Document Clustering For Forensic Investigation

Yogesh J. Kadam[1], Yogesh R. Chavan[2], Shailesh R. Kharat[3], Pradnya R. Ahire[4]

[1]Student, Computer Department, S.V.I.T. Nasik, Maharashtra, India
[2]Student, Computer Department, S.V.I.T. Nasik, Maharashtra, India
[3]Student, Computer Department, S.V.I.T. Nasik, Maharashtra, India
[4]Student, Computer Department, S.V.I.T. Nasik, Maharashtra, India

---***---

Abstract - ***In computer forensic analysis, hundreds of thousands of files are usually examined. Much of the data in those files consists of unstructured text, whose analysis by computer examiners is difficult to be performed. In this context, automated methods of analysis are of great interest. Algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis. We present an approach that applies document clustering algorithms to forensic analysis of computers seized in police investigations.***

Technical keywords: — *Clustering, forensic computing, text mining.*

1. INTRODUCTION

This large amount of data has a direct impact in Computer Forensics, Which can be broadly defined as the discipline that combines elements of law and computer science to collect and analyze data from computer systems in a way that is admissible as evidence in a court of law. In our particular application domain, it usually involves examining hundreds of thousands of and the cluster ensemble algorithm known as CSPA. These algorithms were run with different combinations of their parameters, resulting in sixteen different algorithmic instantiations. In computer forensic analysis, hundreds of thousands of files are usually examined.

Most of these data is unstructured text whose size is enormous; hence analysis of such data is difficult for computer examiner. Here comes the need of automated methods of analysis. Algorithms for clustering can present useful knowledge from the documents under analysis. An approach that applies document clustering algorithms for forensic analysis of computer seized at crime scene.

1.1 Problem Statement

The computers seized at crime scenes might have large data to be examined. The computer analysts are scarce and the data to be analyzed is vast in nature. Here arises the need of automated tools for analysis of this huge data. The system uses two clustering algorithms to form the clusters so that analysis of data becomes effortless.

2. ARCHITECTURE AND MODELING

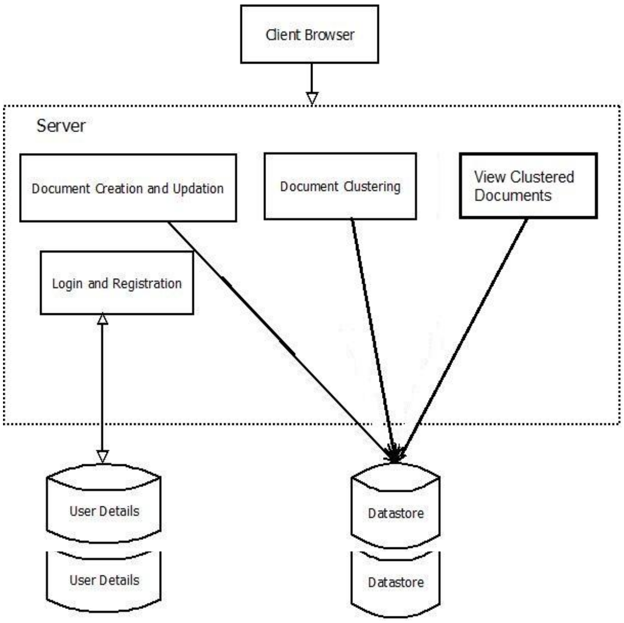


Fig -1: System Architecture

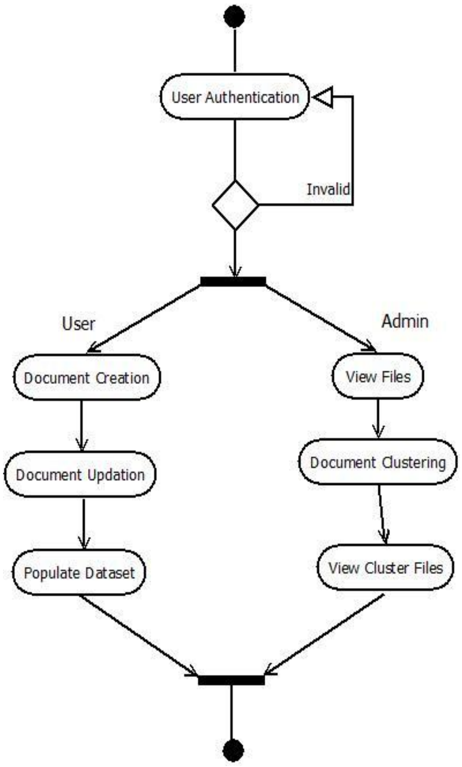


Fig-2: Activity Diagram.

2.1 Modules:

1. User Authentication: This module will enable the system to validate users and enable only valid users to access the application. The users will be facilitated to register themselves and access the application.

2. Document Upload: This module will facilitate the users to upload text documents (txt, rtf formats) which will be analyzed by the system for the clustering mechanism.

3. Admin Module: Here the admin will be facilitated to view the complete documents in the system.

4. Document Clustering : This module will be executed by the admin to segregate the information populated by the user into a set of cluster and will include the proposed mechanism with **cosine similarity matrix** analysis which will include procedures like stemming, tokenization, stopword removal and filtering as well as clustering process which will be finalized later.

5. **View Clusters & Download**: The clusters created will be visible here in segregated format for identifying the output of the above process.

3. MATHEMATICAL MODEL

3.1 Set theory:

A set is defined as a collection of distinct objects of same type on class of objects. The object of a set are called elements or members of the set. Object can be number, alphabet, names etc.

Eg:- A={1,2,3,4,5}

3.1.1 Union of sets:

Union of two sets A & B is defined to be the set of all those elements which belongs to set A or set B or both and is denoted by A U B.

3.1.2 Intersection of sets:

Intersection of two sets A & B is defined to be the set of all those elements which belongs to set A and set B

3.1.3 Difference of sets:

Union of two sets A & B is defined to be the set of all those elements which belongs to set A but do not belong to set B and is denoted by A-B.

4. LITERATURE SURVEY

Clustering algorithm helps to identify the accurate data from the analysis without less knowledge or no prior knowledge of data. Initially, computer forensics has objects which are unlabelled Previous analysis defines data partition from the data and expert examiner only focus on reviewing representative documents from the obtained set of cluster. Firstly, examiner check for investigation, after finding relevant document then the examiner can pass the analysis of the other document for investigation. Text clustering in digital evidence is defined as the data of the investigate value. Digital investigation is much necessary for textual evidence. The present digital forensic tools are used for analyzing many documents, which provided the multiple levels of searching techniques to answer the questions and generate digital evidence related to the specific investigation. However, these technique works improperly which allows the investigator to search for specific documents of certain subject of specified search and grouped the document set based on a given subject

Below are some links for literature survey

1. A Clustering-Based Approach for Integrating Document-Category Hierarchies

Link-
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4459764

2. A Variation on a Nonparametric Clustering Method

Link-
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4766948

3. Surface EMG Decomposition Based on K-means Clustering and Convolution Kernel Compensation

Link-
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6824163

4. Document Clustering for Forensic Computing: An Approach for Improving Computer Inspection

Link-
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6146981&newsearch=true&queryText=document%20cluster%20for%20forensic%20application

4.1 Feasibility Analysis, NP, NP-hard, NP-Complete, Traceability, Satisfy ability.

4.1.1 Feasibility Assessment:

If the running time is some polynomial function of the size of the input**, for instance if the algorithm runs in linear time or quadratic time or cubic time, then we say the algorithm runs in **polynomial time** and the problem it solves is in class **P**.

3 SAT problem is NP Complete. The system can be reduced to 3SAT problem. 3SAT problem takes a Boolean formula S that is in CNF in which each clause has exactly three literals. 3SAT is a restricted form of CNF-SAT problem.

- x1– User login and validating a user for authentication(ex. User name, PAN card no, driving lic no. electricity bill etc).

- x2- After successful validation user is supposed to pass a questionary examination. And accordingly labeled as Fraud or Clean user. And pay initial amount for registration and demo.

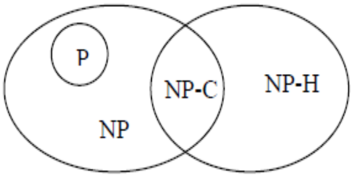
- x3 - Checking of demo model and accordingly allow user to bid on the desired project.

- S=(x1 v x2 v x3) . (x1 v ¬x2 v ¬x3). (x1 v x2 v ¬x3)

S stands true for each and every case because x1 is true in all cases.

Thus proved that problem is NP-Hard.

NP- Completeness:



Where-

NP-C: NP Complete

NP-H: NP Hard

Therefore, since our problem is NP and NP Hard, problem is **NP Complete.**

We know that all problems in P are also in NP. Therefore, since our problem is NP and NP hard (as proved above), problem is NP Complete.

5. ALGORITHM

5.1 Cosine Similarity Algorithm:

Document clustering is particularly useful in many applications such as automatic categorization of documents, grouping search engine results, building taxonomy of documents, and others. For this Hierarchical Clustering method provides a better improvement in achieving the result. Our paper presents two key parts of successful Hierarchical document clustering. The first part is a document index model, the Document Index Graph, which allows for incremental construction of the index of the document set with an emphasis on efficiency, rather than relying on single-term indexes only. It provides efficient phrase matching that is used to judge the similarity between documents. This model is flexible in that it could revert to a compact representation of the vector space model if we choose not to index phrases. The second part is an incremental document clustering algorithm based on maximizing the tightness of clusters by carefully watching the pair-wise document similarity distribution inside clusters. Both the phases are based upon two algorithmic models called Gaussian Mixture Model and Expectation Maximization. The combination of these two components creates an underlying model for robust and accurate document similarity calculation that leads to much improved results in Web document clustering over traditional methods.

HOW THEY WORK? Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is this:

STEP 1 - Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.

STEP 2 - Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less with the help of tf - idf.

STEP 3 - Compute distances (similarities) between the new cluster and each of the old clusters.

STEP 4 - Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

6. FUTURE SCOPE

The utility of system can be extended by adding various file formats. Other clustering algorithms can be added to improve the efficiency. Application areas can be widespread to other domains.

7. CONCLUSIONS

With proposed concept we can estimate the number of clusters automatically to get results. There are several practical results based on our work which are extremely useful for expert working in forensic department.

ACKNOWLEDGEMENT

We take this opportunity to express our hearty thanks to all those who helped us on this topic. We express my deep sense of gratitude to my Project guide Prof. S.A.Gade ,and Prof.S.M.Rokade, (Head Of Computer Engineering Department), Sir Visvesvaraya Institute of Technology, Nashik for their guidance and continuous motivation.
We would also like to thank Prof.S.A.Gade, Project Co-coordinator for his great support and excellent guidance.

REFERENCES

1. 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 throu2010," Inf. Data, vol. 1, pp. 1–21, 2007.

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

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.