IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555
Vol. 2, No.2, April 2012

A Fast Web Transaction Pattern Mining Algorithm

Zaiping Tao
Information Technology Department
Zhejiang Financial College
Hangzhou, China

***Abstract*—In this paper, we explore a new algorithm for obtain valuable web transaction patterns from EC environment. The method is based on the pattern-growth framework. During the process, both the traversing and purchasing behaviors of customers are considered simultaneously. Different prune methods are adopted to increase the mining speed and accuracy, the web site structure is used to optimize the algorithm. The synthetic data is generated to validate our method. The experimental results show that the algorithm WTPM is correct and efficient.**

Keywords-Data mining; Web traversal pattern; Web transaction pattern; pattern-growth framework

I. INTRODUCTION

Due to the rapid growth in the field of electronic commerce (EC), a huge amount of web data has been gathered in many EC systems. How to extract useful information and knowledge efficiently from such huge amount of web data has already been the important issue at present. Web mining [1,2,3] refers to apply data mining techniques in large amount of web data to improve the web services. Web traverse patterns mining[4,5,9,10] and web transaction patterns mining[6,7,8] are two important reasearch topic in web mining issue. Web traverse patterns mining is to find out most of users' access path from web logs. It was initiated by Chen[1] in 1998. The result is valuable to improve the website design, such as provide efficient access between highly correlated objects, better authoring design for web pages, and provide navigation suggestions for web users, etc. However, in EC enviroment, it is important to find out the association rules between purchasing items, which can be used to improve the cross-sellings. Actually the web site managers focus on not only the browsing behaviors but also the purchasing behaviors of customers, that is, combine the web traverse patterns and association rules mining to find out the right information, which is the problem of web transaction pattern mining. It was first raised by Yun and Chen[6] in 2000.

There are several algorithms to deal with the problem of web transaction patterns mining. MTS[6] which is based on Apriori framework, is proposed to deal with both the browsing and purchasing behaviors together at the same time. It cuts web transaction into several web transaction records. And it discovers the web transaction patterns based on these web transaction records, while it maybe lead to inaccurate patterns since the backward references are ignored during the transform from the web transaction into web transaction records. IPA[7] is another algorithm to mine the web transaction patterns without the transform of web transaction records. However it can only process the situation that customers buy one item in a web page, which is not in accord with the real situtation of EC enviroment. Most of proposed algorithms are based on Apriori-like framework, which will generate a large amount of candidate, keeping and counting these candidates is time and space consuming. WTP-4[8] is an algorithm based on pattern-growth framework, actually we can improve the mining efficiency with some modification to it, such as ignoring the infrequent pages and items during the process, as well as sharing with the common projected sub-database etc.

In this paper, we explore a new algorithm called WTPM to mine the web trasanction patterns, which is based on the pattern-growth framework. Some prune methods are adopted to increase the mining speed. The experimental results show that WTPM algorithm is correct and effective.

The rest of paper is organized as follows. Preliminaries are given in section 2. WTPM algorithm is described in section 3. The experimental results are introduced in section 4. And the paper concludes in section5

II. PRELIMINARIES

Let $I=\{i_1,i_2,...,i_n\}$ be a set of items sold in the EC system. Let $W=\{w_1,w_2,...,w_m\}$ be the set of all web pages in an EC environment.

Definition 1 Let $S=\langle w_1,w_2,...,w_p: s_1\{i_1\},s_2\{i_2\},...,s_q\{i_q\}\rangle$ be a transaction pattern, where $i_m \subseteq I$ for $1\le m\le q$, and $\{s_1,s_2,...,s_q\} \subseteq \langle w_1,w_2,...,w_p\rangle \subseteq W$. Then, $\langle w_1,w_2,...,w_p: s_1\{i_1\},s_2\{i_2\},...,s_q\{i_q\}\rangle$ is said to pattern-contain a transaction pattern $\langle n_1,n_2,...,n_y: r_1\{i_1\},r_2\{i_2\},...,r_x\{i_x\}\rangle$ if and only if $\langle w_1,w_2,...,w_p\rangle$ contains $\{n_1,n_2,...,n_y\}$ and $\{s_1\{i_1\},s_2\{i_2\},...,s_q\{i_q\}\}$ contains $\{r_1\{i_1\},r_2\{i_2\}, ... ,r_x\{i_x\}\}$.

Definition 2 A web transaction is said to pattern-contain $\langle n_1,n_2,...,n_y: r_1\{i_1\},r_2\{i_2\},...,r_x\{i_x\}\rangle$ if one of its web transaction records pattern-contains $\langle w_1,w_2,...,w_p: s_1\{i_1\},s_2\{i_2\},...,s_q\{i_q\}\rangle$.

A traverse sequence $S=\langle w_1,w_2,...,w_p\rangle$, $(w_i \in W, 1\le i\le p)$ is a list of web pagses which are ordered by traverse time. If there is a link from w_i to w_{i+1} in the web site structure, then the traverse sequence S is a qualified traverse sequence. A user web transaction sequence $S=\langle w_1,w_2,...,w_p: s_1\{i_1\},s_2\{i_2\},..., s_q\{i_q\}\rangle$, $(w_i \in W, 1\le i\le p, i_j \subseteq I, 1\le j\le q)$ and $\langle w_1,w_2,...,w_p\rangle$ is a

user traverse sequence. In fact, a user web transaction sequence is a traverse sequence with purchasing behaviors, which includes the navigation and purchasing behaviors of customers.

The support count of a traverse sequence *a* is the number of user traverse sequence which contains *a*. The support of a traverse sequence *a* is the ratio of user traverse sequences which contains *a* to the total number of user purchasing sequence in DB, denoted by *Sup(a)*. The support count of a web transaction sequence β is the number of user web transaction sequence which contains β. The support of a web transaction sequence β is the ratio of user web transaction sequences which contains β to the total number of user web transaction sequences in DB, denoted by *Sup*(β). A traverse sequence is a web traverse pattern if it is a qualified traverse sequence and *Sup(a)*>= S_{min}, in which S_{min} is the user specified minimum support threshold. A web transaction sequenceβ is a web transaction pattern if it is a qualified traverse sequence and contains at least one purchase item , as well as *Sup*(β)>= S_{min}.

For instance, in Table 1, if the S_{min} is set to 50%, then the support of < B{5,7}C>=3/6>=50%, and there is a link from "B" to "C" in the web site structure shown in figure1. Therefore, <B{5,7}C> is a web transaction pattern. The reason for using web site structure is to avoid the testing of unqualified web traverse sequences in the process. In fact, those unqualified web traverse sequences or web transaction sequences are meaningless.

TABLE I. SAMPLE DATABASE DB

Tid	Web Transaction Sequences
1	AB{2,5,7}C{3}A
2	B{5,7}CA
3	CAB
4	B{5,7}CAD
5	AC{1,3}
6	A{1,2}DE{3}C{2,3}

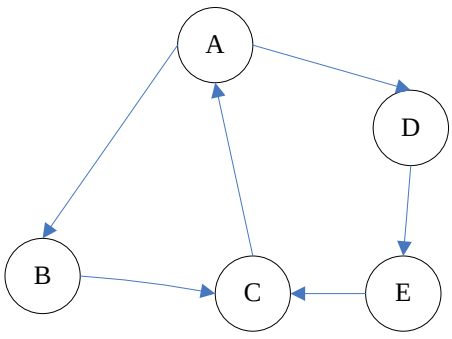

Figure 1. Web site structure

III. WTPM ALGORITHM

In the following, we introduce our new method to discover the web transaction pattern, it is based on the pattern-growth framework. It takes both the traversing and purchasing behaviors of customers into consideration simultaneously. The website structure is considered to decrease the unnecssary check to unqualified web pages. In our method, the forward references as well as the backward references are allowed. In the mining process, all candidate web 1-transaction sequences are generated and then scan the database to calculate the support of candidate sequences. All the frequent web 1-transaction sequences will be generated and correspond projected sub-database will be produced.

WTPM Algorithm: Web Transaction Pattern Mining
Input: A web purchase sequence database DB
The minimum support threshold S_{min}
The web site structure W
Output: All the web transaction pattern WTP
Begin:

1. k = 1, i = 1
2. get all web 1-transaction sequence with length 1 from DB
3. Scan the DB once to get all frequent web 1-transaction sequences
4. ignoring all the infrequent pages and items from DB
5. reconstruct the web site structure W according to frequent web pages
6. construct their projected sub-database for each frequent web 1- transaction sequence db_i
7. while db_i has the next link page
8. {
9. k++
10. calculate the support of next pages *c*
11. if Sup(*c*) >= S_{min} then
12. i++
13. add the web transaction sequence pattern *c* with length k into WTP
14. if not exist the projected sub-database of *c* , construct their projected sub-database db_i
15. else
16. share with the common projected sub-database
17. }

End

We use the example database shown in Tablel and a simple web site structure shown in Figure1 to discover the final mining result of WTPM method. The minimum support threshold is set to 30%.

Step1: At first, we scan the database to calculate the support count of candidate web 1-transaction sequences, which include <A>, <B>, <C>, <D>, <E>, <A{1}>, <A{2}> ,<A{1,2}>, <B{2}>, <B{5}>, <B{7}>, <B{2,5}>, <B{2,7}>, <B{5,7}>, <B{2,5,7}>, <C{1}>, <C{2}>, <C{3}>, <C{1,3}>, <C{2,3}>, <E{3}>. After calculating the support count, we get all frequent 1-transaction sequence pattern are <A>, <B>, <B{5}>, <B{7}>, <B{5,7}>, <C>,<C{3}>. For

each frequent 1-transaction sequence pattern, we build its projected sub-database after ignoring all the infrequent pages and items shown in Table2.

TABLE II. PROJECTED SUB-DATABASE OF FREQUENT 1-TRANSACTION SEQUENCES

Prefix	Postfix database
A	[B,B{5},B{7},B{5,7}][C,C{3}]A B [C,C{3}] [C,C{3}]
B	[_{5},_{7},_{5,7}] [C,C{3}]A [_{5},_{7},_{5,7}]CA [_{5},_{7},_{5,7}]CA
B{5}	[_{7}][C,C{3}]A [_{7}]CA [_{7}]CA
B{7}	[C,C{3}]A CA CA
B{5,7}	[C,C{3}]A CA CA
C	A A A
C{3}	A

Step2: The postfix database of <A> are "[B,B{5},B{7}, B{5,7}][C,C{3}]A", "B", and "[C,C{3}]". Since there is no direct link from page "A" to page "C", which means the traverse sequence <AC> is not a qualified sequence. We need not generate the postfix with the prefix of <AC>. The postfix database of <B> are "[_{5},_{7},_{(5,7)}][C,C{3}]A", "[_{5}, _{7},_{(5,7)}]CA", "[_{5}, _{7},_{(5,7)}]CA". According to the web site structure shown in figure1, there is a direct link from page "B" to page "C", which means the traverse sequence <BC> is a qualified sequence, we have to construct the projected-sub database for <BC>. The postfix database of <B(5)> are "[_{7}][C,C{3}]A", "[_{7}]CA", "[_{7}]CA", after calculating the support count of postfix database, we can get all the frequent 2-web transaction patterns in Table3 and construct the related projected database. As for the prefix <B(5,7)C>, we can find that the projected database is already built after <BC> is processed, which means we can use the common projected sub-database.

TABLE III. PROJECTED SUB-DATABASES OF FREQUENT 2 WEB-TRANSACTION SEQUENCES

Prefix	Postfix database
AB	-
BC	A A A
B{5}C	A A A
B{7}C	A A A
B{5,7}C	A A
CA	A -

Step3:After calculating the support count of postfix database, we can get the frequent 3-web transaction patterns in Table4, which are "BCA","B{5}CA","B{7}CA"and "B{5,7} CA" . We stop the mining process since there is no frequent 4-sequence can be generated.

TABLE IV. PROJECTED SUB-DATABASES OF FREQUENT 3 WEB-TRANSACTION SEQUENCES

Prefix	Postfix database
BCA	-
B{5}CA	-
B{7}CA	-
B{5,7}CA	-

IV. EXPERIMENTAL RESULTS

In this section, we present our performance study in comparison with the IPA and WTP-4 algorithm over various datasets. The main purpose of this experiment is to demonstrate how effectively the WTPM works. First, we show the efficiency in terms of runtime of the IPA and WTPM algorithm. Second, we show that WTPM has good scalability against the number of sequence transactions in the datasets.

TABLE V. SYNTHETIC DATA PARAMETERS

Symbol	Meaning
Pro	Number of products
WP	Number of web pages
|OL|	Average number of out-links per web page
|SP|	Average number of products sold per web page
PSP	Percentage of products sold in the website
|I|	Average size of maximum potentially-frequent sequences
L	Number of potential transaction
|T|	Average size of transaction sequence
D	Number of transactions
S_{min}	Minimum support threshold

Extensive experiments are conducted to assess the performance of the WTPM algorithm utilizing different synthetic data. Synthetic datasets were generated according to that described in [9]. The scalability of the algorithm was also evaluated over different database sizes. The experiments were performed on a 2.2GHz PC with 2048MB memory running the Windows XP.

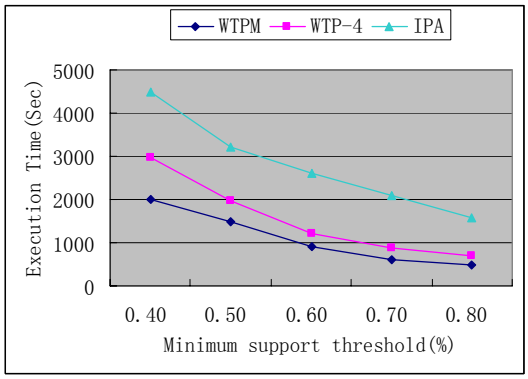

Figure 2. Runtime test results

From the experiment showed in Figure2, WTPM algorithm outperforms IPA with the same minimum support threshold.

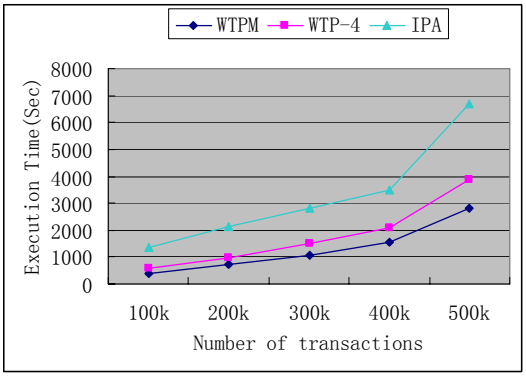

Figure 3. Scalability test results

In Figure3, the dataset is used to test the scalability with the number of sequences in a sequence database. From the performance test, WTPM, as well as IPA has good scalability. The execution time increases linear with the increase of database size.

V. CONCLUSION

In this paper, we examined the issue of mining web transaction patterns that takes both traversing and purchasing behavior into consideration, which is significant to improve the web service, to better design the EC system. To address this issue, we develop an efficient algorithm based on pattern-growth framework. It considers the traversing and purchasing behaviors of customers simultaneously. The comprehensive experiments show that our method is correct and outperforms IPA algorithm. Additionally, the experimental results show that WTPM has good scalability. The mining result is significant to optimize the design of the EC web site structure, as well as the display of items to improve the convenience and efficiency of customers. In the future, we shall improve the process of trimming database and pruning candidates efficiently to speed up the algorithm.

REFERENCES

[1] Ming-Syan Chen, Jong Soo Park, Yu, P.S, “Efficient Data Mining of Path Traversal Patterns In a Web Environment”, IEEE transaction on Knowledge and Data Engineering, Vol.10,No.2, pp.209-221, 1998.

[2] I-Yuan Lin, Xin-Mao Huang, Ming-Syan Chen, “Capturing user access patterns in the Web for data mining”, In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, pp.345-348,1999

[3] Maged El-Sayed, Carolina Ruiz, Elke A. Rundensteiner, “FS-Miner: Efficient and Incremental Mining of Frequent”, In Proceeding of 6th ACM Workshop on Web Information and Data management, pp.128-135,2004

[4] Show-Jane Yen, “An Efficient Approach for Analyzing User Behaviors in a Web-Based Training Environment”, International Journal of Distance Education Technologies, Vol.1, No,4, pp55-71, 2003

[5] Keizo Sato, Akira Ohtaguro, Makoto Nakashima, Tetsuro ITO, “The effect of a website directory when employed in browsing the results of a search engine” International Journal of Web Information Systems, Vol. 1,No.1, pp.43-52, 2005

[6] Ching-Huang Yun, Ming-Syan Chen, “Using Pattern-Join and Purchase-Combination for Mining Web Transaction Patterns in an Electronic Commerce Environment”, In Proceeding of the COMPSAC, pp.99-104, 2000

[7] Yue-Shi Lee, Show-Jane Yen, Ya-Min Chang, Jia-Ching Ying,” New Approaches for Mining Web Transaction Patterns”, In Proceedings of Taiwan Conference on Business and Information, pp.450-456, 2006

[8] I-Yuan Lin, Xin-Mao Huang, Ming-Syan Chen “Mining traveling and purchasing behaviors of customers in electronic commerce environment”, In Proceedings of International Conference on Informatics, Cybernetics, and Systems, pp.1461-1469, 2003

[9] S.J. Yen, Y.S. Lee,”An incremental Data Mining Algorithm for Discovering Web Access Patterns”, International Journal of Business Intelligence and Data Mining, pp288-303,2006

[10] Y.Xiao, M.H. Dunham, “Efficient Mining of Traversal Patterns.” IEEE Transaction on Data and Knowledge Engineering , Vol.39, No.2, pp.192-214, 2001

AUTHORS PROFILE

Zaiping Tao received his B.S. in mechanical design from Jinan university, 1994, the M.S. in mechanical design and manufacturing from Zhejiang university, Hangzhou , 1997 and the PhD degree in computer application from Zhejiang university, Hangzhou, China in 2000. He is an associate professor in information technology department of Zhejiang financial college, Hangzhou ,China. His research interests are in the area of data mining, web service etc .