2008 IFIP International Conference on Network and Parallel Computing

# An Improved Method for Ranking of Search Results Based on User Interest

Hong-Rong Yang, Ming Xu, Ning Zheng
*Institute of Computer Application Technology, Hangzhou Dianzi University, P. R. China.*
*Email: hryang@stu.hdu.edu.cn*

## Abstract

*The most common task for a forensic investigator is to search a hard disk to find interesting evidences. While, the most search tools in digital forensic field fundamentally utilize text string match and index technology, which produce high recall (100%) and low precision. Investigators frequently waste vast time on huge irrelevant search hits. In this paper, we propose an improved method for ranking of search results to reduce human efforts on locating interesting hits. The search results are re-ranked using adaptive user interest hierarchies (AUIH), which considers both investigator-defined keywords and user interest learnt from electronic evidence. Experimental results indicate that the proposed method is feasible and valuable in digital forensic search process.*

## 1. Introduction

The most common activity task for a forensic investigator is searching a hard disk for interesting evidences. The investigator needs to focus on specific evidence and key indicators of suspicious activity (e.g., specific key word searches). The size of hard drives and other storage media today make it extremely difficult and time consuming. Nowadays, many commercial or open source tools have been developed to assist digital investigators to find relevant hits among large amounts of data (e.g., Forensic Tool Kit [5], Encase [3], etc). Nevertheless, huge number of search hits will be returned by search operations with high recall and low precision. Furthermore, many of these tools are unfortunately insensitive to the presentation order of search results. So investigators are forced to perform secondary search in the results set.

There are two ways to solve this problem [12, 13]. One is excluding parts of search hits which are irrelevant to case, while another is the improvement of search algorithm. It is not a good way to decrease the number of hits return, because this may cause information reduction. Therefore, in order to save the investigation time and reduce the complexity, one of the basic approaches is reranking the search results. This method is attracting because it can help investigators locate key hits quickly.

Importantly, search results are relevant to the specific crime case, hence, the proposed re-ranking approach need to be adaptive to fit different cases. Although internet search engines suffer for the similar problem, they personalize the search results with explicit or implicit interest from the user. One of the most effective methods is re-ranking search results using implicit user interest hierarchy learnt from bookmark [7], which is taken in account in our approach. We suppose that user interest can learn from text evidence in his hard disk.

However, the goal of search is not quite the same respectively in digital forensic field and in web search engine. To satisfy the investigator, the keywords proposed by investigator to search can be appropriately incorporate into the investigated user interest hierarchy.

In this paper, an improved method is proposed to rank search results for the purpose of reducing human effort on investigation. Search results are reordered with adaptive AUIH (user interest hierarchy), which combines user interest hierarchy with investigator's keywords. The former is learnt from the electronic evidences, and the latter is defined by investigator. The proposed approach is tested by salted volunteer through three convincing cases. Experimental results indicate that the proposed method is feasible and valuable in digital forensic search process.

The rest of this paper is organized as follow: Section 2 discusses related work in digital forensics field and web personalize field; Section 3 introduces adaptive user interest hierarchy (AUIH) and the clustering algorithm to build it; Section 4 details the proposed approach of reranking search results and file scoring method; Section 5 analyzes experimental results on the efficacy of methodology; Section 6 summarizes our finding and suggests future work.

978-0-7695-3354-4/08 $25.00 © 2008 IEEE
DOI 10.1109/NPC.2008.6

## 2. Related work

Dario Forto illustrated the importance of text searches in digital forensics [1]. He took GREP tool as example, and realized that its power depends on the technical expertise of investigator.

Beebe and Dietrich [11] disclosed a general consensus that industry standard digital forensic tools are not scalable to large data sets. A new, high-level text string search process model was presented. In their following work [12], they proposed and empirically test the feasibility and utility of post-retrieval thematically clustering of digital forensic search results. Also, our method attempts to resort search results for quickly finding important hits. The difference is that we try to learn user interest from evidence and combine it with investigator-defined keyword to build adaptive user interest hierarchy, which is used to score files in the search results.

In Slobodan Petrovic and Katrin Franke's work[13], a new search procedure was presented that 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.

Hyungkeun Jee et al [8] 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, since 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. Jaime Teevan et al [9] 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 general to specific topics.

H R.Kim and P K.Chan's work [6] sufficed this end. Their approach is to learn a user interest hierarchy (UIH) from 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 following study [7], a ranking algorithm was proposed to reorder the results with a learned user profile (UIH). In our search results reranking 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 [10]. 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 interest.

Feng Qin Yang et al [4] also proposed an algorithm for learning hierarchical user interest models according to the Web pages users had browses. 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.

## 3. Adaptive user interest hierarchy

In H R. Kim's user interest hierarchy [6], more general interest is represented by a larger set of words, which are extracted from web pages. Each web page can be assigned to a set of nodes for further processing. According to DHC algorithm, the similarity function and threshold-finding method greatly influence the clustering algorithm [6]. The former measures how close the two words are related, and the latter determines what value of similarity is considered to be "strong" or "weak". Edges with weak weights are removed in SimilarityMatrix [6] (Denoted by SM). In this work, we fix the similarity function to AEMI-SP, and consider threshold-finding method based on MaxChildren. Before discussing our approach for building AUIH, a picture of UIH is drawn to be taken as an example.

To generate a UIH, seven web pages in bookmarks of web browser were collected as input. We firstly parsed the HTML documents and extracted text information from them without considering link or multimedia information. And then, the words were fragmented (Chinese pages) or stemmed (English pages) so that we can get all words in web page. At last, we filtered the words through a stop list, which contains the most common words. The sample data set is shown in Table 1. It should be note that the samples throughout our paper are in Chinese language. To illustrate our idea more intelligibly, we have translated them into English.

**Table 1: Sample data set**

| Page | words |
|---|---|
| 1 | computer academy journal Chinese science |
| 2 | computer academy conference deadline security submit |
| 3 | computer journal submit engineering |
| 4 | computer crime investigate network security forensic technology case |
| 5 | computer confident abuse identify material enterprise |
| 6 | paper submit review revise research study |
| 7 | paper submit review revise research |

In Figure 1, the words in nodes come from the sample data set (Table 1). Each node represents a conceptual relationship if those terms occur together at the same web page. For example, in the left bottom of picture, 'journal' and 'Chinese' can be typed as journal submission, while in its brother node, 'conference' and 'deadline' are brought into conference program, but the exchange is not true. Additionally, these words are all related to some other words, such as 'research' and 'paper', which are contained in the parent node. While investigating the whole tree, it can be easily found that left side represents user interest about research and paper submission, and the right side is related to computer forensics.

In this study, we mainly focus on improving search efficiency of investigator. It seems natural to incorporate investigator's interest into UIH as a new tree, which we call it AUIH. In digital forensic field, it is common that there is hundreds of thousands of files in digital evidence, so a huge UIH will be built by using original cluster algorithm. We attempt to utilize keywords input by investigator to localize original SM. Here, we should pay attention to a hypothesis: there exists an intersection between keywords set and SM words set, that means at least one word in keyword set also occurs in SM words set. It is believed that the investigator can easily seek importance evidence if the input keyword is contained in the SM word set.

Note that we would like to see a small AUIH contains one or more keywords, so a new threshold-finding method should be designed to suffice it. We observe a component of SM, called keyword's similarity set (denoted by C), which contains similarities between keywords and other words. For example, in SM, if there are 10 edges connected between keywords and other words, the member number of C (denoted by n) is 10. We determine the threshold using the formula below:

$$threshold = \left\{ \mathrm{Max}(s, C_i) | i = \mathrm{Min}(n, t) \right\} \qquad (1)$$

Before the threshold is computed, similarity values in C should be arranged in descend order (i=1, 2...n). In equation 1, we choose $C_i$ as the bigger one between the smallest element in C and t, which is defined as a constant, i.e., 10. The role of t is to prevent too many similarity edges are under consideration. s, which is defined in the similar way as t, acts to prevent the situation that too many edges will be cut. The value of s can be determined using MaxChildren method.

After localizing SM with the threshold discussed above, we will build AUIH according to MaxChildren method and AEMI-SP similarity function as the same as original UIH algorithm. A simple example of AUIH is shown in Figure 2.

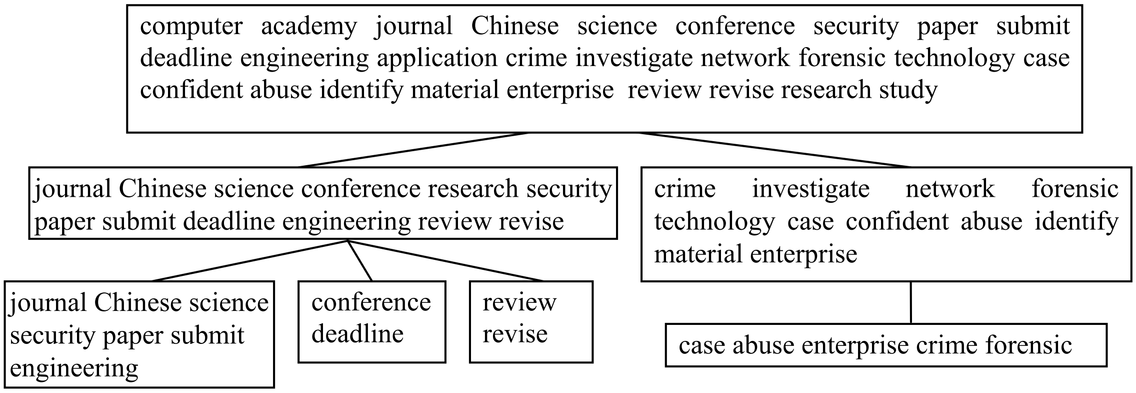
computer academy journal Chinese science conference security paper submit deadline engineering application crime investigate network forensic technology case confident abuse identify material enterprise review revise research study
journal Chinese science conference research security paper submit deadline engineering review revise
crime investigate network forensic technology case confident abuse identify material enterprise
journal Chinese science security paper submit engineering
conference deadline
review revise
case abuse enterprise crime forensic

**Figure1: Sample user interest hierarchy**

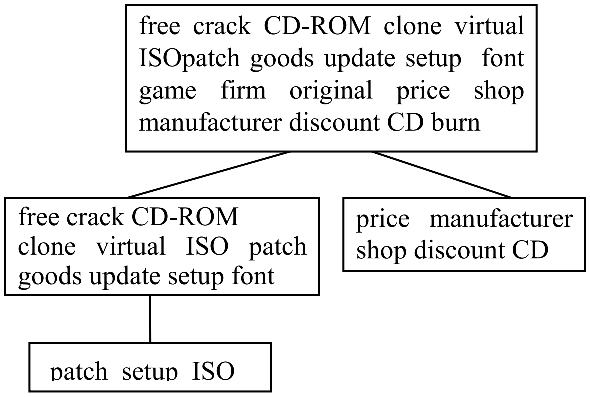
free crack CD-ROM clone virtual ISOpatch goods update setup font game firm original price shop manufacturer discount CD burn
free crack CD-ROM clone virtual ISO patch goods update setup font
price manufacturer shop discount CD
patch setup ISO

**Figure 2: Sample adaptive user interest hierarchy**

In this mock case in figure 2, someone was suspected of making pirate sale of video, audio or games. The input keywords for searching were ‘crack’, ‘manufacturer’ and ‘free’. As drawn in Figure 2, his interest was demonstrated well in the AUIH we build. We are confident that search efficiency of digital investigation would be greatly improved by using AUIH.

To illustrate the proposed approach clearly, the search procedure is shown in Figure 3.

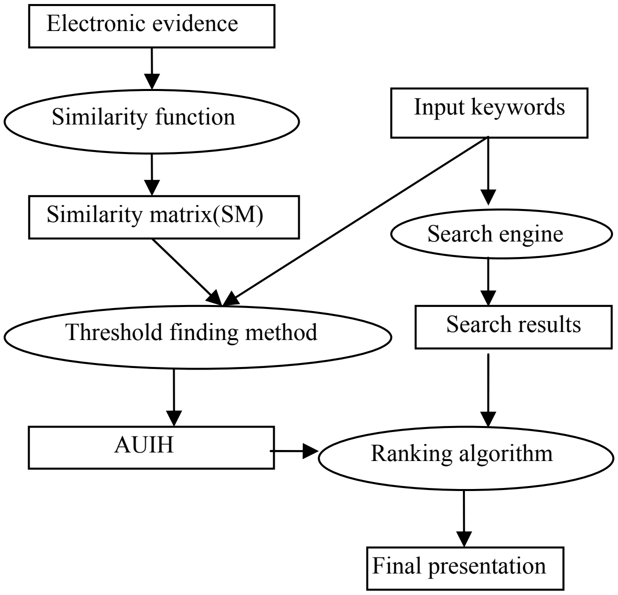

**Figure 3: The proposed search procedure**

## 4. Re-rank search results

In this section, we will discuss the ranking algorithm for reordering search results, which have been returned by traditional search engine. The most important step is scoring each file in search results. Therefore, a scoring method will be designed so that the more interesting file would be assigned a higher score. We fortunately found that H R .Kim’s work has made a good example of how to scoring search results depending on UIH [7], which is referred in our method.

Given a file in search results, we firstly identify the terms both occur in the file and AUIH. The number of distinct terms in AUIH is denoted by m, and the number of distinct terms in the file is denoted by n. For each matching term $t_i$ , we compute the score of it according to three sides: depth of an AUIH node $D_{ti}$, length of a term $L_{ti}$and frequency of a term$F_{ti}$. The first one is related to AUIH structure. The second one is related to the term itself. And the last one is about the importance of the term in file. The emphasis of term [7] is ignored in this work because many of evidences have no visual character. The significance of a term can be measured by estimating the probability. P ($D_{ti}$) represents the probability of marching term $t_i$ at depth $D_{ti}$in AUIH, P ($L_{ti}$) is defined in the similar way of length$L_{ti}$, and P ($F_{t\,i}$) represents the one of $F_{t\,i}$in the file. Assuming independence among these three characteristics, we estimate the score of term $S_{t\,i}$ below:

$$\begin{cases} P(D_{t\,i}) = \dfrac{\text{number of distinct terms at depth } D_{t\,i} \text{ in AUIH}}{m} \\ P(L_{t\,i}) = \dfrac{\text{number of distinct terms of length } L_{t\,i} \text{ in AUIH}}{m} \\ P(F_{t\,i}) = \dfrac{\text{number of distinct terms with frequency } F_{t\,i} \text{ in file}}{n} \\ S_{t\,i} = -\log_2 P(D_{t\,i}) - \log_2 P(L_{t\,i}) - \log_2 P(F_{t\,i}) \end{cases} \quad (2)$$

The score of file is computed by summing the score of each matching term. The final presentation of search results is arranged in a descend order of file score.

## 5. Evaluation

### 5.1. Set up

At the beginning of the experiments, reliable digital evidences are necessary to be collected. Three mock cases were constructed for the purpose of evaluating the feasibility of the proposed methodology in cyber crime cases. Table 2 lists the detail of cases.

**Table 2: The detail of cases**

| Case | Case Description | Evidence Size (hard disk) |
|---|---|---|
| A | Abusing resource inside the enterprise | 40G |
| B | Drug trafficking via Email | 40G |
| C | Spreading sex information via Internet | 40G |

The three cases listed above are all popular in modern society. Another important factor in the evaluation is the set of searching keywords, which will be determined by an experienced volunteer. This volunteer should be well familiar with these cases, and has access to the electronic evidence. Hence, we are confident that the keyword set would be realistic and valuable for investigation.

## 5.2. Measure performance

There are a great many digital forensic software integrated with search tool. For example, FTK (AccessData) and Encase (Guidance Software). However, many of these tools are insensitive to sorting search results other than group them to simple categories, i.e., hits counts. dtSearch[2] , which is a common-used desk search tool, reorder search results by comprehensively computing file score according their relevancies to the keywords. We can estimate the performance of the proposed approach by comparing it with dtSearch. As for the search engine used to produce search results, we would like to select string match or index technology. Specially, dtsearch tool is preferred because it makes our experimental comparison with dtSearch clearer.

Then, we study the measure algorithm. The search efficiency of digital forensic can be measured on whether the search hit is relevant to case or interesting to investigator. Thereby, the investigative relevancy of every hit in search results set produced by our method and dtSearch will be determined by the volunteer. To make a quantitative measurement, we categorize relevancy as 'relevant' and 'irrelevant'. A hit is marked as 'relevant' when this hit is important for investigation. 'irrelevant' hits are those that have no relation to case. Given a fix number of search results, more 'relevant' hits represent better performance. Two specific measures are given as follows:

$$\text{precision}=\frac{\text{relevant hits}}{\text{total hits}} \tag{3}$$

$$\text{recall}=\frac{\text{relevant hits}}{\text{total relevant hits in data set}} \tag{4}$$

Note that we mainly attempted to reduce time spent on analysing search results, so we mainly focus on the precision of search hits in the experiment.

## 5.3. Experiments

To test the availability of the proposed approach, a tool is implemented in C program under Linux operation system. There are several factors that affect the performance of the tool. Other than similarity function and threshold finding method discussed in UIH algorithm [6], four parameters are important in the proposed approach. Those are the value of s&t in equation 1, the size of keyword set, the percentage of reviewed hits and the scope of training set. We will focus on these factors in the following sections.

### 5.3.1. Value of *s&t*

The first parameter we study is s&t shown in equation 1 (section 3). s and t play an important role in threshold finding stage of the proposed method. Simply, s is selected in set {1.5, 2, 2.5, 3, 3.5}, and t is selected in set {5, 10, 15, 20, 25}. The best precision and recall is achieved when s&t pair is (2.5, 10), which is used in the following experiments.

### 5.3.2. The size of keyword set

The second parameter is the size of keyword set, which equals to the number of keyword in this set. In a sense, single keyword is not enough, and the performance will be better when more keywords occur in SM. At the other side, the investigator would be experienced and not necessary to input too many keywords. Hence, more than 2 keywords are considered in our experiments.

### 5.3.3. The percentage of reviewed hits and the scope of training set

The proposed approach is based on learning user interest from electronic evidence. Therefore, the scope of training set is important for building SM and AUIH. The common text evidences in computer encompass IE Favourites, IE browsing history, E-mail archive, desktop files, 'My Document' files and other text files in hard disk. Dissimilar adaptive user interest hierarchies would be built using different scopes of training set and produce dissimilar results. Another factor in experiments is the percentage of search hits reviewed by investigator. Because huge amount of hits are returned by each search process, the investigator would like to find important evidence by reviewing parts of search results. We assumed that the search results are reviewed from end to end. The results using different training sets and different percentages of reviewed hits are illustrated in Table 3.

**Table 3: Precision with different training sets and percentages of reviewed hits**

| Building SM with IE Favourites and IE browsing history | | | | |
|---|---|---|---|---|
| Percentage of search hits | 10% | 20% | 50% | 100% |
| Case A | 80.5% | 82.1% | 84.6% | 90% |
| Case B | 82.7% | 85.5% | 87.8% | 91.4% |
| Case C | **89.5%** | **92.0%** | **92.3%** | **95.1%** |
| Building SM with IE Favourites, IE browsing history and Email archive | | | | |
| Percentage of search hits | 10% | 20% | 50% | 100% |
| Case A | 84.5% | 86.7% | 87.9% | 89.0% |
| Case B | **92.5%** | **93.0%** | **94.4%** | **94.7%** |
| Case C | 85.1% | 86.8% | 86.5% | 88.2% |
| Building SM with all text files in hard disk | | | | |
| Percentage of search hits | 10% | 20% | 50% | 100% |
| Case A | **88.1%** | **89.2%** | **90.7%** | **92.6%** |
| Case B | 84.4% | 85.9% | 87.7% | 90.6% |
| Case C | 87.5% | 88.5% | 90.2% | 92.3% |

As shown in Table 3, best performances (bold) are obtained with different training set for different cases. For example, the suspect of case B, which is about drug trafficking via Email, would left important evidence in Email archive. So the precision is higher than others when Email archive occupies moderate part of the training set. In the similar way, we achieve best performance in case C when IE archive is considered well. In case A, employee might store key evidence in anyplace of hard disk, hence, so the highest precision is obtained when the whole data are considered.

### 5.3.4. Compare with dtSearch tool

In this section, we compare the proposed approach with dtSearch. Different cases are considered singly in this evaluation. The results are drawn in Figure 4-6. X (%) plots the percentage of reviewed hits, and Y plots precision. From these pictures, we can easily found that higher precision is obtained by the proposed approach while the percentage of reviewed hits is low, especially in 10%. It is believed that the investigator can locate relevant hits more quickly than using dtsearch tool.

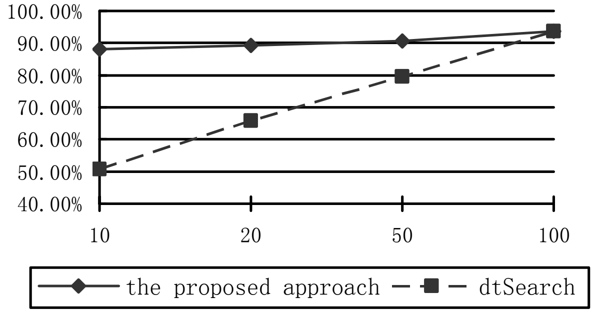

**Figure 4: Precision of the proposed approach and dtSearch (Case A)**

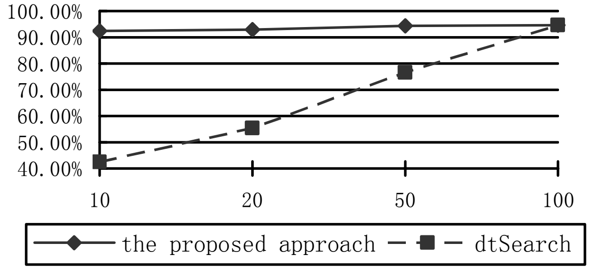

**Figure 5: Precision of the proposed approach and dtSearch (Case B)**

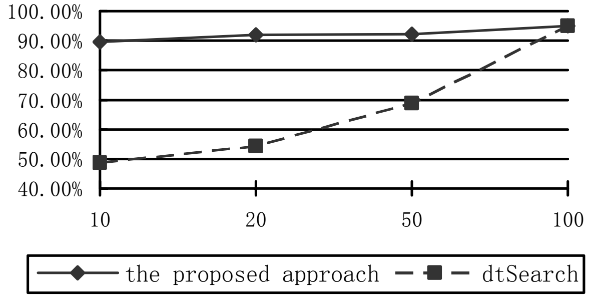

**Figure 6: Precision of the proposed approach and dtSearch (Case C)**

At last, we should realize that the proposed method may not be capable under some circumstances. For example, in some cases, the investigator is interested in credit card NO or ID card No, so the input keywords can be all numbers or letters, instead of meaningful terms. However, the search process would be toilless in these cases since long number (i.e., credit card NO) may occur in a few files than meaningful terms. The proposed method mainly focuses on enhancing search efficiency of investigator while huge hits are ahead.

## 6. Conclusion

The research field of digital forensic search is growing up in recent years. The main contribution of this study is to enhance the search efficiency of digital forensic. There is quite a few works about using user interest to improve the search efficiency before ours. In the proposed approach, the search results are reordered based on the AUIH (adaptive user interest hierarchy), which is learnt from digital evidence and keyword set. Human effort is greatly reduced and experimental results are promising.

However, there are several challenges of the proposed approach. Importantly, there should be more enough text evidence in case that can produce richer user interest. So data recovery is necessary since the criminal may delete important text evidence before investigated. Besides, the investigator's experience is another important factor to affect the performance of this approach. More keywords proposed by investigator occur in SM, a better presentation of investigator's interest can be given by AUIH and better performance would be achieved. Furthermore, the search efficiency can be greatly improved when the scope of training set is limited according to the specific case. Finally, in the proposed approach, a new AUIH should be built for each search process, so additional computer time is required. However, the primary goal of this study is reducing human effort on investigation, and human time spent on analysing search results will greatly eclipse computer time in our approach. We endeavour to find some ways to reduce the complexity of the proposed approach, which is part of our future work. We also plan using other elements (e.g., hit counts of web page, access time of file and so on) to build a richer AUIH for ranking.

## 7. Acknowledgements

This work is supported by the Natural Science Foundation of Zhejiang Province of China under No.Y106176, and the science and technology research Planned Projects of Zhejiang Province of China under Grant No.2007C33058.

## 8. References

[1] Dario Forte. "The importance of text searches in digital forensics", Network Security, Volume 2004, Issue 4, April 2004, pp.13-15.

[2] dtSearch. Available at http://www.dtsearch.com/.

[3] EnCase Forensic. Available at http://www.guidancesoftware.com/products/ef_index.asp.

[4] Feng Qin Yang, Tie Li Sun and Ji Gui Sun. "Learning hierarchical user interest models from web pages", Wuhan University Journal of Natural Sciences, 2006, Vol.11. No.1, pp.6-10.

[5] Forensic Toolkit 2.0. Available at http://www.accessdata.com/Products/ftk2test.aspx.

[6] H Kim and PK Chan. "Learning implicit user interest hierarchy for context in personalization", Proceedings of Intelligent User Interfaces, IUI'03.

[7] H Kim and PK Chan. "Personalized search results with user interest hierarchies learnt from bookmarks", Lecture Notes in Computer Science, Volume 4198, 2006, pp.158-176.

[8] Hyungkeun Jee, Jooyoung Lee, and Dowon Hong. "High speed bitwise search for digital forensic system", Proceedings of world academy of science, engineering and technology, Volume 26, Dec 2007.

[9] Jaime Teevan, Susan T. Dumais and Eric Horvitz. "Personalizing search via automated analysis of interests and activities", SIGIR '05.

[10] M Rogers. "The role of criminal profiling in the computer forensics process", Computers & Security, May 2003, Vol. 22 Issue 4, pp.292-298.

[11] Nicole Beebe and Glenn Dietrich. "A new process model for text string searching", IFIP International Federation for Information Processing, Volume 242, Advances in Digital Forensics III 2007, pp.179-191.

[12] Nicole Beebe and Jan G. Clark. "Digital forensic text string searching——improving information retrieval effectiveness by thematically clustering search results", Digital Investigation 4S.2007, pp. 49–54.

[13] Slobodan Petrovi´c and Katrin Franke. "Improving the efficiency of digital forensic search by means of the constrained edit distance", Third International Symposium on Information Assurance and Security, 2007.