7] expert people conducts the analysis off seized computer or digital devices for crime scene, they faces complicated challenges, which consist path to the acceptance of any potential evidence produced as an outcome. While computer forensics is not an entirely new field, there are still gray areas in its legal definition and application. Therefore it is important to pay crucial attention to the details of the forensics analysis setting and its execution in order to extract the useful information. Consider a company who wants to protect its digital assets might want to reuse various components of an Information Security Management System or a similar framework to protect their sensitive data, destined to prevent the destruction and/or stealth of its corporate information, and to facilitate its recovery, should the worst happen.

Such concerns, though not addressed in this document, are in close relationship with its content, as some of the elements described in existing specifications of such frameworks are part of the general concepts of computer forensics we describe, or applied by the solutions we review.
In the context of a company in the position of building a legally receivable case, the questions we aim to ask and answer in this paper are the following:

- What information is available on a common system (default installation)?
- What information should be collected?
- How can this information be collected?

A computer can be the target of the crime, it can be the instrument of the crime or it can serve as an evidence repository storing valuable information about the crime. Knowing what role the computer played in the crime can of tremendous help when searching for evidence. This knowledge can also help reduce the time taken to package your evidence and data recovery approach is unified modeling approach for clustering methods it includes cluster validation and interpretation final results is useful for students, practitioners and theoreticians of cluster analysis in digital forensic.

III. PROBLEM DEFINITION

The literature on Computer Forensics only reports the use of algorithms that assume that the number of clusters is known and fixed a priori by the user. A common approach in other domains involves estimating the number of clusters from data and analyses the information on the devices clustering is done to huge amount of data here disadvantage is clustering techniques do not address all the requirements adequately, Time complexity is huge.

a) Existing method

Current clustering techniques do not address all the requirements adequately for investigation and to examine the digital devices officially. Dealing with large number of dimensions and large number of data items can be problematic because of time complexity. Effectiveness of the method depends on the definition of "distance" (for distance based clustering). If an obvious distance measure doesn't exist we must "define" it, which is not always easy in document clustering, especially in multidimensional spaces. The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

IV. PROPOSED METHOD

Doing the survey on computer forensic analysis we can say that the clustering on data is not an easy step. There is huge 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, In computer forensic investigation, usually thousands of files are surveyed. The data in those files consists of formless manuscript; it is very touch to accomplish the computer examiners of investigation. Clustering is the unverified organization of designs that is data items, remarks, or feature vectors into groups (clusters). Improved stemming algorithm is used to remove stop words in a document.

Pre-processing step

Before running a clustering algorithm on text dataset pre-processing is performed. In particular to remove stop words in order to improve the clustering methods here stop words like prepositions, pronouns, articles, and irrelevant documents, metadata is to be removed in pre-processing step. Improved stemming algorithm is used and traditional statistical approach is used for text mining in which documents are represented by vector space model here each document is represented by a vector containing the frequencies of occurrences of words, whose number of characters are between 4 and 25. We also used a dimensionality reduction technique known as Term Variance (TV) [8] can increase both effectiveness and efficiency of clustering algorithm in preprocessing step there are three steps such as
a. Fetch a file contents: It is used to check the file content.
b. Stop word removal**:** It is used to remove the stop word like a, an, the etc.
c. Improved Stemming: It is used to stemming on that file which will be removing "ing" and "ed" words from the given statement.

b) Preparing cluster vector

The preprocessing step is already done, for preparing the cluster vector one need to find top 100 words form the file. That document or another way we can say file or data numerical sentences such as the sentence which has numerical word in it, that means the sentence which contains date or any kind on number in it.

c) Forensic analysis

From the forensic data analysis classification matrix need to be made with the help of weighted method protocol. At last one can find accuracy of his work.

d) Removing Outliers to large extent

This approach makes recursive use of silhouette these choose cluster single object only and single objects are removed. Then, the cluster process is repeated over and over again until partition without singletons is found at the end single clusters, data partitions is found

V.EXPERIMENT EVALUATION

The obtained data partitions were evaluated by taking into account that we have a reference partition (ground truth) for every dataset.

Results and discussion

Table I summarizes the obtained ARI results for the algorithms listed in Table I. In general, AL100 (Average Link algorithm using the 100 terms with the greatest variances, cosine-based similarity, and silhouette criterion) provided the best results with respect to both the average and the standard deviation, thus suggesting great accuracy and stability. Note also that an ARI value close to 1.00 indicates that the respective partition is very consistent with the reference partition—this is precisely the case here. In this table, we only report the best obtained results for the algorithms that search for a consensus partition between file name and content (NC100 and NC)—i.e.,

Table I give us clear idea about the clustering algorithm performance which increased to large extent by using new combination of well known clustering algorithms addition to that, improved stemming is used to improve efficiency and efficient of computer inspection for digital devices seized in police investigation Table II

Example of the information found in the clusters

Cluster	Information
C1	7 blank document
C2	12 check receipt
C3	4 daily reports from buying
C4	2 notice about meetings
C5	3 warnings about business hours
C6	2 office applications

Table I results in performance of clustering algorithm used where as Table II is example of information found in cluster there are eight clusters with information about each cluster. As far as the adopted dimensionality reduction technique is concerned—Term Variance (TV) [8]—we observed that the selection of the 100 attributes (words) that have the greatest variance over the documents provided best results than using all the attributes in three out of five datasets (see Table III). Compared to Kms, the worse results obtained from feature selection by Kms100 and Kms100S, especially in the dataset D, are likely due to k-means convergence to local optima from bad initialization. By considering all the results obtained from feature selection, we believe that it should be further studied mainly because of the potentially advantageous computational efficiency gains.

VI. CONCLUSION

We have introduced an approach which can become an ideal application for document clustering to forensic analysis of computers, laptops and hard disks which are seized by police during investigation of police. There are several practical results based on our work which are extremely useful for the experts working in forensic computing department. In our work, the algorithms known as Average Link and Complete Link yielded the best results along with advanced stemming algorithm. In spite of these algorithms having high computational costs, they are suitable for our work domain because dendrograms provides a neat summary of documents which are being inspected. All the textual documents are scanned thoroughly an corresponding output is given. When proper initialization is done, the partitioned K-means and K-medoids algorithms also have satisfactory results.

VII. ACKNOWLEDGEMENT

M.Murali would like to thank Sri Venkateswara College of Engineering And Technology, Head of The Department Dr.J.Janet and guide K.Thyagarajan for their help and support.

REFERENCES

1. N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results," Digital Investigation, Elsevier, vol. 4, no. 1, pp. 49–54, 2007.
2. 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.
3. A. K. Jain and R. C. Daubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.
4. R. Xu and D. C.Wunsch, II, Clustering. Hoboken, NJ: Wiley/IEEE Press, 2009.
5. Daubes,mnemstudio.org/clustering-k-means-introduction.htm
6. K. Jain and R. C. Daubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988
7. Recognition,2010, pp. 23–28.
8. S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, "Text clustering for digital forensics analysis," Computational.
9. Intell. Security Inf. Syst., vol. 63, pp. 29–36, 2009.
10. 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.

.