A Kind of Authentication Method based on User Web Browsing Features

Mengqing Ji, Peihai Zhao, Mimi Wang, Chungang Yan, Zhong Li, and Changjun Jiang

***Abstract*—Traditional authentication methods based on user browsing behaviors consider relatively one-snidely on user browsing habits. They mainly research on the relationships between the sequences of websites or contents without considering user habits comprehensively. So the accuracy when they distinguish different users' web browsing behaviors cannot ensure enough safety, which can be further optimized. This paper introduces a new method which studies from favorite websites, contents and periods of browsing time. It uses Apriori algorithm to mine user's frequent itemsets along with the text classification method and normal distribution to calculate access periods of time. Logic regression algorithm is applied onto user authentication. Experiment shows that detection rate can reach 92.7% while false alarm rate is 6.4%.**

***Index Terms*—Identity authentication, user behavior, web browsing features, frequent itemset.**

I. Introduction

With the development of IT technology, the number of people using online shopping is increasing now. According to the latest IResearch report: "2014 Annual Monitoring Report on China's Online Shopping Industry" [1], China has become the world's largest online retail market. Online shopping is gradually changing Chinese people's consumption habits. But at the same time, the number of fraud cases in online shopping is also growing. "2013 Internet Security Conference" shows that new online shopping frauds have become a primary security threatening to Internet users. In the case of frequent leakage of user password, traditional protecting methods has some weakness, making user authentication one of the key problems of online shopping security. This paper aims to settle the problem of traditional certification methods. The improvement is divided into three parts: replace one-time certification with sustainable certification; introduce a new mining method which analyze user's browsing habits; improve the method for higher safety.

The first part is about authentication type. Researches on user authentication can be divided into two types: one-time certification and sustainable certification [2]. The former can be classified into following methods: Traditional account, password authentication [3]-[5], Smart card-based authentication [6], the authentication based on biological and behavioral characteristics (e.g. fingerprint, users' habits of using the mouse [7], [8] and keyboard input [9], [10]). Since the last one cannot be easily imitated, it attracts more academic attention. Anyway, they are all one-time certification, which means the authentication is done in a moment. Once obtained, the authentication is valid forever. So this kind of method is unable to provide high security level for users. Considering of the fact that one-time certification has above weakness, this paper chooses sustainable certification. Sustainable certification is much more complicated, but can offer higher safety, since it does certification work continuously during user operation.

Secondly, this paper introduces a new mining method into identity certification scene. Currently, research based on user browsing behavior is mainly used in scenes like search engine optimization [11], [12] and web page recommendation [13]-[15]. In [11], in order to optimize search engine, a user browsing path is introduced in the form of access graph to determine the most frequent access sequences. Literature [15] uses preference factor for web recommendation based on the browsing patterns mined out by FP tree. Traditional mining method does not adapt to identity certification scene, since it is a must to do individual authentication for each user. So we need to study personal browsing habits and portray behavior patterns, which differs from traditional methods. According to users' historical browsing records, Literature [2] focuses on the relationship between websites and uses Apriori algorithm to find out potential links among websites. Then it makes identity authentication based on rules combined with Simple CART. Experiment shows that [2] can achieve a detection rate of 91.3% with false positive rate as 10%.

Finally, this paper discusses to the weakness of previous mining methods, and improves it for higher safety. Previous methods mine user browsing feature from the sequences of websites or contents without considering user habits just like [2]. Its accuracy can be further optimized. For instance, there may exists a sequence where some user browses the Internet: *Baidu.com*, *Ifeng.com*-> *Sina.com*, which cannot accurately reflect user's browsing behavior. This is because of the fact that a web site may cover following contents on entertainment, society, science and technology, etc., while each person browses different contents on the same website. For example, user *A* may has such browsing feature: (*Ifeng.com*, society)-> (*Sina.com*, society), while user *B* has a browsing feature like (*Baidu.com*, society), (*Ifeng.com*, entertainment)-> (*Sina.com*, military). Present methods for mining user browsing feature may generalize *A* and *B's* different browsing features into one feature, where the ability to distinguish the two users' features is lost. In addition, different users may view specific content at different time. So

Manuscript received February 15, 2017; revised April 5, 2017. This work was supported in part by Shanghai Science and Technology Innovation Action Plan Project No. 16511100900.

M. Q. Ji, P. H. Zhao, M. M. Wang, C. G. Yan, Z. Li and C. J. Jiang are with the Department of Computer Science and Technology, the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China, 200092 (e-mail: 329875938@qq.com, perry.zhao@foxmail.com, wangmimi2013@hotmail.com, chungangyan@tongji.edu.cn, 260050478@qq.com, cjjiang@tongji.edu.cn).

doi: 10.18178/ijmlc.2017.7.2.613

time is also a very important factor. Considering the aspects above, this paper improves a new mining method. It uses Apriori to mine out the feature of user web browsing habits based on frequent URLs, contents and time, and uses LR for classification and certification. Experiments show the good feasibility and accuracy of this authentication method.

II. Definitions

Definition 1 Visit record set **R** is a set of webpages browsed by users, denoted by users $R=\{r_1,r_2,r_3,....r_i\}$. For each element r_i in R, $\mathrm{r_i}=(url,title,timestamp)$, $r_i.url$ means the URL of website r_i, $r_i.title$ means title of r_i and $r_i.timestamp$ means it's timestamp.

Definition 2 Session **s** [16] is a set of webpages which the request time of adjacent visit record is less than 30min, denoted by

$$s=\{r_i \mid r_m.timestamp-r_{m-1}.timestamp<30\min, r_m\in R\}.$$

And the set of s is denoted as S.

Definition 3 Visit Record Set after Data Processing **R'** is a corresponding set of visit records set R, denoted by $R'=\{r_1^{'},r_2^{'},r_3^{'},....r_\mathrm{i}^{'}\}$. For r_i' in it, $\mathrm{r_i^{'}}=(\mathrm{domain},\mathrm{content},time)$, $r_i'.domain$ is the top domain of r_i , $r_i'.content$ is the classification of $r_i.title$, $r_i'.time$ is the relative time of r_i .

Definition 4 Session after Data Processing S^* is a corresponding set of s in which every visit record is after data processing, denoted by $S^*=\{r_1',r_2',....r_i'\}$. And the set of s^* is denoted as S^*.

Definition 5 Webpage Class **P** is a set of webpages with the same top domain and classification, denoted by

$$P=\{(r'_i.\mathrm{domain},r'_i.\mathrm{content})|r'_m.\mathrm{domain}=$$
$$r'_n.\mathrm{domainsa}\wedge r'_m.\mathrm{content}=r'_n.\mathrm{content}\}.$$

Definition 6 Frequent Access URLs FU_j is a collection contains 15 kinds of top domains which are most frequently accessed by user *j*, denoted by

$$FU_j=\{domain(r_i.url)\mid r_i\in R\wedge domain(r_i.url)\in top(15)\}.$$

Definition 7 Time Page Class **PT** is a set of webpages with the same classification of content, $PT=\{pt_1,pt_2,...,pt_\mathrm{j}\}$. For each element pt_j in PT, $\mathrm{pt_j}=\{(\mathrm{r'}_i.\mathrm{content},\mathrm{r'}_i.time)\mid \mathrm{r'}_m.content=\mathrm{r'}_n.content\}$.

Definition 8 Frequent Access Period of Time **FT** is a frequent period of time for the user to access each classification of content, $FT=\{ft_1,ft_2,...,ft_{12}\}$, where ft_j is user's frequent period of time for content *i*.

Definition 9 Support **fc** [17]. For transaction fc, X, Y in the dataset D, Support (fc) is defined in Function 1, where $\mathrm{count}(X\cup Y)$ is the number of transaction in D containing $X\cup Y$.

$$\mathrm{support}(X,Y)=\mathrm{count}(X\cup Y)/|D| \tag{1}$$

Definition 10 Frequent Itemsets **FC$_\mathbf{j}$** [17] is a set itemset of user j which meets δ , denoted by $FC_\mathrm{j}=\{fc_i \mid \sup port(fc_i)\geq\delta,1<i<m\}$, where δ is the set support threshold [18].

Definition 11 Feature Value Matrix **FV$_\mathbf{j}$**. For a session set after data processing which contains m sessions, each session has a feature value vector $fv_{\mathrm{ji}}=<length,pun,mrn,rpun,mrml,mral,mrms,mras$ $mtn,t\arg et>$ and the meaning of each attribute will be given later.

Definition 12 Set of Frequent Itemset's Length is a set of frequent itemset's length which contained in the session $s^*{}_{ji}$, denoted by $FC_\mathrm{i}.length=\{fc_{in}.length \mid fc_\mathrm{in}\subseteq s_{ji}^*\wedge fc_\mathrm{in}\in FC_\mathrm{i}\}$.

Definition 13 Set of Frequent Itemset's Support is a set of the frequent itemset's support contained in the session $s^*{}_{ji}$, $FC_\mathrm{i}.\mathrm{support}=\{fc_{in}.\mathrm{support}| fc_\mathrm{in}\subseteq s_{\mathrm{j}i}^*\wedge fc_\mathrm{in}\in FC_\mathrm{i}\}$.

For session $\mathrm{s}_{ji}^*\in S_\mathrm{j}^*$, every attributer's meaning in the feature value vector is as follows:

- length: the number of pages contained in the session $s^*{}_{ji}$;
- pun: the number of frequent access URLs contained in the session $s^*{}_{ji}$;
- mrn: the number of frequent itemsets contained in the session $s^*{}_{ji}$;
- rpun: the number of frequent access URLs contained in the session $s^*{}_{ji}$;
- mrml: maximum length of frequent itemset contained session $s^*{}_{ji}$;
- mral: average length of frequent itemset contained session $s^*{}_{ji}$:

$$\mathrm{mral}=\frac{\sum_{FC_i\mathrm{length}}p}{\mathrm{count}(FC_i\mathrm{length})} \tag{2}$$

where $\mathrm{count}(FC_i\mathrm{length})$ means the number of elements in $FC_i\mathrm{length}$;

- mrms: maximum support of frequent itemset contained in session $s^*{}_{ji}$;
- mras: average support of frequent itemset contained in session $s^*{}_{ji}$, its function is as follows:

$$\mathrm{mral}=\frac{\sum_{FC_i\mathrm{support}}p}{\mathrm{count}(FC_i\mathrm{support})}, \tag{3}$$

where $\mathrm{count}(FC_i\mathrm{support})$ means the number of elements in $FC_i.\sup port$;

- mtn: number of frequent access period of time matched by session $s^*{}_{ji}$;
- target: target column, if $s^*{}_{ji}$ is legal, we record it as 1, else 0.

In this paper, the first nine elements of fv_{ji} correspond to the variables vector in $f:X->Y$, and <*target*> means target vector Y. The user browsing feature is shown as a ten-tuple form $w(w_0, w_1, w_2, \ldots, w_9)$, and $(w_1, w_2, \ldots, w_9)$ means the weights of first nine elements in fv_{ji}.

III. Construction and Authentication of UWBF

In this section, we mainly focus on how collected data is used in extracting user web browsing features and identify authentication.

A. Data Processing

This paper collects one month's visit records of 10 users. It uses 30min as the maximum interval of a session [2], [16]. Besides, for the purpose of mining more potential relationships of *(website, content)*, this paper extracts the top domain and classifies the content of websites for generalizing.

1) *Session partition: According to Definition 1,* we collect original records of each user. And then we generate each user's sessions according to Definition 2 to form S. For user *j*, its session set is $S_j = \{s_1, s_2, \ldots\} = \{\{r_1, r_2, \ldots r_x\}, \{r_{x+1}, r_{x+2}, \ldots r_{x+m}\}, \ldots\}$.

2) *Webpage procession:* Here, we convert every record *r* into the structure of *p* in Definition 3:

Step 1: top domain extraction, which means extract top domain of *r.url.*

Step 2: content extraction, which means classify the content of record according to its title. In this step, we reference the major news sites as well as the input methods' thesaurus, define 12 classes in advance and use Sogou lab's text classification samples and some articles on the network as the training text. There are 10 articles under each classification. Then we extract corresponding keywords from every classification's article and classify web pages by matching the keywords in the title.

Step 3: relative time computing. To obtain every day's frequent access period of time of every content, we need to convert *r.timestamp* to the milliseconds away from every day.

For instance, there is a record r stored in the database: (http://news.163.com/14/1103/18/AA58S2T800014SEH.html, the hero in Wenchuan earthquake is involved in the in the fraud trial of 460000 yuan, 36510608). After webpage processing, it changes to (*news.163.com*, society, 36661272). And we generate the final session set $S^*_{\mathrm{j}} = \{s^*_1, \cdots, s^*_n\}$.

3) *Page class merger:* In fact, page class merger is a process to generalize records. After this step, two different records r_i and r_j will become r_i' and r_j' which have the same details except the relative time.

For the purpose of mining user's frequent itemsets, this paper merges r_i' and r_j' into one webpage set **P**, so a webpage set may contain many records. For instance, there is a session set after the step of webpage processing: $S^*_{\mathrm{j}} = \{\{p_1, \cdots, p_x\} \cdots\}$. We merge the same webpages into one webpage class and let it replace the original records, and generate a new session set $S_{\mathrm{j}}' = \{\{p_1, \cdots, p_4\} \cdots\}$.

B. Feature Extraction

In the stage of feature extraction, we extract users' features from three aspects: frequent access URLs, (*website, content*) and (*content, frequent access period of time*).

1) *The mining of frequent itemsets:* We find that each user has different favorite contents in different websites. For instance, user A prefers military news in *news.ifeng.com* after reading social news in *news.163.com* and financial news in *news.qq.com*. Such frequent itemsets are different to each user. We regard the relationship of (*website, content*) as the bases for user identify authentication.

During the mining of frequent itemsets, the correct selection of support threshold δ has a great influence on (false alarm rate (FAR) and detection (DR). So we need to find a proper δ to balance them.

After this step, we may discovery such a frequent set like P_1, P_3, P_5, namely (*domain*$_1$, *content*$_1$), (*domain*$_1$, *content*$_3$), (*domain_*$_5$, *content*$_5$).

2) *The mining of frequent periods of time:* A statistic on frequent access periods of time of different contents for each user is necessary in this step. Since every user is used to browse different contents in specific period of time, we use the *FT* in Definition 6 to dig out them, and save them in the form of Definition 7. For a user, we assume that the period of time of a content he views is subject to a normal distribution process.

For $\mathrm{pt_i} \in PT$, $ft_{\mathrm{j}} \in FT$ and $\mathrm{pt}_i.content = ft_j.content$,

$$ft_j.\hat{\mu} = \frac{\sum_{i=1}^{m} \mathrm{pt}_i.time}{\mathrm{m}} \tag{4}$$

$$ft_j.\hat{\sigma}^2 = \frac{\sum_{i=1}^{m} (pt_i.time - \hat{\mu})^2}{\mathrm{m-1}} \tag{5}$$

where *m* means the number of elements in *FT*.

For example, there is such an element in *FT* of user *A*: $\mathrm{ft}_3 = (content, \hat{\mu}, \sigma^2)$, this means user *A* is accustomed to browse the content in the period of time of $(\hat{\mu} - 3\sigma, \hat{\mu} + 3\sigma)$.

C. Authentication Based on User Browsing Feature

During the phase of constructing the method based on user web browsing feature, we need to choose a proper classification algorithm which is combined with the feature values computed in Section III.B. So we can judge whether the session is legal or not. Through the comparative experiments, we choose the logistic regression algorithm [19] (LR) as the classifier.

LR is a typical binary-class classification algorithm, which is relatively simple and easy to explain, and rarely occurs over fitting. It is actually a process of learning function *f*: *X->Y*. We give n-tuple variable vector *X*= < X_1, X$_2$..., X_n > and an m-tuple target vector *Y*=< Y_1, *Y*2..., Y_m > in advance. LR's responsibility is to learn a function, which can fit *Y* furthest according to *X* we have given before.

We use LR as a binary-class classifier. Once test data is obtained, we will compute *y* according to the function as follows [19]:

$$y = w_0 + w_1 * x_1 + w_1 * x_1 + \cdots + w_n * x_n \tag{6}$$

where $x_1, x_2, \cdots, x_n$ are the features of a sample data. Then we use the Function *sigmoid* to do the following things:

$$\sigma(y) = \frac{1}{1+\exp(y)} \quad . \tag{7}$$

Domain of Function *sigmoid* is (-INF, +INF) and domain of value is (0, 1).

In order to eliminate the influence of authentication caused by the difference of feature values' range, we normalize feature values before using the LR first so that every value can be transformed into the range of [0, 1]. Its function is as follows:

We assume that $x = (x_1, x_2, ..., x_n)$ and build mapping *f*:

$$\mathrm{x_k} \mapsto f(x_k) = \frac{x_k - x\min}{x\max - x\min}, \tag{8}$$

where $x_{\max} = \max(x_1, x_2, ..., x_n)$, $x_{\min} = \min(x_1, x_2, ..., x_n)$.

For the authentication algorithm of this paper UWBF (a kind of authentication algorithm based on user web browsing features), data is the train session set after normalization. It uses LR and frequent itemsets, frequent access URLs, frequent access periods of time of legal user to learn a matrix on feature values' weights. The specific process of the authentication method is shown in algorithm 1.

Algorithm 1: UWBF

```
Input: session set S*, a legal user's frequent itemsets FC, the legal user's
frequent access URLs and    frequent access periods of time FT
Output: the matrix of feature values' weights w, array score_legal
For all  s*_i in S*  do
        Calculate each element value in FV_i
                $mrtl \leftarrow 0,\ mrts \leftarrow 0$
                $hum \leftarrow fromFU(FU)$
                $(mrn, mrtl, mrts, mrms, mrml) \leftarrow fromFU(FU)$
                $(mras, mral) \leftarrow fromFUAver(FU)$
                $mtn \leftarrow fromFT(FT)$
        Save all attributes into FV_i
end for
for all  element in ten-tuple FV_i  do
        Create and set value matrix
                $dataMatrix[] \leftarrow creatDataMatrix(FV_i)$
                $labelMatrix[] \leftarrow creatDataMatrix(FV_i)$
                $weightMatrix[] \leftarrow creatDataMatrix(FV_i)$
end for
 for all  FV_i  in FV_s  do
        Calculate session's score according to $\omega$
                $score_{legal}[i] \leftarrow calc(\omega)$
end for
```

D. Identify Authentication

We actually uses LR to train and learn the authentication method based on user's browsing features. In this section, we need to compute a score's threshold for classification. If the computed score is in the classification threshold, we regard it as a legal session, otherwise we regard it as illegal. So the identify authentication will be divided into two parts.

1) *The choice of classification threshold:* Before choosing the classification threshold, we get the matrix of weights $w(w_0, w_1, w_2, \ldots, w_9)$ and feature value vector $\mathrm{fv_j}$ of session *j* to compute its *score_j*, the function is as follows:

For $\mathrm{fv_i} \in FV$,

$$\mathrm{score} = w_0 + w_1 * \mathrm{fv_i}.length + w_2 * \mathrm{fv_i}.pun + ... + w_9 * \mathrm{fv_i}.m\mathrm{tn} \tag{9}$$

By using Function 9 and Definition 11, we can get the train set's feature value set FV_{train} and compute the legal sessions' scores: $\mathrm{score_{train}} = \{score_{legal1}, \ \mathrm{score}_{legal2}, ..., score_{legalm}\}$.

Then we compute the threshold of classification and the function is as follows:

$$\overline{\mathrm{score}} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{score_{legali}}}{\mathrm{m}} \tag{10}$$

$$\partial = \frac{\sum_{\mathrm{i=1}}^{\mathrm{m}} (\mathrm{score_{legali}} - \overline{\mathrm{score}})}{\mathrm{m}} \quad , \tag{11}$$

where *M* means the length of array *score_{legal}*.

In this paper, the classification threshold is set to $[\overline{score} \quad \gamma, 1]$, which means that if the session's score is in this threshold, we regard it as legal, otherwise we regard it as illegal.

2) *The calculation of session's score:* After using the first step of the construction of the authentication method based on user browsing feature, we generate every session's set of feature values FV_{test}. Then we compute the matrix of weights according to UWBF and every session's score in the test session set according to Function 9. Finally, we judge whether every session is legal or not according to the classification threshold.

IV. EXPERIMENT RESULT

Experimental environment: Intel (R) Core (TM) i3CPU (3.23GHz), 4GB memory, Windows8 operating system, Java development language, eclipse development environment.

Experimental data: web browsing records collected from 10 users within 4 weeks. For each webpage, we collect its title, URL and timestamp. According to Section III.A, each user can form a collection of about 200 sessions after data processing. We suppose that one of 10 users is legal while the other users are not.

Since the effect of identify authentication may be impacted by δ and the algorithm of classification. For the purpose of having a better authentication effect, this paper does the experiments from three aspects: (1) choose a proper δ; (2) choose a suitable algorithm of classification according to the contrast experiments; (3) use the computed δ and algorithm of classification to authenticate.

A. Choice of Support Threshold

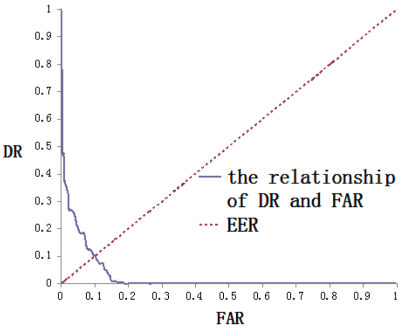

Fig. 1. The relation of FAR and DR.

We use FAR as abscissa and DR as ordinate. In Fig. 1, the blue curve shows the relationship of DR and FAR when the support threshold changes, and the red dotted line is EER which means *y*=*x*. Through Fig. 1, we can find that when FAR=DR=0.102, two curves intersect, and δ at this moment is the optimal *threshold* [20]. Then by Fig. 2 which shows the relation between FAR and δ , we can find that when FAR=0.0102, $\delta = 0.0257$. We set it as the support threshold for mining frequent itemsets.

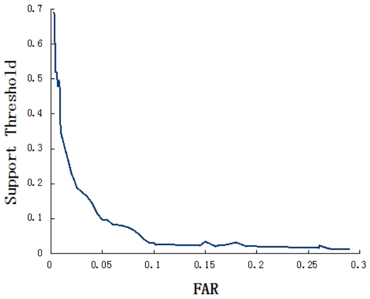

Fig. 2. The relation of FAR and support threshold.

B. Choice of Classification

In the period of authentication, there is a need to select a suitable algorithm of classification. We use this paper's data and compare performances of LR, decision tree and KNN by showing their authentication effects in ROC. As shown in Fig. 3:

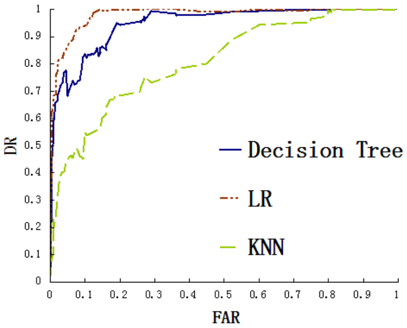

Fig. 3. ROC of each algorithm.

From Fig. 3, we can see that performance of LR is the best, so we choose LR as this paper's algorithm of classification.

C. Identify Authentication

In order to compare UWBF with WS (the algorithm of [2]) which based on sequences of webpages, we conduct two experiments with the support threshold δ of 0.0257 and present the experimental results in the form of ROC curves.

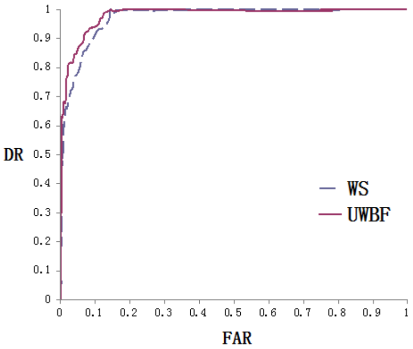

Fig. 4. The contract authentication effects of UWBF and WS in different data.

Experiment 1: Based on their own data, two algorithms' contrastive performances in authentication are shown in Fig. 4. UWBF's FAR is 6.4%, DR is 92.7% while FAR of WS is 10%, DR is 91.3%. What's more, the *AUC* [20] (Area under the Curve of ROC) of UWBF is 0.983 while the AUC of [2] is 0.952.

Experiment 2: Based on this paper's data, the authentication effects of UWBF and WS are shown in Fig. 5. UWBF's FAR is 10%, DR is 93.6% while FAR of WS is 10%, DR is 87%. And UWBF's AUC is 0.983 while the AUC of [2] is 0.932.

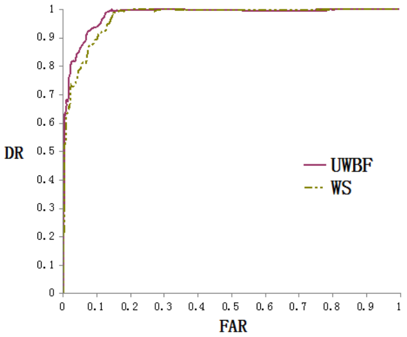

Fig. 5. The contract authentication effects of UWBF and WS in same data.

The two experiments indicates that UBFAA algorithm has lower FAR and higher DR. This is because that WS only makes analysis on website sequences, ignoring of the fact that user's browsing habit is random and variable, which results in a higher FAR. UBFAA algorithm, on the other hand, considers websites, contents and time altogether, which covers a wide range and the model is more complete. Thus, UBFAA can reflect more characters of particular user, and reaches a better effect in identity authentication.

V. Conclusion

This paper proposes a kind of authentication method based on user web browsing feature. It combines user's favorite websites, content and period of time and regards them as a whole entity to analyze. We use Apriori to mine out every user's browsing features and use LR to construct a classifier to do authentication. Experiments show that compared with method which considering the links between websites, this paper has a better performance in identify authentication and reflecting different users' web browsing features. It can achieve higher DR under smaller FAR.

In future work, we will study more on sequence association rule mining. Multipliers can be especially confusing.

References

[1] iResearch (2014). Annual monitoring report of china's online shopping industry in 2014. [Online]. Available: http://www.iresearchchina.com/content/details828268.html

[2] J. Zhong, C. Yan, W. Yu, P. Zhao, and M. Wang, "A kind of identity authentication method based on browsing behaviors," in *Proc. 2014 Seventh International Symposium on Computational Intelligence and Design (ISCID*, 2014*)*, pp. 279–284, vol. 2.

[3] L. Gong, J. Pan, B. Liu, and S. Zhao, "A novel one-time password mutual authentication scheme on sharing renewed finite random sub-passwords," *Journal of Computer and System Sciences*, vol. 79, pp. 122–130, Feb. 2013.

[4] F. F. Moghaddam, N. Khanezaei, S. Manavi, M. Eslami, and A. Samar, "Uaa: User authentication agent for managing user identities in cloud

computing environments," in *Proc. 2014 IEEE 5th Control and System Graduate Research Colloquium (ICSGRC)*, Malaysia, Aug. 2014, pp. 208–212.

[5] M. E. Haque and T. M. Alkharobi, "Adaptive hybrid model for network intrusion detection and comparison among machine learning algorithms," *International Journal of Machine Learning and Computing*, vol. 5, pp. 17-23, Feb. 2015.

[6] X. Li, Y. Xiong, J. Ma, and W. Wang, "An efficient and security dynamic identity based authentication protocol for multi-server architecture using smart cards," *Journal of Network and Computer Applications*, vol. 35, pp. 763-769, March 2012.

[7] C. Shen, Z. Cai, X. Guan, Y. Du, and R. A. Maxion, "User authentication through mouse dynamics," *IEEE Trans. on Information Forensics and Security*, vol. 8, pp. 16-30, Jan 2013.

[8] D. Wang, D. He, P. Wang, and C.-H. Chu, "Anonymous two-factor authentication in distributed systems: certain goals are beyond attainment," *IEEE Trans. on Dependable and Secure Computing*, vol. 12, pp. 428–442, Jul-Aug 2015.

[9] C. H. Jiang, S. Shieh, and J. C. Liu, "Keystroke statistical learning model for web authentication," in *Proc. the 2nd ACM symposium on Information, Computer and Communications Security*, pp. 359-361, 2007.

[10] A. S. Nathiarasan and M. Manikandan, "Performance oriented mining of utility frequent itemsets," in *Proc. 2014 International Conference on Circuits, Communication, Control and Computing (I4C)*, 2014, pp. 317–321, vol. 1.

[11] C. J. Carmona, S. Ram´ırez-Gallego, F. Torres, E. Bernal, M. J. del Jesus, and S. Garc´ıa, "Web usage mining to improve the design of an ecommerce website: OrOlivesur.com," *Expert Systems with Applications*, vol. 39, pp. 11243–11249, Sep. 2012.

[12] M. A. Awad and I. Khalil, "Prediction of user's web-browsing behavior: Application of markov model," *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, pp. 1131–1142, Aug. 2012.

[13] S. G. Esparza, M. P. OMahony, and B. Smyth, "Mining the real-time web: a novel approach to product recommendation," *Knowledge-Based Systems*, vol. 29, pp. 3–11, May.2012.

[14] N. N. Liu, M. Zhao, E. Xiang, and Q. Yang, "Online evolutionary collaborative filtering," in *Proc. the Fourth ACM Conference on Recommender Systems*, Sep, 2010, pp. 95-102.

[15] E. Poovammal and P. Cigith, "Mining web path traversals based on generation of FP tree with utility," *Trends in Computer Science, Engineering and Information Technology*, pp. 94-100, vol. 204, 2011.

[16] L. C. X. K. W. Liming and X. Jun11, "A web content topic combined session identification and segmentation method," *Computer Applications and Software*, vol. 6, pp. 167-169, Sep. 2011.

[17] T. Herawan and M. M. Deris, "A soft set approach for association rules mining," *Knowledge-Based Systems*, vol. 24, pp. 186-195, Feb. 2011.

[18] A. W. Ding, S. Li, and P. Chatterjee, "Learning user real-time intent for optimal dynamic web page transformation," *Information Systems Research*, vol. 26, pp. 339-359, June 2015.

[19] A. B. Musa, "Comparative study on classification performance between support vector machine and logistic regression," *International Journal of Machine Learning and Cybernetics*, vol. 4, pp. 13-24, Feb. 2013.

[20] J. M. Lobo, A. Jim ́enez-Valverde, and R. Real, "Auc: A misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, vol. 17, pp. 145–151, Mar. 2008.

Mengqing Ji received the B.S. degree from Taiyuan University of Technology, Taiyuan, China, in 2013. She is currently pursuing M.S. degree at the Department of Computer Science and Engineering, Tongji University, Shanghai, China.

Her current research interests include behavioral analysis, web security and data mining.

Peihai Zhao received the B.S. degree from Tongji University, Shanghai, China, in 2011. He is currently pursuing Ph.D. degree at the Department of Computer Science and Engineering, Tongji University, Shanghai, China.

His current research interests include behavioral analysis, web security and data mining.

Mimi Wang received B.S. degree from the Department of Mathematics and computational mathematics, Xinyang Normal University, Xinyang, China, in 2009, and the M.S. degree from the Department of Computer Science and Engineering, Anhui University of Science and Technology, China, in 2013. She is currently working towards the Ph.D. degree at the Department of Computer Science and Technology, Tongji University, Shanghai, China.

Her research interests include formal methods, consistency analysis, model checking, and Petri net theory.

Chungang Yan received the B.S. degree in applied mathematics from the Shandong University of Science and Technology, Taian, China, in 1986, M.S. degree in operational research and cybernetics from the Shandong University of Science and Technology, Taian, China, in 1993, and the Ph.D. degree from Tongji University, Shanghai, China, in 2006.

She is currently a Professor with the Department of Computer Science and Technology, Tongji University. She has published more than 30 papers in domestic and international academic journals and conference proceedings. Her current research interests include concurrent models and parallel algorithms, Petri net theory, formal verification of software, and trusty software process theory.

Zhong Li received her B.S. and M.S. degrees on computer science and technology from Shandong Normal University in 2007 and 2010, respectively. She is currently a Ph.D. student in The Department of Computer Science at Tongji University in Shanghai, China.

Her research interests include wireless communication, social network analysis and distributed computing.

Changjun Jiang received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1995 and conducted post-doctoral research at the Institute of Computing Technology, Chinese Academy of Sciences, in 1997. Currently he is a Professor with the Department of Computer Science and Engineering, Tongji University, Shanghai. He is also a council member of China Automation Federation and Artificial Intelligence Federation, the vice director of professional committee of Petri Net of China Computer Federation, and the vice director of professional committee of Management Systems of China Automation Federation. He was a visiting professor of Institute of Computing Technology, Chinese Academy of Science; a Research Fellow of the City University of Hong Kong, Kowloon, Hong Kong; and an Information Area Specialist of Shanghai Municipal Government. His current areas of research are concurrent theory, Petri net and formal verification of software, concurrency processing and intelligent transportation systems.