Volume 8, No. 3, March – April 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

Thematically Clustering In Digital Forensics Text String Searching: A Survey

Nensi Kansagara and Shubham Singh
Institute of Technology
Nirma University, Ahmedabad, Gujrat, India

Abstract: These days' digital crimes are increasing more rapidly than earlier days. This is because now all data is on computer and small data store devices. For text string search forensic expert use text search tool which is initially design for giving 100 percentage query recall but it fails because it leads to high number of result which is irrelevant. The similar problem is faced by search engines too, but they use ranking algorithm to get precise result. Here we try to thematically clustering our text search result. It will improve investigators ability to search more reverent result in minimum search hits.

Keywords: Investigator, recall, clustering, thematically, neural network, koheneh self-organizing map.

I. INTRODUCTION

In digital forensic textual evidence is most important thing because by this evidences investigator comes on some conclusion about crime or any other analysis. This textual evidence are reside in server and available online or they are in your computer's hard disk. Example of textual evidence are e-mail, calendar, some web page, documents in hard disk, file system etc. When investigator wants to find some pattern, they enters a relevant query to retrieve relevant result. Tools which are used by investigators are typically design for getting 100 percentage query recall, but it is nearly irrelevant to say this that it achieve 100 percentage recall. This text string search result in to very noisy data means in result
(1)Decrease the irrelevant search hits.
(2)Represent your search hits in a manner such that, you find relevant his or result more quickly.
The second approach is more suitable for our forensic investigators because it suggest the ranking algorithm for evidence so more recent evidence is on the top.

II. LITERATURE SURVEY

For improving IR effectiveness we have to focus on improving text mining search. Digital forensic means not

A. Search Engines

From past two decades search engine technology is very useful and efficient. By using search engine we can get our result in seconds. This working of search engine can be categorize in to information retrieval. Search engine uses prioritize approach so most reverent hit on top for these it uses five ranking variables:
(1)Page Rank
(2)Query coincidence with anchor text (3)
Proximity measure
(4) Query term order
(5) Visual presentation.

Using these five variables we can improve our search results but we are talking about digital forensics for text searching so it is hard to extend some variables to digital text string search. These set there are many irrelevant results are present. Other problem for investigator is these tool design for giving each and every result so they give very large number of result which is very tedious for investigator [1].
For solution of these approach there are two solution or class can be define.

Different, so we cannot say that these particular tool is follow all standards[2]. Now a days text mining is very popular area in research, the entire field is known as text mining. These text mining is sub part of information retrieval. We can describe retrieval in various field, some of them are given below.

variables are query coincidence with anchor text and query term order. Here we want to achieve revelent result more quickly [3].

B. Desktop Search Engine

These days computer or personal device has hard drives which have very numerous amount of storage capacity. It will store the entire file system. For text string search forensic tool will examine all structured file information. The example of desktop search engines are Google desktop, Copernic desktop and open source are Eureka and Semantic file system[4].

These fie system search technically very fast then Internet search engines. Query processing speed is also more but it cannot extend to digital text string search. Because it takes a lot of time to prepare index for each and every document of file system and second is forensics require text string search independent of file system[5].

© 2015-19, IJARCS All Rights Reserved

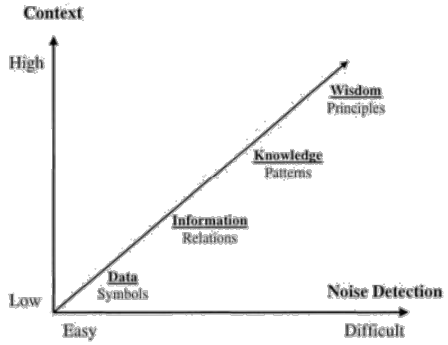

Fig. 1. Knowledge Management System (KMS) Understanding Hierarchy [6]

C. Text mining search

Information retrieval has many sub parts one of them is text mining. As name suggest it is about to retrieve information from data base but it has lot more useful than this it use to predict current information by having past data in textual form. Text mining has several processing task are given below [6].

Information extraction: it identifies relationship between two text using pattern recognizing.
Topic tracking: it provides automatic filtering of subject of user interest. It provides domain in which user is interested.
Information visualization: represent text data in graphic format.
Question answering: we can get specific text result by putting question to the system.
Concept linkage: detail conceptual relationship between pattern and text to extract relevant information.
Text categorization/classification: it thematically predefine categorize all text documents in classes.
Text clustering: after identifies theoretic categories of text documents it will give clusters of similar kind of documents.

The topic tracking process could be extends to digital forensic because it gives the relevancy feedback to user which allow forensic expert to get relevant result more quickly.

Text clustering and text categorization fits it to second class because categorization of text document is generated. From the semantically pattern of text documents they are clustered in various categories. So when user searches for query it will show more relevant hits and it will be very easy for user to retrieve data [6].

One thing is that for text clustering we use machine learning approach and text clustering is derived from text classification, text clustering is form of unsupervised learning so it does not require training set or any other thing.

In other approach text categorization needs supervised Approach of machine learning. Here training set is already availed. In thematic clustering two terms are important. One is preretrival method in which documents are clustered already and user have to enter relevant query to retrieve result.

Other is post retrieval goal is optimize query by thematically grouping query results. In this approach user enters relevant query and similar result of those query make clusters. Thus query can be optimize.

III. PROPOSED APPROACH

The research is ongoing for clustering of document sets for digital forensic text string research. Here researcher tries to prove feasibility and thematic clustering capability of text string search tool. For this purpose they have proposed a new algorithm.
When investigator tries to search particular text string then there are two types of data set one is thematically cluster structure and other one is un-structure by this investigators draw a conclusion, time between getting relevant results from clustered and unstructured data.

A. Algorithm Selection

Generally for data clustering we use five approach namely: partitioning, density base, grid base, hierarchical, model based. But problem with them is they are not capable of reducing noisy search result completely. Every one of them have problems.

Best of this approach is model based. The computational expenses of model base with respect to input size is O (square of n) but this is too expensive. For this other method is kohenen self organizing map approach which progress linearly with input size O(n) and some time it goes in logarithmic O(log n) thus investigator able to reduce much amount of noise.

The main use of SOM is to categorize following types of textual documents: Internet homepage, document abstract and newsgroup posting. Kohenen SOM is use for post retrieval, therefore we can say that it is unsupervised type of learning. Other new approach which improves scalability and performance we use scalable self organizing map algorithm (SSOM) [7]. Which uses sparse matrix multiplication and reduce computational expense.

IV. PROPOSED METHODOLOGY

To better understanding of text string search we worked on some useful tool for retrieving text information from database.AS we mentioned we used The sleuth kit(TSK) tool and modified version of autopsy are used to retrieve textsearch hits effectively [8].This experiment evaluation has begin.

A. Sample

In text string search we used tool/process for development over real-world dig-ital evidence. For commercial digital forensics we used 40 GB hard drive. Here we determine experimental volunteer to provide access to digital evidence. the digital investigator is very expert and experienced in

© 2015-19, IJARCS All Rights Reserved

forensic text string search do-main. So we can say that he has enough experience to deal with this kind of text search.

B. Performance Measurement

In this field various text string search tool is used. Which provides very good and optimized result of text document. These tool are follow. First type of tool we use for development during research. Second category of tools are industry stand-ard digital forensic. As they use at industry level and also research level so they use high algorithms like string matching and indexing/Boolean algorithm.

In our domain IR effectiveness is depends on both our query presentation and result getting from there. Suppose we enter a query and we have to find IR effectiveness so we can say that the time consume by irrelevant result hits from relevant hits is define as IR effectiveness [9].

Here we mentioned three tool for IR effectiveness. Here our main goal is to prioritize tools. So we determine IRefectiveness tool and rank that tool. By this we can determine feasibility of tool for post retrieval [10].

By using three types of different algorithm we can analysis there performance. Suppose we enter a same query and three algorithm gives different answer. So comparing their output we can use two algorithm EnCase and FTK. By help of these we can analyse the performance of three tools.

With the use of these two algorithm we can examine the output of three tools. We get the output in terms of investigating recall and investigated precision and their cut off points are 10 and 20 percentage of hit review [1].

So based on above two formulas.

Query precision= Relevant hits retrieved/ hits retrieved

Query Recall= Relevant hits retrieved / Total relevant hits in dataset

V. CONCLUSION

In this paper we proposed the text string search approaches in digital forensics. We tried to optimize our query precision and recall. Here we use different methods like kohenen self-organize map (SOM) by using machine learning approach for clustering the documents. Here we achieve very good results by document clustering. This field is very emerging field in digital forensic.

VI. REFERENCES

[1] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results," Digital investigation, vol. 4, pp. 49–54, 2007.

[2] V. SaiKrishna, A. Rasool, and N. Khare, "String matching and its applications in diversified fields," International Journal of Computer Science Issues, vol. 9, no. 1, pp. 219–226, 2012.

[3] G. Cheng and Y. Qu, "Searching linked objects with falcons: Approach, implementation and evaluation," International Journal on Semantic Web and Information Systems (IJSWIS), vol. 5, no. 3, pp. 49–70, 2009.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Computer networks and ISDN systems, vol. 30, no. 1,
pp. 107–117, 1998.

[5] D. Bhagwat and N. Polyzotis, "Searching a file system using inferred semantic links," in Proceedings of the sixteenth ACM conference on Hypertext and hypermedia, pp. 85–87, ACM, 2005.

[6] N. Beebe and G. Dietrich, "A new process model for text string searching," in IFIP International Conference on Digital Forensics,
pp. 179–191, Springer, 2007.

[7] N. L. Beebe, J. G. Clark, G. B. Dietrich, M. S. Ko, and D. Ko, "Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies," Decision Support Systems, vol. 51, no. 4, pp. 732–744, 2011.

[8] S. L. Garfinkel, "Digital forensics research: The next 10 years," digital investigation, vol. 7, pp. S64–S73, 2010.

[9] N. Beebe, "Digital forensic research: The good, the bad and the unaddressed," in IFIP International Conference on Digital Forensics,
pp. 17–36, Springer, 2009.

[10] G. L. Garcia, "Forensic physical memory analysis: an overview of tools and techniques," in TKK T-110.5290 Seminar on Network Security,
pp. 305–320, 2007.

© 2015-19, IJARCS All Rights Reserved