Efficacious IR System for Investigation in Digital Textual Data

S. Gowri* and G. S. Anandha Mala²

¹Sathyabama University, Chennai, India; gowriamritha2003@gmail.com
²St. Joseph’s College of Engineering, Chennai, India; gs.anandamala@gmail.com

Abstract

There are numerous imperative digital text based proofs, some of which are SMS (Short Message Services), messages, mails, chat logs, etc. The person who researches a case is fundamentally overflowed with information and he needs to invest all his profitable examination time, examining through the loud indexed lists and experiencing unimportant query items. Hence a system is initiated using digital textual data mining standards for configuration and execution, which enhances IIR (Intelligent Information Retrieval) viability in digital forensics. The framework analyzes the input corpus data with domain specific keywords after which search and ranking of the SMSs, based on the weight of the keywords of forensic interest is computed. This software is developed as a proof of concept with data mining and weighted search concepts.

Keywords: Digital Forensics, Digital Textual Data, IR, Keyword Ranking, Knowledge Mining, String Search

1. Introduction

In digital investigations the textual evidence is very important. It has a vast majority during investigations. All this is due to a very great deal of stored digital data which is linguistic in nature (e.g. human languages, programming languages). Some of the most important text-based evidences are SMS, chat logs, emails, word processing documents, spreadsheets, address books, calendar appointments and system logs. Here, the one who investigates is flooded with data and he has to spend valuable investigation time, scanning through the noisy search results and going through irrelevant search results.

The current digital forensic text string search tools use matching or indexing algorithms for searching digital evidences at a physical level to locate specific text strings. All these are designed in such a way that it achieves 100% query recall (i.e. it finds all instances relating to text strings). Here the nature of the data set is given, which leads to a highly extreme incidence of hits that are totally irrelevant to investigative objectives. There is a text string search tool that fails to group or order search hits in a manner that considerably improves the investigator’s ability to get to the relevant hits first or at least in a quick manner.

Thus, the text mining has been taken up as a new initiative for digital forensics. This type of text mining approach will enhance the IIR (Intelligent Information Retrieval) effectiveness of digital forensic text string searching. Henceforth, the technology for text mining can be scalable up to large datasets in Gigabytes or Terabytes. Here the software that has been developed consists of text analysis, data extraction, and Meta data correlation and visualization features and Information retrieval. The main aim of the system that was developed is to analyze the various transcripts like SMS, Email and word letters, event logs and chat transcripts. There are specific keywords that are searched by the system, which are weighted by the user in compliance with domain specific analysis. The correlated data are ranked by the system and hence is displayed by the system. For the investigators to analyze further this helps it by providing user graphs and charts about the ranked data.

*Author for correspondence
2. Proposed System

The extent of this work is mainly to develop and design a Forensic Analysis and Inference of Correlated Data tool, which will be used as a method to detect any tendency or activities that may compromise of security.

![Architecture diagram]

2.1 Extraction

The data from the SMS corpus is extracted using the data extraction module. Format conversion, storage Meta data and Meat data extraction information are all performed by this module. This type of information and the kind of file format present in this will be completely different for various SMS service systems. The file in various formats converting to a simple text format is done by the Format conversion module for further processing. The parsing of text file and the extraction of Meta information are done by a module named Meta data extraction module. This includes Message-ID, Source-ID, Collection Method, Year, etc. The Meta information will be stored in relational database MYSQL.

2.2 Message Indexing

The index of message items is constructed by this module. There are various functionalities that are present in this module, such as creation Lucene documents, Tokenization Message refining and inverted index Management.

- The Create Lucene document function has a specific format for representing a document as keywords that creates a data type called a Lucene document.
- Message refining and tokenization function removes the stop words first. Once this is done, the remaining words are being refined for base words. Tokens are then created by base words.
- Inverted index Management module performs various index management functions such as Index Compression, Index Creation, Update Index, and Index Merging. For the entire SMS corpus, the Index creation creates an index of keywords. Here each token will be a keyword and a data structure with document id and no of times the keyword present in a document, etc. The details of the keyword index for each SMS document are updated using the Update Index module.

2.3 Message Scrutiny

The functionalities for this are Search, Ranking, Query analysis and expansion. Query analysis performs
query pre-processing by taking the user query as its input and expands the query. To search query keywords in the inverted index of message documents the term Search is used and hence it generates the list of documents in which the query words are present. There are different methods for search such as Boolean search (AND, OR, NOR), Semantic search (semantic meaning). Rearranging of the index of an SMS corpus with defined parameters as the criteria is done by an important module called the Ranking. Term frequency for the query keyword and weight of terms for query keywords are the parameters based on which ranking has to be made. Based on the proximity of the query keywords the ranking can be done in this.

2.4 Visualization

The functionalities based on visualization are Displaying of Output, Query input and Graphic display. The module that allows the investigator to feed the input query is the GUI module i.e. a Query Input. The query may be based on keywords of interest or on message header information like recipients, sender id, time, day, etc. The displaying of the message information on html page based on ranking criteria is done by Output Display. The actual message that has to be displayed in a new window should be done when the mouse is clicked on a particular message id.

The Chart comprises of a user interactive map and also the date and time frequency edges in the map. The user integration map will reveal the sender id and the recipient interaction mapping based on a given data and time interval. Hence the graph constitutes of sender id and receiver id as nodes. The SMSs that are the interfaces among themselves are represented as the number of lines. The time/date frequency map is shown by the Date/Time frequency with sender id or receiver id. The nodes in the map represent the users and the edges as the conversation SMSs between the users. The edges will show the entire Message details which include the time, date, message, etc., which are represented by the keywords for and effective retrieval of SMSs from a corpus.

3. Technologies

In the history of Information technology, the volume of the information and the usage of information are growing day to day. The capacious data are generated by various mechanisms in this informative history. In many fields of work like banking, communication, network security and forensics the need to retrieve textual data from a large corpus of information store is required a lot these days. Referring to all this information manually and performing analysis of this information will be very difficult. Therefore the information retrieval requires few automated mechanisms to search and rank in accordance with user requirements. The Desktop search engines, Text Mining, Internet search engines are those that are used by the text string search approaches in an information retrieval system.

3.1 Desktop Search Engines

In digital forensic investigations a computer hard drive is a widespread evidence type, it seems real to explore the file system search engine technology. To “browse” once computer just as one would “browse” the Internet for information, this allows the users by the File system search engines. Google Desktop, Yahoo! Desktop, Spotlight, and X1 are some of the current commercial desktop search engines. Eureka, Connections, and Semantic File System are some of the academically developed or open source desktop search engines that are included in this. There are two main reasons for File system search engine technology not being able to be extensible to digital forensic text string searching.
The first reason is that the high startup costs for the index creation process. It can take days for Initial indexing of one’s file system. The second reason is that digital forensic text string searches, seek to find out digital evidence which are independent of the file system. They aspire to find data at a physical level, as opposed to that of the logical file system level. Therefore, any approach that relies on the file system is inextensible.

### 3.2 Internet Search Engines

A new Internet search engine that prioritized search hits for the user was introduced by Google during the 1990s. Google used five ranking variables:
1. Page Rank
2. Query coincidence with anchor text
3. Proximity measures
4. Query term order
5. Visual presentation characteristics

To expand the ranked list approach to digital forensic text string searching, new ranking variables must be speculated and empirically validated.

### 3.3 Text Mining

From the larger field of Information Retrieval (IR) a new technology area known as text mining, had been emerging in recent years. The text mining and IR fields rejoice over 30 years of research into locating and retrieving textual artifacts. Here the web-based information retrieval and knowledge discovery field also brag extensive research and technological advances in text retrieval. The text mining is the technology that is used in this work to develop software for searching and analyzing the SMS corpus to help the investigative agencies and IT forensic labs. The influence of text mining includes several information processing tasks. They are content summarization, visualization, information extraction, topic tracking, question answering, concept linkage, text categorization, and text clustering.

### 4. Design Overview

The scope is to design and develop a Forensic Analysis and Inference of Correspondence Data tool which will be used to detect any trend or activities that lead to affecting the system. In any of the message passing communication system Information Retrieve is the domain for analysis in forensic Department. The SMSs transacted through an SMS server can be collected as an SMS corpus and is the input for this tool. The text data mining is the key technology area upon which the system is going to be designed and developed. The goals of the system to be developed are:
- To design a Forensic analysis and inference system.
- The system should be capable of analyzing various transcripts like SMS, Emails, and word letters, chat transcripts, event logs.
- The system should have search methodology to keep information retrieval system faster.
- The system should have modules to decide on the ranking of documents so that users will get the exact information what he wants.
- The ranking modules / tools should be domain specific.
- The system should have a user friendly GUI module for input and visualizing the desired outputs.
The SMS corpus is a file and will be stored in the hard drive. The data extracted will be stored in MySQL database. The Meta data will be stored as relational databases. It is planned to use Lucene tool for indexing the SMS corpus. Lucene doc is the index data structure of the SMS corpus. The data extraction is to extract the header information and store them in a relational database. The information stored like message id, sender id, receiver id will be used for analysis depending on the query. Message-id is the primary key. The message indexing generates an inverted index for an SMS corpus. This inverted index is a specific data structure. Message analysis module is dependent on this Lucene inverted index. The data structure hits docs are the output of message analysis. Visualization module depends on message analysis module output hits doc and relational table of the header information.

### 5. Initiated Procedure

Here it is first and foremost that the suppositions of the users are gratified by providing a pertinent result to them. Results generated based on only user query terms may not be relevant to the user need. So it is required to have a ranking based on many parameters and values. Forensic Search should be much more effective and should give minimum results for easy analysis.

The Meta Data of the SMS messages or documents are as important as the content of the messages in Forensic Analysis. The experts working forensics prepare evidence and analyze a case with due importance and weight to Meta Data. In this we are presenting a customized re-ranking profile. The utter knowledge related to the case that is being investigated is very familiar to the Expert. Therefore, he can provide the ranking profile with specific attributes such as date, time, messages in other than the inbox folder, etc. The weightage factor will be assigned to the input values given by the investigator in the input field, with which the ranking profile in the Meta data score will be computed. The computed score for the keyword using the inverted index along with the Meta data score will be used to compute the new rank score for all the messages. With this scoring methodology, it is believed that the user will get relevant retrieval results on search, which will serve his need.

The formula used to calculate the new Score is: $f(x) = \alpha^*KW + \beta^*AW$ Where,

- **KW**: Keyword weighted
- **AW**: Attribute weighted
- **\(\alpha\)**: Weight factor for keyword based score
- **\(\beta\)**: Weight factor for Meta data based score

#### 5.1 KW- Keyword Weighted

$$KW = \sum_{i=0}^{n} (kwf_i \cdot kww_i)$$

**n** - No of keywords in the query

**kwf** - keyword weighted factor

**Example:** Delhi ^1 operation^2  

Here **n** is equal to 2 and Delhi has weighted factor 1 while the keyword operation has weighted factor 2

**kww** - keyword weight value

$$kww = (1+\log_{tf} d) \cdot (\log_{10}(N/dft))$$

Where,

- **tf** - Term frequency or no of occurrences of term ‘t’ in document ‘d’
- **df** - No of documents consisting the term ‘t’
- **N** - Total no of documents in the corpus

#### 5.2 AW - Attribute Weighted

$$AW = \sum_{i=0}^{n} (awf_i \cdot awv_i)$$

**n** - No of meta-data or attributes

**awf** - Attribute weighted factor

**awv** - Attribute weighted value

**Example:** The Attributes considered here are Date, time, SMS folder, message type and user message id. User can define the factor for each of the attributes with the condition that the sum of the weighted factor of each attribute should be equal to 1.

For each attribute user can give the specific weighted value for different ranges of attribute values. The attribute Date has the weighted factor of 0.4 and message folder has 0.3 and message type has 0.4 and other attributes has 0 factors. The date ranges can have different weighted values. 1.12.12 to 10.12.12 has 0.5 as weighted value and 11.12.12 to 25.12.12 has 0.3.

Then $aw$ for date range 1.12.12 to 10.12.12 = 0.4 * 0.5

The $aw$ for date range 11.12.12 to 25.12.12 = 0.3 * 0.5

Then $AW$ for the attribute Date will sum of the above. The visualization is basically a GUI for giving user inputs and to present the output results in text or graphical formats.

It is planned to implement this module using Java server pages.
6. Conclusion

A software is elaborated for retrieving messages pertinent to the user’s query from the SMS corpus and to present about all the concealed knowledge that are accommodated in those messages. The NUS SMS corpus was utilized as input for testing our system, which is made public by US government. The resultant output snapshots of the designed system are given in the appendix section below.

7. Future Works

Research is currently ongoing and as a part of future work this system would be facilitated with the aspect of investigation of suspicion in images as well by embedding cryptographic technique as well. This enhancement would enrich the application to be utilized in any sector.

8. References