Modeling and Intelligent Analysis of Web user behavior of WEB User Behavior

Xuezhi Lei
Liaoning JianzhuVocational University*; Liaoyang 111000,China*

***Abstract*—The core research of Web mining is interest association rule in Web logs and clustering algorithm of user browsing behavior. Traditional association mode and browsing path have certain advantage in browsing path for the user, while they cannot provide accurate recommendation in the important area at the same page. Therefore, based on the association rules of users' interest, an intelligent user interest association rule is proposed in this paper, integrated with Web area partition. It comes from the choice of area of current network users and different interest degree of user browsing on the web. Then, related mining algorithm is put forward based on interest area. The algorithm improves the accuracy of recommendation of single page area by interest degree of page browsing and weight computation of click-stream data. Finally the effectiveness of the intelligent system is verified by the experiments.**

Keywords-user behavior; data mining; interest degree; intelligent analysis

I. INTRODUCTION

With development and popularization of Internet technology, the Internet provides people with vast knowledge resources, and it becomes the best way for people to learn, educate and communicate with each other. Network has such characteristics as huge amount of data, diversity of types, dynamic, equality and virtualization. It permeates all levels of the network itself, network services and network applications. But in this huge network system, how to extract the knowledge of interest quickly and efficiently has become focus of people's attention. The network user analysis will become an important tool in the network service. It is to improve the website service quality, to improve network efficiency, ensuring the security of the network, providing multiple aspects of personalized service and playing a very important role, to meet the two aspects of network users and service providers demand. The classification of network user behavior and the analysis based on it become more and more urgent to be put on the agenda.

This paper analyzes the key technology of current Web user behavior analysis of Web usage mining. Web mining is the core research content of interest association rules in the Web log and Web user browsing pattern clustering algorithm. Based on user interest association rules, and combined with the regional division of Web, we propose a new user interest association rules, with different level of interest in interest association rules from the network users and the users in the choice of regional page browsing is shown. Then a mining algorithm using region of interest is proposed. Finally, the experimental analysis is carried out, and the key indexes of clustering algorithm are verified by several sets of experiments.

II. KEY TECHNOLOGIES OF USER BEHAVIOR ANALYSIS

A. Data Processing

Data preprocessing is a very important step in data mining, and it is also the most time-consuming part. The quality of data preprocessing has a direct impact on the effect of data mining algorithms. The data preprocessing mentioned in this paper mainly includes data cleaning, user identification, session identification, path supplement, and transaction identification. Data cleansing is the basis of the whole data preprocessing. As the first step of data preprocessing, data processing directly affects the quality of data and guarantees the success of data mining. The data cleansing module is shown in figure 1:

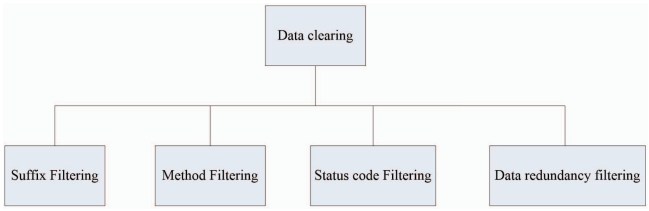

Figure 1. Data cleaning

User identification is mainly used to identify specific users from the log so the log files can be divided into several separate user access groups according to different users. This article adopts the following user identification rules:

(1) The user IP address is used to distinguish the user, and the IP is different from the user.

(2) If the IP is the same, but the browser, the Agent, or the client operating system is different, the user is also different. With above recognition rules, the accuracy of the recognition can basically meet the requirements of the algorithm even when current PC popularity is getting higher and higher.

Each user session represents a set of consecutive page access sequences for users in a session and access behavior and access interest are obtained through this access sequence. The system reads the log file circularly during the preprocessing process, and each reading a row of records determines whether the IP already exists. If not, this is a new user session. Then the system creates a new user session and stores in the database; if the IP already exists, it determines the difference between the user and the access to

978-1-5386-8028-5/18/$31.00 ©2018 IEEE
DOI 10.1109/ESAIC.2018.00052

last visit to the time difference and time threshold size. If the time is greater than the threshold it indicates this is a different session from the same user, so the system creates a new session and the session is recorded to the session table; if the time is less than threshold which shows the same session of the same user, it updates the user session table. Above steps are repeat until the end of the log scanning.

User identification separates different users from the log according to user classification log collection. It will separate each user session identification according to the sequence of width and specified time division to form many shorter sequences in conversation. After a few steps in front of the data quality that has been greatly improved, it is not very accurate to reflect a user's behavior, since the user access process may be caused by the browser's "back" and "forward" on the client cache function. The object also results in that the page cache access records lost, and the log can not reflect the user's browsing behavior, so it is necessary to delete the path of recovery.

B. User Behavior Analysis System Framework

Before making user analysis, we must make clear the objectives and directions of the research, and refine the whole system into several small targets, to implement them one by one. The user behavior analysis system needs to establish corresponding behavior database, prepare data and analyze data. These three steps constitute the core of behavior analysis. Data preparation work occupies a lot of time and energy of the whole process. The main work is to select variables and records, and put them into the variables and data structures need analysis of behavior. Analyzing data is to clean, integrate and calculate the acquired data. The establishment of user behavior analysis model is an iterative process that requires careful study of different models, and find out the design of more effective model from which is to achieve every module of the model through the environment and programming. Then according to the results to the interpretation and evaluation of the value of the model, the model is applied to the environment of actual application through the test of practice.

The object of network user behavior analysis is the individual Internet users. We name a behavior as "user access events". Whenever the event occurs, the user behavior analysis system to do is to obtain corresponding user online behavior through a certain method, including information on the user, to distinguish other parts of calculating the interest of users. But the collection of browsing information is often out of order, and there are some interference information that can not fully reflect the user's interests and needs to be carried out in the basic information input after pretreatment in data mining. They can get the web page for a topic correlation calculation or clustering after classification, combined with several parameters of user access for the theme correlation and the setting of the definition formula to calculate each correlation of users for a topic. The topic will set the user for a subject correlation as user interest degree. Finally it provides web page recommendation and personalized service for users, or to other purposes, according to the user on the subject of the relevant degree.

Based on above description, this paper presents the basic framework for network user behavior analysis as shown in figure 2:

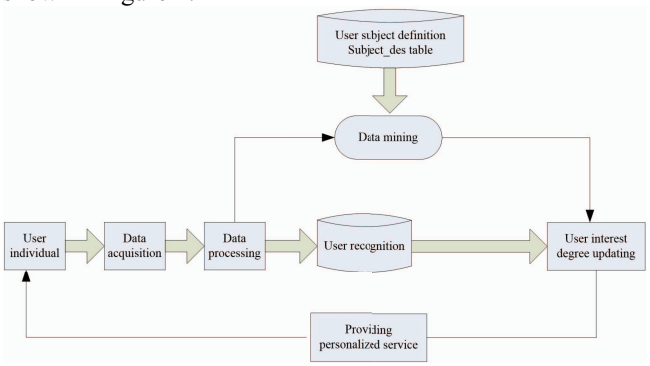

Figure 2. Principle structure of user behavior

III. ASSOCIATION RULES MINING OF USER BEHAVIOR

A. User Browsing Behavior Classification

A large number of studies have pointed out that the user's interest in web pages is closely related to their browsing behavior on the web page. Most of user browsing behavior implies user preferences and interests, such as inquiries, browsing the page, marking bookmarks, feedback information, click the mouse, drag the scroll bar, forward, backward and so on. When users access the page, they also show user interest, such as stay time, access number, save, edit and modify.

If the behavior of Web user when browsing a page is saved as a behavior of history, we can determine the level of interest of the user on a page in a period of time by the analysis of various behaviors of the users for a period of time browsing, by and information mining. Here, only 5 types of general browsing behaviors are taken into account:

(1) If Web user saves the page that he has browsed, it shows that the user has a strong interest in the page.

(2) If Web user prints the page that he has browsed, it shows that the user has a strong interest in the page.

(3) If Web user collects the pages that he has browsed, it shows that the user has a strong interest in the page.

(4) If Web user spends a lot of time browsing the page, it shows that the user has a strong interest in the page.

(5) If Web users browse the same page for several times over a period of time, it shows that the user has a certain interest in the page.

B. Computation of Interest

Web interest refers to the degree of interest in a web page, using real numbers of 0 and1. 0 denotes no interest, and 1 means the maximum interest. Obviously, the user's interest is closely related to the interest of the web browser. The five minimum browsing behaviors for web pages are

depicted as follows: save pages $S(L)$, print pages $P(L)$, store pages in bookmarks $B(L)$, repeat access times $R(L)$, and dwell time $T(L)$ on page (L). These actions reflect the interest of users of L pages in different degrees, different level of interest in order to distinguish these behavior expressed. W_v denotes a weight for each V, so the user from the user behavior of these reasoning L pages to the extent of interest can be calculated using the following formula

$$I(L)=\sum_{v\in F}W_v f_v(L) \quad (1)$$

where $\sum_{v\in F}W_v=1$ and $F=\{S(L),P(L),B(L),R(L),T(L)\}$. W_v is the weight assigned to behavior V. If user has behavior on page L, its function value is 1; otherwise, the value is 0.

Related research points that the interest is

$$Int(P)=at(P)+bv(P)+C$$

$t(P)$ is the browsing time of page and $v(P)$ is the number of times to pull scroll bar. In this paper it is believed that the two key behaviors revealing the user's interest in web pages are:

the browsing time on the web page $t(L)$ and the number of visits to the web page $R(L)$. To find the quantitative relationship between T and R as well as the interest degree of web pages, the detailed analysis and experiment must be carried out. After the experimental analysis, the interestingness quantification estimation equation can be obtained as

$$I(L)=at(L)+br(L)+C \quad (2)$$

a, b and c are all unknown parameters which are not related with $t(L)$ and $r(L)$.

C. Algorithm Realization

The interest degree page is defined as

$$V=\begin{cases}0; & t<t_1, t>t_2\\ \dfrac{tS_i}{L_j}; & t_1\le t\le t_2\end{cases} \quad (3)$$

Among them, t_1 is the minimum reading time. When user time is less than t_1, it means the user does not have time to read the page. t_2 is the maximum reading time not increasing when exceeding the interest, to avoid the user brought by the time delay due to effect of other things. When $V<0.2$, it is considered that the page is not a page of interest to the user, that is, deleting the page. The average page browsing speed mentioned is closely related to the user's browsing behavior and the amount of information on the page. L_j is the length of the first J pages.

The interest mining algorithm based on Web behavior mining is described as follows:

For each user session s_i;

Define the page type, the minimum browsing time and the maximum visiting time of the web site;

Set the interest degree of user as $V_0=0.2$

While (s_i is not null)

{

While (s_iUrls is not null)

{

If (Url is not the last one) compute the value of V as the regulation of interest degree

Else $V=1$

If ($V<V_0$) delete this page

}

}

IV. SIMULATIONS

We select interested web pages browsed by user in a week on Amazon as the data processing. Clustering algorithm is adopted to cluster them into 5 themes and contains the text content. We randomly select 10 articles of each subject as the experimental data. According to the user's browsing behavior and operation we adopt *dyna Trace AJAXEdition* software. Then through the method of user browsing behavior in front of the introduction of access to the user for each web page interest, we can obtain each subject interest. A week later, statistical investigation is performed by the browse online user group. The user is made to evaluate each subject of interest, and the subjective evaluation results will be calculated according to the users interest degree for comparison.

Figure 3 includes the aforementioned required browsing behavior, that is, the number of bytes when user browsing the page, the residence time of users browsing the web, the calculated browsing speed and the preservation and collection operation of users when browsing the web page .

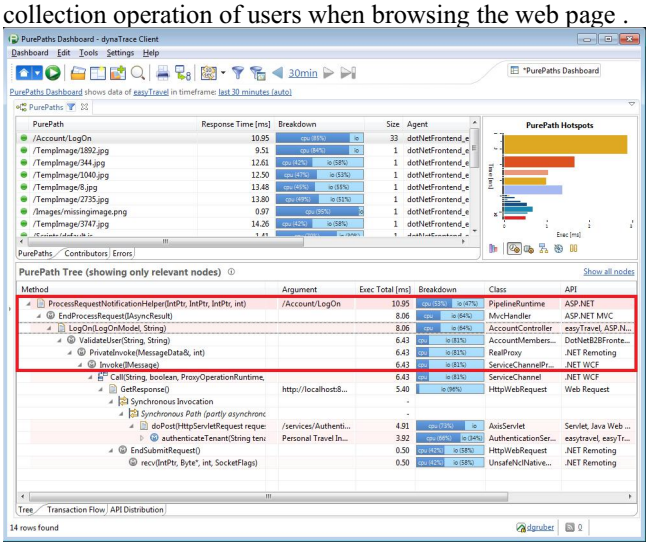

Figure 3. Experimental results of user behavior through related software

In original recording data we choose 100 different subject representative pages. Through Web log mining the data collection 100 pages are collected for calculation, based on the number and browsing action based on the browsing speed under the interest rate and interest times. By the integration with neural network, we get the final interest degree of web page. At the same time, through the survey of specific user groups, users evaluation of the interest and the subjective evaluation results will be calculated according to the web page interest comparison. Table 1 describes the comparison results of certain typical data.

Table 1. A sample of comparison between user interest degree computation and self evaluation

Page URL	Interest degree	Self evaluation	Absolute error
https://www.amazon.cn/gp/product/B00Y20UI1K/ref=pd_rhf_dp_s_cp_1?i.....	0.9867	0.967	0.0197
https://www.amazon.cn/dp/B01EMMYEC6?_encoding=UTF8.....	0.7445	0.788	0.0435
https://www.amazon.cn/gp/product/B00O1F8F98/ref=s9_acsd_al_bw_tc_CHANGEME_2_i_r?.....	0.9998	1	0.0002
https://www.amazon.cn/dp/B01MTVDZRL/ref=lp_813830051_1_1?s=kitchen&i.....	0.7238	0.725	0.0012
https://www.amazon.cn/dp/B00NH9YDP4/ref=lp_863872051_11?s=softw.....	0.6635	0.603	0.0605
……	……	……	……

V. CONCLUSIONS

Users' usage mining in Web mining is an important method to analyze user behavior, and the analysis of this behavior plays a very important role in the field of current business website. This paper is based on the user browsing patterns and user association patterns in Web usage mining. According to the steps of Web mining, we find out the user's interests and excavate the browsing patterns of most users. We takes into account not only the traditional path interest association patterns in the analysis, but also the click Page area to increase the dimension to design a new Web mining algorithm. From the experimental results it indicates that , the algorithm puts forward by area interest model algorithm in recommendation domain is a good one, and the interest calculation is more accurate thatn before.

REFERENCES

[1] Ghose S, Jenamani M, Mohapatra P K J. Design benchmarking, user behavior analysis and link-structure personalization in commercial web sites. Internet Research, 2006, 16(3):248-266

[2] Phoa F K H, Sanchez J. Modeling the Browsing Behavior of World Wide Web Users. Open Journal of Statistics, 2013, 03(2):145-154

[3] Punera K, Merugu S. The anatomy of a click:modeling user behavior on web information systems/ Proceedings of ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October. DBLP, 2010:989-998

[4] Bali T G, Neftci S N. Disturbing extremal behavior of spot rate dynamics. Journal of Empirical Finance, 2002, 10(4):455-477

[5] Lee I, Kim J, Choi B, et al. Measurement development for cultural characteristics of mobile Internet users at the individual level. Computers in Human Behavior, 2010, 26(6):1355-1368

[6] Head M, Ziolkowski N. Understanding student attitudes of mobile phone features: Rethinking adoption through conjoint, cluster and SEM analyses. Computers in Human Behavior, 2012, 28(6):2331-2339

[7] Taddy M, Gardner M, Chen L, et al. A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation. Journal of Business & Economic Statistics, 2014:193-211