An Automated Forensic Analysis Approach in Financial Domain

Nitin S. Kharat
Department of Computer Engineering,
MMCOE, Pune,

Harmeet.K . Khanuja
Department of Computer Engineering,
MMCOE, Pune

ABSTRACT
Recently due to increase in digitalization most of documents are stored electronically. So whenever any criminal case is reported then it is often required to do forensic analysis of such document dataset. Forensic analysis refers to analyzing such document dataset stored in computer seized device in order to resolve reported case. Nowadays numbers of crimes are reported related to financial domain. So we need to do forensic analysis of such financial documents, but it is very complex and time consuming task to do forensic analysis of such financial document dataset. Hence, the proposed system facilitates framework towards automated forensic analysis of financial documents. The proposed system does forensic analysis of listed financial documents in automated manner. The obtained result also shows significant improvement that leads to forensic analysis of such documents within quick period of time.

Keywords
Clustering, Forensic Analysis, DST, Data Mining

1. INTRODUCTION
Recently due to digitalization in world, all the documents are stored inside computer in electronic form. As digitalization increases there is also increase in crime inside digital world. Also survey [1] notifies that there is also rapid increase of financial crimes in every year. So when such case is reported in forensic department to investigate financial crime, they need to adopt forensic analysis of such financial documents. Forensic analysis refers to analyzing set of documents acquired from computer seized devices. But document dataset acquired from computer seized device is composed of relevant as well as non-relevant documents in accordance with reported case. But to find such relevant and non-relevant documents from acquired document dataset requires analysis of each and every documents individually which leads to time consuming and complex approach. Hence ultimately such forensic analysis causes too much delay in delivering results to court of law. If the crime is in financial domain then again it is very complex task to find relevant and non-relevant document from acquired document dataset. After detection of relevant and non-relevant documents, forensic analysis tends to perform individual analysis of such document dataset. So there is need of an automated approach to find relevant and non-relevant financial documents acquired from computer sized devices which further performs forensic analysis of those relevant financial documents to investigate reported case. So proposed system provides an approach to detect relevant and non-relevant financial documents and proceeds it towards further forensic analysis of such relevant financial documents in automated manner.

2. LITERATURE SURVEY
Document clustering algorithms plays vital role in the field of computer forensics. The Self Organizing Maps (SOM) algorithm is used in [2] to enhance decision making ability of computer forensic investigator. They proposed SOM based architecture to support decision making by computer forensic investigators. Specific pattern can be detected by using SOM. Hence SOM is used by forensic examiner to search the interested pattern in huge forensic data set. Here the forensic examiner is only interested in particular pattern set from huge data set. But clustering was based on different parameters. In [3] first clustering was applied and also it divides documents into different sections to enhance searching. But there was limited work on textual data for further analysis of financial documents as well as it also needs to specify number of clusters a priori. In FCM based techniques [4], they applied clustering on raw data and then extraction of membership function from data, but in real time it is very hard to fix membership function according to number of clusters. They have also made use of fuzzy inference system and fuzzy clustering, but limitation of this approach is to specify number of clusters for dimensions so that meaningful member function can be generated. In [5], they have employed clustering based text mining techniques for analysis of data set in order to simplify the job of forensic examiner.

They proposed framework for digital text analysis which makes use of clustering based text mining techniques for the analysis of reported case. Hence it works as baseline for analysis of acquired data by the forensic examiner. The framework integrates two phase's first textual information extraction and second textual data analysis via clustering based text mining tools. They made use of kernel K-means algorithm for clustering. The clustering was based on extraction of frequent words. But limitation of this system is user need to specify number of clusters before applying clustering. But in real time it is hard task to specify number of clusters because user does not know similarity of underlying dataset. After clustering the forensic examiner has to elaborately analyze relevant individual document to resolve reported case. But after the clustering they have limited attention on actual content of document during forensic analysis phase, it will need to be performed by forensic examiner manually after clustering and also they expect number of clusters from user before clustering. In [6], they propose post-retrieval clustering of digital forensic text string search results. Basically current digital forensic text string search tools use match and indexing algorithms to search digital evidence at the physical level which locate specific text strings .But this leads to an extremely high incidence of hits that are not relevant to investigative objectives. This approach works on to group or order the search results hits in a manner that improves the investigators ability to get the relevant hits first in quick time. They proposed post –retrieval clustering of digital forensic text string search results specifically by using Self Organizing Maps (SOM).The objective of paper is to improve the information retrieval effectiveness in digital forensic text string searching. But they have less worked on post processing of clustered results. In [7], they have made survey of clustering algorithms adopted

in financial fraud detection, but clustering was bounded to definite domain. Document clustering was not adopted before analysis to classify documents from acquired document dataset. But, in computer forensics it deals with analyzing large dataset for investigational purpose; hence it is a prior task to detect financial documents via document clustering. After document clustering, forensic examiner only analyzes relevant financial cluster for further analysis rather than analyzing irrelevant documents. Recently, in [8], they introduces technique that applies document clustering on computer seized devices which produces different clusters such that specific cluster contains same type of documents. They tested document dataset with six well known clustering algorithms and compares six well known clustering algorithms in order to analyze performance of different clustering algorithm. Since there is no need of analyzing irrelevant document, this approach reduces overhead of forensic examiner while analyzing computer seized devices for forensic analysis. They have also achieved dynamic clustering of different document and reduces overhead of specifying number of clusters by user. But, they have less concentrated on actual data within documents of different cluster and no post processing is performed on content of clustered documents. From literature it is also observed that there are fewer attempts in design of automatic forensic analysis framework for financial documents in accordance with forensic examiner criteria. Literature also notifies that, there are very few forensic analysis frameworks employing real time banking rules for suspicious transaction [9], [10] in identifying suspicious transaction from log file of banking application.

Additionally, proposed system also integrated banking rules along with Dempster Shafer Theory (DST)[11],[12] for combining multiple evidences as well as forensic examiners own rule in forensic analysis of listed financial documents.

3. PROPOSED SYSTEM

As shown in fig.1, dynamic document clustering is applied on dataset acquired for forensic investigation which outputs different document clusters according to content similarity of document. So user need not view irrelevant document clusters instead user can only view relevant financial cluster. Another beauty of this approach is that user need not to specify number of clusters before clustering which is very hard task, instead proposed system automatically generates clusters. After clustering, proposed system also works on content of specific cluster for listed type of financial documents, so user need not to analyze each and every document individually that resides in cluster.

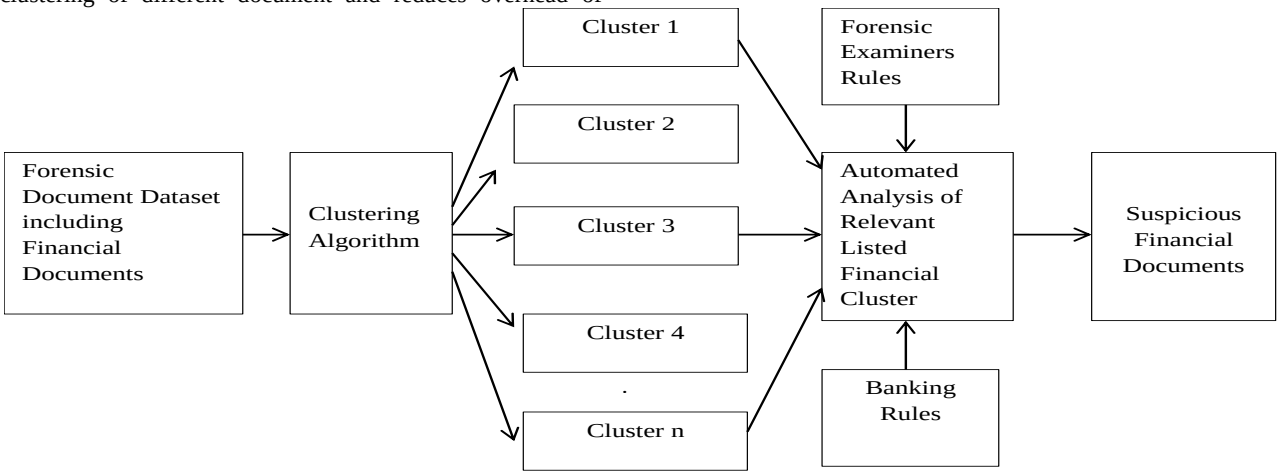

Fig 1: Proposed System Architecture

The proposed system performs automated forensic analysis of listed financial documents including financial receipts, transactions captured from log file of banking application and financial reports. The forensic analysis of financial receipts is done based on keyword total and amount entered by forensic examiner. The forensic analysis of transactions captured from log file of banking transaction is done by using RBI (Reserve Bank of India) and DST (Dempster Shafer Theory) Rules [11],[12] which lists suspicious transaction. The forensic analysis of financial annual reports is performed based on financial subdirectories keywords such as 'total', 'total asset' which helps to speed up forensic analysis system. The results are also shown in further section. Hence, proposed system provides framework for automated forensic analysis of listed type of financial documents. Ultimately, proposed system enhances the process of forensic analysis for financial documents.

4. PERFORMANCE ANALYSIS

4.1 Cluster Generation

Proposed system will clusters documents into different clusters based on result of vector generation model according to content similarity. Also, beauty of this approach is no need to specify number of clusters by user; instead numbers of clusters are estimated automatically from content of document dataset. The result is shown in fig.2.

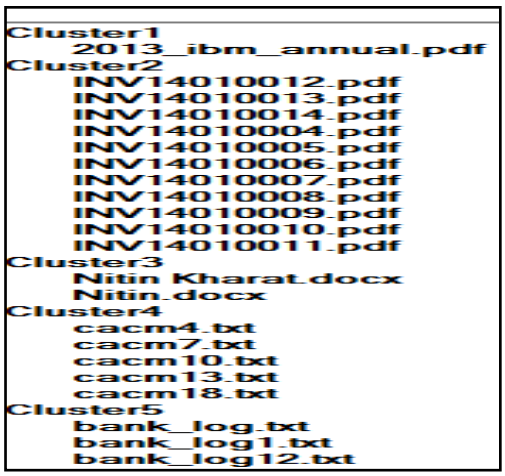

Fig 2: Generated Clusters

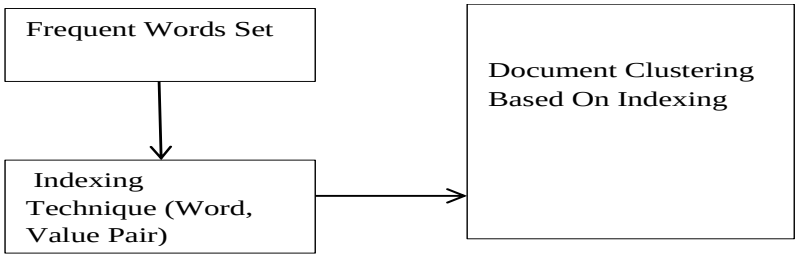

Fig 3: Proposed Indexing Technique

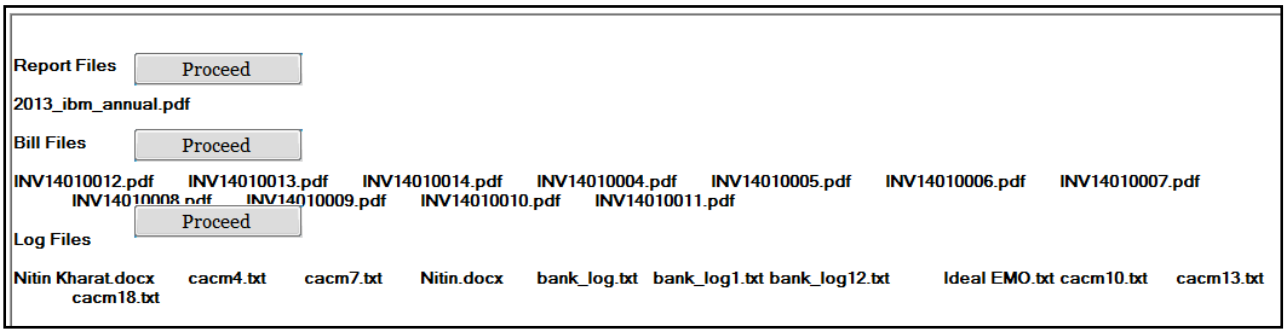

Fig 4: Selection of Clusters based on Provided Criteria

4.2 Use of Indexing in Clustering to Achieve Improved Performance

In order to improve performance of clustering, we also aimed to make use of indexing techniques. Indexing in document clustering plays vital role to achieve significant performance. The aimed indexing is as shown in fig.3.

4.3 Result of Criterion of Clusters

After cluster generation, we aimed automated forensic analysis of listed financial documents. As, shown in fig.4. those listed financial clusters are parsed for deciding suspicion based on banking rules, private banking rules (e.g. Syndicate bank, Canara bank), DST rules[11],[12] as well as forensic examiners own rules.

4.4 Result of Automated Financial Receipt Analysis

Proposed system also aimed automated forensic analysis of financial receipts as shown in fig.5. It will list suspicious financial receipts that exceeds threshold total amount specified by forensic examiner.

Fig 5: Result of Automated Financial Receipt Analysis

4.5 Result of Automated Annual Report Analysis

In automated forensic analysis of financial annual reports, as annual reports are usually huge in size. So to minimize overhead of forensic analysis proposed system highlights financial subdirectories keywords in order to enhance the process of forensic analysis as shown in fig.6.

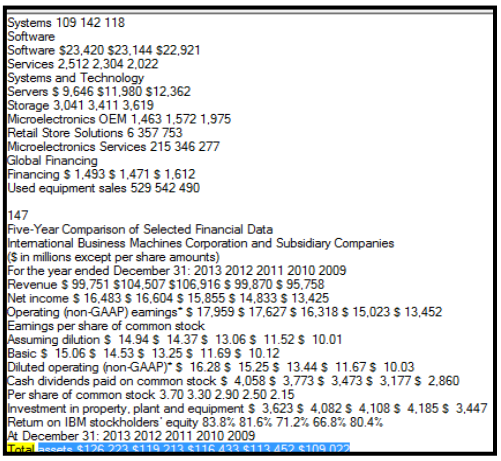

Fig 6: Result of Automated Annual Report Analysis

4.6 Performance Analysis

Here, we used F-measure for performance evaluation as given below.

- Now C= {c1, c2, c3..cn} is set of generated clusters by proposed system.
- K= {k1, k2, k3, k4…..Kn} is natural classes of documents.
- Now calculates Precision and Recall:
- Cj $\in$ C and Ki $\in$ K
- **Precision:**

Precision (ki, Cj)
= (true_ positive) / (true_positive+false_positive)
= (Ki $\cap$ Cj) / |Cj|

- **Recall:**

Recall (ki, Cj)
= (true_ positive)/(true_positive+false_negative)
= (Ki $\cap$ Cj)/|Ki|

- **F-measure:**

F (Ki, Cj)

=2*Precision (ki,cj)*recall(ki,cj) /
(Precision(ki,cj)+Recall(ki,cj)

Example 1:
Documents in dataset D are 92 i.e. |D|=92,

Here, we have 4 natural classes from above document dataset as stated below, i.e. |K|=4,

K={K1, K2, K3, K4} such that these document classes have following document descriptions,

K1: Log file cluster,

K2: Financial receipts cluster,

K3: Financial reports cluster,

K4: Textual other file cluster.

So, here when we provide these document dataset to proposed system, it will give 4 set of clusters as stated below. i.e. |C|=4,

C={C1, C2, C3, C4}

C1: Log file cluster,

C2: Financial receipts cluster,

C3: Financial reports cluster,

C4: Textual other file cluster.

Precision(Ki, Cj)= (Ki ∩ Cj)/|Cj|

= (4)/ (4)

=1.

Recall (Ki, Cj)= (Ki ∩ Cj)/|Ki|

= (4)/(4)

=1

F-measure:

F(Ki,Cj)=2*Precision(ki,cj)*recall(ki,cj)/

(Precision(ki,cj)+Recall(ki,cj)

= (2* 1* 1)/(1+1)

=2/2

=1.

The following Table 1 shows F-Measure Score values obtained by proposed system against sets of financial documents and subsequently its corresponding performance graph is shown in fig. 7.

Table 1. F-Measure Score Values

Sr.No.	Data Set Size	F-Measure Score
1	Dataset - 92 Documents	1.00
2	Dataset - 37 Documents	0.87
3	Dataset- 190 Documents	0.89

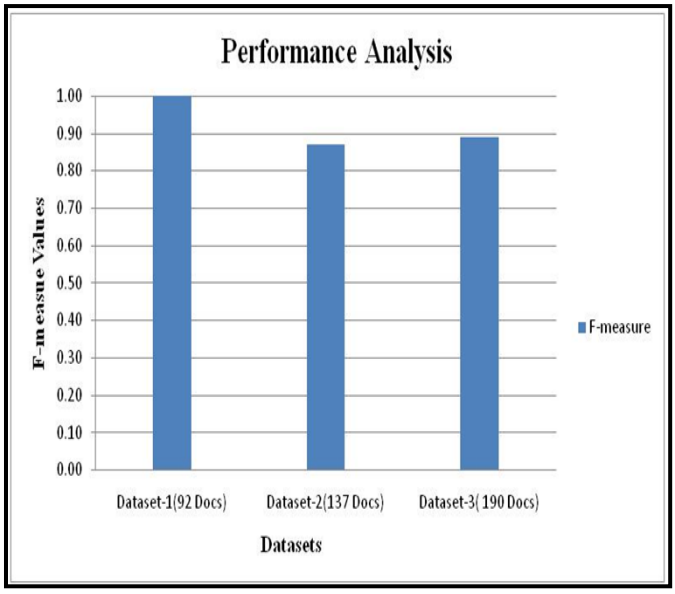

Fig 7: F-Measure Score Obtained by Proposed System Evaluation on Document Data Sets

5. CONCLUSION and FUTURE WORK

This system, first applies the document clustering on the dataset for forensic investigation to classify the documents into the different clusters including financial clusters. Hence the forensic examiners will only lookup the clusters of financial documents which are of certain interest and relevant to investigative case. It also reduces overhead of manual analysis of individual financial document by forensic examiner. Hence it enhances the forensic process significantly. The court of law requires the quick result of investigative cases.

So, additionally the proposed system also designed a framework aimed to automate the forensic analysis process in order to provide the quick results to the court of law. The forensic examiner just provides listed financial documents based on certain parameters, and then proposed system automatically lists the suspicious documents on dashboard in accordance with specified parameters. Hence this automated approach helps to enhance the forensic analysis process significantly.

Proposed system also facilitates automatic forensic analysis system for listed type of financial documents including financial annual reports, financial transactions captured from log file of banking application and purchase invoice of shop or organization. The further research can be to work on different type financial documents based on different parameters in order to design generalized framework for automated forensic analysis system for financial documents.

6. REFERENCES

[1] Price Waterhouse Coopers 2011 UK, Global Economic Survey , PriceWaterhouse.

[2] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver 2005, Exploring forensic data with self-organizing maps In Proc. IFIP Int. Conf. Digital Forensics.

[3] Patrick Grafe. Topic Modeling in Financial Documents, Department of Computer Science Stanford University.

[4] K. Stoffel, P. Cotofrei and D. Han 2010 Fuzzy methods for forensic data analysis In Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition.

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

[6] N. L. Beebe and J. G. Clark, 2007 Digital forensic text string searching, improving information retrieval effectiveness by thematically clustering search results Digital Investigation, Elsevier, vol. 4, no. 1.

[7] Anderi Sorin SABU 2012 Survey of Clustering based Fraud Detection Research In Informatica Economia vol 16, no. 1

[8] Luis Filipe da Cruz Nassif and Eduarado Raul Hruschka 2013 Document Clustering for Forensic Analysis An Approach for Improving Computer Inspection In Ieee Transactions On Information Forensics And Security,Vol.,No. 1.

[9] RBI Rules Available: http://rbidocs.rbi.org.in/rdocs/content/Pdfs/68787.pdf

[10] RBI Rules Available: http://rbidocs.rbi.org.in/rdocs/ notification/ PDFs/ PM2212A_II.pdf

[11] Suvasini Panigrahi, Amlan Kundu, Shamik Sural, A.K. Majumdar2011 Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning.

[12] Harmeet Kaur Khanuja and Dr D.S. Adane.‟Forensic Analysis of Databases by Combining Multiple Evidences‟ International Journal of Computer and Technology , Vol 7,No 3.

[13] H. Lee, T. Palmbach, M. Miller 2001 Henry Lee's Crime Scene Handbook, San Diego: Academic Press.

[14] K. M. Hammouda and M. S. Kamel 2004 Efficient phrase-based document indexing for web document clustering In IEEE Transactions on knowledge and data engineering.

[15] Alessandro Moschitti and Roberto Basili 2004 Complex linguistic features for text classification: A comprehensive study, In ECIR '04: 27th European conference on IR research, Sunderland, UK.

[16] L. Garfinkel 2010 Digital forensics research: The next 10 years, Digital Investigation.

[17] L. Liu, J. Kang, J. Yu, and Z. Wang 2005 A comparative study on unsupervised feature selection methods for text clustering, In Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Engineering.

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

[19] Luiz G. P. Almeida, Ana T. R. Vasconcelos and Marco A. G. Maia, 2007 A Simple and Fast Term Selection Procedure for Text Clustering In Seventh International Conference on Intelligent Systems Design and Applications.

[20] M.R. Clint, M. Reith, C. Carr, G. Gunsch 2002 An Examination of Digital Forensic Models.

[21] B. D. Carrier, E. H. Spafford 2004 An event-based digital forensic investigation framework In Proceedings of the 4th Digital Forensic Research Workshop

[22] A. Miller. 1995 Wordnet: a lexical database for English. Common. ACM.

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

[24] Prof. K. Raja, C. Prakash Narayanan 2010 Clustering Technique with Selection for Text Documents.

[25] Computer Forensics Avialable: http://www.computerforensicsworld.com/

[26] E. Casey 2000 Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet with Cdrom, 1st ed., Academic Press, Inc., Orlando, FL, USA.

[27] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem 2009 Towards an integrated e-mail forensic analysis framework In Digital Investigation, Elsevier, vol. 5.

[28] A. Weigend 1997 Data Mining in Finance, Computer Intensive Methods for Financial Modeling and Data Analysis.

[29] Forensics Available http://www.us-cert.gov/sites/default/files/ publications / forensics.pdf

[30] Harmeet Kaur Khanuja and Dr. D. S. Adane. 2012. A Framework For Database Forensic Analysis. Published in Computer Science & Engineering: An International Journal (CSEIJ), Vol.2, No.3.

[31] G. Palmer, M. Corporation 2001 A Road Map for Digital Forensic Research, in: Proceedings of the 1st Digital Forensic Research Workshop.