Improving Intelligent IR Effectiveness in Forensic Analysis

S. Gowri[1] and G.S. Anandha Mala[2]

[1] Sathyabama University, Chennai, India
[2] St. Joseph's College of Engineering, Chennai, India
gowriamritha2003@gmail.com

Abstract. In the era of Information technology textual evidence is important to the vast majority of digital investigations. Important text-based evidence include Email, Internet browsing history instant messaging, system logs and so on. The investigator is flooded with data and has to spend valuable investigative time scanning through noisy search results and reviewing irrelevant search results. Current digital forensic text string search tools use match and/or indexing algorithms to search digital evidence at the physical level to locate specific text strings. The text string search tools fail to group and/or order search hits. This research uses text data mining principles and technologies for design and implementation which improves IIR (Intelligent Information Retrieval) effectiveness in digital forensics. The proposed system can analyze the corpus of mail data or SMS data with domain specific keywords. The searching and ranking of the mails in the proposed system is based on the weight of keywords of forensic interest.

Keywords: Digital forensics, text string search, text data mining.

1 Introduction

Current digital forensic text string search tools use match and/or indexing algorithms to search digital evidence at the physical level to locate specific text strings. The text string search tools fail to group and /or order search hits. Text mining is the new approach in digital forensics. The text mining approach will improve the IIR (Intelligent Information Retrieval) effectiveness of digital forensic text string searching. The text mining technology can be scalable to large datasets in GBs or TBs. The system searches specific keywords, weighted by the user in accordance with domain specific analysis. The system then ranks the correspondence data and displays them. It also provides user graphs and charts about the ranked data which will help the investigators to analyze further.

2 Proposed System

The scope of this research 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 may compromise security. The architecture of the proposed system is shown in fig.1.

V.V. Das and J. Stephen (Eds.): CNC 2012, LNICST 108, pp. 448–451, 2012.

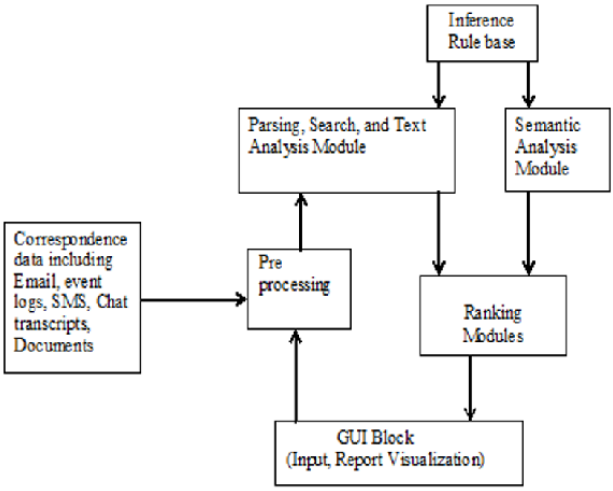

Fig. 1. System Block Diagram

The development of the system includes the implementation of software modules listed below.

(1)Inference and Text analysis module which include pre processing and query processing.(2)Ranking module which include syntax and semantic analysis and pattern matching (3)Reporting and visualization module include the GUI. It is assumed that the Email service provider will provide the mail corpus of specific users and for specific duration. This mail corpus data is parsed to extract meta-data of the mails and content of the mails. The meta-data is stored in RDBMS. The body content is converted to a data structure called inverted index of mail corpus. This data structure is then analyzed for presence of keywords and ranked against weighed factor. The list of mails more relevant to the user query is shown as output. The investigator can browse the mail for further analysis and conclusions.

3 Analysis Tool

The functional requirements of the tool are described below:

3.1 Data Extraction

The data extraction extracts the data from mail corpus. The data extraction will perform Format conversion, header extraction and store header information. The file format and type of information will be different for different mail system. Format conversion module converts the file in a simple text format for further processing. Header extraction module parses the text file and extracts header information. This includes sender details, receiver details, time, date etc. The header information will be stored in relational database mysql.

3.2 Message Indexing

The message indexing constructs the index of message items. This message indexing has functionalities such as creation of lucene documents, Message refining and Tokenization and Inverted index Management.

- The Create Lucene document function creates a data type called lucene document which is specific format of representing a document as keywords.
- Message refining and tokenization function first removes the stop words. After that remaining words are refined for base words. The base words are then created as tokens.
- Inverted index Management module performs the index management functions like Index Creation, Update Index, Index Compression, and Index merging. Index creation create index of keywords for the entire mail corpus. Each token will be a keyword and a data structure with document id, no of times the keyword present in a document etc, Update Index module data extraction extracts the data from mail corpus. The data extraction will perform Format conversion, header extraction and store header information. The file format and type of information update the details of the keyword index for each mail document.

3.3 Message Analysis

The functionalities are Query analysis and expansion, Search and Ranking. Query analysis takes user query as input and performs query preprocessing and expands the query. Search is to search query keywords in the inverted index of message documents and generate the list of documents in which the query words are present. Different methods of search like Boolean search (AND, OR, NOR), Semantic search (semantic meaning) and Pharse queries. Ranking is the important module which will rearrange the index of mail corpus with defined parameters as the criteria. The parameters based on which ranking has to be made are term frequency of the query keywords, weight of terms of query keywords.

3.4 Visualization

Functionalities of visualization are Query input, Output display and Graphic display. Query input is a GUI module which will allow the investigator to feed the input query. The query may be keywords of interest, or based on message header information like sender id, recipients, time day etc. Output display is display the message information in html page based on ranking criteria. When mouse is clicked on particular message id the actual message has to be displayed in a new window. Graphic/Chart has user interaction graph and date/time frequency graph. User Integration graph will show the sender id and recipient interaction graph based during given date or time interval. This graph will represent sender id and receiver id as nodes. The number of lines will show the mails communicated between them. Date/Time frequency will show the time/date frequency graph with sender id or receiver id. The X axis will show the sender / receiver id and y axis will show the number of messages sent or received.

4 Conclusion

The framework of the proposed system is developed to retrieve email documents relevant to the user query from the email corpus and to present the hidden knowledge containing in those emails to the user in an easily understandable form. The searching and ranking of the mails in the proposed system is based on the weight of keywords of forensic interest. The Email mining software requires a corpus to do analysis and testing. Enron mail corpus is made public by US government and this mail corpus is used for our testing.

References

1. Smith, Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
2. Beebe, N.L., Dietrich, G.: A new process model for text string searching. In: Shenoi, S., Craiger, P. (eds.) Research Advances in Digital Forensics III, pp. 73–85. Springer, Norwell (2007)
3. Beebe, N.L., Clark, G. J.: Digital forensic text string searching: Improving information retrieval effectiveness by thematicallyclustering search results, The University of Texas at San Antonio, Department of IS&TM, One UTSA Circle, San Antonio, TX 78249, United States
4. Naqvi, S., Dallons, G., Ponsard, C.: Applying Digital Forensics in the Future Internet Enterprise Systems - European SME's Perspective, pp. 89–93 (May 20, 2010) 978-0-7695-4052-8
5. Schmerl, S., Vogel, M., Rietz, R., Konig, H.: Explorative Visualization of Log Data to Support Forensic Analysis and Signature Development, pp. 109–118 (May 20, 2010) 978-0-7695-4052-8