http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (4): 746-753, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.746.753

Web Access Patterns Mining for Individuals with Timing and Link Sequence

[1,2]Yonglong Ge, [1,2]Chungang Yan, [1,2]Zhijun Ding and [1,2]Wangyang Yu
[1]Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China
[2]Key Laborotory of Embedded System and Service Computing, Ministry of Education,
Tongji University, Shanghai, 201804, China

Abstract: The issue of user identity trustiness has become more and more serious in e-commerce recently. Researches on evaluating user identity trustiness by user behaviors have received great attentions recently. In this work, proposed a method of constructing the individual access patterns based on the relationship among web pages they have visited by web log mining. The web pages were clustered based on the contents and timing frequent sequence. Then, Individual Access Pattern (IAP) was proposed to describe the patterns of individual access behaviors. IAP stored highly compressed, critical information for individual access behavior and habits in a fine-grained way. IAP also could be easily transformed to Markov Model, which could be used to compute the confidence of various users. Thus, IAP could be applied to e-commerce transaction for identifying users in real time. The average accuracy of individual pattern recognition can reach 88.5% after the user browse 5 webpages. FAR (false accept rate) or FRR (false reject rate) can be lower than 10% by adjusting different strategies. The experimental results showed that IAP can reflect the differences of various individual access patterns both from the timing sequence and the hyperlink relationship precisely.

Key words: User identity trustiness, individual access pattern, timing sequence, link sequence, Markov model

INTRODUCTION

With the rapid development of the internet, e-commerce has become one of the most important modes of making trading. However, recognizing the real identity of the consumers accurately is the key security issue of e-commerce. Due to the prevalence of fishing websites, traditional approaches of identity authentication, such as digital certificate, cannot solve the problem of user identity trustiness totally. Thus, researches on evaluating user identity trustiness by individual access patterns get great attention recently.

Regarding the difference of each user's age, hobbies, education degree and surfing habits, each user's focus of the Internet, including the websites and the webpages he/she browses, is different. i.e., individual Web access pattern is unique. Therefore, evaluating user identity trustiness can help recognize the user identity effectively according to individual access behavior. Obtaining the behavior pattern, which can describe the user's specific behaviors, the first priority to recognize user identity.

Web logs record all of user browsing behaviors and can mine the user access patterns, which are hidden within the logs by Web log mining. For example, the authors proposed efficient incremental and interactive data mining algorithms to discover web traversal patterns, which could optimize web structure persistently (Lee and Yen, 2008). The study provided the method to construct the TSPs based on the linkage relationship among webpage access frequencies (Wang and Lee, 2011). Also, the author proposed to a new similarity computational method to cluster web users (Ding and Ma, 2009). Additionally, ant colony algorithm is employed for clustering individuals with similar access pattern (Loyola *et al*., 2012). This study proposed a frequent and utility-based web traversal pattern mining algorithms, which considered the browsing time that a user spend on a web page (Thilagu and Nadarajan, 2012).

And other works (Motegi *et al*., 2012; Shan and Sun, 2011) are targeting at a group of people and used for personal recommendation and website structure optimization. Zhou *et al*. (2005) proposed to the approach to find users' interest based on manual classifying of webpages and dividing one day into several periods for personal recommendation.

However, these methods mainly focus on personal recommendation or website structure optimization, not suitable for individual identity recognition and cannot apply to real-time identify individual identity. Here the proposed method performs clustering based on the web page content accessed by individuals. According to the relationship among timing and link sequence, individual

Corresponding Author: Chungang Yan, Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China

access patterns can be mined, to reflect the habit of individual focuses changing with timing and links. Therefore, proposed Individual Access Pattern (IAP) to describe the individual access pattern precisely. This model can reflect the web access differences among various individuals in detail and comprehensively. It is also provided the method to transform the IAP to corresponding Markov model based on the attributes of number of webpages, the webpages linkage. According to the IAP, this can identify a user's identity in a few webpages accessed, which achieves high real-time and recognition efficiency.

MATERIALS AND METHODS

The proposed method is described in detail within this section. Only certain properties of web page are concerned, such as URL, domain, content, referer and request time. The webpage is denoted by p and the URL, content, referer and request time stamp are represented by p. Domain, p. Content, p. Referer, p. Timestamp, respectively.

Related concepts and definition: Below are some related concepts and definitions for the proposed method.

Definition 1: Class of webpages (P). A set consists of timing series pages, which have similar domain and content, i.e., $P = \{p_i|p_i.\ Domain = p_m.\ Domain \wedge p_i.\ Content = p_m\ Content,\ m = i = n\}$. Where p. Domain = p_m. Domain, p. Content = p_m. Content, p. Name = (p. Domain, p. Content), p. RefererSet = $\{p_i.\ Referer|p_i \in P\}$, P. Timestamp = p_m. Timestamp, p. Count = n-m+1 represent the URL, content, referer set, start time stamp and number of webpages, respectively.

Definition 2: Individual Access Sequence (IAS). The sequence of webpages that individuals have accessed within a span is $S = \{<P_1, P_2, P_3,\ldots, P_n>|P_n.\ Timestamp\text{-}P_i.\ Timestamp = TS\}$, where TS represents a time span. The set formed by these individual access sequence is called the individual access sequence database, where $IASD = \{S_1, S_2, \ldots, S_2\}$.

Definition 3: Frequent webpages class (FWC). If the support in IASD is larger than the threshold TH, for a webpage class P, P is defined as the frequent webpage class, where:

$$sup_{IASD}(P) = \frac{|\{S_i \mid P \in S_i, S_i \in UASD\}|}{m} > TH$$

Definition 4: Link jump. Within the webpage classes P_1, P_2, if $\exists p_1 \in P_1$, $p_2 \in P_2$ and p_2. Referer = p_1, which means $P_2 \in P_1$. RefererSet, then between P_1, P_2 there is the link jumping from P_1 to P_2, which is noted as l.

Definition 5: Timing sequence jump. Within the webpage classes P_1, P_2, if P_1. Timestamp<P_2. Timestamp and $P_2 \notin P_1$. RefererSet, then there is the timing sequence jump from P_1 to P_2, which is noted as t.

Definition 6: Individual user access pattern (IAP). Five tuple IAP = (Σ, L, T, K, W) called individual user access pattern graph, if f:

- $L \cup T \neq \Phi$, $L \cap T = \Phi$
- $L \cup T \subseteq \Sigma \times \Sigma$
- K: P→Q, called page class scale function:

$$w.r.t.P \in S, K(P) = \frac{S_{P_i.Name=P.Name} P_i.Count}{|\{P_i \mid P_i.Name = P.Name\}|}$$

- W: L∪T→Q is the weighted function:

$$w.r.t.jump\ P_1\ to\ P_2,$$
$$j \in L \cup T, W(j) = S_{p \in P_2\ and p.Referer \in P_1} 1$$

Here, Σ represents the set of webpage classes; L corresponds to the set of link jumps; T stands for the set of timing link jumps; the function of webpages scale K(P) represents the expectation of the number of webpages within P, which reflects the attention that users paid for P; the weighted function W describes the relationshiop among webpages; and Q is the set of rational number.

Mining for individual access pattern: The traditional access pattern mining can be catergorized into two classes. One is represented by Apriori (Agrawal and Srikant, 1994) and PrefixSpan (Pei *et al*., 2004), which generate the candidate sets. The other does not generate the candidate set, like WAP algorithm (Pei *et al*., 2000). The first class of algorithms needs to generate large number of candidate sets during mining and needs to parse the candidate sets to calculate the supportive degree. Therefore, such algorithms are not very efficient. The other class of algorithms employs the dictionary tree structure to store the access sequence, which improves the efficiency of storage and mining.

Based on the advantages of the dictionary tree for storage and mining, also apply this structure to store the access frequency sequence. However, as the information of Internet updates very fast, the webpages accessed by individuals are almost unique at each time except some main web pages. So the traditional WAP algorithm, which constructs a WAP-tree based on the access pattern of single webpage for individuals and stores each single webpage as a node to store the access sequence, is not suitable in this case.

Thus, in order to improve the WAP-tree (Pei *et al*., 2000) storage structure, it is proposed to use a new timing sequence WAP-tree. According to the characters of the individual access activities, the webpages accessed by them seem quite unique, the content within these webpages are highly related. Construct clusters based on the webpages accessed by individuals and make the class of webpages as the node of WAP-tree. As the webpages accessed do not have link jumping relationship, the timing sequence stored in WAP-tree is timing jumping rather than the link jumping relationship.

Considering that link jumping among different webpages can also reflect the access habit of individuals. However, after clustering this link jumping relationship among the webpages is hidden in webpage classes. Therefore, the webpage link jumping mining can reflect the individual access pattern from both the timing sequence and link jumping relationship.

The ideal of this access pattern mining algorithm is that first preprocess the web logs of individuals and retrieve useful information to cluster the webpages based on the content. Then construct the timing sequence WAP-tree to analysis and mine out the timing frequency sequence for individuals. Finally, mine the frequent link jumping relationships of access sequence by individuals based on their access frequency.

Data preprocessing: Data within this study is obtained by the client web log collector, which can capture the HTTP packages of the client. Though the collector can generate an item according to every HTTP requirement, the items are all the original data, where most useful information is hidden. The HTTP requirement webpages are distributed in other failure post, pictures and video resources. Therefore, before do the access sequence analysis, it is needed to collect all these records distributed in the logs for the same access sequence. The preprocessing is illustrated as below:

- **Prune redundant logs:** Sort the webpage request in the access web logs by timestamps. Filter out the picture, sound and video request, which can be achieved by checking the extension of file name. Check if the status of HTTP is 200 to indicate success, which can be retrieved resource of webpage from
- **Conversation recognition:** According to the definition 1 and 2, form the set of sessions by storing webpages based on the Timestamp order
- **Webpage clustering:** According to definition 1, cluster the webpages within the same session based on the attributes of Domain and Content of the webpages and form the class of webpages

After the data preprocessing, the individual access sequence can be obtained for individual, as shown in Table 1. The sequence order is represented by number from 1 to 5, respectively, which means that there are 5 access sequences. In the access sequence, the classes of webpages are represented from a to z. Assume that the classes of webpages of these five access records are {a, b, c, d, e, f, g}.

Access timing frequency mining: The process of mining access timing frequency sequence is the one constructing the timing sequence for WAP-tree. The attributes contained within each node of the timing sequence WAP-tree are label, count, weight, parent and children. The meanings of these attributes are shown in the Table 2.

From Table 1, pointer tJump represents both the relationship for child node and the frequency timing jumping relationship. Weight is one of the most important criterions to indicate the access pattern for individuals, which can be calculated from the expectation of numbers of the webpages within the webpage class. I Jump is the preserved word section to store the access frequency link jumping which is mined later. According to the approach of constructing WAP-tree and computing the weight of every node, the timing sequence WAP-tree can be obtained.

Back to the example of data preprocessing, take the access sequence in Table 1 into account, suppose that the threshold of supportive factor is TH = 0.8, which means that frequent access webpages should appear at least in four of the access records. Thus, through parsing the access sequences, can obtain the frequency webpages class {a, b, c, d}. Additionally, it can obtain the access frequency sequence, which is shown as in Table 3. The webpages are represented with tuple

Table 1: Five access sequences of a user. Each lower case letters represents a webpage and various individual should have different corresponding access sequences

Sequence order	Access sequence
1	a b d a f c
2	a d c e
3	a b g d c
4	a b g d c b e
5	a b g a f c

*a,b,..., g: Each represents a webpage

Table 2: Each node has the following attributes to represent webpage class properties so as to describe individual access pattern

Attributes	Meaning
Label	Label the class of webpages
Count	Supportive factor for node
Weight	Weight of the node, expectation of the No. of webpages
Parent	Pointer of parents
tJump	Pointer of child, which is the timing frequency jumping
lJump	Reserved, frequency link jumping

Table 3: Show the access frequency for each webpage class indicating the specific individual pattern according to Table 1. Each tuple consists of web class name and the No. of webpages inside class

Order	Access frequency sequence
1	(a, 1) (b, 7) (d, 1) (a, 1) (c, 7)
2	(a, 1) (d, 1) (c, 9)
3	(a, 1) (b, 2) (d, 1) (c, 12)
4	(a, 1) (b, 3) (d, 1) (c, 6) (b, 9)
5	(a, 1) (b, 8) (a, 2) (c, 11)

*tupe (P.Name, P.Count) represent webpage classes, P.Name indicates the name of certain webpage class, P.Count shows the number of webpages in the class

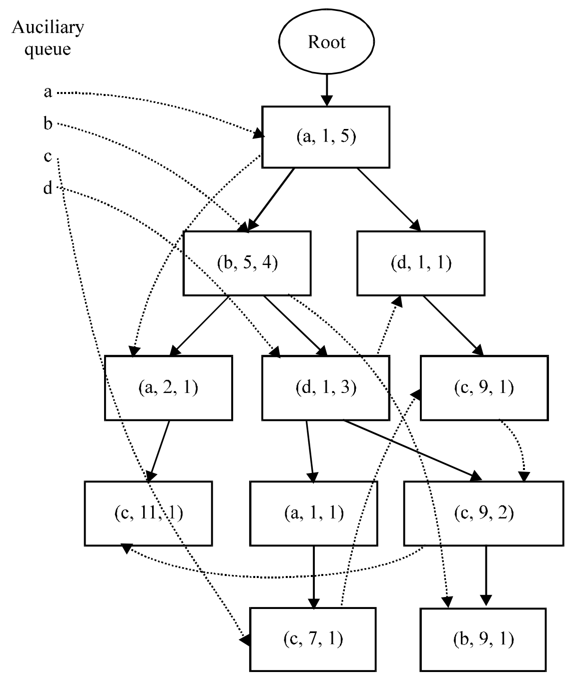

Fig. 1: The timing sequence WAP-tree based on data in Table 3

(P.Name, P.Count). According to Table 3, it can construct the corresponding timing sequence WAP-tree, which is shown in Fig. 1.

Explanotion of Fig. 1; Every node is represented by a 3-triple (Node.Label, Node.Weight, Node.Count). Each node represents the access state of users. Within each access state, Node.Label denotes the current webpage class, Node.Weight and Node.Count are the properties of the corresponding webpage class. The directed arc shows the timing sequence of states. Auxiliary queue stores the same webpage class of Node.Label for the convenience of computation. Construction of the timing sequence WAP-tree and the process is shown as below. Initialize the timing sequence WAP-tree and construct a new node (a, 1, 1) as the child node of the root and build up the linkage relationship between them. Then, insert the sub sequence bdac in order to form the branch starting with node a, such as (a, 1, 1)→(b, 7, 1)→(d, 1, 1)→(a, 1, 1)→(c, 7, 1). Insert the second sequence adc from the root. When the root already has a sub node a, only plus 1 for a.Count instead of creating a new node a. Recalculate the expectation with a.Weight, resulting to (a, 1, 2). Loop until all the access frequency forms the timing sequence WAP-tree.

Frequent web link jumping mining: After finishing constructing the WAP-tree, can obtain the users' timing frequency access sequence, from which it can be discovered the interested webpages of the client in order. Obtain the accurate interests and access patterns related to the timing sequence of clients. However, it is still unclear on the webpages jumping patterns. Thus, need to mine the high frequent link jumping in the timing frequency sequence in depth to get the access patters of the clients.

Mining the frequent link jumping is to find the referer set within the frequent accessed class of webpages FPC and project each of the linked webpage class to FPC in order to obtain the frequent referer set of webpage class. Afterwards, add frequent link jumping I Jump to each node and form the individual access pattern. The mining frequency link jumping algorithm is as follow.

Algorithm 1: Mining frequency link jumping:

Input: Timing sequence WAP-tree (TWT)
Output: Individual access pattern (IAP)

- For each class of webpages P in the set of frequent access webpages FPC:
  - For each webpage p in P, find the class of webpages P', that webpage p.Referer belongs to
  - If P' does not belong to P.RefererSet, add P' to P.RefererSet
- Parse the TWT in depth first method for each node
  - Delete the class of webpages in Node. label.RefererSet, that does not belong to FPC
  - Parse Node.label.RefererSet and for each class of referer P:

    - Node'←Node;
    - While Node'.label is not equal to P and Node' is root, do Node'←Node'.parent
    - If Node' is not root then insert Node into Node'.l Jump. Set the number of links that has jumped from one to another

Else check the auxiliary queue where Node belongs to and assign the next element to the Node and go back to the previous step:

Table 4: List an example of access frequency sequence for user in Table 1. Each tuple consists of the web class name and the set of webpages of before that class

Order	User access frequency sequence
1	(a, Ø) (b, Ø) (d, {a}) (a, {a}) (c, {d})
2	(a, Ø) (d, {a}) (c, {d})
3	(a, Ø) (b, Ø) (d, {a}) (c, {d})
4	(a, Ø) (b, Ø) (d, {a}) (c, {d}) (b, Ø)
5	(a, Ø) (b, Ø) (a, Ø) (c, {d})

*2-tuple (P.Name, P.RefererSet): represents webpage classes, P.Name: represents the name of the webpage class, P.RefererSet: indicates the set of pre-webpage class

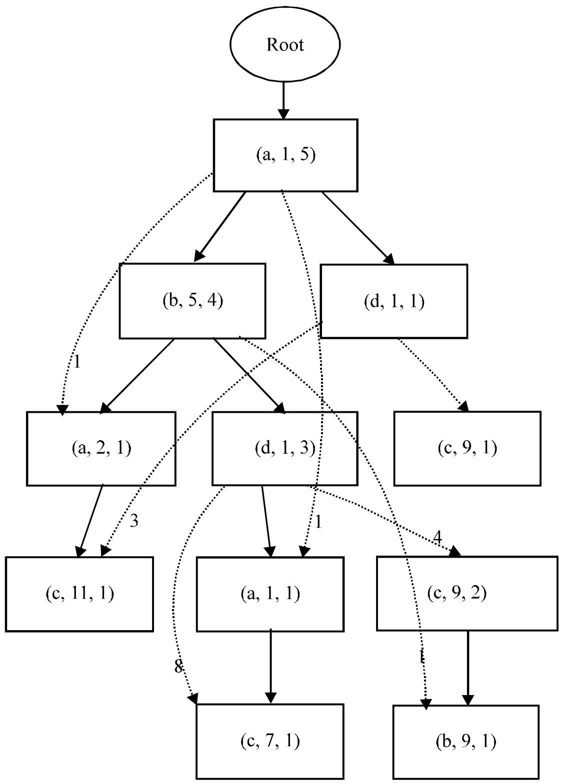

Fig. 2: Individual access pattern derived by algorithm 1 and data from Table 3 and 4

- Adjust TWT to form IAP
- Return IAP

Back to the example, through algorithm 1, the frequent referred set RefererSet can be obtained from each class of frequent webpages. The detailed mining results are shown in Table 4, where the class of webpages is represented as a 2-tuple (P.Name, P.RefererSet). According to Algorithm 1, finally obtain the individual access pattern graph as Fig. 2 shows.

Explanation of Fig. 2; in the IAPG, every node is represented by a 3-triple (Node.Label, Node.Weight, Node.Count). The circle represents the start state, the square frame represents other access states. Within each access state, Node.Label shows the current webpage class and Node.Weight, Node.Count is the property of the corresponding webpage class. IAPG contains the transition of frequency timing as well as the frequency link, which are represented as the solid and dash line, respectively. The weights of arcs describe the number of jumping times among classes of webpages.

Table 5: The web access sequences three users at one time. Each lower case letter represents the class name of the specific webpage

User ID	Access sequence
1	a b b b a f f d c c
2	a g g b g g g i i i
3	m n n m e a h j k

*a,b,..., g: Each represents a webpage

EXPERIMENT RESULTS

Experimental environment: Intel(R) Core(TM) i3 CPU (3.2GHz), 4GB RAM, Windows 8 operating system, C# development language, Visual Studio 2010 IDE. The data set: 31 thousands of web logs from client collected from 10 clients within 4 months. Each people visit about 30-40 webpages per day. The related information collected includes the webpage URL, the link time and linked in webpage URL.

In the experiment, use Markov model to verify the validity of IAP. According to the webpage content attribute and the jumping weights of the individual access pattern, transform the individual access pattern into Markov model.

Corresponding to individual access pattern graph derived from section 3, transfer Fig. 2 to the corresponding Markov model as shown in Fig. 3.

Firstly, choose three users' access sequences for one day randomly, where user 1 is the owner of the model. Select the top 10 webpages and they are shown in Table 5.

As shown in Table 5, each webpage in the access sequence is represented by the name of the webpage class it belongs to. Then, according to the webpage the user access, compute the current confidence of the user, using the forward algorithm, as follow:

$$C(n)=\begin{cases} C_0, n<1 \\ C(n-1)\times A(n), n\geq 1\wedge A(n)\neq 0 \\ C(n-1)\times ?, n\geq 1\wedge A(n)=0 \end{cases}$$

where C_0 is the initial confidence, here C_0 = 1, C(n) represents current confidence of the user, λ is the smoothing coefficient, A(n) is the current transition probability to prevent the problem of zero probability. Then, the probability of user access sequence S is as follows:

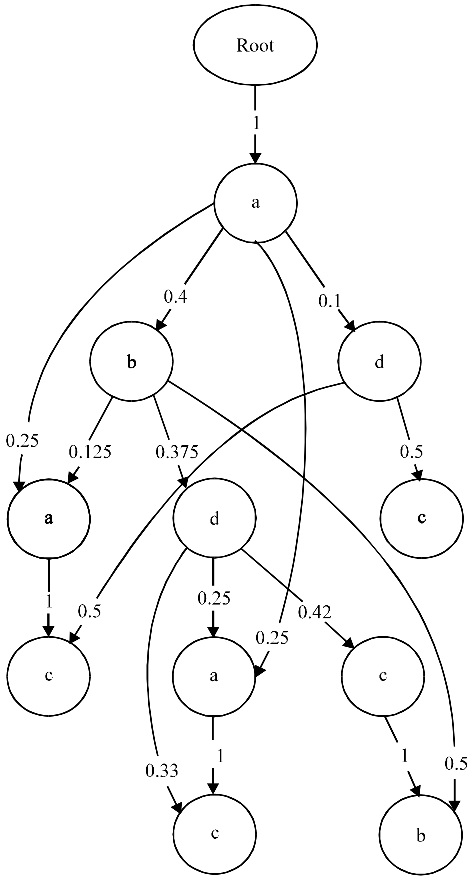

Fig. 3: Individual access behavior model derived from individual access pattern graph based on Fig. 2. Each node represents the user access state, which is the webpages that current user access frequently; and the weighted directed arc indicate the probability of state transitions among webpages

$$Pr[S] = C(n)$$

Now choose the first quartile Q to compute the threshold as follows:

$$TH = \varepsilon \times Q^n$$

where, ϵ is the weighting factor, TH is s the threshold value of probability used to decide that the acceptance of the user access sequence S is confirmed if following expression is true:

$$Pr[S]>TH$$

The weighting factor ϵ can be specified with respect to different level of security strength.

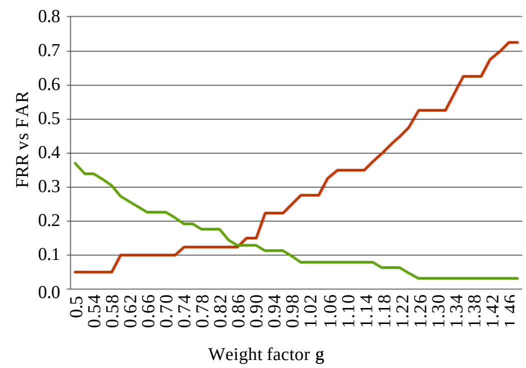

Fig. 4: The changing trend for FAR and FRR based on ε. The data set is collected from 10 clients within 4 months. Each people visited about 30-40 webpages per day. Minimum length of the access sequence is 5, average length = 10, EER = 11.5% (FRR: False reject rate, FAR: False accept rate, EER: Equal error rate)

Table 6: Analyzed the actual results of identification for various individuals according to Table 5. The access pattern of user 1 is used to validate the result

User ID	Pr[S]	TH	Result
1	1.25e-6	9.5e-7	True
2	4e-8	9.5e-7	False
3	1e-10	9.5e-7	False

*Pr[S]: User trustiness calculated based on user access sequence, TH: Threshold of user trustiness, Result: User recognition result

According to the Table 5, assume that λ is 0.1 and ε is 1.0. And get the result of identification shown in Table 6.

Secondly, evaluate the value of threshold weighting factor ϵ between 0.5 and 1.5 with interval of 0.02. The Equal Error Rate (ERR) with the minimum target length of 5 and the reference sample size of 10 could be 11.5% (Fig. 4).

From Table 6 and Fig. 4, it shows that the IAP can distinguish its owner to other users with high detection rate. And the FAR (false accept rate) or FRR (false reject rate) can be lower than 10% by adjusting different strategies. According to the second experiment, identify a user in 5.3 pages average. It shows that recognizing user identity using the IAP performs high real-time. However, the EER is a little higher. There are some factors could affect the equal error rate. Only focus on the frequent web page classes during the process of mining, which may cause the loss of information.

Comparison and analysis of the results: According to Table 7, conclusion can be made that:

Table 7: User behavior considered by related methods and the proposed method are listed. Each kind of behavior represents a component of behavior pattern. The proposed method considers the cross-site browse behavior and cluster the webpages based on content

Variables	Lee and Yen (2008)		Wang and Lee (2011)	Zhou *et al*. (2005)	Proposed method
Webpage class	No	No	No	Yes	Yes
Behavior pattern and habits	Yes	Yes	Yes	Yes	Yes
Attribute of time	No	Yes	No	Yes	Yes
Cross-site browse behavior	No	No	No	Yes	Yes
Most of user's access pattern	Yes	Yes	Yes	No	No
Individual identification	No	No	No	Yes	Yes
Real time	Yes	No	No	No	Yes

- Lee and Yen (2008) and Wang and Lee, (2011): The common character of these three methods is that they both build up the model based on a single website or webpages. Therefore, they can precisely describe the structure of websites and this is suitable for optimizing website structure and user navigation. The proposed method builds up the cross-site model based on the webpage classes, which are clustered according to the content. This way, the proposed method is able to describe the individual user behavior precisely based on their distinct habit. Compared with the existing methods, the proposed method can describe the information of the habit, access priority based on individual habits and website preference. This makes the model more suitable for user identification
- Zhou *et al*. (2005) applies manual classification to webpages. They also build up the model taking into account the webpage classes and time characters to better model the individual habit or preference. However, the proposed method focus on the time sequence of the webpage classes accessed by users, rather than the on-line time interval as (Zhou *et al*., 2005). This makes the method more proper for individual identification. Besides, the automatic clustering strategy used by the proposed method makes it applicable for networks with large amount of users involved

The proposed consider the most kinds of user behavior and the content-based clustering strategy can efficiently reflect online user behavior, such as searching, shopping and checking out. Therefore, this proposed method is more suitable for user identification and risk management for e-commerce in practical. As time goes on, user may generate some new concerns, which results in accuracy decline of the model. These factors will also be considered in future work.

CONCLUSION

This study proposed a novel method for mining individual access patterns based on the web log analysis. This method takes all the webpages which users surfing Internet and mines patterns based on the interest of users changing with timing features and link sequences. Thus, the IAP can reflect the individual surfing hobby comprehensively. For the validation of this method, this study shows the method for user identity recognition with experiments. By the experiments results, it is persuasive that the individual access patters discovered in this work can reflect the different access habits of users better and provides the support for user identity trustiness recognition.

For the user behavior mining in this work, only consider the timing sequence and link jumping features among web classed. In the future work, would include the mouse and keyboard behaviors of users and model the user behaviors in detail. In addition, also solve and apply this work to the real world application to efficiently make use of the individual access pattern to achieve user identity recognition and benefit more related areas.

ACKNOWLEDGMENTS

This study was supported in part by the National Natural Science Foundation of China under Grants 91218301 and 61173016, by the National Basic Research Program of China (973 Program) under Grants 2010CB328101, by the Shanghai Science and Technology Research Plan of China under Grant 10DJ1400300 and by the International Science and Technology Cooperation Program of China (2013DFM10100).

REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago, pp: 478-499.

Ding, X.M. and X.Y. Ma, 2009. A web users clustering model based on users' browsing path. Proceedings of the International Conference on Computational Intelligence and Software Engineering, December 11-13, 2009, Wuhan, pp: 1-4.

Lee, Y.S. and S.J. Yen, 2008. Incremental and interactive mining of web traversal patterns. Inform. Sci., 178: 287-306.

Loyola, P., P.E. Roma and J.D. Velasquez, 2012. Predicting web user behavior using learning-based ant colony optimization. Eng. Appl. Artif. Intell., 25: 889-897.

Motegi, M., Y. Ichikawa and T. Kobayashi, 2012. Identifying purchase-stage by applying pattern mining to web access logs. Proceedings of the International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, August 8-10, 2012, Kyoto, pp: 740-745.

Pei, J., J. Han, B. Mortazavi-Asl and H. Zhu, 2000. Mining access patterns efficiently from web logs. Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining: Current Issues and New Applications, April 18-20, 2000, Kyoto, Japan, pp: 396-407.

Pei, J., J. Han, B. Mortazavi-Asl, J. Wang and H. Pinto *et al*., 2004. Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Trans. Knowledge Data Eng., 16: 1424-1440.

Shan, X. and H. Sun, 2011. The research of web users? behavior mining based on association rules. Proceedings of the 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce, August 8-10, 2011, Deng Leng, pp: 7415-7418.

Thilagu, M. and R. Nadarajan, 2012. Efficiently mining of effective web traversal patterns with average utility. Proc. Technol., 6: 444-451.

Wang, Y.T. and A.J.T. Lee, 2011. Mining Web navigation patterns with a path traversal graph. Expert Syst. Appl., 38: 7112-7122.

Zhou, B., S.C. Hui and A.C.M. Fong, 2005. Discovering and visualizing Temporal-based web access behavior. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, September 19-22, 2005, Compiegne, France, pp: 297-300.