Chapter 6
A Re-Ranking Method of Search Results Based on Keyword and User Interest

Ming Xu
Hangzhou Dianzi University, P. R. China

Hong-Rong Yang
Hangzhou Dianzi University, P. R. China

Ning Zheng
Hangzhou Dianzi University, P. R. China

ABSTRACT
It is a pivotal task for a forensic investigator to search a hard disk to find interesting evidences. Currently, most search tools in digital forensic field, which utilize text string match and index technology, produce high recall (100%) and low precision. Therefore, the investigators often waste vast time on huge irrelevant search hits. In this chapter, an improved method for ranking of search results was proposed to reduce human efforts on locating interesting hits. The K-UIH (the keyword and user interest hierarchies) was constructed by both investigator-defined keywords and user interest learnt from electronic evidence adaptive, and then the K-UIH was used to re-rank the search results. The experimental results indicated that the proposed method is feasible and valuable in digital forensic search process.

INTRODUCTION
The most common activity task for a forensic investigator is to search a hard disk for interesting evidences. The investigator needs to focus on specific evidence and important indicators of suspicious activity (e.g., specific key word searches). Unfortunately, the large size of modern hard disk makes it extremely difficult and wastes investigator’s vast time on huge irrelevant search hits. Many commercial or open sources tools have been developed to assist investigators to find relevant hits among large amounts of data, e.g., Forensic Tool Kit (AccessData, 2009), Encase (Guidance Software, 2009), etc. Nevertheless,
A Re-Ranking Method of Search Results Based on Keyword and User Interest

A huge number of search hits will be returned by search operations with high recall and low precision. What’s more, these digital forensic text string search tools fail to group and/or order search hits in a manner that appreciably improves the investigator’s ability to get to the relevant hits first.

In the works of Petrovic, and Franke (2007), they presented a new search procedure which makes use of the constrained edit distance in the pre-selection of the areas of the digital forensic search space that are interesting for the investigation. They divided the whole search space into several fragments and then computed constrained edit distance between each fragment and the query. However, our approach focuses on the entire hard disk instead of dividing it into small search spaces. Jee, Lee, and Hong (2007) also tried to improve search efficiency of digital forensic. Pattern matching board was used to build high speed bitwise search model for large-scale digital forensic investigations. This approach is different from ours, because we attempt to re-rank search results to reduce human efforts, and no additional hardware is used in the search process. It is not a new issue to personalize search results, which has been successfully applied in web information retrieval field. Kim and Chan (2008) learnt implicit interest from user to reorder search results. Various files on user’s computer were used as the training set of user interest. Unfortunately, their user profile did not focus to represent from general to specific topics. The works of Kim and Chan (2008) sufficed this end. Their approach is to learn a user interest hierarchy (UIH) from the web pages visited by user. A divisive hierarchical clustering (DHC) algorithm was designed to group words into hierarchy where higher-level nodes are more general and lower-level ones are more specific. In their study (Kim, & Chan, 2006), a ranking algorithm was proposed to reorder the results with a learned user profile. In our search results re-ranking algorithm, large amounts of data from digital evidence can be used to learn user interest, but the primary goal of digital forensic search is to satisfy the investigator, which is different from web personalization.

However, during the digital investigation, developing a profile of the offender can help focus the search. Armed with a better understanding of the possible motivation, modus operandi (MO), and signatures, the investigator can be able to derive specific search criterion for forensic analysis (Rogers, 2003). After all, our approach attempts to automate extract user interest from digital artifact, no human effort act in this process. So we believe that identifying user interest is important in digital forensic search process, and the UIH method can be extending to digital forensic field after combined with investigator’s focus.

Yang, Sun, and Sun (2006) also proposed an algorithm for learning hierarchical user interest models according to the Web pages which users had browsed. But they attempted to update user interest according to dynamic document set, while the dataset of the proposed method is based on static electronic evidence.

RELATED WORK

Dario Forto illustrated the importance of text searches in digital forensics (Forte, 2004). He took GREP tool as example, and realized that its power depends on the technical expertise of investigator. Beebe, and Dietrich (2007) disclosed a general consensus that industry standard digital forensic tools are not scalable to large data sets. In their following work (Beebe, & Clark, 2007), a new and high-level text string search process model was presented. They proposed and empirically tested the feasibility and utility of post-retrieval thematically clustering of digital forensic search results. Our method also attempts to resort search results for quickly finding important hits. The difference is that we try to learn
Related Content

Semi-Automatic Design of Spatial Data Cubes from Simulation Model Results
Hadj Mahboubi, Sandro Bimonte, Guillaume Deffuant, Jean-Pierre Chanet and François Pinet (2013). *International Journal of Data Warehousing and Mining* (pp. 70-95).
[www.igi-global.com/article/semi-automatic-design-spatial-data/75616?camid=4v1a](www.igi-global.com/article/semi-automatic-design-spatial-data/75616?camid=4v1a)

Discovering Hidden Concepts in Predictive Models for Texts’ Polarization
[www.igi-global.com/article/discovering-hidden-concepts-in-predictive-models-for-texts-polarization/130665?camid=4v1a](www.igi-global.com/article/discovering-hidden-concepts-in-predictive-models-for-texts-polarization/130665?camid=4v1a)

Dynamic Itemset Hiding Algorithm for Multiple Sensitive Support Thresholds
[www.igi-global.com/article/dynamic-itemset-hiding-algorithm-for-multiple-sensitive-support-thresholds/202997?camid=4v1a](www.igi-global.com/article/dynamic-itemset-hiding-algorithm-for-multiple-sensitive-support-thresholds/202997?camid=4v1a)

Data Mining in the Investigation of Money Laundering and Terrorist Financing
[www.igi-global.com/chapter/data-mining-investigation-money-laundering/73541?camid=4v1a](www.igi-global.com/chapter/data-mining-investigation-money-laundering/73541?camid=4v1a)