Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A PRIVACY PRESERVING DATA MINING SCHEME BASED ON NETWORK USER'S BEHAVIOR

[1]LI-FENG WU, [2]JIAN XIAO
[1] Department of Computer Science and Technology, South Central University for Nationalities, Wuhan, China, 430073
[2]Hunan University of Commerce Beijin College, Changsha, China, 410219
Email: wulifeng2009@163.com

ABSTRACT

The privacy preserving data mining has become a research hot issue in the data mining field. The server log of the Web site has preserved the page information visited by users. If the page information is not protected, the user's privacy data would be leaked. Aiming at the problem, it discusses the privacy preserving problem based on the user's behavior in the Web data mining, and then introduces a method which converts the Web server log information into the relational data table, produces the information which disturbs the data sheets with a randomized responding method. It offers the data users a discovery algorithm of the frequent itemsets and the strong association rules, then it receives the real private association rule among the online shopping basket goods. The experiments prove that the introduced privacy of the privacy preserving association rule mining algorithm in the Web data mining is good, and it has definite applicability.

Key words: *Data Mining, Conversation Identification, Privacy Preserving; Association Rule, Web Log*

1. INTRODUCTION

The computer network's high openness has caused the great trouble to the user privacy preserving. According to the statistics, 52％ of the web users give up shopping online for they are afraid of their individual privacies. What's worse, they have the anti-shopping online sentiment; 86％ of netizens are very or a little worried about their privacies acquired by others. Therefore, the privacy preserving problem has become the important obstacle to puzzle the network development.

Due to the rise of the data warehouse and data mining technology, driven by profits, several enterprises process utilize the individual data without limit in recent years, and it violates the individual privacy. In the electric commerce, it is called a secondary exploitation and utilization to the individual data, namely, the individual data collected by the merchants from the internet would be stored in the specialized data base. The information with the commercial value collected from the data mining methods is used in the production and management processing. For example, if a consumer had purchased the fitness equipment, the sports dress, the sports shoes and other commodities on the internet, but the purchasing intentions are not disclosed during the shopping process. At this time, according to the consumer's previous purchasing records, the merchants would track and analyze it in terms of the established data base, and then they would send the advertise e-mails related to the sports series products. Worse still, several merchants would hawk all kinds of the diet tea, the diet pills and other products. It seriously disturbs user's daily life, that is, it is the secondary exploitation and utilization to the individual data.

In the Web sites, all page records browsed by users are preserved in the Web server log without reservation. These browsing records reflect user's purchasing habit, purchasing ability and other behaviors. If it intends to protect the web user's behavior from being revealed, it should consider the privacy preserving problem in the Web data mining.

The privacy preserving association rules mining try to find its frequent itemsets accurately under the inaccurate visiting original transaction sets condition. The support degree and the confidence degree produced by the association rules are not lower than the definite threshold value respectively. Therefore, its privacy and its accuracy is a contradiction. At present, according to the basic strategy, the privacy preserving data mining method is mainly divided into two classifications: data interference[1-3] and query limitation[4-5]. Rizvi and Haritsa introduce a representative privacy preserving association rule mining method MASK (Mining Associations with Secrecy Konstraints)

[1]. It is the application of one kind of the data interference strategy. With the Probability P，this method makes the item value remain the same, and the opposite with 1-P. If the item is changed from 1 to 0, it means that it deletes the items; otherwise, it adds the noise items. In fact, it deletes, adds or keeps the items in the data collection unchanged situation with the certain probability, and then it protects the information in the original data collection. However, it uses the Warner model in the statistics. All changed data is directly related to the real original data, and it makes the protecting degree of the privacy data not ideal. The choices of the randomized parameters are limited, that is, it must deviate 0.5[6].

The paper discusses the privacy preserving problem based on the user's behavior in the Web data mining, and then introduces a method which converts the Web server log information into the relational data table, produces the information which disturbs the data sheets with a randomized responding method. It offers the data users a discovery algorithm of the frequent itemsets and the strong association rules, then it receives the real private association rule among the online shopping basket goods.

2. DATA PRE-PROCESSING

Each click in the Web sites is recorded in the Web log. The original Web logs include abundant browsing information. The logs can be produced automatically by the pictures and scripts in the web pages. The Web spiders or agents used in the network searching engine can also produce the accessing logs. These logs are useless to analyze the user's browsing situations, so it should be filtered. In addition, all users' browsing logs are mixed in the log sequence, and the logs should be divided into groups in terms of the users, namely, user's identification. As to the same users, there are many browsing times to the Web sites, so it should be divided into conversation groups in the whole browsing logs of the same users. All of them belong to the data pre-processing work. Specifically speaking, the data preparations before mining are as follows[7]:

2.1 Filter

The logging data capacity of the original network servers is amazing. Thousands of computers are at the service of the web pages. The different requests of the same users can be dealt with by the different servers and be recorded by the different methods. These records have been merged before the click stream is not meaningful. Once it collects data, the primary work is filtering the surplus records, which can make preparations for the analysis.

2.2 Anti-spidering

The so-called spider belongs to the indexed semi-automatic procedure which is established by the searching engine to the world-wide web scanning. The behavior of the spider is different from the human's behavior. The behavior of the spider should be distinguished from the user's behavior in the data processing, and you should filter the records received from the spider's behavior in the server.

2.3 User confirmation

It must identify the users before the conversation. On the one hand, it should identify the page requests sent by the same users whose purpose is to establish the conversation in a browsing; on the other hand, it should identify the same users in many browsing, and then it can analyze the user's behavior in a few days, a few months and a few years. At present, the user's usual identification methods are IP address identification, embedding SessionID, Cookies and so on. The IP address identification assumes that each address corresponding with a user is the most easy behavior, but there is a situation in which many users have the same computer and surf the internet by the agent IP, so there is the largest deviation. The embedding SessionID is often used in the electric commerce which mainly records the goods in the user's shopping baskets. The ID number is produced by the use of the dynamic methods. When the embedding is in the user's browsing requests, that is, the same user's requests remark the same SessionID number during a period of time. Currently, the paper adopts the embedding SessionID technology.

2.4 Conversation Identification

Conversation refers to all Web pages visited by the users in a browsing. It can reflect browser's interest in the web sites by conversations. In order to establish conversations, it should find out all page requests from the same users, and then divide them into the conversation groups by the enlightenment method.

2.5 Routing completion

When users browse the web sites, it may appear the page back phenomenon, and then it causes the routing loss, so it should make a illation and complete the browsing route in terms of user's before and after pages.

3. THE PRIVACY PROTECTION IN THE WEB APPLICATION MINING

It would extract data related to the mining in the Web server logs after the data pre-processing. In order to analyze it, the data needs to be conversed and aggregated in the different extraction layers. The most basic layer of the data extraction is the page browsing in the Web log record mining. A page browsing shows the collection of the Web objects which affect the browser pages produced by user's behavior. In the user's layer, the most basic layer extracted by the behavior is conversation.

3.1 User's Conversation Exploration Method In Face Up With Time

Considering the table 1 case, each line represents a user's conversation in the table, and identifies it by the ID. The IP represents user's address. Different addresses represent different users. URL represents the pages browsed by users in the conversation. Time represents the user's browsing time. The ideal user routing exploration method can rebuild user's browsing real procedure in the conversation[8]. It adopts the user conversation exploration method used in the object orientation.

Table 1 Original Conversation Case

ID	IP	UPL	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

The vector M{M1， M2， …， Mi， …， MN}represent the log data sets, including Mi={IPi， URLi， Datei， Timei， Methodi， Codei， Bytesi}, N is the number of the log files. Web log mainly contains the IP address. The browsed URL, the browsing data and time, the browsing method, the browsing structure, the size of the browsing information and so on. The vector S{S1，S2，…， Si， …， Sk} represent the conversation data sets, k is the number of conversation, including Si={IPi，Pagei，ti}. Each conversation preserves the IP address, browsing pages and timestamp, Pagei={Pagel，Page2，…，Pagei，…，PageL), L is the number of the pages in the conversation.

```
Algorithm 1
For each Mi of M
  If Methodi is 'GET' AND Urli is 'WEBPAGE'
    If Sk ∈ public conversation sets and IPk=IPi
      Set to for the first requested timestamp in the conversation
      If(ti-t0≤20 min)// set the time limitation reach to 20 mins
        Sk=(IPk, Pagesk U Pagesi, tk)
      Else
        Delete Sk from public conversation sets
        Add Sk into S
      End if
    Else
      Add Sk into S
    End if
End for
```

Algorithm 1 generates the conversation sets to the user's conversation in the table 1, seen in the table 2.

Table 2 User's Conversation Sets Generated In The Algorithm 1

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

3.2 Transform The User's Conversation Sets Into The Two-Dimensional Relation table.

As to the user's conversation sets in the table2, it can use the two-dimensional relation table to deposit user's conversation route. The structure of the table is (TID，IPID，SESSIONID，Pi) for each conversation just preserve the IP address and the browsing pages. Therefore, it can deposit user's IP address, conversation number and the browsing situations of pages in the two-dimensional table. If the value is 1 in the Pi line, it shows that users have browsed the page and purchased the goods in the page. If the value is 0, it shows that users have not browsed the page.

The following is algorithm 2, transforming user's activation records into the two-dimensional relation table easily solved by the privacy protection.

Algorithm 2

Input: user conversation set S

Output: protect the two-dimensional table S table in the user conversation route

```
Create table S_table(TID，IP，SESSIONID，Pi)
For each Si of S
Insert into S_table(IPID，SID)values(IPi，SIDi)
If there is Pagei of Si
Insert into S_table(Pi) Values('1');
Else
Insert into S_table(Pi) Values('0');
End for
```

Regarding each conversation as an affair, the user's conversation order can be transformed into the following relation table. If you have purchased the goods of the corresponding page number, the value is 1; otherwise, the value is 0.

Table 3 adds a new field TID. The field is identity column, and the system can be automatic assigned, and it represents the affair number.

Table 3 Preserve The Two-Dimensional Table Of User's Conversation Route

TID	IPID	SID	P_1	P_2	P_3	P_4	P_5
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

3.3 Privacy preserving algorithm

It needs conducting the privacy preserving in the user route of the two-dimensional table generated in the previous chapter, and the row are P1、P2 and P3. These rows are called the sensibility in the privacy preserving technology[9]. In here, the sensitive attribute values can not reveal its privacy, but the connection of these sensitive attribute values can disclose user's purchasing habit and ability. Therefore, the sensitive attributes are not protected by adopting the traditional anonymity method, but they are protected with the use of improving randomized answer method. In order to improve the data transforming efficiency, it conducts the secondary randomization in term of the certain probability to the x. The algorithm is called the secondary randomization answer row replacement algorithm 3——SRRCR

Algorithm 3

Give the definite randomized parameter I, 0≤p1，p2，p3，p4≤1, and there is p1+p2+p3+p4=1;

Give the definite randomized parameter II，0≤r1，r2，r3，r4≤1, and there is r1+r2+r3+r4=1;

Construct the randomized function R(x). Transform the original row C(i，j), and then hide F(i,j) after transformation. In the F(i,j)∈{0,1)， C(i,j)∈{0,1}, i represents i tuples in the data table, j represents j attributes.

As to the x=F(i,j)∈{0,1), the transformation of the randomized function seen in the table4.

Table 4The Changing Diagram Of The Randomized Function

	P_1	P_2	P_3	$P_{445555555}$
χ	χ	0	1	y
	r_1	r_2	r_3	r_4
y	χ	0	1	null

It sets i as an item, π represents i support degree in the C, and πnull represents the null support degree in the C. λ represents i support degree in the F. It assumes that the affair Tin the C is processed by the SRRCR method, the affair T will be transformed into the affair T', the formula is as follows:

$$\pi=\lambda\times(p_1+p_4\times r_1)+p_3+p_4\times r_3 \qquad (2)$$

The formula is received by the formula (1) and (2):

$$\lambda=\frac{\pi - p_3 - \pi_{null}\times\frac{r_3}{r_4}}{p_1+\frac{r_1}{r_4}\times\pi_{null}}=\frac{\pi - p_3 - k\times\phi}{p_1+y\times\phi} \qquad (3)$$

$k=\frac{r_3}{r_4}$ is the escalating factor, $y=\frac{r_1}{r_4}$ is the descending factor, $\phi=\pi_{null}$ is the null support degree.

$$\pi_{null}=p_4\times r_4 \qquad (1)$$

It shows that the advantage of the algorithm is to bring in the adjustment factors in the process of counting the support degree. It can further add the privacy protection by using a null or an illegal value in the actual scrambling procedure.

Table 5 Data Mapping Probability In The SRRCR Algorithm

No	*C(i, j)*	*F(i, j)*	Probability
1	0	0	$p_1+ p_2+ p_4\times(r_1+r_2)$
2	0	1	$p_3+ p_4\times r_3$
3	1	0	$P_2+ p_4\times r_2$
4	1	1	$p_1+ p_3+ p_4\times(r_1+r_3)$
5	0	null	$p_4\times r_4$
6	1	null	$p_4\times r_4$

Take the previous relation table for example, if the support degree threshold value is 0.4 in the M, it chooses P1=0.2， P2=0.2， P3=0.1， P4=0.5 and r1=0.3，r2=0.2，r3=0.2，r4=0.3 according to the user conversation sets transformed by the SRRCR algorithm, seen in the table 6 to table 8.

The pseudo affair set transformed by the SRRCR algorithm not only includes the values: 0 and 1, but also the null value. Null shows the user's interest to the page's goods, and also shows that users have browsed the page without purchasing any goods. It counts the null rows in the following relation rule mining algorithm, and obtains the potential useful relation rule. It proves that the i itemsets support degree in the pseudo affair sets and the j itemset support degree in the original affair sets are the same after being transformed by the SRRCR algorithm.

3.4 The Relation Rule Mining Algorithm Of The Privacy Protection

It improves it on the basis of the Apriori algorithm [10]. Improving the Apriori algorithm row by row can realize the purpose of mining the relation rule of the pseudo affair sets transformed by the SRRCR algorithm. By giving the minimum support degree minsup and the data set, the mainly specific algorithm steps are as follows:

Algorithm4 improves the data processed by the Apriori algorithm in the opposite sequence, and the data generates the frequent itemsets.

Build the exhaustive itemset φ in the affair set, the φ includes itemsets whose total number is n probable itemsets M_n,$n=\sum_{i=1}^{k}C_k^i$, k is the maximum affair number.

Table 6 Data Set Added In The Pseudo-Line

SID	P_1	F_1	P_2	F_2	P_3	F_3	P_4	F_4	P_5	F_5
1	1	1	1	null	0	1	1	1	0	null
2	1	1	0	1	1	0	1	1	1	0
3	0	null	0	0	0	null	0	0	1	1
4	0	1	1	1	1	0	1	1	0	1
5	1	0	1	0	0	0	1	null	1	1

Table 7 Replace The Line Information

Raw	Fake
p_1	F_2
p_2	F_3
p_3	F_1
p_4	F_5
p_5	F_4

Table 8 Ultimate Generated Pseudo-Data Set

SID	P_1	P_2	P_3	P_4	P_5
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

(2) It searches the maximum itemsets row by row, and matches each line with all itemsets. If several itemsets are matched with the row, it adds 1 in each position of the φ sequence.

(3) It partly adopts the Apriori pruning principle in the matching itemsets procedure.

(4) When the proportion of the searching lines are reaching to proportion of the support degree, check the ratio k between the exhaustive set series φ and the whole rows m. When the situation is k<l-2×(1-minsup), the corresponding itemsets of the value position in the φ is not satisfied with the minimum support degree, it should give up counting the value in the position. Later, each cycle will scan the remaining lines and count whether the remaining itemsets data is 1, namely, the whole itemsets is 1. It is impossible to regard it as minsup. If it was, it should give up counting the value in the position.

(5) Obtain the whole itemsets satisfied with the minsup request

The counting procedure is as follows:

Input: the affair set D’ handled by the SRRCR method, the threshold value of the minimum support degree is s.

Output: the frequent itemset L in the D.

Set φ, φ={M1，M2，⋯，Mn}={(L1)，(L2)，⋯，(Lk), (L1,L2),⋯(L1,Lk)} is the whole set of the subset based on the I itemset,

```
for (i=n; t>0; i--) , where n=$\sum_{i=1}^{k} C_k^i$
set Ai=index(i) // set up the searching matrix matched with φ(i)
for each transaction t∈D’ //scan all lines
Ri=Ci×Ai
if Ri≠0,then //there is line match
if i=n
 (c.count=c.counto+l ， c.countl=c.count1+1 ⋯ c.countn=c.countn+1) //give each counting template
else
select_add c.count // chose the factorial method
end if
end if
end for
Fk←{c ∉Ck c.count/n⩾minsup}
end for
return Fk←UkFk
```

When a certain value is 0 in the select_add function, it should prune the item in the previous generated frequent itemsets (it will not explain it the same as the pruning principle on the Apriori algorithm). The maximum affair number after being pruned is p=$\sum_{i=1}^{k\text{-}u} C_{k\text{-}u}^{i}$ （u is the element number of the 0 item in the line）. It counts it after pruning each item in the (k-i) itemsets, that is, c.count=c.count0+1 ， c.countl=c.countl+1…c.countp=c.countp+1. After obtaining the frequent itemsets, it counts the Support(A→B)=Support_count(AUB)/all_count in terms of rage support degree and the confidence degree. Data provider can count the relevant degree among the shopping basket goods.

4. ALGORITHM ANALYSES

4.1 The complex degree analysis of the relevancy algorithm

In theory, the Apriori algorithm is the exponential order algorithm. It assumes that the size of the affair set is Ⅳ, it contains n items in each affair on average, and the scale of the collection space of the whole itemsets will reach to O (2”). When the situation is k<l-2× (1-minsup), the corresponding itemsets of the value position in the φ is not satisfied with the minimum support degree in the improving mining algorithm, it should give up counting the value in the position. Later, each cycle will scan the remaining lines and count whether the remaining itemsets data is 1, namely, the whole itemsets is 1. It is impossible to regard it as minsup. If it is, it should give up counting the value in the position.

The complex degree of the time algorithm: if it does not consider the line optimizing, the complex degree of the time algorithm is $O(N\times\sum_{i=1}^{n/2}C_n^i)=O(N\times 2^{n/2})$; if it considers probability prognosis in the line judging procedure, the complex degree will descend.

4.2 Experimental results

It obtains the web server log files operated by a certain Web site for a month successively. It transforms into the affair sets F by the algorithm 1 、 2. The affair number is 1000, the total item number is 11 and the average affair length ATL is 2.8. It analyzes it with the minimum support degree is 3%. The table 9 shows the minimum support degree is 3%. It shows that the algorithm privacy preserving situation in the different values of the randomizing parameters.

Table 9 The Privacy Preserving Degree In The Different Parameters

No	I(p_1, p_2, p_3, p_4)	II(r_1, r_2, r_3, r_4)	Privacy ratio
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

figure 1 is the contrast effect when SRRCR method and MASK method conducts the privacy preserving association rules mining, and explains the relations between the accuracy of the data privacy and mining result and the random parameters

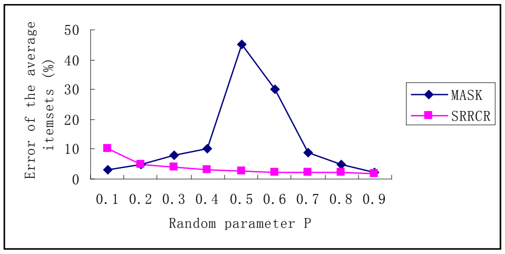

Figure 1 Error Figure Of The Average Itemsets Between The SRRCR Method And MASK Method

From the above figure, it shows that the error transformation in the MASK method is relatively large. When P is approaching 0 or 1, the mining result is comparatively accurate, but the protection degree to the privacy is bad; When P is from 0 or 1 to 0.5, privacy damaging factor is gradually diminishing, and the protection degree to the privacy is gradually improving, but the accuracy of the mining result will descend remarkably. The SRRCR method introduced by the paper shows that the error transformation in the SRRCR method is relatively stable. With the P value, the method to the privacy preserving degree is gradually descending with the proportion of the real data from 0 to 1 and the privacy damaging factor from 0 to 1, and the accuracy of the mining result is gradually improving.

5. CONCLUSIONS

The paper discusses the privacy preserving problem in the user's behavior of the Web sites, and introduces a kind of the secondary randomized answer column replacement algorithm SRRCR, which mainly transforms and hides the data. Then it puts forward a frequent itemsets generating algorithm whose data is handled by the SRRCR method. Therefore, it realizes a new privacy preserving relational rule mining method. In the future, it will improve the operating efficiency of the mining algorithm and extends the range of solving problems by using the good application of the SRRCR method.

REFERENCES:

[1] Rizvi S J, Haritsa J R., "Maintaining data privacy in association rule mining", *Proceedings of the 28th Int'I Conf on very Large Data Bases*. Hong Kong: Morgan Kaufmann Publishers, 2002: 682-693.

[2] Evfimievski, A., Srikant, R., Agrawal, R. & Gehrke, J., "Privacy Preserving Mining of Association Rules", *Proceedings fo 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton*, AB, Canada, 2002, pp. 217-228.

[3]Evfimievski A., "Randomization in privacy preserving data mining", *SIGKDD Explorations*, Vol. 4, No. 2, 2002, pp. 43-48.

[4] Sayfin Y. Verykios V S, Clifton C., "Using unknowns to prevent discovery of association rules", *ACM SIGMOD Record*, Vol. 30, No. 4, 2001, pp. 45-54.

[5] Oliveira S, Zaiane O., Privacy preserving frequent itemset mining Clifton C, Estivill Castro V", *Proceedings of the IEEE Int'l Conf on Data Mining Workshop on Privacy, Security and Data Mining*, IEEE Computer Society, 2002: 43-54.

[6]Zhao J K., "Theory and methods of sampling design in statistical survey", Beijing: China Statistics Press, 2002.

[7]Linoff G S, Berry M J A., "Mining the Web: transforming customer data into customer value", Publishing House of Electronics Industry, 2004, pp. 32-37.

[8] Ngai, E. W. T., Xiu, L., & Chau, D. C. K., "Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification", *Expert systems with applications*, Vol. 36, 2009, pp. 2592-2602.

[9 Thomas, L. C., "Consumer finance: challenges for operational research", *Journal of the Operational Research Society*, Vol. 61, 61, pp.41-52.

[10]Agrawal R, Srikant R, "Fast algorithms for mining association rules", *Proceedings of the 20th Intl Conf on very Large Data Bases(VLDB'94)*, 1994, pp. 487-499