A Method for Privacy Preserving Mining of Association Rules based on Web Usage Mining

Wang Yan
Glorious Sun School of Business and Management
DonghuaUniversity
Shanghai,China
yanwang@mail.dhu.edu.cn

Le Jiajin
School of Computer Science and Technology
DonghuaUniversity
Shanghai,China
lejiajin@dhu.edu.cn

Huang Dongmei
College of Information Technology
Shanghai Ocean University
Shanghai,China
dmhuang@shou.edu.cn

Abstract*: *Data mining basing on Privacy preservation has become a research hot point now. Web usage mining is one kind of data mining applications, and how to prevent data leakiness in web usage mining is also an important issue. In this paper, we present an effective method for privacy preserving association rule mining in the web usage mining, Secondary Random Response Column Replacement (SRRCR) to improve the privacy preservation and mining accuracy. Then, a privacy preserving association rule mining algorithm based on SRRCR is presented, which can achieve significant improvements in terms of privacy and efficiency. Finally, we present experimental results that validate the algorithms by applying it on real datasets.

Keywords: web usage mining; privacy preservation; randomized response; association rules

I. INTRODUCTION

With the continuous development of data mining technology, its appliance is more and more widely. Web data mining is an application hot point of data mining in recent years. The fast development of web makes itself to be the world's largest public data source, Web data mining is based on public data sources, but there are also privacy leakiness problem. For example, on e-commerce site, all records of customers visited the website are stored in Web server log without reservation. These records reflect the consumers' buying habits and purchasing powers，but the customers do not want these data to be disclosed, so privacy protection should also be considered in the web data mining.

Privacy preserving association rule mining is to discover the frequent itemsets as accurately as possible and generate the minimum support and confidence by accessing the original transaction sets of conditions. There is a contradiction between data privacy and data accuracy. At present, the privacy protection of data mining methods can be divided into data perturbation [1-3] and query restriction [4-5]. In 2002, S. J. Rizvi and J. R. Haritsa presented a representative association rules mining method, MASK (Mining Associations with Secrecy Konstraints) [1] that is a data perturbation strategies. Because Warner model in statistics is used in MASK, all the data transformed have the directly relations with the real raw data, which makes the preserving less than satisfactory, and the selection of randomized parameters should deviate from 0.5 [6] is another limitation.

In this paper, we give the general framework for mining association rules in the web usage mining and use a user-oriented and time-oriented session explore method to generate sequence sets firstly. Then we propose a method to transform user session information into a relational data table, and perturb data by using Secondary Random Response Column Replacement (SRRCR) algorithm. Based on the pseudorandom data set, frequent itemsets and strong association rules discovery algorithm are presented.

II. THE GENERAL FRAMEWORK

As e-commerce, Web services and Web-based information system's sustainable development and growth, there are a large number of user data collected in web organizations. The principal data sources in web usage mining are web server log files, and other Data sources including the web site files and metadata, operational databases, application templates and domain knowledge. In order to solve privacy preservation issues of web usage mining, we proposed Secondary Random Response Column Replacement **(**SRRCR) algorithm, and a privacy protection association rule mining methods based on it. The overall structure and framework shown as follow Fig.1.The frameworks were divided into 2 stages: data preprocessing stage and privacy preserving &mining stage. The data preprocessing stage preprocessing the data by five steps and generate the session translate table. The privacy preserving stage is the key stage including the privacy preservation methods and association rule application.

978-0-7695-4224-9/10 $26.00 © 2010 IEEE
DOI 10.1109/WISM.2010.66

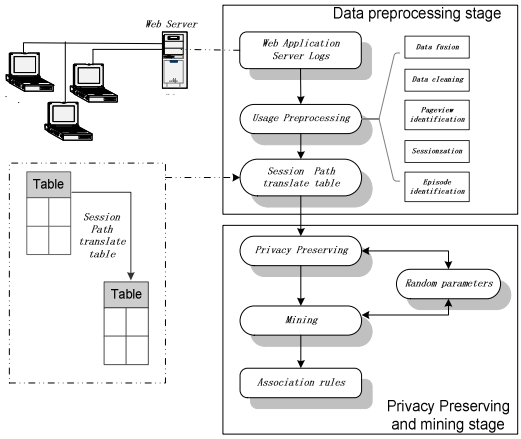

Fig. 1 Web usage mining framework

III. PRIVACY PRESERVATION IN WEB APPLICATIONS MINING

A. Data Preparation

Each click action of web user in e-commerce sites is recorded in the web log, so the original web log records contain a lot of browsing information. Before web data mining, data should be preprocessed [8], where the process includes filtering, despidering, user identification, sessionization and path completion.

B. Time-oriented user session exploring method

After data preprocessing, we can mine and extract relevant data in web server logs. For the purposes of analysis, these data need to be converted and aggregated in the different levels. In web usage mining, the page is the basic level of data extraction and the session is the basic action in the data extraction for users.

TABLE1 ORIGINAL SESSION SAMPLE

ID	IP	URL	Time
1	2.3.4.5	Page1	0:01
2	2.3.4.5	Page2	0:05
3	2.3.4.5	Page4	0:16
4	3.4.5.6	Page1	0:10
5	3.4.5.6	Page3	0:12
6	3.4.5.6	Page4	0:15
7	3.4.5.6	Page5	0:19
8	5.6.7.8	Page2	0:11
9	5.6.7.8	Page5	0:13
10	9.6.7.8	Page2	0:11
11	9.6.7.8	Page4	0:17
12	5.6.7.9	Page1	0:15
13	5.6.7.9	Page3	0:19
14	5.6.7.9	Page5	0:25

Table1 shows a sample of user's original session. In the table each row represents a user session identified by using uniquely ID , and IP represents the user's IP address, where different addresses for different users. URL means the page accessed by users in the session. The ideal method of user path restoring can reconstruct the real process of user session. Here we use time-oriented user sessions explore method.

Let the vector M $\{M_1, M_2, \ldots, M_i, \ldots, M_N\}$ represent the log data set, where $M_i = \{IP_i, URL_i, Date_i, Time_i, Method_i, Code_i, Bytes_i\}$, and N is the number of log files. Web logs contain the IP addresses, visited URLs, date and time records, access methods (GET or POST), access structure and the size of access information. the vector S $\{S_1, S_2, \ldots, S_i, \ldots, S_K\}$ represents the session data set, where k is the number of sessions, and $S_i = \{IP_i, Page_i, t_i\}$,this structure contains the IP address , visited page and the timestamp in each session. $Page_i = \{Page_1, Page_2, \ldots, Page_i, \ldots, Page_L\}$, L is the number of pages for the session.

The algorithm 1 is described below:

For each M_i of M
If Methodi is 'GET' and Url_i is 'WEBPAGE'
If $\exists S_k \in S$ and $IP_k = IP_i$
Set t_0 // set the first time stamp of request in the session
If($t_i - t_0 =< 20min$) // set the limitation of access time is 20 minutes
$S_k = (IP_k, PAGE_{Sk} \cup PAGE_{Si}, t_k)$
else
Delete S_k from S
Add S_k into S
Endif
else
Add S_k into S
Endif
Endif
Endfor

Table2 is the session sets generated by the path exploring algorithm.

TABLE2 SESSION SETS GENERATED BY PATH EXPLORING ALGORITHM

SID	IP	Page	Time
1	2.3.4.5	Page1,page2,page4	0:16
2	3.4.5.6	Page1, page3,page4,page5	0:19
3	5.6.7.8	Page2,page5	0:13
4	9.6.7.8	Page2,page4	0:17
5	5.6.7.9	Page1,page3,page5	0:25

C. Transforming user sessions set into two-dimensional relation table

For the user session set in Table 2, we can use relations table S_table to store the user session path, the table structure is (TID, IPID, SESSIONID, P_i),

where we store the IP address, session ID, and the visits of each page in a two-dimensional table. The value of Pi column means whether the user visits the page and buys the products or not .in the column 1 means an access to the page and 0 means no access to the page.

The algorithm 2 is described below; it transfers the user activity records into the relation two-dimensional table that can be handled easy.

Algorithm 2

Input: user session set S

Output: two-dimensional user session table S_table

Create table S_table（TID,IP,SESSIONID,P_i）

For each S_i of S

Insert into S_table(IPID,SID) values (IP_i,SID_i)

If $\exists P_i$ of S_i

Insert into S_table（P_i）Values('1');

else

Insert into S_table（P_i）Values('0');

Endfor

Each session will be seen as a transaction, the customer's session sequence can be converted to relational tables where the page numbers represent the corresponding purchase action. We added a new field TID in the table 3, which is the identity column and can be assigned values automatically by the system.

Table3 2D table saving the user session path

TID	IPID	SID	P1	P2	P3	P4	P5
10001	2.3.4.5	1	1	1	0	1	0
10002	3.4.5.6	2	1	0	1	1	1
10003	5.6.7.8	3	0	0	0	0	1
10004	9.6.7.8	4	0	1	1	1	0
10005	5.6.7.9	5	1	1	0	1	1

D. Privacy Protection Algorithm

In the two-dimensional table the columns which need to privacy protected are P1, P2 and P3, P4 and P5, those columns are known as sensitive attributes in the privacy protection technology [7].The attribute values itself are not to disclose privacy, but the attribute group will leak customers privacy information. Therefore we protect the sensitive attributes by using the improved randomized response method. In order to improve the efficiency of data preservation, we randomize the value twice at a certain probability, and this algorithm is named the secondary random response column replacement algorithm (SRRCR). Specific methods are as follows:

Set random parameter I,

$0 \le p_1, p_2, p_3, p_4 \le 1$,and $p_1 + p_2 + p_3 + p_4 = 1$,

Set random parameter II,

$0 \le r_1, r_2, r_3, r_4 \le 1$,and $r_1 + r_2 + r_3 + r_4 = 1$,

Create randomization function R (x), the original column C (i, j) is transformed to hiding column F (i, j), where F(i,j)∈ {0,1},C(i,j)∈ {0,1}.

For x= F(i,j) ∈ {0,1}, random function convention is as follows:

TABLE4 CONVENTION OF RANDOM FUNCTION

	p_1	p_2	p_3	p_4
x	x	0	1	y
	r_1	r_2	r_3	r_4
y	x	0	1	null

Let i is an item , where π is the support of i in C, π_{null} is the support of null in C, λ is the support of i in F. The transaction T of C is converted by SRRCR process to the transaction T′ of F, and then there is:

$$\pi_{null} = p_4 * r_4 \quad (1)$$

$$\pi = \lambda * (p_1 + p_4 * r_1) + p_3 + p_4 * r_3 \quad (2)$$

By the (1) and (2):

$$\lambda = \frac{\pi - p_3 - \pi_{null} * \frac{r_3}{r_4}}{p_1 + \frac{r_1}{r_4} * \pi_{null}} = \frac{\pi - p_3 - \kappa * \phi}{p_1 + \gamma * \phi} \quad (3)$$

Here $\kappa = \frac{r_3}{r_4}$ is the top coefficient, $\gamma = \frac{r_1}{r_4}$ is the bottom coefficient，and $\phi = \pi_{null}$ is the support of null value.We can see that the benefit of the algorithm is the introduction of the adjustable factors to calculate of the support. In t the actual perturbation process, we can use a null value to further enhance privacy preservation.

TABLE 5 PROBABILITIES OF DATA MAPPING BY SRRCR METHOD

No.	C(i, j)	F(i, j)	Probability
1	0	0	$p_1 + p_2 + p_4 * (r_1 + r_2)$
2	0	1	$p_3 + p_4 * r_3$
3	1	0	$p_2 + p_4 * r_2$
4	1	1	$p_1 + p_3 + p_4 * (r_1 + r_3)$
5	0	null	$p_4 * r_4$
6	1	null	$p_4 * r_4$

Assuming that support threshold is 0.4 in the M, and choosing p_1 =0.2,p_2=0.2,p_3=0.1,p_4=0.5, while select r_1=0.3,r_2=0.2,r_3=0.2,r_4=0.3 , according to SRRCR algorithm customer session set is converted as table6:

TABLE6 FAKE COLUMN DATA SET GENERATION AND FINAL FAKE COLUMN DATA SET

SID.	P1	F1	P2	F2	P3	F3	P4	F4	P5	F5
1	1	1	1	*null*	0	1	1	0	0	*null*
2	1	1	0	1	1	0	1	1	1	0
3	0	*null*	0	0	0	*null*	0	1	1	1
4	0	1	1	1	1	0	1	0	0	1
5	1	0	1	0	0	0	1	*null*	1	1

Raw	Fake
P1	P2
P2	P3
P3	P1
P4	P5
P5	P4

SID.	P1	P2	P3	P4	P5
1	*null*	1	1	*null*	0
2	1	0	1	0	1
3	0	*null*	*null*	1	1
4	1	0	1	1	0
5	0	0	0	1	*null*

After conversion algorithm SRRCR, fake sets include not only the value of 0 and 1, also null value. The meaning of Null can be defined by the user.

Mining association rules can take into account the column of null value and the support of itemset i in fake set after converted from SRRCR algorithm can matches the support of j in the original transaction set.

IV. PRIVACY PRESERVING ASSOCIATION RULE MINING ALGORITHM

Apriori algorithm is an influential algorithm for mining frequent itemsets and Boolean association rules. Based on the Apriori Algorithm [9], we presented an improved algorithm to achieve association rule mining for fake transaction sets converted by SRRCR. Given the minsup and data set, main implementation actions of the algorithm are described as follows:

Algorithm 4. Progressive scan association rule mining algorithm.

1. Setting frequent itemsets φ , and φ is collection of itemset M_n whose item number is n，where $n=\sum_{i=1}^{k} C_k^i$ ，and k is the maximum number transactions.
2. Searching the largest itemset progressively. Algorithms scan the data row by row, and matching each line with all items set, if the row matches with some itemsets, then plus 1 for each location of sequence φ .
3. Using Apriori pruning principle in the process of match the itemset in the row data. Algorithm first match 1-itemset, then 2- itemset, 3-itemset ….to reduces the computer load.
4. Circulation in turn，when the ratio of the number in k- itemset of the searched rows to the numbers of all rows has reached the minimum support of k-itemset, stop this k-itemset match process .If the ratio o f the number n in k- itemset of the searched rows to the numbers of all rows m has not reached the minimum support of k- itemset, only reached the number of x , when n/m> minsup , algorithm calculate r=((n-m)+x)/m, if r> minsup, continues circulating, else stop this itemset search.
5. Getting all the frequent itemsets which satisfy the minsup.

Processing of Algorithm is described below :

Input: itemset D′ after conversion algorithm SRRCR, minimum support threshold s.

Output: frequent itemset F of real dataset D.

Set φ ,

$\varphi = \{M_1, M_2, \ldots, M_n\} = \{(L_1), (L_2), .. (L_k), (L_1, L_2), .. (L_1 ... L_k)\}$ is the full itemset

for (i=n; i>0； i--) ,and n= $\sum_{i=1}^{k} C_k^i$

set A_i=index(i) // set the searching matrix matching φ (i)

for each transaction t∈ D′ // scan each row

$R_i = C_i * A_i$

if $\exists R_i \neq 0$, then // exit matched row

if i=n //counting every item calculate template

（c.count= $c.count_0$+1 , $c.count_1$= $c.count_1$+1….

$c.count_n$= $c.count_n$+1）

else

select_add(c.count) // function select_add

endif

endif

endfor

$F_k \leftarrow \{ c \in C_k | c.count / n \geq \min sup\}$

endfor

return $F_k \leftarrow U_k F_k$

Function select_add is to count each itemset after pruning (ki) , that is $c.count = c.count_0 + 1$, $c.count_1 = c.count_1 + 1 .... c.count_n = c.count_n + 1$. In the counting process, algorithm prune the frequent itemsets generated (Pruning methods is in Apriori principle, no explanation).

After obtained frequent itemsets, based on the support and confidence formula below, data providers can find the assoication rules of web log item.

Support (A $\Rightarrow$ B) = Support_count (AUB) / all_count

V. PERFORMANCE ANALYSIS

A. Algorithm Complexity Analysis

Theoretically, Apriori algorithm is an exponential algorithm. let N be the size of transaction, each transaction includes n items averagely, and then the whole itemset space will reach O （ 2^n ）.The mining algorithm we have improved, when k <1-2 * (1-minsup), the itemset corresponds with the position of the value in sequence φ does not match the minimum support, the last calculations for this itemset will be given up. If we do not consider this kind of row optimization, time complexity of the algorithm is

O（N* $\sum_{i=1}^{n/2} C_n^i$ ）=O（N*2n/2）.

If the probability pre-judgment mentioned above is taken into account during the row judge, the complexity will decrease continually.

B. Experimental results

We obtained the web server log files in some C2C website in 30 days and converted transaction sets T by the algorithm 1and 2. The transact number is 1000, the item number is 11, and the ATL (the average transaction length) is 2.8. We analyzed the privacy protection with minimum support of 3%, and the following table shows the privacy protection intensity of the minimum support of 3% with different values of randomized algorithm parameters for different.

TABLE 7 INTENSITY OF PRIVACY PROTECTION

No.	I(p_1,p_2,p_3,p_4)	II(r_1, r_2, r_3, r_4)	Privacy intensity
1	(0.2, 0.3, 0.4, 0.1)	(0.2, 0.3, 0.2, 0.3)	91.6
2	(0.1, 0.5, 0.2, 0.2)	(0.2, 0.3, 0.4, 0.1)	92.8
3	(0.2, 0.2, 0.3, 0.3)	(0.2, 0.4, 0.2, 0.2)	92.6
4	(0.4, 0.1, 0.2, 0.3)	(0.3, 0.1, 0.4, 0.2)	93.5
5	(0.2, 0.4, 0.3, 0.3)	(0.4, 0.2, 0.3, 0.1)	92.3

The Fig.2 shows the average itemset error of SRRCR method and MASK method, and states the relationships between the random parameter, the data privacy intensity and the accuracy of mining results.

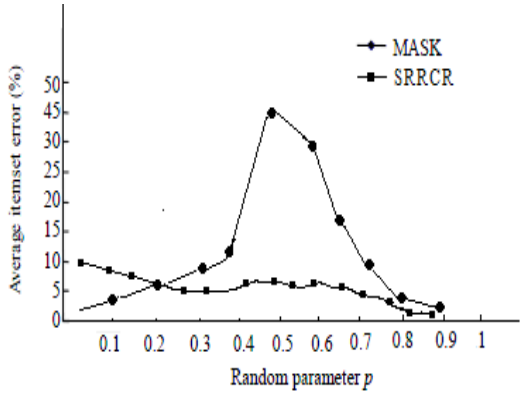

Fig. 2 Average itemsets error of SRRCR method and MASK method

As can be seen, MASK method has wide error by comparison. When the P is close to 0 or 1, the mining results is more accurate, but the intensity of privacy protection is poor. When the value of P gradually approaches 0.5, privacy protection intensity has been improved, but the accuracy of mining results will be significantly decreased. Error changes of the method SRRCR proposed in this paper is relatively stable, according to the P value, that is, when the proportion of real data vary from 0 to 1, the privacy damage coefficient will increased from 0 to 1 simultaneously, so it lead to the decrease of the intensity of privacy protection, and continuously improve the accuracy of mining results.

VI. CONCLUSIONS

Privacy preserving in web data mining has arisen worldwide concerns with the promotion of network technology and the demand of application. But there are many drawbacks and open questions.

In this paper, we have presented a method for privacy preserving mining of association rules based on web usage mining. First, we gave the general framework for mining association rules in the web usage mining, generated session sets by exploring user sessions and transfer session sets to relation two-dimension table. Second, we proposed secondary random response column replacement (SRRCR), a simple and effective privacy preserving algorithm, and achieve privacy protection association rule mining based on SRRCR. Finally, we presented experimental results that validated the algorithm (SRRCR) in practice by simulation.

In the future, we will enhance the efficiency of mining algorithm further by parallelization and other methods, and combine SRRCR with other privacy preserving ways to achieve more significant improvements in terms of privacy, accuracy, efficiency, and applicability.

ACKNOWLEDGMENT

The study is funded by key project of Shanghai Scientific Committee, subject number: 08dz1204802 .

REFERENCES

[1] Rizvi SJ, Haritsa JR. Maintaining data privacy in association rule mining. In: Bernstein PA, Ioannidis YE, Ramakrishnan R,Papadias D, eds. Proc. of the 28th Int'l Conf. on Very Large Data Bases. Hong Kong: Morgan Kaufmann Publishers, 2002, pp. 682-693.

[2] Agrawal S, Krishnan V, Haritsa JR. On addressing efficiency concerns in privacy-preserving mining. In: Lee YJ, Li JZ, Whang KY,Lee D, eds. Proc. of the 9th Int'l Conf. on Database Systems for Advanced Applications. LNCS 2973, Jeju Island: Springer-Verlag,2004, pp.113-124.

[3] Evfimievski A. Randomization in privacy preserving data mining. SIGKDD Explorations, 2002,4(2), pp. 43-48.

[4] Saygin Y, Verykios VS, Clifton C. Using unknowns to prevent discovery of association rules. ACM SIGMOD Record, 2001,30(4),pp. 45-54.

[5] Oliveira SRM, Zaiane OR. Privacy preserving frequent itemset mining. In: Clifton C, EstivillCastro V, eds. Proc. of the IEEE Int'lConf. on Data Mining Workshop on Privacy, Security and Data Mining. Maebashi: IEEE Computer Society, 2002, pp. 43-54.

[6] Zhao JK. Theory and Methods of Sampling Design in Statistical Survey. Beijing: China Statistics Press, 2002 (in Chinese).

[7] Sweeney L Achieving k-anonymity privacy protection using generalization and suppression[J] . International Journal off Uncertainty. Fuzziness and Knowldege-Based Systems. 2002. 10(5), pp. 57l-588

[8] Gordon S. Linoff Michael J.A.Berry Mining the Web: Tansforming Customer Data into Customer Value[J]. Publishing House of Electronics Industry. 2004, pp. 32-37

[9] R.Agrawal and R.Srikant. Fast Algorithms for Mining Association Rules. In Proc.of the 20th Intl.Conf.on Very Large Data Bases(VLDB'94),1994, pp. 487-499