ISSN 2395-1621

Secured Forensic Analysis with document Clustering : For Improving Computer Verification

[#1]Saurabh Singh Negi, [#2]Bhagwan Patil, [#3]Rahul Baccchav, [#4]Pratik Patil

[1]Sauravnegimkvn@gmail.com,
[2]patilbhagwan93@gmail.com,
[3]rahul.bacchav.22@gmail.com,
[4]pratikpatil933@gmail.com

[#1234]Department Of Computer Engineering, TSSM's P.V.P.I.T Collage Of Engineering Pune, India

ABSTRACT

Document clustering has shown to be very useful for computer inspection for forensic analysis. In computer forensic investigation, thousands of files are usually served. Clustering algorithm is typically used for the exploratory data analysis, Where there is little or no prior knowledge about the data. Those data of these files are not structured format, it is very complicated to cluster that data or analyze that data. We can use clustering algorithm for analyze that data, these algorithm are clustered all files and discovered a new knowledge from document under the analysis. We are use these algorithm to document clustering for forensic analysis of computer seized by investigators. Those all computer seized the various tools are used to abstract the valuable information from that devices. In this paper we are proposed approach by carrying out extensive experimentation two known clustering algorithm (K-means,K-mediodes) applied to five real world datasets obtained from computer seized in real world investigations. In this project we have performed with various combinations of parameters, resulting in 16 different instantiations of algorithms, addition is two relative validity indexes were used to automatically build the numbers of clusters. Average link and complete link algorithm provide the good result for application domain and k-means & k-mediodes are also providing a very good result.In this paper we can facilate the security of the cluster data with the help of ASE algorithm. Here the work is inspired by motivation of finding conceptual cluster among the given document set. And finally we can present technical result that can be useful for investigator for investigation.

ARTICLE INFO

Article History

Received :15th April 2016

Received in revised form :

17th April 2016

Accepted : 19th April 2016

Published online :

23rd April 2016

Keywords- K-Means & K-Mediodes,Text Mining,AES Algorithm,Clustring,Forensic Computing

I. INTRODUCTION

It is prove, that todays digital world volume of data is increased from 161 hex bytes in 2006 to 988 hex bytes in 2010.[1].So these large unstructured data has direct impact of computer forensic investigation. Machine learning and data mining is important methods for automated data analysis. In particular, algorithms for pattern recognition from information present in text document are promising, as it will hopefully become evident later in these paper.

Clustering analysis are used for unstructured data analysis where small knowledge or information of that data.[2][3].Clustering is division of data into groups of similar objects each group is called as clusters ,our dataset is consist of un-balanced objectthe classes or category of document that can be found are unknown. In these paper the use of clustering algorithm, which is Supporting to find latent pattern from text document found in seized

computers, can enhance the analysis performed by expert investigator.After the clustering of all documents, we are providing the security of that analyzed data with the help of advance encryption standard; we are providing the security of that data. AES algorithm requires a 128-bits round key block.
For implementation of AES there are 4 basic steps are important.

A)Sub bytes
B)Shift rows
C)Mix columns
D)Add round key
E)Final round (Mix column columns)

Finally these paper describe the issues related to the forensic analysis system and what actions to be performed by the development.

II. RELATED WORK

In previous all Researches ,There are very few studies reporting the use of clustering algorithm in the computer forensic ; but every research explains the traditional algorithm for clustering data. e.g.(EM) Expectation-Maximization for unsupervised learning of Gaussian mixture models, k-means , fuzzy c-means(FCM) & self organising in maps (SOM).

K-means & FCM can be explained in a very few cases of EM algorithm like , in their turn generally have inductive basics similar to k-means but are usually less computationally efficient.

K-Means Algorithm:

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.
These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bary centers of the clusters resulting from the previous step.
After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done.
In other words centroids do not move anymore finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function ,Where is a chosen distance measure between a data point and the cluster centre , is an indicator of the distance of the *n* data points from their respective cluster centres.
The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.

Advanced Encryption Standard (AES)

Advanced Encryption Standard (AES) Published by NIST in Nov 2001: FIPS PUB 197 FIPS PUB 1970 Based on a competition won by Rijmen and Daemen .

1.**KeyExpansions**—round keys are derived from the cipher key using Rijndael's key schedule. AES requires a separate 128-bit round key block.
2.AddRoundKey—each byte of the state is combined with a block of the round key using bitwise xor.
3.Rounds
SubBytes—it's substitution step where each byte is replaced with a subByte using s-box. S box is matrix given by rijndael.
ShiftRows—a transposition step where the last three rows of the state are shifted cyclically a certain number of steps.
MixColumns—a mixing operation which operates on the columns of the state, combining the four bytes in each column.

a. AddRoundKey
b. Final Round (no MixColumns)
c. SubBytes
d. ShiftRows
e. AddRoundKey.

1] **In the SubBytes step**: each byte in the state matrix is replaced with a SubByte using an 8-bit substitution box, the Rijndael S-box.

2]**Shift-row step :**
The first row is left unchanged. it cyclically shifts the bytes in each row.

3] **MixColumn step:**
The four bytes of each column of the state are combined using an invertible linear transformation.

4] **AddRoundKey step:**
In the this step, the subkey is combined with the state. For each round, a subkey is derived from the main key using Rijndael's key schedule;

III. EXISTING SYSTEM

Clustering algorithms are typically used for exploratory data analysis, where there is little or no prior knowledge about the data. This is precisely the case in several applications of Computer Forensics, including the one addressed in our work. From a more technical viewpoint, our datasets consist of unlabelled objects—the classes or categories of documents that can be found are a priori unknown. Moreover, even assuming that labelled datasets could be available from previous analyses, there is almost no hope that the same classes (possibly learned earlier by a classifier

in a supervised learning setting) would be still valid for the upcoming data, obtained from other computers and associated to different investigation processes. More precisely, it is likely that the new data sample would come from a different population. In this context, the use of clustering algorithms, which are capable of finding latent patterns from text documents found in seized computers, can enhance the analysis performed by the expert examiner. The rationale behind clustering algorithms is that objects within a valid cluster are more similar to each other than they are to objects belonging to a different cluster. Thus, once a data partition has been induced from data, the expert examiner might initially focus on reviewing representative documents from the obtained set of clusters. Then, after this preliminary analysis(s) he may eventually decide to scrutinize other documents from each cluster. By doing so, one can avoid the hard task of examining all the documents (individually) but, even if so desired, it still could be done.

IV. PROPOSED SYSTEM

Clustering algorithms have been studied for decades, and the literature on the subject is huge. Therefore, we decided to choose a set of (six) representative algorithm in order to show the potential of the proposed approach, namely: the partitioned k-means k-medoids, the hierarchical single/complete/average link, and the cluster ensemble algorithm known as cspa. these algorithms were run with different combinations of their parameters, resulting in sixteen different algorithmic instantiations. Thus, as a contribution of our work, we compare their relative performances on the studied application domain—using five real-world investigation cases conducted by the Brazilian Federal Police Department.

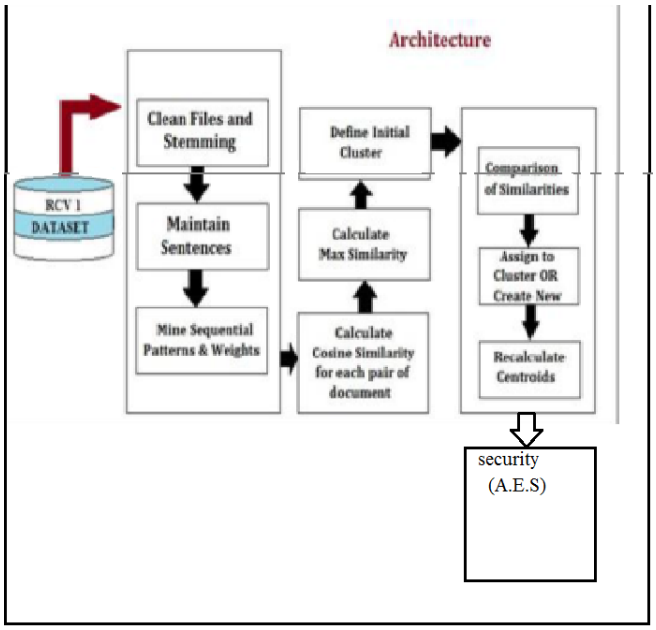

Fig 1. System Architecture

In order to make the comparative analysis of the algorithms more realistic, two relative validity indexes have been used to estimate the number of clusters automatically from data.

V. MODULES IMPLEMENTED

- GUI modules
- File Upload and Security
- Clustering
- Search and download

1.**GUI modules:-**

GUI modules represents the outlook of our project, In computer science, a **graphical user interface** or **GUI**, pronounced is a type of interface that allows users to interact with electronic devices through graphical icons and visual indicators such as secondary notation, as opposed to text-based interfaces, typed command labels or text navigation. GUIs were introduced in reaction to the perceived steep learning curve of command-line interfaces (CLIs) which require commands to be typed on the keyboard.

2. **File Upload and Security**

File uploads module is basically used for taking a file from our computer hard drive, and we analyse also that what is our file id no and that its pop a message regarding "file uploads successfully".and then it checks for the security by encryption and then it encrypted by 128 bit AES algorithm ,we can done our encryption and this store temporarily inside admin folder by the .des extension and after downloading it is use for the user information .By this data remains safe and only used by the admin.

3.**Clustering**

Clustering is a process of cluster the data into indexes which we fixed such as in our projects we make 3 folder of robbery ,drugs ,murder.so clusters the data can be done when we uploading the data and data then only get changes into clusters of this particular indexes which we fixed .
As we know there is use of k-mean algorithm by which we can done this clustering successfully.

4. **Search and download**

After we done with the admin then now the user comes into pictures and it search for the query ,related query which can be search by the user can be download and then this information can be fetched by the user.

VI. FUTURE WORK

The more effective technique such machine learning algorithm for finding sequential patterns can be adopted. Dictionary of found sequential patterns may be created for Sub-sequent processing to achieve more time efficiency. As the number of clusters is not user defined, the number of cluster may increase as sparse document increase, so we can move towards predefined number of cluster method.

VII. CONCLUSION

Clustering has a number of applications in every field of life. We are applying this technique whether knowingly or unknowingly in day-to-day life. One has to cluster a lot of thing on the basis of similarity either consciously or unconsciously. Clustering is often one of the first steps in data mining analysis. The partitioned K-means algorithm also achieved good results when properly initialized. Considering the approaches for estimating the number of clusters, the relative validity criterion known as silhouette has shown to simplified version. It identifies groups of related records that can be used as a starting point for exploring further relationships. In addition, some of our results suggest that using the file names along with the document content information may be useful for cluster ensemble algorithms. Most importantly, we observed that clustering algorithms indeed tend to induce clusters formed by either relevant or irrelevant documents, thus contributing to enhance the expert examiner's job.

ACKNOWLEDGEMENT

The authors would like to thanks Prof. N.D.Kale and anonymous referees for their constructive remarks and valuable comments.

REFRENCES

[1] Ieee Transection on information Forensics And Security ,vol.8,No.1,January 2013 Luis filipe da cruz ,Nassif and eduardo Raul Hruschka.

[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] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[4] L. Kaufman and P. Rousseeuw, Finding Groups in Gata: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.

[5] R. Xu and D. C.Wunsch, II, Clustering. Hoboken, NJ: Wiley/IEEE Press, 2009.

[6] A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions," J. Mach. Learning Res., vol.
3, pp. 583–617, 2002.

[7] 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

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

[9] 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.

[10] 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.

[11] 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.

