International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Survey on Forensic Document Clustering Using High Performance Clustering Algorithm

Asmita V. Mane[1], **Gitangali. R. Shinde**[2]

[1]Smt. Kashibai Navale College of Engineering, Department of Computer Engineering, Pune, India

[2]Professor, Smt. Kashibai Navale College of Engineering, Department of Computer Engineering, Pune, India

Abstract: ***Computer forensics is a new and fast growing field which involves carefully collecting and examining electronic evidence. This evidences not only assesses the damage to a computer as a result of an electronic attack, but also to recover lost information from such a system to prosecute a criminal. Document clustering can simplify the browsing large collections of documents by reorganizing them into a smaller number of manageable clusters. Hundreds and thousands of files are usually examined in computer Forensic analysis. Most of the data in unstructured format. Due to the unstructured format and explosive growth of amount of data that humans want to analyze, fast algorithms are necessary, but they can often give poor quality results. The methodology that compare and evaluate the quality of clustering algorithms is investigated. Present an approach that applies clustering algorithms for forensic analysis of seized computer documents in investigations. Not a one specific algorithm is Cluster analysis itself but the general task to be solved. Various algorithms are used that significantly differ in their notion of what they constitutes a cluster. Automatic clustering and cluster labeling helps to improve efficiency.***

Keywords: Clustering, forensic analysis, text mining, crimes, Unstructured Document.

1. Introduction

In Forensic Investigation process in which the digital device like computers are used to analyze the digital evidence those are facts which are under the investigation. The digital evidences are the digital data which supports the incident hypothesis. Volume of data of digital world increased from 161 hex bytes to 988 hex bytes about 18 times the amount of information present in all the books ever written—and it continues to grow exponentially. Analysis of this documents is difficult task of the digital forensic the analysis of other documents belonging investigation process. This large amount of data directly impact on computer forensics. It usually more complex to examining because it is unstructured. It is more complex to examining hundreds of thousands of files if the number of documents are large. The analysis of large amount of data exceeds the expert's ability of analysis and interpretation of data. Therefore, automated data analysis, like those widely used for machine learning and data mining, are of paramount importance. Recently clustering algorithms are used in the process of forensic analysis. These methods are used to convert unstructured documents to structured documents for further investigation. This is precisely the case in many applications of Computer Forensics From a more technical viewpoint, the classes or categories of documents that can be found are a priori unknown.

Moreover, even assuming that labeled datasets available from previous analyses, there is almost no hope that the same would be still valid for the incoming data, obtained from other computers and associated to different investigation processes. More precisely, the new data come from a different population. In this context, the use of clustering algorithms, which is capable of finding latent format from text documents found in seized computers, which enhance the analysis performed by the expert examiner. The clustering method is a valid cluster are more similar to each other than objects belonging to a different cluster. Thus, once data partition has been created from database, the expert examiner might initially focusing on reviewing representative documents from the obtained set of clusters. Examiner may eventually decide to explore other documents from each cluster after this preliminary analysis. By doing this, one can avoid the hard and difficult task of examining all documents but, even if so desired, it still could be done. In reality domain experts are scare and have a limited time to perform examinations so, it's reasonable to assume that, after finding a relevant document, the examiner could prioritize to the cluster of interest, because it is likely that these are also relevant to the investigation. Such approach of document clustering, to support to take a decision by conducting data analysis which can indeed to improve the computer forensic investigators and assist them in analysis of seized computers.

During this paper, following different sections presented. In section II, we are discussing about Forensic Investigation process and its associated problems. In section III we will discuss the document or text clustering algorithms, finally in section IV conclusion.

2. Related Work

Clustering is a method of grouping of similar objects in the same group are more similar to each other than to objects in other groups. Most of the studies in the Computer Forensics field describe the use of classic algorithms for clustering data—e.g., Expectation- Maximization (EM) for unsupervised learning of Gaussian Mixture, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM), K-means,K-meloid.

A. Cluster Ensembles - The Knowledge Reuse Framework for Combining Multiple Partitions.[7]. This paper present the problem of combining multiple partitioning of set of objects into a single consolidated clustering without accessing the features or algorithms that deter-mined these partitioning's. Introduced the cluster ensemble problem and to solve this problem provided a three effective and efficient algorithms. It define a mutual information-based objective function that enables to automatically select the best solution from several algorithms and to build asupra-consensus function as well.

B. Text Clustering for Digital Forensics Analysis.[8] Present an effective digital text analysis strategy, based on clustering based text mining techniques, is introduced for investigational purposes. It gives an overview on the possibilities offered by textual clustering when applied to Digital Forensics analysis.

C. Term-Weighting Approaches In Automatic Text Retrieval[8]. This article shows the insights gained in automatic term weighting and provides single term indexing with which other more elaborate content analysis procedures can be compares.[9]

D. Fuzzy methods for forensic data analysis [11]: In this paper describe a methodology and an automatic procedure for inferring accurate and easily understandable expert-system-like rules from forensic data. This methodology is based on the fuzzy set theory.

E. Exploring forensic data with self-organizing maps: SOM discusses the application of a self-organizing map (SOM), an unsupervised learning neural network model, in a more efficient manner. The paper explores that how a SOM can be used as a basis for further analysis and also, it demonstrates how SOM visualization can provide investigators with greater abilities to interpret and explore data generated by computer forensic tools.[12]

3. Clustering Algorithms and Preprocessing

A. Pre-Processing Steps

We performed some preprocessing steps before running clustering algorithms on text datasets. In particularly, *stopwords i.e.*prepositions, pronouns, articles, and irrelevant document metadata have been removed. And also, the Snowball *stemming* algorithm for Portuguese words has been used. Then, adopted a traditional statistical approach for text mining, in which documents are represented in a vector space model. In VS model, each document is represented by a vector containing the frequencies of occurrences of words, defined as delimited alphabetic strings, whose number of characters is between 4 and 25. We also used a Term Variance technique which dimensionality reduction technique that can increase both the effectiveness and efficiency of clustering algorithms. TV selects a number of attributes (in our case 100words) which have the greatest variances over the documents. To compute distances between the documents, two measures have been used, namely: cosine-based distance and Levenshtein-based distance. The later has been used to calculate distances between file /document names only.

B. Estimation of Number of Clusters From

Data:

In order to calculate the number of clusters, widely used method is that consists of getting a set of data partitions with different number of clusters and then selecting partition that provides the best result according to a specific quality criterion. Such a set of partitions may result in directly from a hierarchical clustering dendrogram or, alternatively, from multiple runs of a partitioned algorithms starting from different numbers and initial positions of the cluster prototypes.

For a moment, let assume that the set of data partitions with different number of clusters is available, from which is the best one according to some relative validity criteria. Note that, from choosing such data partition, we are performing model selection and, as an intrinsic part of this process, we are also estimating the number of clusters. A widely used relative validity index is silhouette.

Let consider an object belonging to cluster P. Average dissimilarity of i to all other objects of X is denoted by $p(i)$. Now take into account cluster Q . The average dissimilarity of i to all objects of Q will be called d$(i\ ,Q)$. After computing d$(i\ ,Q)$ for all clusters Q$\neq P$, the smallest one is selected, i.e b(i) = min d$(i\ ,Q), Q \neq P$. This value represent the dissimilarity of its neighbor cluster, and the silhouette for a give object, (i) , is:

$$S_{(i)}=\frac{b(i)-p(i)}{max\{p(i),b(i)\}}$$

It verified that$-1 \leq s(i) \leq 1$. That is , the higher $s(i)$the better the assignment of object to a given cluster. If it is equal to zero, then it is not clear that whether the object should have been assigned to its current cluster or to a neighboring one. if cluster P is a singleton then$S_{(i)}$ is not defined and the most neutral choice is to set $S_{(i)} = 0$. Once we have completed the computation , where is the number of objects in the dataset, then we take the average over these values, and the resulting value is then a quantitative measure of the data partition in hand. The best clustering corresponds to the data partitions are that has the maximum average silhouette. This average silhouette just addressed depends on the computation of all distances among all objects in cluster.

C. Clustering Algorithms

The clustering algorithms adopted in study—the partitional K-means and K-medoids the hierarchical Single/Complete/Average Link, and cluster ensemble based algorithm known as CSPA are popular in the data mining and machine learning fields, and therefore they havebeen used in our study. Some of our choices nevertheless regarding their use deserve further comments. For instance, K-means is similar to K- medoids. So, instead of computing centroids, it uses medoids, which are the representative objects of the clusters andthis property makes it particularly interesting for applications in which (1) centroids cannot be computed; and (2) distances between pairs of objects are available, as for computing dissimilarities between names of documents with the Levenshte in distance .

Considering the partitional algorithms, that both K-means and K-medoids are sensitive to initialization and usually converge to solutions that represent local minima. Trying to minimizes these problems, we used a nonrandom initialization in which distant objects from each other are chosen as starting prototypes . Unlike the partitional algorithms such as K-means/medoids, hierarchical algorithms such as Single/Complete/Average Link provide a hierarchical set of nested partitions, usually represented in the form of a dendrogram, from which the *best* number of clusters can be estimated.

The CSPA algorithm finds essentially a consensus clustering from a cluster ensemble formed by a set of different data-partitions. More precisely, a similarity matrix is computed after applying clustering algorithms to the data .Each element of this matrix represents pair-wise similarities between objects. The fraction of the clustering solutions in which those two objectslie in the same cluster is the similarity between two objects is simply. Later, this similarity measure used by a clustering algorithm that can process a proximity matrix. The sets of data partitions/ clusterings were generated in two different ways: (a) by running K-means 100 times with different subsets of attributes (in this case CSPA processes 100 data partitions); and (b) by using only two data partitions, namely: one obtained by K-medoids from the dissimilarities between the file/document names, and another partition achieved with K-means from the vector space model. In this case, each partition have different weights, which have been varied between 0 and 1 (in increments of 0.1 and keeping their sum equals to 1). For the hierarchical algorithms, we simply run them and then assess every partition from the resulting dendrogram by means of the silhouette. Then, the best partition is selected and taken as the result of the clustering process. For each partitional algorithm, i.e. K-means/medoids, we execute it repeatedly for an increasing number of clusters. For each value of , a number of partitions achieved from different initializations are assessed in order to choose the best value of and its corresponding data partition, using the Silhouette. In this experiments, we assessed all possible values of in the interval ,where is the number of objects to be clustered.

D. Dealing With Outliers

Assess a simple approach to remove *outliers in* this approach makes recursive use of the *silhouette*. Fundamentally, singletons i.e., clusters formed by a single object only, these are removed if the best partition chosen by the silhouette has singletons. Then, repeated process of the clustering over and over again until a partition without singletons is found. At the end of the process, all singletons are incorporate into the resulting data partitions as single clusters.

4. Conclusion

In presented approach that applies document clustering methods to forensic analysis of seized computers in police investigations. Also, reported and discussed several practical results that can be very useful for researchers of forensic computing. More specifically, in this experimental approach the hierarchical algorithms known as Average Link and Complete Link presented the best results. Clustering has been used in number of applications in every field of life. One has to cluster a lot of thing or objects on the basis of similarity either consciously or unconsciously. Clustering is the one of the first steps often in data mining. When the partitioned K-means algorithm is properly initialized is also achieved good results. Considering the approaches for calculating the number of clusters, the relative silhouette has shown to simplified version. For exploring further relationships it identifies groups of related records that can be used as a starting point. In addition, some of our results suggest that document with file names contain information may be useful for cluster ensemble algorithms. Most importantly, observed that clustering algorithm indeed tends to induce clusters formed by either relevant or irrelevant documents or objects, so contribution to enhance the expert examiner’s job. Furthermore, our experimental evaluation of proposed approach is in five real-world applications show that it has the potential to speed up the computer inspection process.

References

[1] Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka, “Document Clustering for Forensic Analysis: AnApproach for Improving Computer Inspection” , IEEE Transactions On Information Forensics And Security, Vol. 8, No. 1, January 2013

[2] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz, “The expanding digital universe: A forecast of worldwide information growth through 2010,” *Inf. Data*, vol. 1, pp. 1–21, 2007.

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

[4] A. K. Jain and R. C. Dubes, Algorithms for ClusterinData. EnglewoodCliffs, NJ: Prentice-Hall, 1988.

[5] L. Kaufman and P. Rousseeuw, Finding Groups in Gata: AnIntroduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.U.K.: Arnold, 2001.

[6] R. Xu and D. C.Wunsch, II, Clustering. Hoboken, NJ: Wiley/IEEEPress, 2009.

[7] A. Strehl and J. Ghosh, “Cluster ensembles: A knowledge reuse frameworkfor combining multiple partitions,” J. Mach. Learning Res., vol.3, pp. 583–617, 2002.

[8] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, andR. Zunino, “Text clustering for digital forensics analysis,” Computat.Intell. Security Inf. Syst., vol. 63, pp. 29–36, 2009.

[9] G. Salton and C. Buckley, “Term weighting approaches in automatictext retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, 1988.

[10] Aggarwal, C. C. Charu, and C. X. Zhai, Eds., “Chapter 4: A Surveyof Text Clustering Algorithms,” in Mining Text Data. NewYork:Springer, 2012.

[11] K. Stoffel, P. Cotofrei, and D. Han, “Fuzzy methods for forensic dataanalysis,” in Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition,2010, pp. 23–28.

[12] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, andM. S. Oliver, “Exploringforensic data with self-organizingmaps,” in *Proc. IFIP Int. Conf. DigitalForensics*, 2005,pp. 113–123.