ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Forensic Analysis Using Document Clustering

Ansari Saad[1], Prof. R.B. Wagh[2]

P.G. Student, Department of Computer Engineering, SES's RCPIT, Shirpur, India[1]

Associate Professor, Department of Computer Engineering, SES's RCPIT, Shirpur, India [2]

ABSTRACT: Now days, Forensic analysis is doing crucial job in crime investigation. The crimes for investigation are embrace hacking, drug trafficking, erotica, numerous larceny crimes etc.In forensic Analysis, thousands of files are typically examined and those files consist of unstructured text so it is a difficult task for forensic examiner to do such analysis in short time periods. Algorithms for clustering documents will facilitate the invention of new and useful information from the documents under analysis. Clustering algorithms are partitional (k-mean, k-medoids) and hierarchical (Single/Complete/Average) clustering for finding relevant documents from huge amount of data and relative validity index is use to automatically estimate the number of cluster, result will be display in dendrogram; results of dendrogram is useful for expert examiner.

KEYWORDS: Forensic analysis, clustering algorithms, text mining.

I. INTRODUCTION

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document-clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances [4].

The Forensic Investigation is the process in which the digital devices like computers are used to analyze the digital evidence from a particular computing device in a way that is proper for presentation in a court of act. It also deals with the preservation, identification, extraction as well as documentation of digital evidences. It is task of analyzing huge number of files from computer seized devices. But in computer forensic procedure all the information and files are stored in digital form. This digital information stored in computer seized devices has an important factor from an investigative point of view which treated as evidence in the court of law to prove what occurred based on such evidences [2].

Clustering algorithms are generally used for exploratory data analysis, where there is very little or no previous data about the data. From a lot of technical viewpoint, data contains unlabeled objects—the categories or classes of documents that will be found are a priori unknown. Moreover, even assuming that labelled datasets may be available from previous analyses, there is almost no hope that a similar category would be still valid for the upcoming data, obtained from different computers and associated to different investigation processes. More exactly, it is possible that the new knowledge sample would come from a different population. In this context, the utilization of cluster algorithms, which are capable of finding latent patterns from text documents found in seized computers, will enhance the analysis performed by the professional examiner [1].

The information knowledge partition has been induced from data; then expert examiner may initially target reviewing representative documents from the obtained set of clusters. Then, after this preliminary analysis, he could eventually commit to alternative documents from every cluster. By doing so, one can avoid the hard task of examining all the documents but, even if so desired, it still may be done [1].

Clustering algorithms have been studied for decades, and the literature on the subject is huge. Therefore, we decided to choose a set of representative algorithms in order to show the potential of the proposed approach, namely: the partitional K-means [4] and K-medoids [5], the hierarchical Single/ Complete/Average Link [6]. These algorithms were run with different combination of their parameters, and compare their relative performances on the studied application domain using five real world investigation cases conducted by Brazilian Federal Police Department. Use relative validity index (silhouette) to estimate the number of cluster automatically from data [1].

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

II. RELATED WORK

L.F.C Nassif has been proposed an approach that applies document clustering algorithms for the forensic analysis of computer devices. They illustrated an approach by carrying out wide experimentation with six well known clustering algorithms (K-mean, K-medoids, Single Link, Average Link, complete Link and CSPA) applied to five real world datasets obtained from computer seized. They were also studied uses of the comparative validity index criteria for the estimating the number of clusters in an automated manner which overcomes the limitations of previous techniques [1].
There are studies regarding use of clustering algorithms in the field of Computer Forensics and other fields related to text analysis of text documents. Most of the studies describe the use of algorithms for clustering data e.g., Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, K-means, Fuzzy C-means (FCM), and Self- Organizing Maps (SOM). K-means and FCM can be seen as particular cases of EM [12]. Algorithms like SOM [13]generally have inductive biases similar to K-means but are usually less computationally efficient [9].
A. Maind has been proposed approach the forensic analysis was done very scientifically i.e. retrieved data is in unstructured format get particular structure by using high quality well known algorithm and automatic cluster labelling method. Two relative validity indexes were used to repeatedly approximation the amount of clusters with automatic labeling to it; which makes it very easy to retrieve most relevant information for forensic analysis. They proposed hybrid hierarchical algorithm such as density based clustering such as DBSCAN algorithm which had many features such as Discover clusters as random shapes, Handle noise and one scan. Which was better to achieve fast and efficient analysis [7]?
S.Oliver have proposed SOM-based algorithms were used for clustering files with intend of making the decision-making process performed by the examiners more efficient. The files were clustered by taking into account their creation dates/times and their extension. That kind of algorithm has also been used to cluster the results from keyword searches. The underlying hypothesis was that the clustered results can increase the information retrieval efficiency, because that could not be necessary to review all the documents found by the user [8].
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. Aimed at relaxing this assumption, which is often unrealistic in practical applications, a common approach in other domains involves estimating the number of clusters from data. Essentially, one induces different data partitions (with different numbers of clusters) and then assesses them. With a relative validity index in order to estimate the best value for the number of clusters. This work makes use of such methods, thus potentially facilitating the work of the expert examiner—who in practice would hardly know the number of clusters a priori [2], [3], [9].

III. METHODOLOGY

Pre-processing Steps :

Before execution of clustering algorithms on text datasets, weperformed some preprocessing steps. Especially, stop words (prepositions, pronouns, articles, and irrelevant documentmetadata) are removed. Also, the Snowball stemming algorithm for Portuguese words has been used. Then, weadopted a conventional statistical approach for text mining, in which documents are described in a vector space model [14]. During this model, every document is described by a vector containing the frequencies of occurrences of words. Use a dimensionality reduction technique known as Term Variance (TV) that can increase both the effectiveness and efficiency of clustering algorithms. TV selects a number of attributes (in our case 100 words) that have the greatest variances over the documents. In order to compute distances between documents, namely: cosine-based distance [14].

Estimate Number of Cluster :

First to get a set of data partitions with different numbers of clusters and then selecting that particular partition that provides the best result according to a specific quality criterion called silhouette. Such a set of partitions may result directly from a hierarchical clustering dendrogram or, alternatively, from multiple runs of a partitional algorithm (e.g., K-means) starting from different numbers and initial positions of the cluster.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Let us consider an object i belonging to cluster A. The average dissimilarity of i to all other objects of A is denoted by a(i). Now let us take into account cluster C. The average dissimilarity of object i to all other objects of C will be called d(i,C). After computing d(i,C) for all clusters C ≠ A, the smallest one is selected, i.e. b(i)=min d(i,C), C ≠ A This value represents the dissimilarity of i to its neighbour cluster and the silhouette s(i) is,

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

S(i) is always lies in between -1 to 1, Thus the higher s(i) is better assignment of object i to a given cluster. If s(i) is equal to zero then it is not clear whether the object should have been assign to current cluster or to neighbour. Compute silhouette of all object in cluster and the find average. The best partition is has the maximum average silhouette.

Clustering Algorithms :

The clustering algorithms is popular in data mining fields is partitional k-means[4] and k-medoids[5], and hierarchical clustering Single/Complete/Average link algorithm[6].
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 mining.

K-medoids [5] is similar to K-means. However, instead of computing centroids, it uses medoids, which are the representative objects of the clusters. This property makes it particularly interesting for applications in which (i) centroids cannot be computed; and (ii) distances between pairs of objects are available.
K-means and k-medoids are best if they are sensitive to initialization and usually converge to local minima. To solve this problem use non-random initialization in which distant objects from each other are chosen as starting prototypes [4].
Hierarchical clustering algorithm provides a nested partitioned, and output of hierarchical algorithm is represented by using dendrogram, in dendrogram each pair of object is clearly shown. Single link use a smallest value between object and assign to cluster, complete link use a largest value and average link find average value of both object and then assign to cluster. To estimate number of cluster get result of hierarchical algorithm and then find silhouette [5]. Run clustering algorithm with all attributes and 100 attributes having greatest variance over documents.
For the hierarchical algorithms (Single/Complete/Average Link), simply run them and then assess every partition from the resulting dendrogram by means of the silhouette [5]. Then, the best partition is taken as the result of the clustering process. For each partitional algorithm (K-means/medoids), execute it repeatedly for an increasing number of clusters. For each value of K a number of partitions achieved from different initializations are assessed in order to choose the best value of K and its corresponding data partition, using the Silhouette [5], use all the possible value of k in the interval [2,N] where N is number of objects to be clustered.

IV. EXPERIMENTAL RESULTS

Dataset

In explicit, any kind of content that is digitally compliant may be subject to investigation. Within the datasets assessed in our study, for example, there are textual documents written in several languages (Portuguese and English). Such documents are originally created in several file formats, and a few of them are corrupted or are literally incomplete within the sense that they have been partially recovered from deleted information [10].
Five dataset of real world investigation case conducted by Brazilian federal police department obtain from author of our base paper Luis Filip da cruz Nassif, the dataset is in csv file format that contain relative frequency of words per documents. Due to privacy issue detail information about the document names and their respective bags of words are not provided.
The datasets contain varying amounts of documents (N), groups (K), attributes (D), singletons (S), and number of documents per cluster (#), as reportable in below table.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Table 1: Dataset Characteristics

Dataset	N	K	D	S	# largest cluster
A	37	23	1744	12	3
B	111	49	7894	28	12
C	68	40	2699	24	8
D	74	38	5095	26	17
E	131	51	4861	31	44

N contain number of documents in particular dataset and k is a number of cluster computed by silhouette and D is attributes of file during preprocessing get bags of word or attribute of file and S is Singleton, cluster contain only one object those cluster is called as singleton. #largest cluster, after performing clustering algorithm number of cluster is found then calculate which one is largest and number of object in it.

Results

Perform algorithm on dataset all characteristics of dataset is shown in table 1. Perform partitioning algorithm on data with different number of cluster varies from 2 to number of document in dataset and then calculate silhouette of each partition, the maximum value of silhouette is best number of cluster for dataset. Another way to perform hierarchical clustering on data we get nested partition use those partition to calculate silhouette and then get a result of number of cluster and best partition.

Clustering algorithms is performing on all attributes (Count All) as well as perform on 100 greatest variance attributes (TV>100) and calculate which one is give a best result, the following figures show result of Average link algorithm with 100 attributes uses a similarity distances.

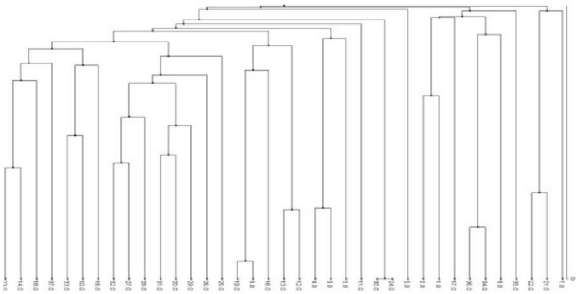

Fig: Dendrogram of AL100 – Dataset A

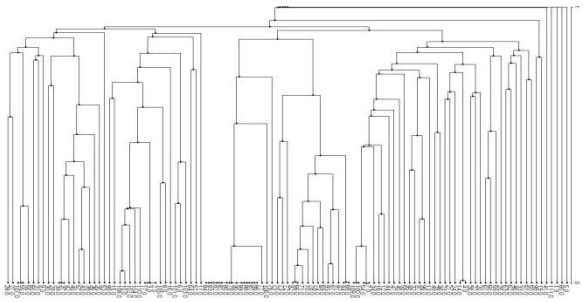

Fig: Dendrogram of AL100 – Dataset B

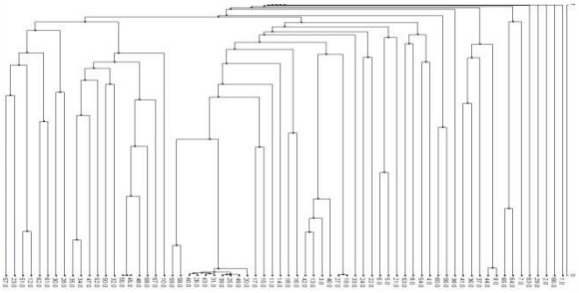

Fig: Dendrogram of AL100 – Dataset C

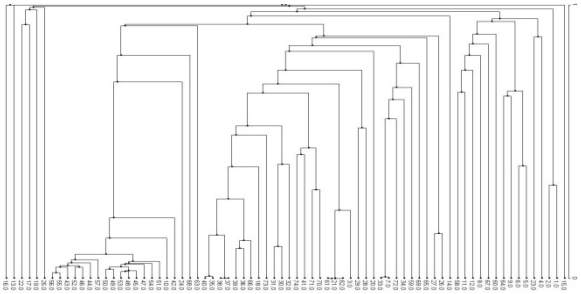

Fig: Dendrogram of AL100 – Dataset D

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

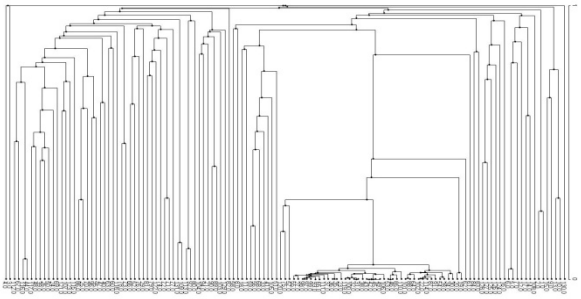

Fig: Dendrogram of AL100 – Dataset E

V. CONCLUSION

The document clustering methods for forensic analysis of computers seized in police investigations. Also reported and discussed several practical results that can be very useful for researchers and practitioners of forensic computing. More specifically, hierarchical algorithms known as Average Link and Complete Link present a best result. The dendrograms provide offer summarized views of the documents being inspected, thus being helpful tools for forensic examiners that analyse textual documents from seized computers. As already observed in other application domains, dendrograms provide very informative descriptions and visualization capabilities of data clustering structures.

Most importantly, the clustering algorithms indeed tend to induce clusters formed by either relevant or irrelevant documents, thus contributing to enhance the expert examiner's job. Furthermore, the evaluations of approach in five real-world applications show that it has the potential to speed up the computer inspection process.

REFERENCES

[1] L.F.D.C Nassif and E.R. Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection", IEEE Transactions on Information Forensics and Security, Vol. 8, No. 1, January 2013.
[2] 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.
[3] 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.
[4] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[5] L. Kaufman and P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.
[6] R. Xu and D. C.Wunsch, II, Clustering. Hoboken, NJ: Wiley/IEEE Press, 2009.
[7] R. Mundhe, A. Maind and R. Talmale, "Information Retrieval Using Document Clustering for Forensic Analysis", International Journal of Recent Advances in Engineering & Technology (IJRAET), Vol. 2, 2014.
[8] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver, "Exploring forensic data with self-organizing maps", Internatinal Conference Digital Forensics, 2005.
[9] 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.
[10] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 6, pp. 835–850, Jun. 2005.
[11] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, "Relative clustering validity criteria: A comparative overview," Statist. Anal. Data Mining, vol. 3, pp. 209–235, 2010.
[12] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer-Verlag, 2006.
[13] S. Haykin, Neural Networks: A Comprehensive Foundation. Engle- wood Cliffs, NJ: Prentice-Hall, 1998
[14] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988
[15] S.S. Ansari, T.M. Pattewar, "Forensic Analysis Using Document Clustering: A Survey" in Proc. IJMTER Conf. Nasik, India ,Volume 3, Issue 4, April 2016 Special Issue of ICRTET'2016.