An Approach for Document Clustering in Forensic Analysis Using Labeling.

Sahil Vishnu Bansal [1], Hem Singh [2] , Sahil Pratap Jagdale [3], Rupali Shishupal [4]

[1,2,3]*Student, Computer Department, Sinhgad Institute of Technology, Lonavala, India*

[4]*Assistant Professor, Computer Department, Sinhgad Instituteo f Technology, Lonavala, India*

Abstract- **We present an approach that applies document clustering algorithm to forensic analysis of computer seized in police investigations. We also present a better approach in the field of forensic computing using automatic labeling and avoiding overlapping of clusters to improve response time of the search.**

Keywords- Stemming, Document clustering, Forensic computing, Text mining, Labeling.

I. INTRODUCTION

A. Fundamental Concepts On (Domain)

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information-information that can be used to increase revenue, cuts costs, or both.

Computer forensics includes analyzing huge number of files from seized computer devices[1]. Clustering is the process of organizing objects into groups whose members are similar in some way. Labelling is the problem of picking descriptive human readable labels for the clusters produced by a document clustering algorithms.

B. Contribution

1) *Forensic Analysis*: Forensic Analysis is the use of controlled and documented analytical [1] and investigative techniques to identify, collect, examine and preserve digital information.
2) *Clustering*: Clustering [2] is the process of organizing objects into groups whose members are similar in some way.
3) *Labeling*: Labeling[3] is the problem of picking descriptive human readable labels for the clusters produced by a document clustering algorithms.
4) *Computer Forensics*: Computer forensic is a branch of digital forensic science pertaining to legal evidence found in the computers and digital storage media.

II. LITERATURE SERVEY

A. *Existing System*

The literature on computer forensic only reports the use of algorithms for clustering [1] the document in digital devices and examined officially by police departments where cluster is known and fixed according to the priority set by the user. Essentially, one includes different data partitions and accesses them with relative validity index in order to estimate the best value for the number of clusters.

Current clustering techniques [4] do not address all the requirements adequately for investigation and to examine the digital devices officially. Dealing with the 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 [7], especially in multidimensional spaces. The result of the clustering algorithm (that is in many cases can be arbitrary itself) can be interpreted in different ways.

III. PROPOSED SYSTEM

Doing the survey on computer forensic analysis we can say that the clustering on data is not an easy step. There is a huge data to be clustered in compute forensic so to overcome this problem, this paper presents an approach that applies document clustering methods to forensic analysis of computers that are seized in police investigations. In computer forensic investigations, usually thousands of files are surveyed. The data in those files consists of formless manuscript; it is very tough to accomplish for the computer examiners of investigation. Clustering is the unverified organization of designs that is data items, remarks, or feature vectors into groups (clusters)[2]. Improved stemming algorithm is used to remove stop words in a document. An approach of automatic labeling is also proposed which makes a system more smart and speeds up the process of searching.

Here, we decided to choose a set of representative algorithm in order to show the potential of the proposed

approach, namely; Partitional K-means, K- medioids, cluster ensemble algorithm known as CSPA[7].

IV. RELATED WORK

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. Different algorithms are used in previous work for clustering and searching desired file, but response time, accuracy and overlapping of files are some issues that are to be overcome.

V. IMPORTANT COMPONENTS (MODULES)

A. Preprocessing Steps And Stemming

Before running a clustering algorithm on text, dataset pre-processing is performed particularly 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. Each document is represented by a vector containing the frequencies of occurrences of words, whose number of characters are between 4 and 25. We have also used a dimensionality reduction technique known as Term Variance (TV) which can increase both effectiveness and efficiency of clustering algorithm in pre-processing step.

B. Cluster Vectors

After the preprocessing step gets processed, next step is preparing the cluster vector in which one needs to find out the top 100 words from the file. Those words or a file can contain numbers, characters (like date, place, and address)[1].

C. *Forensic Database*

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. *Automatic Labeling*

Labeling is a technique in which various keywords are assigned to cluster vectors according to the priority of most frequent words occurring in the searching process[3]. Therefore, making it easy to identify the proper cluster and its information present in the system.

VI. DISCUSSION AND FUTURE WORK

Support for audio and video files can also be processed but require super-fast algorithms which can handle large database and gives results accurately in desired amount of time. Choosing right algorithms and collaborating some clustering techniques, it can be achieved.

VII. 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 investigations. There are several practical results based on our work which are extremely useful for the experts working in forensic computing departments. 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 and corresponding output is given. When proper initialization is done, the partitioned K-means and K-mediods algorithms also givesatisfactory results.

REFERENCES

[1] Luís Filipe da Cruz Nassif and Eduardo Raul 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] Mr.Nitin S. Kharat , Prof. HarmeetKhanuja; "*Role of Text Clustering and Document Clustering Techniques in Computer Forensic Analysis: A Review*"; *International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 International Conference on Industrial Automation and Computing (ICIAC- 12-13th April 2014)*

[3] "*Automatically Labeling Hierarchical Clusters*"; http://www.firstgov.gov/

[4] "*Text Clustering for Digital Forensics Analysis*"; Sergio Decherchi1, Simone Tacconi2, Judith Redi1, Fabio Sangiacomo1, Alessio Leoncini1 and Rodolfo Zunino1.

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

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

[7] *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.*