Advanced Computer Forensics
Ruiyang Liu 85C576
Conference Paper Summary – Group #3

Paper selected: Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results
Paper author: Nicole Lang Beebe*, Jan Guynes Clark
Word count: 650 words

In digital forensics science, text searching is a huge part which using indexing algorithms. However, current text searching tools have two major problems. The tool always designed to return all possible matching results, and most of evidences are not relevant to what the investigator wants. Easy to imagine, when investigator examines all the forensics evidence, it cost huge amount of time. Secondly, most of search results contain lots of noisy, which cause high level of information retrieval(IR) and information overload. In conclusion, a general consensus in the digital forensics text searching that standard tools are not able to process large data sets and not able to deal with IR overhead and information overload.

This paper gave a detailed analysis and comparison of previous solutions to this problems, internet search engines, desktop search engines and text mining researches.

Currently, there are two main solution to the problem of digital forensics text searching. First, decreasing the number and amount of result. But, this solution will cause information reduction which prohibit the investigator from getting the important evidence. Second, using a search hit ranking to support investigator get a more relevant result. This second approach is more attractive than the first one, because it provide the ability to obtain important information without sacrificing fidelity. According to the second solution, the paper attempt to get inspiration from other text search engines.

As internet search engines, they use ranking variables, such as PageRank, query term order, proximity measures, visual presentation characteristics, etc. If we want to extent internet engine solution to forensics text string searching, we must theorize and empirically validate new ranking variables.

Desktop search engines: they creates document indices and executes queries against inverted files. This may be a possible extend solution, because it is a very high efficiency solution. However, this solution is not realistic regarding with forensics, because this has a high startup costs of the index creation process, and also, digital forensics want to find digital evidence independent of the file system.

Text mining research: There are lots of techniques in text mining area, such as information extraction, text clustering, etc. The paper analyzes whether different techniques can be extended to digital forensics text searching. There are two possible approaches which fit well into the second solution class, because they are both the automatic and probabilistic way to process string data. First, text categorization approaches need to supervise the learning process. But, it is impossible to have time and ability to train and refine the dataset in digital forensics. Second, text clustering is a realistic approach to offer digital forensics relevant results because it is an unsupervised machine learning.

Based on previous analysis, the purpose of this on-going research in this paper is to test the feasibility and utility of thematically clustering digital forensic text string search results. Considering the cost of computation and complexity, the paper chooses a model-based algorithm called Self-Organized Map(SOM) approach for the solution. Further, in order to compare the new approach with previous standard approach, the same text search will be conducted on the same evidence using three tools: SOM new approach, EnCase and FTK, which EnCase and FTK are industry standard digital forensics tools. To evaluate, the first set of hits reviewed is presumed to be that corresponding to the investigator’s highest priority text search string. Also, a professional digital forensics volunteer will give final evaluation of the approach.

In summary, this paper give a theoretical performance superior of clustered search hits than other kinds of method. However, further research will determine whether this conclusion holds true in the digital forensic text string search context using real-world digital evidence.