Subject based Clustering for Digital Forensic Investigation with Subject Suggestion

Sweedle Mascarnes
Department of Computer Engineering
St. Francis Institute of Technology
Mumbai, India

Joanne Gomes
Department of Information Technology
St. Francis Institute of Technology
Mumbai, India

ABSTRACT
Recently digital forensics has become a prominent activity in crime investigation since computers are increasingly used as tools to commit crimes. During forensic investigation the digital devices such as desktops, notebooks, smart phones etc. found at the crime scene are collected for further investigation. Investigators have to go through humongous amount of data stored on these devices to gather evidence. This activity exceeds the expert's ability of analyzing and interpreting the data. In this context data mining techniques such as clustering are used for automated data analysis. This research work focuses on a novel document clustering model that allows an investigator to semantically cluster the documents stored on a suspect's digital devices with the help of subject suggestions initially provided to him. Providing subject suggestion improves the accuracy and speeds up the process of searching the evidence. Without subject suggestion, the investigators are heedless about the suspect's dataset and fail to give appropriate search query which may delay the process of investigation.

General Terms
Digital Forensic Investigation

Keywords
Crime Investigation; Digital Forensic; Semantic Clustering; Subject Suggestion

1. INTRODUCTION
With the boom of Information and Communication Technologies, computer fraud and digital crimes are increasing at an alarming rate. Thus, analyzing computers and other digital devices has become a necessity in the field of criminal investigations. The dramatic increase in computer-related crimes requires the development of advanced techniques to systematically search digital devices for significant evidence. Now a day in crime investigation, digital forensics has become a prime activity. Digital forensics is a process of examining digital media with the aim of generating digital evidence related to an incident under investigation.

The model of Digital Forensic Investigation has been explained in [1]. Figure 1 illustrates the six phases involved in Digital Forensic Investigation (DFI) process as defined by Digital Forensics Research Workshop (DFRWS) [1]. Identification phase detects all items, devices, and data associated with the incident under investigation. The Preservation phase preserves the crime scene by stopping or preventing any activities that can damage digital information being collected. In the Collection phase digital information is collected by copying files or recording network traffic that might be related to the incident under investigation. Examination phase involves an in-depth systematic search of evidence from the data collected from previous phases. Analysis phase draws conclusions from the evidence found. Finally, in Presentation phase the physical and digital evidences are presented to a court or corporate management.

Out of all the six phases, Examination phase is most the complex phase. In Examination phase, investigators have to wade through number of unstructured documents stored on suspect's computer to gather credible and convincing evidence. The continuously increasing size of storage devices makes the task more cumbersome. Due to the complexity of this information acquisition activity automated methods of data analysis such as data mining are of paramount importance.

The use of clustering algorithms attempts to address these problems, which are capable of finding latent patterns from text documents found in seized computers and can thus enhance the analysis performed by the forensic investigator. The rationale behind clustering algorithms is that objects within a same cluster are more similar to each other than the objects belonging to a different cluster [2]. Forensic investigators also often rely on numerous forensic tools such as Guidance Encase [3], Access Data Forensic Toolkit (FTK) [4], Sleuth kit / Authopsy [5] etc. to search and analyze digital evidence. The search techniques provided by current DFI tools include keyword search, regular expression search and approximate matching search. Unfortunately, these tools and clustering algorithms are applied directly against all of the stored documents without any advance knowledge about the topics discussed in each document. Hence, the results based on these techniques give large number of false positives and false negatives [6].

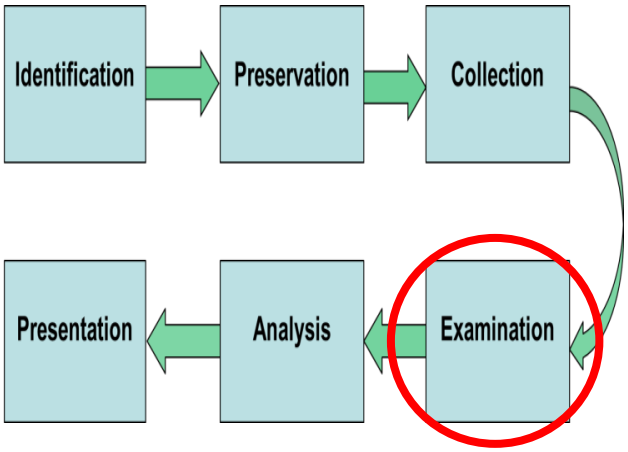

Figure 1: DFI process as defined by DFRWS [1]

The research work presented in [7] focuses on a new subject-based semantic document clustering model that allows an investigator to cluster documents stored on a suspect's computer, according to the subject initially defined by the investigator. The limitation of this method is that, the investigators may fail to give an appropriate search query, as they are unlikely to have the prior knowledge of all criminal events that have already occurred on a suspect's computer.

The research work presented here overcomes this limitation by introducing a novel method of subject suggestion to help investigators efficiently identify relevant information from the unknown dataset in order to improve the accuracy and the performance of existing DFI framework.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 explains the proposed subject suggestion model. Section 4 assimilates the implementation results. Finally, Section 5 concludes the paper.

2. RELATED WORK

Data mining is a powerful technique that enables forensic investigators to explore large databases quickly and efficiently. There are few research articles reporting the use of data mining techniques in the DFI field.

Clustering is a well-known data mining technique. The objective of clustering algorithms is to partition the dataset into clusters such that objects within a valid cluster are more similar to each other than the objects belonging to a different cluster [8]. Once a data partition has been induced, the forensic examiner might only focus on reviewing representative documents from the obtained set of clusters. In [9], SOM-based algorithms have been used for clustering files by taking into account their creation dates/times and their extensions, with the aim of making the decision-making process performed by the examiners more efficient. The research presented in [10] uses Kohonen Self-Organizing Maps to cluster digital forensic text string search results. This approach presents search hits in a manner that helps investigators locate hits relevant to investigative objectives more quickly. In [11], the authors have presented a comparative study of six well-known clustering algorithms (K-means, K-medoids, Single Link, Complete Link, Average Link, and CSPA) by applying them to five real-world datasets obtained from computers seized in real-world investigations. Authors conclude that Average Link and Complete Link algorithms provide the best results for this application domain. In [12], Kernel based variant of K-means algorithm has been used to arrange unstructured documents into content based homogeneous groups. In [7], authors have proposed a subject based document clustering model that allows an investigator to cluster all documents on a suspect's computer according to certain subjects he is interested in (e.g. hacking, child pornography etc.). The research work presented in the current paper enhances the concept presented in [7] by providing the subject suggestions in order to the investigator to accelerate the process of finding the evidence.

3. PROPOSED WORK

Figure 2 shows an overall architecture of the proposed system for DFI, based on subject suggestion. The proposed system works in various stages. In the first stage, the proposed model pre-processes all the data stored on suspect's hard disk. The goal of pre-processing is to reduce the high dimensionality of the documents while keeping the necessary information to do text mining. Pre-processing is done in three steps, namely stop-words removal, stemming and indexing. In stop-words removal, common words such as pronouns (he, she, it), conjunctions (and, or, but) and prepositions (a, an, the) are removed as they do not convey any meaning and have no effect on the text mining process. Stemming reduces the words to their root form. For example the words 'thinker', 'thinking', 'thinks' are all stemmed to 'think'. The most commonly used Porter Stemmer algorithm [13] for Natural language processing (NLP) is implemented here for stemming process. In indexing a set of distinct terms acquired after stemming process is chosen and the weight of these terms is computed for every document. Thus each document is transformed into a vector of weighted terms. The two main metrics involved in computing the weight of a term are Term Frequency (TF) factor and Inverse Document Frequency (IDF) factor [14]. TF is a frequency of occurrence of each term *t* in a document *d*. IDF is the frequency of the occurrence of each term *t* in the document set *D*. Each weighted term frequency is calculated as

$$W_{t,d} = TF * IDF \quad (1)$$

and

$$IDF = 1 + log\ (|D| / (1 + Freq_{t,d})) \quad (2)$$

where $|D|$ is the number of documents in the document set *D*, and $Freq_{t,d}$ is the number of documents in *D* that contain the term *t*

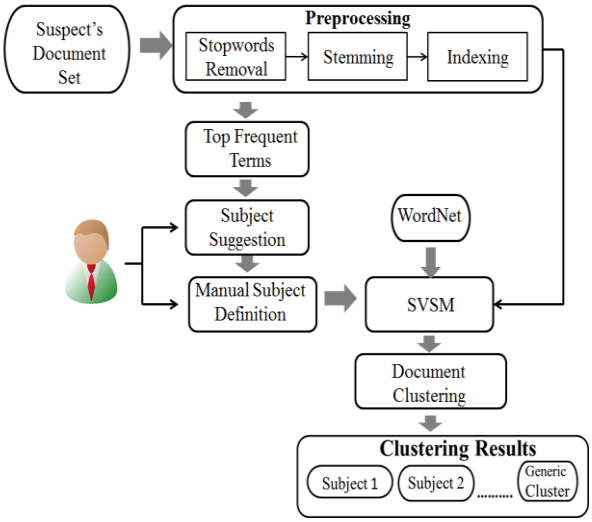

Figure 2: The proposed DFI system, based on subject suggestion

After pre-processing, the proposed framework identifies top frequent keywords that repeatedly appear in the dataset. The top frequent keywords are identified based on weighted term frequency $W_{t,d}$ given by Eq. (1). The keywords whose $W_{t,d}$ is more than the threshold, are highlighted. In this work threshold is 10 keywords and top 10 keywords whose $W_{t,d}$ is more are highlighted.

The key contribution of this research work is subject suggestion/recommendation module. Subject Suggestion module recommends the top frequent keywords to the investigator for further investigation. The investigator can either select the subject from suggested list of keywords or manually enter his own keyword as the search query. By suggesting these subject keywords it aids the investigator for formalizing the search query as they are unaware of the suspect's dataset.

The subject keywords that are suggested by the system or entered manually by the investigator are further expanded by finding its synonyms through Wordnet [15]. Wordnet is a lexical database for the English language used to establish semantic relations between terms. Wordnet database gives synonyms, definition, several senses of a word and sometimes example for the word searched in the database. The proposed framework of Figure 2 integrates Wordnet to find out semantic relations among words and thus makes an attempt to improve the effectiveness of information retrieval.

In the next stage, terms and documents are represented by Subject Vector Space Model (SVSM) [7]. SVSM is an algebraic model based on Vector Space Model (VSM) [16] and Topic-based Vector Space Model (TVSM) [17]. Similar to TVSM, all axis coordinates in SVSM are positive, and all axes are orthogonal to each other. In SVSM, each dimension (axis) represents a subject $s_i \in S$ where S is the set of subjects. Each term $t_i \in T$ is represented by a *term-vector* t_i and has a *term-weight* between one and zero i.e. $|t_i| \in [0:1]$. Similarly each document $d_i \in D$, where D is a collection of documents, is represented by a document vector d_i in the space. Since the similarity value between any two term vectors is between zero and one, the similarity value between any two document vectors is also between 0 and 1.The cosine similarity measure [18] is to gauge the similarity between two term vectors and document vectors.

Figure 3 illustrates an example of the SVSM. Subject related terms (e.g. 'cocaine', 'drugs', 'kill') point into the same direction as their fundamental topic which they are related to. Terms which are not subject specific (e.g. 'a', 'the') are aligned near 45^0 towards all the fundamentals subjects.

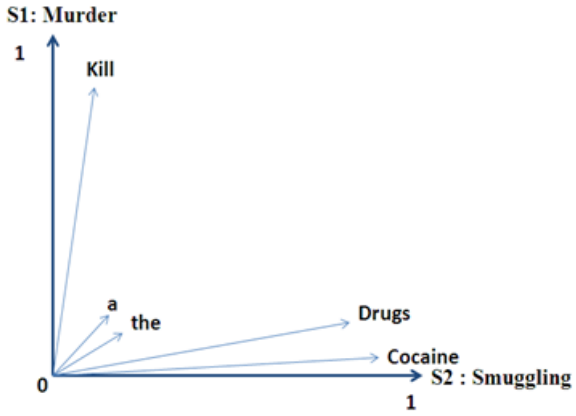

Figure 3: An example of Subject Vector Space Model

Once terms and documents are represented in SVSM, the proposed framework uses a subject-based semantic document clustering algorithm [7] to cluster the documents. This algorithm allows a forensic examiner to cluster all documents stored on a suspect's hard disk by grouping them into a set of overlapping clusters, each corresponding to a subject of interest initially defined by the investigator or suggested by the model (e.g. hacking, child pornography). Once the documents are clustered corresponding to a subject, the investigator can search for documents that belong to a certain subject and do not have to search entire document set.

In this way the proposed framework is designed to quickly find the relevant evidence from large dataset found at the crime scene by suggesting the keywords or the subjects on which the investigator can search for the evidence.

4. EXPERIMENTAL EVALUATION

4.1 Implementation Details:

The proposed framework is implemented using java with help of Net Beans IDE 7.4. For validating our proposed system, experiments were conducted on demo dataset consisting of 100 documents from 4 different categories: 25 kidnapping-related documents, 25 robbery related documents, 25 sexual assault-related documents and 25 murder-related documents. The demo dataset consists of the documents of .txt, .doc, .pdf, .ppt, .pptx, .xls and .xlsx extension. All experiments were conducted on an Intel(R) Pentium(R) CPU N3510@ 1.99 GHz PC with 4 GB RAM.

The Snapshots of core modules of proposed framework are explained below. Figure 4 illustrates user interface of the proposed system. The investigator has to enter the folder path in which he/she has to search for evidence. In Figure 4 the 'C:\Demo' folder is selected for testing the system. By clicking "Start Pre-processing" button all the documents of the 'Demo' folder are converted into text (.txt) documents and are stored in 'Plain Text' folder within 'Demo' folder.

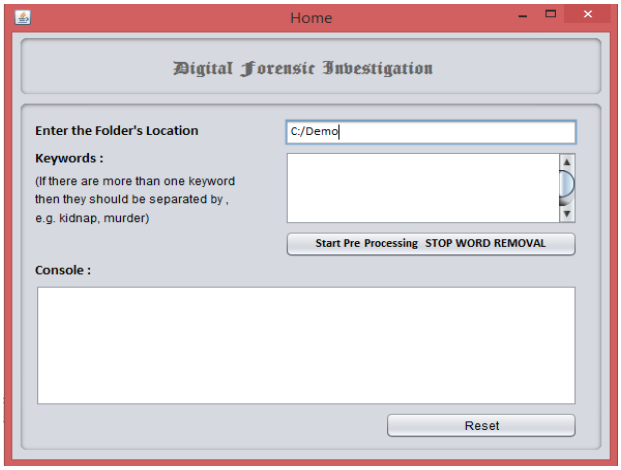

Figure 4: GUI of the proposed DFI system

The pre-processing techniques namely stop words removal, stemming and tokenization as discussed in section 3 are applied to plain text i.e. .txt documents. A Sample pre-processing input file and sample output file obtained after applying pre-processing techniques is shown in Figure 5 and Figure 6 respectively.

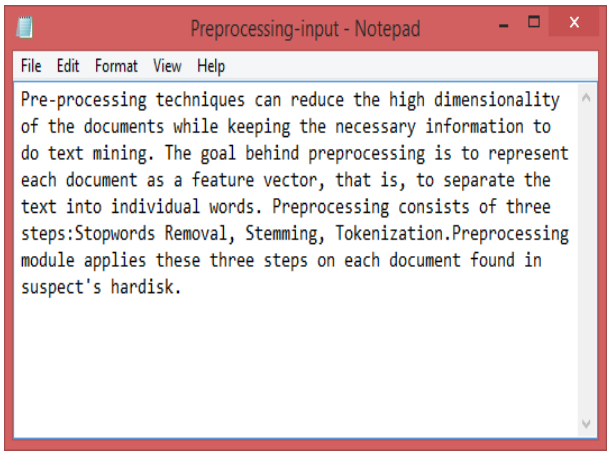

Figure 5: Sample file before preprocessing

All the stopwords (can, the, of, while, to, do etc.) from the sample file shown in Figure 5 were removed. Each word was reduced to its root form. (For example, the word 'Pre-processing' was reduced to 'preprocess'. Similarly 'dimensionality' was reduced to 'dimension' and so on). The preprocessed output file is shown in Fig. 6.

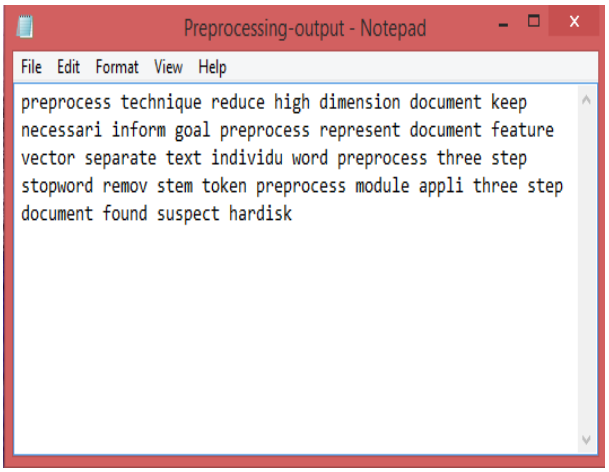

Figure 6: Sample file after preprocessing

After pre-processing, the system calculates the frequency of occurrence of each keyword in each preprocessed documents. The top frequent keywords that are commonly appearing in the dataset are identified. For experimental purpose top 10 keywords whose frequency of occurrence is high are recommended to the user as shown in Figure 7. Figure 7 lists the top 10 words and its percentage of occurrence in descending order of their frequency.

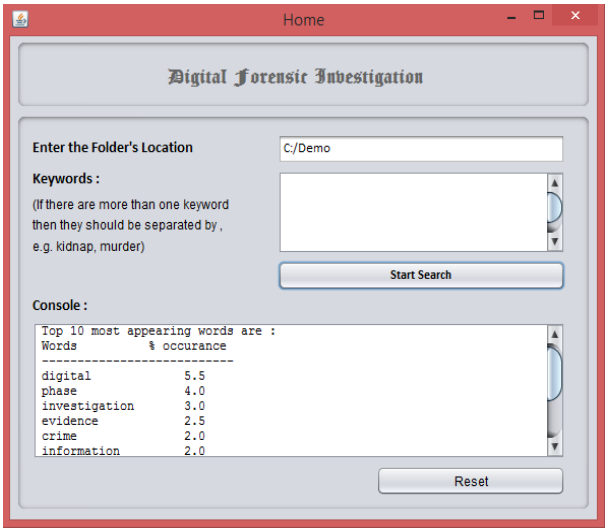

Figure 7: Frequent keywords suggested to investigator

4.2 Experimental Results:

The performance of our proposed system is evaluated based on two parameters: Accuracy and Scalability.

Accuracy:

F-measure [19] is used to measure the accuracy of the clusters produced by our method. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0.

Precision: Precision is the percentage of results retrieved that are relevant to the search objectives.

$$Precision = \frac{Number\ of\ Relevant\ returned\ results}{Number\ of\ Returned\ results} \quad (3)$$

Recall: Recall is the percentage of relevant results that are retrieved in comparison to the total number of relevant hits in the data set.

$$Recall = \frac{Number\ of\ Relevant\ returned\ results}{Total\ number\ of\ Relevant\ results\ in\ the\ system} \quad (4)$$

The traditional F-measure is the harmonic mean of precision and recall:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

Accuracy of the system is tested with and without subject suggestion module. The proposed system is tested to search for different keywords and the precision, recall and f-measure are calculated for individual results obtained. To test the accuracy of the system experiments are conducted on demo dataset consisting of 100 documents from 4 different categories: kidnapping, robbery, sexual assault, murder as described above. Table 1 shows the results obtained by taking five sample keywords i.e. Kidnap, Sex, Money, Knife, entered as search queries for finding the evidence.

Table 1 Precision, Recall and F-measure for sample keywords

Keywords/ Parameters	Kidnap	Sex	Money	Knife
No. of relevant returned results	9	13	5	13
No. of returned results	9	14	7	15
Total no. of Relevant results in the system	12	16	8	17
Precision	1	0.92	0.70	0.86
Recall	0.75	0.81	0.62	0.76
F-Measure	0.85	0.86	0.66	0.80

In similar way, Precision, Recall and F-measure has been calculated for 30 different keywords and the average of these results has been computed. Table 2 depicts the values of Average Precision, Average Recall and Average F-measures of the results obtained with subject suggestion module and without subject suggestion module.

Table 2 Accuracy of the proposed DFI system

Accuracy	With subject suggestion	Without subject suggestion
Average Precision	0.95	1
Average Recall	0.75	0.54
Average F-measure	0.8	0.68

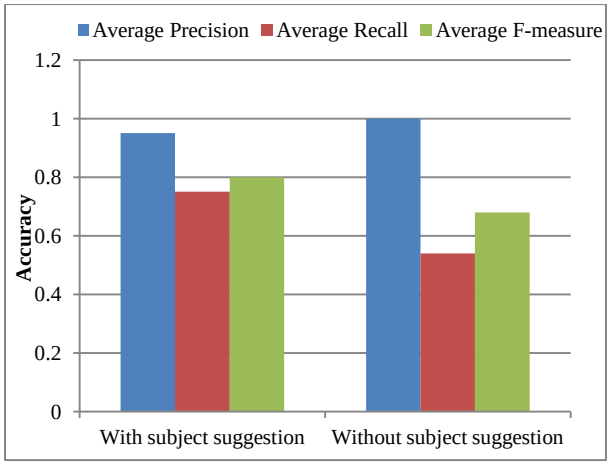

Figure 8: Accuracy of system with and without subject suggestion

Figure 7 shows the graphical representation of the Accuracy of our proposed system in terms of Average Precision, Average Recall and Average F-measure. Here the Average F-measure with subject suggestion module is 0.8 whereas Average F-measure without subject suggestion module is 0.68.

Scalability:

Table 3 describes the scalability of the proposed system. To test the scalability of the proposed system, experiments have been conducted on demo dataset consisting of 100 documents from 4 different categories, i.e. kidnapping, sexual assault, robbery and murder-related documents. Further, the files in the data set are duplicated so that the scalability can be measured starting from 10 documents, going up to 1,00,000 documents. The implemented system consists of two main phases: Preprocessing and Clustering. The objective of this scalability experiment is to measure the runtime of each phase to ensure it does not grow proportionally as the data set sizes increases. The overall system runtime remains approximately same with subject suggestion and without subject suggestion as providing subject suggestions have negligible effect on runtime of the system. The scalability results shown in table 3 are independent of subject suggestion approach.

Table 3 Scalability of the proposed DFI system

Number of documents	Time(Seconds)		
	Preprocessing	Clustering	Total
10	18	2	20
100	21	3	24
1000	34	6	40
10000	62	13	75
100000	109	18	127

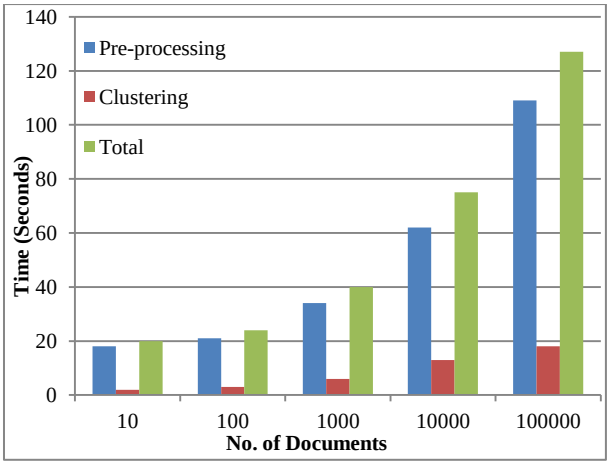

Figure 9: Scalability of the proposed DFI system

Figure 9 graphically demonstrates the runtime of the two phases with respect to the total number of documents being clustered. The total runtime for processing 10 documents is 20 sec, which includes 18 seconds for pre-processing documents and 2 seconds for clustering the document set. Likewise, the total runtime for processing 1, 00, 000 documents is 127 seconds, where 109 seconds are spent in pre-processing whereas 18 seconds are spent in clustering the document set. It is observed that most of the system runtime is spend in pre-processing phase. The runtime scales linearly with respect to the data set's size. Since each phase of the algorithm grows linearly with respect to the total number of documents, hence the experimental results suggest that the system presented in this work is scalable.

4.3 Comparative Analysis:

The results presented in Table 2 are compared with the results presented in [7] to validate the obtained results. The Average F-measure of subject based semantic clustering algorithm tested on different experimental setup in [7], without subject suggestion, is 0.65. From the overall results obtained here, it is observed that the Average F-measure with subject suggestion module is 0.8 whereas without subject suggestion module is 0.68. Hence the accuracy of system with subject suggestion module is comparatively more than the system without subject suggestion module. Thus the subject suggestion improves the accuracy of the system in finding the evidence in suspect's computer.

5. CONCLUSION AND FUTURE WORK

In many crime investigations, digital devices owned by the suspect, such as desktops, PDAs, and smart phones are seized for forensic analysis with intend to find evidences relevant to the case under investigation. This task is very complex since the forensic investigator has to go through very large amount of extraneous data not germane to the case. This research work discusses DFI system that allows an investigator to semantically cluster the suspect's dataset based on subjects initially specified by the investigator or subject suggestions given by the proposed framework. Thus the investigator only has to search within limited documents rather than going through the entire dataset. Experiments were performed on Demo crime dataset to test the accuracy and scalability of the proposed DFI framework. The experimental tests show that subject recommendations improve the accuracy of the system. The experimental results prove that the model presented in this work is scalable.

Currently, the presented framework mainly uses WordNet to find the synonyms. In future, the framework can be improvised by using online sources such as Wikipedia or online dictionary/thesaurus to find synonyms. For future work it would be interesting if the DFI system framework supports image mining/detection i.e. enabling the tool to extract images on the hard drive that are related to a topic or another image provided by the investigator.

6. REFERENCES

[1] G.L. Palmer. 2001. A Road Map for Digital Forensics Research. Technical Report. First Digital Forensics Research Workshop (DFRWS).

[2] J. Han and M. Kamber. 2006. Data mining: Concepts and Techniques. Second Edition. Elsevier.

[3] Tool: Guidance Encase. http://www.guidancesoftware.com/computer-forensics ediscover software-digitalevidence.htm.

[4] Tool: Access Data Forensic Toolkit. http://www.accessdata.com/forensictoolkit.html.

[5] Tool: Sleuth Kit & Authopsy. http://www.sleuthkit.org.

[6] S.L. Garfinkel. 2010. Digital forensics research: The next 10 years. Digital Investigation. pp. 64–73.

[7] G. Dagher and B. Fung, "Subject-based semantic document clustering for digital forensic investigations", Journal of Data & Knowledge Engineering, Vol. 86, pp. 224–241, 2013.

[8] A. K. Jain. Data Clustering: 50 Years Beyond K-Means. 2010. Pattern Recognition Letters, Vol. 31, pp. 651-666.

[9] 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 International Conference on Digital Forensics. pp. 113–123.

[10] N. L. Beebe and J. G. Clark. 1997. Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results. Digital Investigation. Elsevier. Vol. 4. pp. 49–54.

[11] L. Nassif and E. Hruschka. Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection. 2013. IEEE Transactions on Information Forensic and Security, Vol. 8. pp 46-54

[12] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo and R. Zunino. 2009. Text clustering for digital forensics analysis. Computat. Intell. Security Information System. Vol. 63. pp. 29–36.

[13] M.F. Porter. 1997. An Algorithm for Suffix Stripping, Morgan Kaufmann Publishers Inc. USA. pp. 313–316.

[14] Wu HC, Luk RWP, Wong KF and Kwok KL. 2008 Interpreting tf–idf term weights as making relevance decisions. ACM Transactions on Information Systems. pp. 1–37.

[15] G.A. Miller. 1995. WordNet: a lexical database for English, Communications of the ACM, pp. 39–41.

[16] G. Salton, A. Wong and C. S. Yang. 1975. A vector space model for automatic indexing. Communications of the ACM, Vol. 18. pp. 613-620.

[17] J. Becker and D. Kuropka. . 2003. Topic-based Vector Space Model. In Proc. of the 6th International Conference on Business Information Systems.Colorado Springs.

[18] G. Salton and M.J. 1986. Introduction to Modern Information Retrieval. McGraw-Hill.

[19] http://en.wikipedia.org/wiki/F1_score 8/07/2014