Chapter 5

WEB USER PROFILING BASED ON BROWSING BEHAVIOR ANALYSIS

Xiao-Xi Fan, Kam-Pui Chow, and Fei Xu

Abstract Determining the source of criminal activity requires a reliable means to estimate a criminal's identity. One way to do this is to use web browsing history to build a profile of an anonymous user. Since an individual's web use is unique, matching the web use profile to known samples provides a means to identify an unknown user. This paper describes a model for web user profiling and identification. Two aspects of browsing behavior are examined to construct a user profile, the user's page view number and page view time for each domain. Four weighting models, based on the term frequency and term frequency – inverse document frequency weighting schemes, are proposed and compared. Experiments involving 51 personal computers demonstrate that the profiling model is very effective at identifying web users.

Keywords: Web user profiling, browsing behavior, term frequency

1. Introduction

Due to the widespread use of computers and networks, digital evidence plays an important role in criminal investigations. The large amount of digital evidence that resides on computer systems can help investigate and prosecute criminal actions [3]. The evidentiary artifacts are varied and include chat logs, agendas, email, application information, Internet history and cache files. In many cases, investigative leads are found in a criminal's web browser history [8].

Investigators often face situations where a computer associated with a cybercrime has been found, but the suspect is unknown. In such a scenario, the investigator can create the suspect's profile from the digital evidence residing on the computer, and match the profile with the profiles of other individuals to identify the suspect.

G. Peterson and S. Shenoi (Eds.): Advances in Digital Forensics X, IFIP AICT 433, pp. 57–71, 2014.

Identifying a user from his/her web browsing history requires an efficient model that associates individuals with their web activities. This paper proposes a web user identification model that creates a user profile based on web browsing activities and identifies relationships between different users. The model provides web user identity information based on the assumption that a given user has consistent preferences. Experiments involving 51 computers demonstrate that the profiling model can be used to identify web users.

2. Related Work

User profiling has been used in the context of e-commerce and personalization systems, the two primary applications being content-based filtering and collaborative filtering. Grcar, *et al.* [6] have developed a topic-ontology-based system that creates dynamic profiles of users from their browsed web pages. The viewed pages are represented as word vectors and hierarchical clustering is performed to construct a topic ontology. A similarity comparison maps users' current interests to the topic ontology based on their browsing histories.

Fathy, *et al.* [4] have proposed a personalized search that uses an individual's click-history data to model search preferences in an ontological user profile. The profile is then incorporated to re-rank the search results to provide personalized views.

Some research has been conducted in the area of behavioral profiling. Bucklin, *et al.* [2] have modeled the within-site browsing behavior of users at a commercial website. Their model focuses on two browsing behaviors: a user's decision to continue browsing the site or to exit, and how long a user views each page during a website visit.

Forte, *et al.* [5] conducted a web-based experiment to observe how people with different backgrounds and levels of experience make decisions. The experiments analyzed the types of actions, sequences of actions and times between actions using significance testing, regression modeling and evidence weighting. The results suggest that recording the actions taken and the time spent on each decision can help create statistical models that can discern demographic information.

Similarly, Hu, *et al.* [7] have presented an approach that predicts user gender and age from web browsing behavior. Gender and age are predicted by training a supervised regression model based on users' self-reported gender, age and browsing history. Then, based on the age and gender tendency of the web pages that a user has browsed, a Bayesian framework is used to predict the user's gender and age.

Research in the areas of user profiling and behavior analysis have not explicitly targeted digital forensic investigations. Indeed, criminology profiling techniques have primarily focused on violent criminal activities.

One area in which criminal profiling has been applied to computer crimes is the insider threat. McKinney, *et al.* [10] have proposed a masquerade detection method, which creates a user profile using a naive Bayesian classifier that analyzes running computer processes. Bhukya and Banothu [1] engage GUI-based behavior to construct user profiles; a support vector machine learning algorithm is used to train the system to classify users. To our knowledge, there is little, if any, research that has explicitly focused on identifying web users based on their browsing history.

3. Methodology

Figure 1 presents a schematic diagram of the proposed web user profiling and identification model. The investigator uses evidence from the target computer to identify the user. The suspect computers are the candidate computers that are compared with the target computer. The candidate computer with the highest similarity score is assumed to have been used by the same user as the target computer.

Browsing activity data is extracted from all the browsers installed on the target and candidate computers. After preprocessing the data, the user's domains of interest are extracted and the page view number (PVN) and page view time (PVT) are calculated for each domain. Each computer is then represented as a weighted vector of the domains, where each weight is the term frequency (TF) or term frequency – inverse document frequency (TFIDF) of the PVN or PVT. Finally, a cosine similarity measure is applied to calculate the similarity scores between the target computer profile and the candidate computer profiles to identify the candidate computer with the highest similarity score.

3.1 Data Extraction

Web browsers can be classified based on their web browser engines, as shown in Table 1 [15]. Different browsers record browsing history in different ways. Table 1 shows that most browsers are based on the Trident engine (IE core). Every web page visited by a Trident engine browser is stored in the `index.dat` file, along with the access time, Windows user name, etc. Analyzing the `index.dat` file can help understand a user's browsing behavior. Some browser forensic tools can automatically parse the `index.dat` file; example tools are Index.dat Viewer, Pasco, IE History View, Web Historian and NetAnalysis. We used IE History View

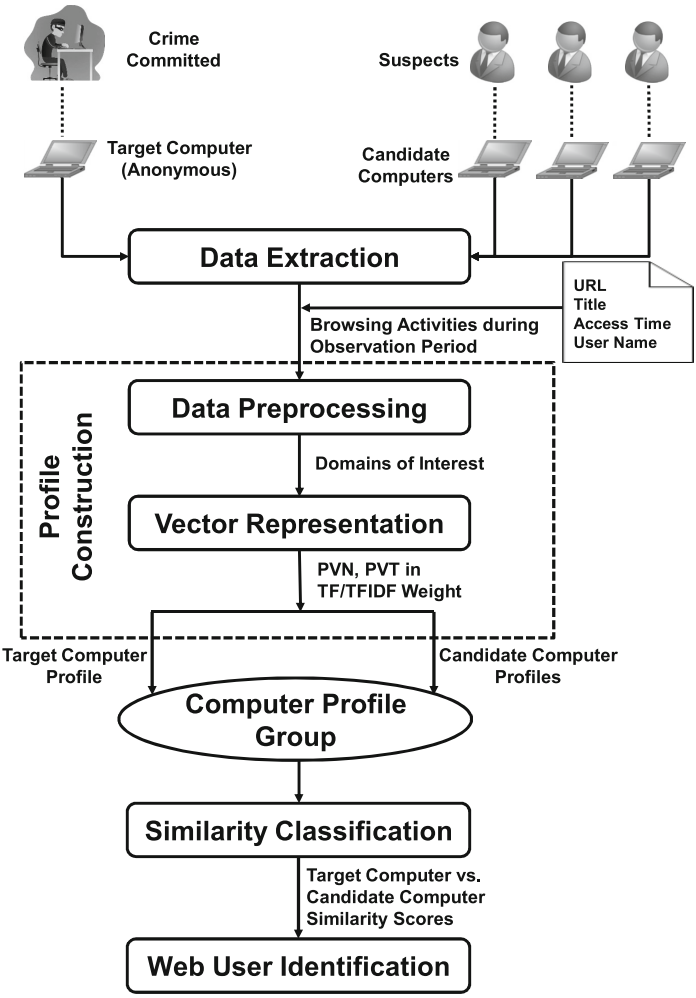

Figure 1. Proposed web user profiling and identification model.

v1.70 [12] to extract Internet history records from `index.dat` files in this research. Also, Chrome History View v1.16 [13] and Mozilla History View v1.51 [14] were used to extract history records from Google Chrome and Mozilla Firefox browsers, respectively.

For each computer, the history records from all the installed browsers were extracted and combined together. Each record contained a URL, title, access time and Windows user name. When a computer had multiple users, a separate profile was created for each user.

Table 1. Web browser categories.

Engine	Corresponding Browsers
Trident (IE Core)	Internet Explorer, MyIE9, The World, Maxthon, Tencent Traveler, 360 Secure Browser, Sougou Explorer, 360chrome, MenuBox, GreenBrowser
Gecko	Mozilla Firefox
Webkit	Google Chrome, Safari
Presto	Opera

History records from the target and candidate computers were extracted for the last 40 days. The reason for using a 40-day observation period is that users can easily change their browsing topics over short periods of time, but it is much harder for users to change their browsing interests over the long term.

3.2 Data Preprocessing

Several preprocessing steps were applied to the history data that was collected. For Trident engine browsers, duplicate records may exist in the primary `index.dat` history file as well as the daily/weekly `index.dat` file. Duplicate records with the same URLs and access times were deleted. Also, pop-up desktop news notices were removed because, if the users did not actively click on the pop-ups, they were likely not of interest.

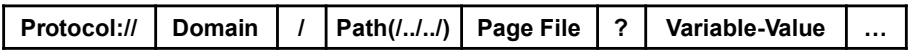

Figure 2. URL structure.

Next, the domain information was extracted from each history record. Most URLs have the structure shown in Figure 2. The structure includes the protocol, domain, path, page file and variable-value pairs [11]. For example, the domain of the URL `http://en.wikipedia.org/wiki/Domain` is `en.wikipedia.org`. Every domain name is a hierarchical series of character strings, with the different levels separated by dots proceeding from the top-level domain at the right end to specific host names on the left (Table 2).. The top-level domain is expressed using a generic code (e.g., `.com`, `.edu` or `.gov`) or a country code (e.g., `.cn` or `.uk`). The second label from the right denotes the second-level domain. Labels to the left of the second-level domain correspond to sub-domains of

Table 2. Examples of domain names.

Domain Example	Host Name	Sub-Domain	Second-Level Domain	Top-Level Domain	Extracted Domain Part
www.hku.hk	www	–	hku	hk	hku.hk
www.cs.hku.hk	www	cs	hku	hk	hku.hk
www.discuss.com.hk	www	–	discuss	com.hk	discuss.com.hk

the second-level domain. The top-level and second-level domains (referenced as domain in later sections) were extracted from every history record.

Note that an extracted domain visited only once during the observation period was not included in a profile. This is because one-page visits were assumed to be primarily due to user input errors.

3.3 Vector Representation

A vector space model was used to build computer profiles. Each computer (c_j) was represented as a vector $c_j = (w_1, w_2, ..., w_N)$, where N is the number of domains. Each weight domain pair (w_i) corresponds to a domain extracted from the history records of the target computer. The weight was calculated using TF and TFIDF over the PVN or PVT of the domain. Domains were ranked according to their weighting values. The top N domains from the target computer were chosen to form the profile vectors of the target and candidate computers. This resulted in four web user profile models: TF-PVN, TF-PVT, TFIDF-PVN and TFIDF-PVT.

TF-PVN Model. Term frequency (TF) is the number of times that a term occurs in a document. In this study, the TF-PVN model assigns a weight to the page view frequency $tf\_pvn_{d,c}$ of domain d in the browser history of computer c according to the equation [9]:

$$tf\_pvn_{d,c} = \frac{pvn_{d,c}}{\sum_k pvn_{k,c}} \tag{1}$$

where $pvn_{d,c}$ is the number of page views of domain d from computer c during the observation period and $\sum_k pvn_{k,c}$ is the total number of page views from computer c during the observation period.

For a given computer c, the set of weights determined by TF-PVN may be viewed as a quantitative digest of the computer. Table 3 shows the top seven domains from a target computer sorted by their TF-PVN

Table 3. TF-PVN weighting results.

Domain	pvn	tf_pvn (3,067 Visits)
google.com	535	0.174
hku.hk	366	0.119
taobao.com	235	0.077
chinamac.com	202	0.066
hsbc.com.hk	102	0.033
hkst.com	68	0.022
youtube.com	60	0.020

values. The exact ordering of domains can be ignored, but the number of occurrences of each domain is material. However, certain domains such as google.com are visited very frequently by users. These domains have high TF-PVN values for most users and, therefore, have little or no user discriminating power.

TFIDF-PVN Model. TFIDF-PVN attenuates the effect that frequently visited domains have on domain weighting by giving high weights to domains that occur frequently for one computer but rarely for the other computers.

Computing the TFIDF-PVN of a domain for a given computer involves two sub-calculations: how often the domain is visited by the computer (TF) and how many computers in the collection visit the domain (DF). The DF is inverted to yield the IDF, which is then combined with the TF. IDF is calculated as the logarithm (base 2) of the quotient of the total number of computers ($|C|$) and the number of computers containing the domain ($|\{c \in C | d \in c\}|$) in order to scale the values [9]:

$$tfidf\_pvn_{d,c} = tf\_pvn_{d,c} \times \log_2 \frac{|C|}{|\{c \in C | d \in c\}|}. \quad (2)$$

Table 4 shows the same seven domains from the target computer as in Table 3 sorted by their TFIDF-PVN values. Note that some new domains have emerged among the top seven domains. Also, the weight of chinamac.com has increased substantially because only two out of fifteen computers shared this domain (high IDF). Furthermore, the weight of google.com has decreased because it appeared in almost all the computers (low IDF). The reason for these differences is that TFIDF-PVN assigns a high value to a domain when it occurs many times in a small number of computers (high discriminating power for the computers).

Table 4. TFIDF-PVN weighting example.

Domain	pvn	tf_pvn (3,067 Visits)	idf (15 Computers)	tfidf_pvn
chinamac.com	202	0.066	2.907	0.191
taobao.com	235	0.077	1.100	0.084
hku.hk	366	0.119	0.447	0.053
hkst.com	68	0.022	1.585	0.035
hsbc.com.hk	102	0.033	0.907	0.030
youtube.com	60	0.020	0.907	0.018
google.com	535	0.174	0.100	0.017

The value is lower when a domain occurs few times on a computer or occurs on many computers (low discriminating power for the computers). The value is the lowest when a domain appears on virtually all the computers.

TF-PVT Model. The PVN indicates a user's preferred (i.e., frequently visited) websites, but the same weight is assigned to each visit and the weight does not reflect the amount of time the user spent viewing the website. For example, a user who watches videos on a website would spend more time on the website, but the click frequency would be low. The page view time (PVT) is used to capture this important aspect of a user's browsing behavior.

The PVT calculation is based on a view session [2]. A view session starts when a website is requested and ends after an idle period of at least 30 minutes or when a new domain is requested. The idle period accounts for the fact that a browser does not record when a user leaves a page. The PVT is computed as the difference in access times between two consecutive page views based on the assumption that a user browses the new page after clicking a link. Thus, the TF-PVT weight is the time spent on domain d based on the browser history of computer c:

$$tf\_pvt_{d,c} = \frac{pvt_{d,c}}{\sum_k pvt_{k,c}} \tag{3}$$

where $pvt_{d,c}$ is the total page view time for all the pages corresponding to domain d for computer c during the observation period, and $\sum_k pvt_{k,c}$ is the total page view time for computer c during the observation period.

Table 5 shows the same seven domains and their weights. Note that TF-PVT gives different results than TF-PVN. For example, the weights of youtube.com and taobao.com are high because the user watches

Table 5. TF-PVT weighting example.

Domain	pvt (sec.)	tf_pvt (207,021 sec.)
google.com	42,506	0.205
youtube.com	16,805	0.081
taobao.com	14,474	0.070
hku.hk	14,112	0.068
chinamac.com	11,343	0.055
hkst.com	3,815	0.018
hsbc.com.hk	859	0.004

videos and shops online. Like TF-PVN, this weighting scheme does not abandon non-discriminating domains such as google.com that may lead to incorrect user identification.

TFIDF-PVT Model. Like TFIDF-PVN, TFIDF-PVT attempts to provide more meaningful weights. If a domain is viewed for a long time from most of the computers, the discriminating ability of this domain is weaker. Consequently, TFIDF-PVT ranks the website lower than TF-PVT. The TFIDF-PVT weight is computed as:

$$tfidf\_pvt_{d,c} = tf\_pvt_{d,c} \times \log_2 \frac{|C|}{|\{c \in C | d \in c\}|}. \quad (4)$$

Table 6. TFIDF-PVT weighting example.

Domain	pvt (sec.)	tf_pvt (207,021 sec.)	idf (15 Computers)	tfidf_pvt
chinamac.com	11,343	0.054	2.907	0.159
taobao.com	14,474	0.070	1.100	0.077
youtube.com	16,805	0.081	0.907	0.074
hku.hk	14,112	0.068	0.447	0.031
hkst.com	3,815	0.018	1.585	0.029
google.com	42,506	0.205	0.100	0.020
hsbc.com.hk	859	0.004	0.907	0.004

Table 6 shows the same trend as TFIDF-PVN (Table 4) in that TFIDF-PVT enhances the domain weight due to a high IDF. Specifically, TFIDF-PVT assigns a high weight to a domain when it consumes a lot of browsing time on a few computers (high discriminating power).

A low weight is assigned when the page view duration for the domain was short or the domain was browsed by many computers (low discriminating power). The lowest weight is assigned when a domain was browsed by all the computers.

3.4 Web User Identification

The cosine similarity measure is commonly used to assess the similarity of two documents. We use this measure to assess the similarity of two browsing histories:

$$Similarity_{\cos} = \frac{c_1 \cdot c_2}{||c_1||||c_2||} = \frac{\sum_{i=1}^{N} c_{1,i} \times c_{2,i}}{\sqrt{\sum_{i=1}^{N} (c_{1,i})^2} \times \sqrt{\sum_{i=1}^{N} (c_{2,i})^2}}. \quad (5)$$

In the cosine similarity calculation, $c_1 \cdot c_2$ denotes the intersection of computers c_1 and c_2, and $||c_i||$ is the norm of vector c_i. Candidate computers that share domains with the top N domains of the target computer have higher similarity scores, while candidate computers that share no domains are assigned similarity scores of zero. The candidate computer with the highest similarity score has the highest probability that it was used by same web user as the target computer.

4. Experimental Design and Results

The web user profiling model was tested using 34 participants from the University of Hong Kong. The participants, all of whom were students between the ages of 20 to 31, provided their personal computers for model testing. All the computers had been in use for at least two months.

A total of 51 computers were collected from the 34 participants; seventeen of the participants had two computers. Thirty four computers, one from each participant, were assigned to Group I (i.e., computers that can be used as the target computer or candidate computers). The seventeen remaining computers were placed in Group II (i.e., computers that can only be used as candidate computers).

Browsing history records from July 1, 2013 through August 9, 2013 (40 days) were extracted from each computer. The time settings and time change logs were reviewed before data extraction. Table 7 presents the browsing history statistics of the collection of 51 computers.

Table 7. Browsing history statistics.

July 1, 2013 to August 9, 2013 (40 days)	
Number of computers	51
Number of sessions	36,801
Total page views	131,709
Total page view time (seconds)	8,822,703
Mean sessions per computer	721.588
Mean page views per computer	2,582.529
Mean page view time per computer (seconds)	172,994.177
Mean extracted domain per computer	63.980
Mean page views per session	3.579
Mean page view time per session (seconds)	239.741
Mean page view time (seconds)	66.986
Mean sessions per domain	11.278
Mean page views per domain	40.364
Mean page view time per domain (seconds)	2,703.862

4.1 Evaluation Metric

The performance of the profiling model was evaluated using an accuracy metric, which is defined as the proportion of the correctly identified examples out of all the examples to be identified. The accuracy was calculated using the equation:

$$accuracy = \frac{\sum_j^J tp_j}{K} \tag{6}$$

where tp_j is the number of true positives for participant j chosen in Group I, J is the number of different participants chosen in Group I and K is the total number of computers chosen in Group I.

4.2 Experimental Design

Two evaluations of the web user profile model were performed. The first evaluation compared the impact of TF-PVN, TFIDF-PVN, TF-PVT and TFIDF-PVT. The second examined the impact of the size of the profile group (M) on model accuracy.

The target and candidate computers were selected from the pool of $M = 51$ computers. Computers from Group I were successively selected as the target. Three samplings of candidate computers were used ($M = 15, 33, 51$). The samplings included ten computers from Group I and five computers from Group II ($M = 15$); 22 computers from Group I and eleven computers from Group II ($M = 33$); and all the computers from

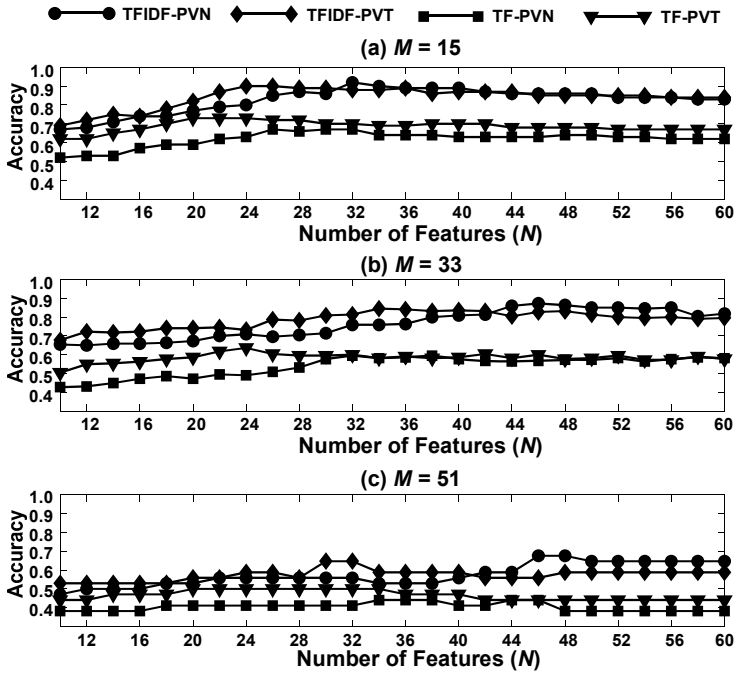

Figure 3. Influence of feature number for the four weighting models.

the two groups ($M = 51$). When $M < 51$, ten repeated experiments were performed for each M and the average accuracy was computed.

After the weight of each domain was computed for the target computer, the top N domains were chosen. When $N < 10$, the possibility existed that these domains could not be found on any of the candidate computers if the domains were too special. Therefore, we decided to choose more than ten domains on the target computer as features.

4.3 Feature Influence

Since the average number of extracted domains per computer was 63, we tuned the parameter N from 10 to 60 to observe the influence on web user identification accuracy for the four weighting models: TF-PVN, TFIDF-PVN, TF-PVT and TFIDF-PVT.

Figure 3 presents the results obtained for the three computer group profile sizes ($M = 15, 33, 51$). For all four weighting models, when user identification was performed on a small group ($M = 15$), the performance improves and then flattens as N increases. In the case of the TF-PVN model, the best user identification result occurred for $N = 26$ (67%). For TFIDF-PVN, the best user identification result was 92%

Table 8. Web user identification accuracy for the four weighting models.

M		TFIDF-PVN	TFIDF-PVT	TF-PVN	TF-PVT
15	Min	0.670 (N=10)	0.690 (N=10)	0.520 (N=10)	0.62 (N=10)
	Max	0.920 (N=32)	0.900 (N=24)	0.670 (N=26)	0.730 (N=20)
33	Min	0.650 (N=12)	0.677 (N=10)	0.427 (N=10)	0.505 (N=10)
	Max	0.873 (N=46)	0.845 (N=34)	0.595 (N=32)	0.636 (N=24)
51	Min	0.471 (N=10)	0.529 (N=10)	0.382 (N=10)	0.441 (N=10)
	Max	0.676 (N=46)	0.647 (N=30)	0.441 (N=34)	0.500 (N=20)

when $N = 32$, which indicates that TFIDF is more effective than TF. The best accuracy for TFIDF-PVT was 90% when $N = 24$, which is a slightly lower than TFIDF-PVN, but this result was obtained with fewer domains. Also, TF-PVT yielded better results (73%, $N = 20$) than TF-PVN, but worse results than TFIDF-PVT.

The trends for each feature remain consistent as M increases. However, Table 8, which presents the complete results, shows that better performance was obtained for smaller computer profile groups.

4.4 Profile Group Size Influence

Figure 4 presents the analysis of the influence of the computer profile group size M. For all the features, as the size of computer profile group increases, the web user identification accuracy gradually decreases. Also, when the number of domains (N) is 20 and 30, the performance relation TFIDF-PVT > TFIDF-PVN > TF-PVT > TF-PVN holds for all M. However, when $N = 40$, TFIDF-PVN and TFIDF-PVT have about the same accuracy. When $N = 50$, TFIDF-PVN has better performance than TFIDF-PVT for all M, which implies that if high accuracy is required, TFIDF-PVN should be used, but more domains would be required; otherwise TFIDF-PVT is a good choice.

5. Conclusions

The model proposed for web user profiling and identification is based on the web browsing behavior of users. Experiments involving 51 computers show that the profiling model can be applied to identify web users. Evaluation of the performance of the four weighting models, TF-PVN, TFIDF-PVN, TF-PVT and TFIDF-PVT, demonstrates that TFIDF-PVN is the most accurate at identifying web users.

Our future research will attempt to combine PVN and PVT to create a more accurate identification model. Additionally, the dataset will be

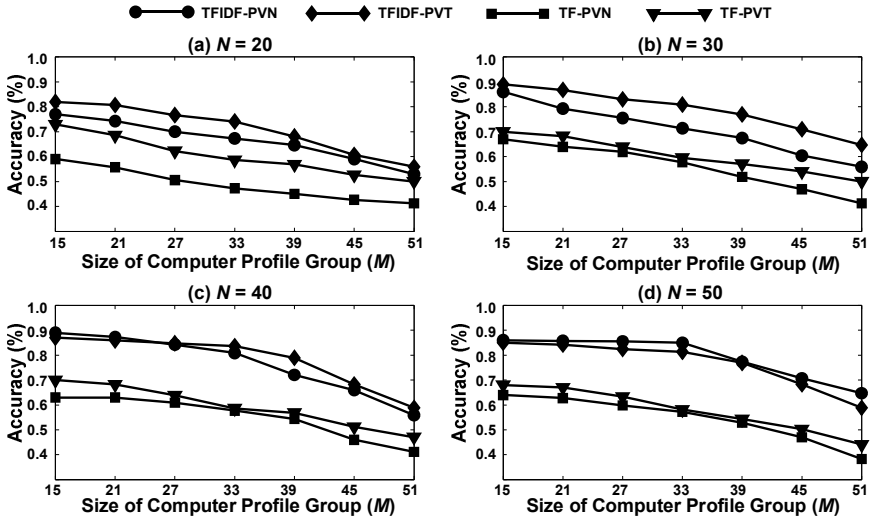

Figure 4. Influence of computer profile group size using the four weighting models.

expanded by collecting computers from a broader population in a real-world scenario. Our research will also focus on more complex scenarios, such as identifying all the users who have used a public computer.

Acknowledgement

This research was supported by the University of Hong Kong and by a Conference Grant.

References

[1] W. Bhukya and S. Banothu, Investigative behavior profiling with one class SVM for computer forensics, *Proceedings of the Fifth International Conference on Multi-Disciplinary Trends in Artificial Intelligence*, pp. 373–383, 2011.

[2] R. Bucklin and C. Sismeiro, A model of website browsing behavior estimated on clickstream data, *Journal of Marketing Research*, vol. 40(3), pp. 249–267, 2003.

[3] C. Colombini and A. Colella, Digital profiling: A computer forensics approach, *Proceedings of the IFIP WG 8.4/8.9 International Cross Domain Conference on Availability, Reliability and Security for Business, Enterprise and Health Information Systems*, pp. 330–343, 2011.

[4] N. Fathy, N. Badr, M. Hashem and T. Gharib, Enhancing web search with semantic identification of user preferences, *International Journal of Computer Science Issues*, vol. 8(6), pp. 62–69, 2011.

[5] M. Forte, C. Hummel, N. Morris, E. Pratsch, R. Shi, J. Bao and P. Beling, Learning human behavioral profiles in a cyber environment, *Proceedings of the IEEE Systems and Information Engineering Design Symposium*, pp. 181–186, 2010.

[6] M. Grcar, D. Mladenic and M. Grobelnik, User profiling for the web, *Computer Science and Information Systems*, vol. 3(2), pp. 1–29, 2006.

[7] J. Hu, H. Zeng, H. Li, C. Niu and Z. Chen, Demographic prediction based on user's browsing behavior, *Proceedings of the Sixteenth International Conference on the World Wide Web*, pp. 151–160, 2007.

[8] K. Jones and R. Belani, Web Browser Forensics, Part 1, Symantec, Mountain View, California (`www.symantec.com/connect/articles/web-browser-forensics-part-1`), 2010.

[9] C. Manning, P. Raghavan and H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, United Kingdom, 2008.

[10] S. McKinney and D. Reeves, User identification via process profiling: Extended abstract, *Proceedings of the Fifth Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies*, art. no. 51, 2009.

[11] J. Oh, N. Son, S. Lee and K. Lee, A study for classification of web browser logs and timeline visualization, *Proceedings of the Thirteenth International Workshop on Information Security Applications*, pp. 192–207, 2012.

[12] N. Sofer, IE History View v1.70 – View Visited Web Sites of Internet Explorer, NirSoft (`www.nirsoft.net/utils/iehv.html`), 2011.

[13] N. Sofer, Chrome History View v1.16, NirSoft (`www.nirsoft.net/utils/chrome_history_view.html`), 2012.

[14] N. Sofer, Mozilla History View v1.52 – Mozilla/Firefox Browsers History Viewer, NirSoft (`www.nirsoft.net/utils/mozilla_history_view.html`), 2013.

[15] Wikipedia, List of web browsers (`en.wikipedia.org/wiki/List_of_web_browsers`), 2014.