The Research of Web Users' Behavior Mining Based on Association Rules

Xiaohong Shan
The College of Economics and Management
Beijing University of Technology
Beijing, China

Huamei Sun
School of Management
Harbin University of Technology
Harbin, China

Abstract—**When accessing websites, users usually leave a lot of access information, which can be mined reasonably to help the managers of website to get accessing patterns of users. This article first introduces the preprocessing procedure of web logs, which includes the tasks of data cleaning, Data Discretization and their implementation. On the basis of preprocessing the analysis method of requested resource "URL" and "referrer" which is the webpage before users browse the URL in web log by the use of association rules is proposed to find the accessing patterns of users. Finally the experiment is accomplished. The result shows that the method is feasible, and it can help the manager in making decisions about the analysis of website users' behavior and the optimization of website.**

Keywords- Association Rules; Web Log; Web mining

I. INTRODUCTION

Web using schema mining is one of the research directions in web usage mining. By collecting the web logs data when users interact with the website, the massive data about users' behavior can be mined to get the knowledge of users' accessing behavior and access patterns, thus to improve the website structure and help enterprises in making business decisions. So using web logs data that reflects the detail information of users' access behavior from all aspects effectively has important significance for enterprises in providing intelligent information services to users.

In the research of web usage mining, the methods, such as statistics, cluster analysis [1], association rules [2-3] and genetic algorithm [4] were often used to analyze the web logs data and obtain knowledge of users' accessing behavior. A lot of research focused on doing cluster analysis and association analysis about users' browsing behavior using customer IP, the resource, and the remotehost and so on. The association analysis of the requested resource "URL" and "referrer" which is the webpage before users browse the URL is relatively less, while their relationship can help us to analyze the jump principle among webpages in this website and that between different wetsites, thus provide the important basis for website optimization. So this article aims to find the potential relationship between the webpages that users browse using association rules.

This article research work obtained the State Natural Sciences Foundation project subsidization （70971032）

II. RELATED WORK ABOUT ASSOCIATION RULES

To find the association rules in transaction database was first proposed by Agrawal. R[5], its formal description is as follows:

Let I= $\{i_1,i_2,i_3,\ldots,i_m\}$ be a set of m different items, the element of which is called *item*. Let D be a set of transactions, where each transaction T is an itemset that $T\subseteq I$. Associated with each transaction is a unique identifier, called its *TID*. A set of items $X\subset I$ is called an *itemset*. We say that a transaction T contains an itemset X, if $X\subseteq T$. An association rule is an implication of the form X=>Y, where $X\subseteq I$, $Y\subseteq I$, and $X\cap Y=\varnothing$. The rule X=>Y holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The rule X=>Y has support s in the transaction set D if s% of transactions in D contains XUY.

The classical algorithm of association rules is Apriori algorithm proposed by Agrawal.R in 1993, which is used to analyze the supermarket basket data, where rules like "34% of all customers who buy fish also buy white wine" may be found. The guiding idea of Apriori is to mine frequent itemsets by use of iterative method of level sequential search. Each iteration consists of a two phase process. One is to generate candidate set C_k that tend to bring frequent itemset. The other is to compute support and decide the frequent itemset L_k based on candidate set C_k. Apriori has been improved in different degree in Literatures [6-8], for example, [6] extracted frequent itemset from candidate itemset; [7] used vector operation to implement the count of frequent itemset; [8] proposed the algorithm that integrated top-down and bottom-up search strategy to reduce the number of candidate sets.

III. THE RESEARCH OF WEB USERS' ACCESSING BEHAVIOR

Like most web log mining, web users' accessing behavior mining is also divided into three stages, that is, data collection, data preprocessing, web accessing behavior knowledge mining and analysis.

A. Data collection

In the stage of data collection the web log files are collected at the server side, client side, and proxy servers.

978-1-4577-0536-6/11/$26.00 ©2011 IEEE

B. Data preprocessing

First the webpages that have no correlation with web users' accessing behavior are filtered out, such as pictures, multimedia, getway, Flash. The concrete method is: delete from Weblog where url like '%.gif%' or URL like '%.jpg%' or URL like '%.ico%' or URL like '%.bmp%' or URL like '%.cgi%' or URL like '%.css%' or URL like '%g.swf%'.

Then the activities of "HEAD" and "POST" are filtered out, only the activity of "GET" is preserved. The concrete method is: delete from Weblog where Method ='POST' or Method ='HEAD'.

The third, the failure status of request are filtered out. The concrete method is: delete from Weblog where status like '3%' or status like '4%' or status like '5%'.

Finally, requested resource "URL" and "referrer" which is the webpage before users browse the URL are discretized. There exists hierarchical relation in website structure, for example, http://www.smsync.com/order/?ref=030 and http://www.smsync.com /order/?ref=003 are both belong to the category of http://www.smsync.com/order/. This hierarchical relation can help us find more meaningful association rules. So in data preprocessing the discretization of requested resource "URL" according to sitemap is necessary, which can derive a new column "URL-Type". Similarly the discretization of "referrer" according to the website that web users come from can derive another new column "Referred-Type".

C. web accessing behavior knowledge mining and analysis

The web accessing behavior knowledge mining based on association rule is to find the rules like "Referred-Type=>URL-Type" in web log database that satisfies the minimum support and minimum confidence set in advance. Here we adopt Apriori algorithm. In order to find the web users accessing pattern more clearly, we divide the analysis into two parts, that is Referred-Type that its URL and referrer belongs to the same website, and Referred-Type that its URL and referrer do not belong to the same website. We call them inside Referred-Type and outside Referred-Type respectively. Thus the association rule of "inside Referred-Type=>URL-Type" can help us to find the webpage' jump principle when users browse our website and the result can be the basis for website optimization. While the association rule of "outside Referred-Type=>URL-Type" can help us to find from which website that the users come to our website, and then take measures to increase the probability that our website can be retrieved or accessed in those websites.

IV. EXPERIMENT

The experiment uses the sample data of WebLog Expert Lite[9]. Which is the web log file of the website of "www.smsync.com" from the interval of 2007.12.8.00:00:41 to 2007.12.14.23:59:56, which has 30474 web log records.

A. Data preprocessing

First, after filtering out the invalid URL according to the previous three steps in Section 3.2 there are 6351 records left。

Then according to the sitemap of "www.smsync.com" shown is Figure 1, we discretize the URL in web log files to derive a new column "URL-Type". The results are shown in Table I.

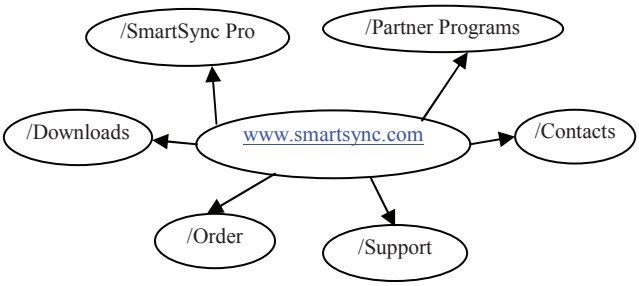

Figure 1. **the sitemap of www.smsync.com**

TABLE I. **MAPPING BETWEEN URL-TYPE AND URL**

URL-Type	URL
1	www.smartsync.com
2	www.smartsync.com/Contacts
3	www.smartsync.com /Downloads
4	www.smartsync.com /Order
5	www.smartsync.com /Partner Programs
6	www.smartsync.com /SmartSync Pro
7	www.smartsync.com /Support
0	others

Finally to discretize Referred-URL according to the website that users come from, Where Referred-Type starts with "1" when users come from the website of "www.smsync.com", and "2" when users come from other websites or searching engine. Part of the mapping between Referred-URL and Referred-Type are shown in Table II.

TABLE II. **MAPPING BETWEEN REFERRED –URL AND REFERRED-TYPE**

Referred-Type	Referred -URL
10001	www.smartsync.com%
10002	www.smartsync.com/Contacts%
10003	www.smartsync.com/Downloads%
10004	www.smartsync.com/Order%
10005	www.smartsync.com/Partner Programs%
10006	www.smartsync.com/SmartSync Pro%
10007	www.smartsync.com/Support%
20036	http://%google%
20073	http://%01net.com%
20122	http://www.listsoft.ru%
……	……

B. *The establishment of Web users' accessing behavior model*

Using the data mining software Spss Clementine 12.0, we establish the Apriori association rule models of "jump between webpages in the website of www.smsync.com" and "jump between webpages from other websites to the website of www.smsync.com" respectively shown in Figure 2.

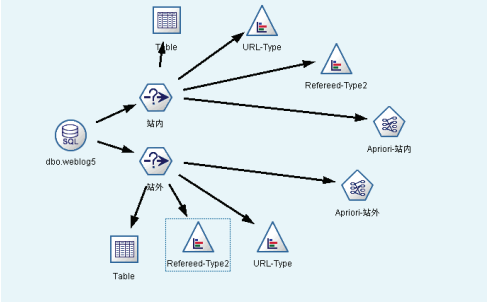

Figure 2. **model construction**

Suppose minimum support is 5%, minimum confidence is 70%. We get three strong association rules from "jump between webpages in the website of www.smsync.com" model and 2 strong association rules from "jump between webpages from other websites to the website of www.smsync.com" model. The results are shown in Figure 3 and 4 respectively.

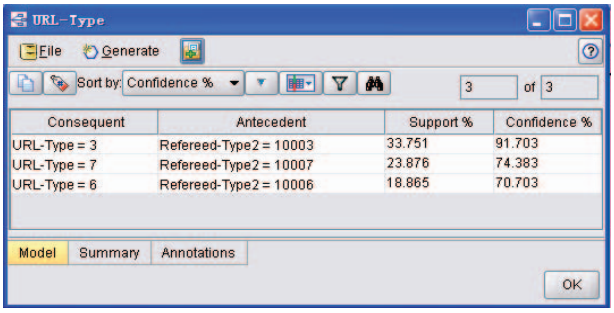

Figure 3. **association rules of jump between webpages in the website of www.smsync.com**

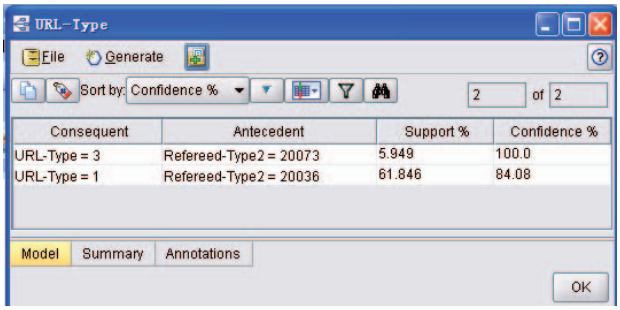

Figure 4. **association rules of jump between webpages from other websites to the website of www.smsync.com**

(1) The analysis of jump between webpages in the website of www.smsync.com

From the three association rules in Figure 2, the users of www.smsync.com usually jump between webpages of the same category, while the case of the jump from the subordinate webpage to the homepage of the website or the contrary is very scarce. Furthermore, the webpages under the three categories of smarsyncpro, support and downloads are paid more attention by users. So in designing the website the link between the webpages of these categories should be considered to make the users visit the website effectively.

(2) The analysis of jump between webpages from other websites to the website of www.smsync.com

From the two association rules in Figure 3, we can see that the users jump from other websites to the website of "www.smsync.com" can be divided into two groups. One group is the users entering the homepage of "www.smsync.com" by the searching engine "Google", the other group is the users that want to download some software, who often search some software in "http://www.01net.com" and then directly jump to the webpages of "download" of "www.smsync.com".

These behavior patterns can be used to service spread of the website. First, the operation target of promoting the reputation of the website and attracting more new users can be obtained by the methods such as improving the Google Search Rank or occupying the Google advertisement place. Secondly, for popularizing the new software we can effectively use the website of "http://www.01net.com". From the jumps after users search the resources in "http://www.01net.com" we can see that users trust the resources of "http://www.01net.com". So it is feasible to popularize the future software in this website.

V. CONCLUSIONS

The relationship of the requested resource URL and referrer which is the webpage before users browse the URL in web log files reflects some accessing pattern when users accessing website. Based on the association rule mining, this article proposes the data preprocessing method of web log files, and the web users' accessing behavior analysis method. The method can not only find the web users' behavior pattern and interests, but also help the managers of the website to improve the website structure and optimize the website construction.

REFERENCES

[1] Ya-Xiu Yu, Xin-Wei Wang. Web Usage Mining Based on Fuzzy Clustering. 2009 International Forum on Information Technology and Applications. Chengdu, China, 15 May, 2009: 268-271.

[2] Pan Lei,Su Jing, Xu Tingrong. Research on Mining Multi-Dimensional Association Rules About Network Accessing Behavior. Computer Applications and Software.2008, 25(3): 189-191(in Chinese).

[3] Xia Min-jie, Zhang Jin-ge. Research on Personalized Recommendation System for e-Commerce based on Web Log Mining and User Browsing Behaviors. 2010 International Conference on Computer Application and System Modeling (ICCASM 2010). Taiwan. 2010,10,22-24: 408-411.

[4] Ozel.S. A. A Web page classification system based on a genetic algorithm using tagged-terms as features. Expert Systems with Applications. 2010, 08: 1-9.

[5] Agrawal.R., Imielinski, T., Swami, A. Mining Association Rules between Sets of Items in Large Databases. In Proceedings, ACM SIGMOD Conference on Management of Data, Washington, D.C., 1993: 207-216.

[6] Wanjun Yu, Xiaochun Wang, Fangyi Wang.The research of improved Apriori algorithm for mining association rules. In Proceedings of the

11th IEEE International Conference on Communication Technology Proceedings, 2008: 513-516.

[7] Yuan Jian, Wang Wenhai. An improved Apriori Algorithm Based on Mining Association Rule. Journal of Qingdao University of Science and Technology(Natural Science Edition). 2008, 29(5):448-451(in Chinese).

[8] Lu SF, Lu ZD. Fast Mining maximum frequent itemsets. Journal of Software, 2001, 12(2):293-297.

[9] http://www.weblogexpert.com/download.htm, 2011,5,15.