

Study and Implementation of User Behaviour Analysis

Jingling Zhao*, Xia Wang*, Yue Zhou*

* School of Computer, Beijing University of Posts and Telecommunications, Beijing, China, (100876)

jing_lingzh@sina.com, wangxia312@sina.com, zhouyue@mail.ustc.edu.cn

Abstract— With the development of the Internet, network users urgently hope the service providers to provide the services what they need. The paper introduces the basic structure based on the analysis of user behaviour. In the structure, different contents of the user behaviour are achieved. After dealing with these contents, the knowledge of data mining is used to analyze the degree of the user interest. And the paper introduces a method to calculate the values. At last the service providers can give the personal service to the users according to the values of the user interest.

Keywords— user behaviour analysis; data acquisition; data mining; interest degree; SVM.

I. INTRODUCTION

In the past few years, with the growth of information content on the Internet, the number of Internet users increases rapidly. Until now the total number of Internet users has reached up to 300 million in China[1]. People are more inclined to get information through the Internet. But in the huge network society, the information that each user is really interested in is very limited because of the differences of people. The simple transfer of information and resource sharing can no longer meet the needs of users. So they hope that the service organizations can provide the personalized services after sorting the services out. For this reason, service providers must understand the users' behaviour and consider the differences between individual users. And then they can provide personalized services to improve service quality and service efficiency based on each user's interests and needs.

II. CONCEPTS OF THE USER BEHAVIOR ANALYSIS

The user is defined as a visitor who surfs the internet through one or more servers [2]. And an action from visitor is named to be an event. The different factors such as gender, age, occupation, living environment and cultural backgrounds determine the uncertainty of user's behaviours. User information behaviour includes information inquiry activities, information choice behaviour, and information usage behaviour [3]. User behaviour analysis (UBA) aims to monitor, predict and even control users' behaviours by finding out the rules users show when using the network [4].

User interest (UI) reflects the user's preferences in the subject needs and content forms. The paper represents the UI in some subject as a certain number which is called interest-degree, and proposes a formula for calculating the interest degree. The interest degree is defined as the number between 0 and 1. And 0 indicates that the user isn't interested in the subject. Larger values indicate greater interest that the user shows in this subject. At last the service providers provide users with personalized services based on the number of the interest degree. When an event occurs, the system of UBA is composed of five steps: data acquisition, data pre-processing, data mining, the calculation of interest degree and the provision of the personalized service. Figure 1 focuses on the analysis of user behaviour architecture. And the five steps will be introduced in detail.

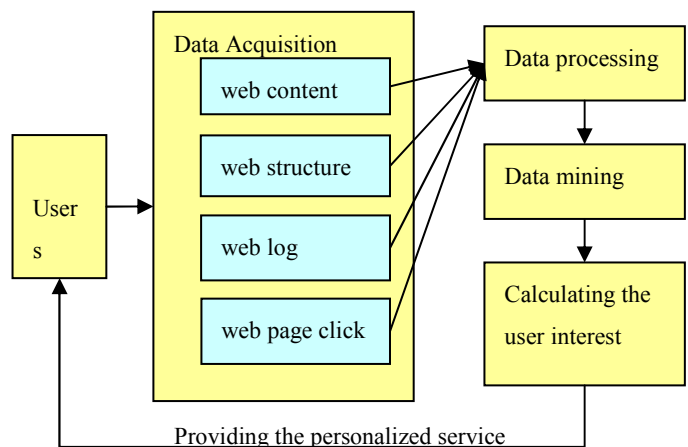


Figure 1. The basic framework of user behavior analysis

III. DATA ACQUISITION

Before the analysis of user behaviour, it is necessary to collect related operations and information of every user. The purpose of data acquisition is to find and identify the operation of the object. We collect the raw data according to the user's needs and actual behaviours from servers, user terminals, databases, application middleware and so on[5]. then we can find the user interest by analysing and summarizing the original and haphazard data. This paper considers collecting data as following:

1) The collection of web content data: the real content of web page. The process is to obtain useful knowledge from web contents and the description of the contents. For example: extracting title, description, keyword and other related information according to some subject in the web documents.

2) The Acquisition of web structure: As it is known, Web page is written by HTML and semi-structured text files. In addition to plain text content, the page itself also contains some organizational structure and hyperlinks pointing to other pages. They can reflect the relationship of inclusion, reference or subordination between pages. Through these relationships, we can sort these pages and find an important page [6].

3) The Acquisition of web log: Analyse the user behaviour through the information of proxy server log. The log file from Web server records the data about the web file which the user visits [7]. These data in the file include: the user's IP address, date and time when the user visit the web, browser type, response status, the user web site, some configuration information and so on. We can use the information obtained to carry out user identification. In normal circumstances, IP address can be used to identify each user here.

4) The method based on the page click: We trace the user behaviour by making use of the basic operation of each page. The basic operation includes: Saving, printing, copying or pasting part of the contents and so on. And hook procedure can achieve the above functions. Hook is a platform of processing Windows message. The application program can set up sub-processes to monitor some message of the specified window, and the monitoring window can be created by other processes.

In this framework, it's a good idea to distinguish users by their IP address. The advantages are shown as following: First of all, the IP addresses can be gotten easily. For the user who connects to the internet directly and has a unique IP address of the computer, IP address can accurately identify the computer and user. The computer of some user can be tracked by the IP address, and the cookies track the browser. The computer that has a IP address can run a variety of browsers and retain a number of trace files, So it is a better way to use IP addresses to identify the individual user.

IV. DATA PRE-PROCESSING AND DATA MINING

A. Data Pre-processing

Since the amount of data generated every day by many large sites is amazing, and the database is vulnerable to noise data, vacancy data, and a number of inconsistencies in the data. The data pre-processing refers to the analysis of web pages, filtering data, eliminating duplicate records, transformation of data type and collating the necessary information. So the data mining process can be made to be more effective and easier. The main task of data pre-processing are as follows: data processing, data integration, data transformation and data reduction.

For example: the data what we get is generally incomplete, noisy and inconsistent [8]. When the user visits a Web page, images, audio and video files will be sent to the browser device too.

In most cases, what we need to analyse and deal with is just the corresponding HTML file, rather than images or audio files which constitute the noise data, or known as the "dirty data". These "dirty data" often fail to reflect the user's actual behaviour. And in terms of user behaviour analysis these dirty data are of no value. So these data will be cleared, converted or unified into a form suitable for data mining. In this paper, HTMLs which contain these strings: .ani, .cab, .conf, .crl, .cur, .class, .dll, .embed, .flv, .ini, .gif, .GIF, .jpg, .jpeg, .JPG, .png can be deleted.

The text should be divided into signal words after pre-processing. The basic algorithm concludes: (1) Mechanical sub-lexical, the main maximum matching method, reverse maximum matching method, word for word matching method, frequency statistics of word and so on. (2) The sub-lexical method based on semantics. (3) sub-lexical method of understanding[9].

B. Data Mining

Existing techniques on data mining can be divided into two categories: direct data mining and indirect data mining. In the paper we can use the support vector machine (SVM) technology to calculate the values that a web page is related to some subject.

Support Vector Machines (SVM) is a relatively new supervised learning method approach introduced by Vapnik in 1995 for solving two-dimensional pattern recognition problems [10]. SVMs are a set of related supervised learning methods used for classification and regression. Based on the principle of Structural Risk Minimization principle, SVM tries to find a separating hyper-plane with maximum margin to separate positive examples and negative examples from the training set [11]. A SVM constructs a hyperplane or set of hyper planes in a high-dimensional space, which can be used for classification, regression or other tasks.

SVM maps the vectors to a higher-dimensional space. In the space, a maximum interval hyper-plane is established. On each side of the hyper-plane separating the data there is one hyper plane. The interval hyper-plane maximizes the distance between the two parallel hyper-plane. The greater the distance is the smaller the error of the classifier.

SVM shows a better performance than the other methods. It defines a n-dimensional hyper plane and tries to calculate a decision line (solid lines in Figure 2). The solid line is the classification line. The dotted lines are two parallel lines that are nearest to the solid classification line. The distance between the two dotted lines is called margin. The optimum classification line claimed is to ask the solid classification line to not only separate two kinds correctly, but also make classification interval biggest. And figure b) shows a maximized margin. The vectors which are crossed by the margin lines are the support vectors.

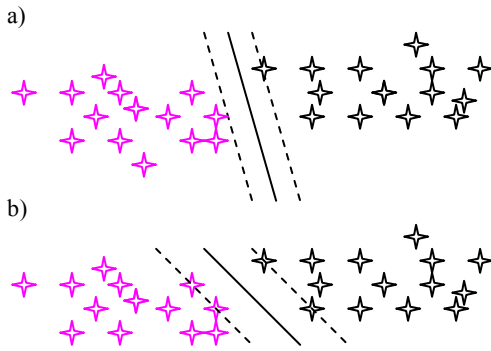


Figure 2. SVM vector plane with two different decision lines

V. THE ESTABLISHMENT OF A DATABASE

After the pre-processing of raw data and data mining, we can establish the database which contains six tables. The relationship of these six tables is shown in Figure 3. The specific description for each table is as follows:

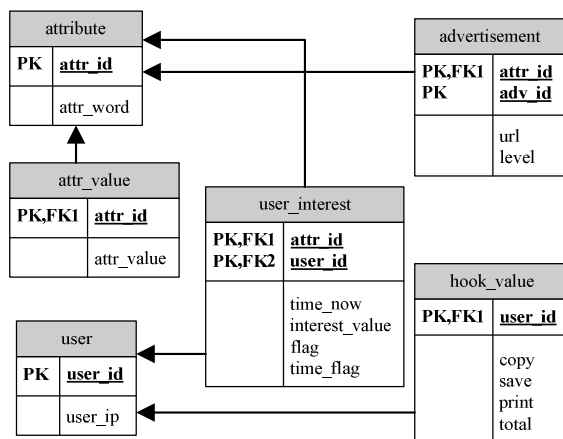


Figure 3. The relationship of the tables in the database

1) attribute table: the attribute table defines a keyword for a particular theme. Table 1 is the contents of this table. We predefine some keywords such as book, computer, and movie.

Table 1. Content in the attribute table

attr_id	attr_word
1	book
2	computer
3	movie

2) attr_value table: After data mining we can know the relationship between the pages and a theme. SVM is used to calculate the percent of three major categories occupying in the entire page, namely the attr_value: k1, k2, k3.

3) user table: Distinguish the user according to the IP address.

4) hook_value table: Get the numbers of the user's operations such as copying and printing which are achieved by the hook procedure. They will be used to calculate the degree of the user interest in some theme.

5) user_interest table: Record the degree of the user interest in a topic. mark the number of advertisements which are send to the users, record the time when the advertisements for the last time are send to the users.

6) advertisement table: Register ads that the advertisers provide. And these ads are set to be different classes.

VI. CALCULATE THE DEGREE OF USER INTEREST

Degree of the UI refers to the level of the user interest in some topic, represented by the number between 0 and 1. In the paper we propose the numerical method such as equation (1). The values of interest_value which stand for the degree of the user interest in a certain topic are gotten by calculating the data from acquisition and data mining. They are put in the user_interest table that is shown in Table 2. And interest_value is expressed in percentage in this table.

Table 2. Content of User_interest Table

user_id	attr_id	time-now(s)	interest_value (%)	flag	time_flag
1	1	1229067208	24.1	0	1229067208
1	2	1229067208	60.4	0	1229067208
1	3	1229067208	90.5	1	1229066149
2	1	1229067251	58.8	0	1229067251
2	2	1229067251	47.2	0	1229067251
2	3	1229067251	80.6	0	1229067251

$$F_{\text{new}}(n) = F_{\text{old}}(n) - p_1 * (\text{time}_{\text{new}} - \text{time}_{\text{old}}) + (p_2 * v_{\text{save}} + p_3 * v_{\text{print}} + p_4 * v_{\text{copy}} + p_5) * k_n \quad (1)$$

In the formula, $F_{\text{new}}(n)$ represents the interest_value shows in the nth ($n=1, 2, 3$) attr_word. The initial value is 0; subscript old represents the related values of the last event which are stored in the table. And subscript new represents the related values of the new event. time_{new} and time_{old} separately represents the arrival time of the new events and the arrival time of the front event, and the time is shown in seconds.

$v_{\text{save}}, v_{\text{print}}, v_{\text{copy}}$ separately represents the number of operations to save pages, print pages, copy and paste by the hook procedures. The value is 0 or 1 normally. k_n represents the percentage of data mining corresponding to the three themes, $n = 1, 2, 3$. p_1, p_2, p_3, p_4, p_5 are the corresponding parameters.

$p_1 * (time_{new} - time_{old})$ shows the reduction of interest over time. And the value of p_1 can be set to $1 / (3600 * 24)$, $3600 * 24$ means we take one day as standard configuration. p_2, p_3, p_4 are set to be in the same range, and we can set their values according to the actual needs.

We set a threshold named Th for interest_value. The value of Th is between 0 and 1. Here Th is set to be 0.85. If interest_value is higher than Th, the program will send the appropriate advertising to the user. From the advertisement table, relevant URLs of five ads are chosen randomly according to the advertisement hierarchy. Maybe the user is tired of the ads if the system sends the ads frequently. So it's a good idea to record the time when the ads are sent. And when the interest_value is up to 0.85 for the next time, calculate the time interval and compare it with the user tolerance range that can be set by the user. Only if the time interval is bigger than the range, relative ads will be sent. So according to the different needs of individual customers, we can provide personalized services to them so that the user and the service provider can achieve the win-win purpose.

VII. CONCLUSIONS

To provide personalized service is the trend in the network. In this paper we can calculate the user's degree of interest in some topic by mining and analysing the data from data acquisition. And then the user's personal interests and hobbies come to be obvious. After knowing what the users

hope to do and what are they needed, the service providers obtain the potential commercial value. And the providers can connect to the customers closely to obtain greater economic benefits.

REFERENCES

- [1] China Internet Network Information Center. The 24th China Internet Development Statistics Report. [Online]. Available: <http://www.cnnic.cn/html/Dir/2009/07/15/5637.htm>, July 15, 2009.
- [2] Hu Yong. "Study and realization of user behaviour in the Network Environment". Beijing: CNKI. Nov 21, 2007.
- [3] Hu ChangPing, Shao Qigan, Sun Gaoling. "User preferences and behaviour analysis in the personalized information service". Intelligence Theory and Practice. Vol.1. pp:4-6. 2007.
- [4] Li Huijuan. "Analysis and Study of Campus Network User Behaviour". Beijing: CNKI. Sep 14, 2007.
- [5] Jiang Hong, He En. "Behavioural analysis technology and the prospects for application in the trusted network. Information security and confidentiality of communications". vol.2. pp:67-69. 2009.
- [6] Wang Li. Literature link analysis based on user browsing behaviour. Now intelligence. vol.29. pp:150-152.
- [7] Dong Fuqiang. "Analysis and Application of Internet User Behaviour". China Beijing: CNKI. April 26, 2005.
- [8] Liu Mingyue. "Analysis of user behaviour based on web log". China Beijing: CNKI. July 10, 2008.
- [9] Xiaodan Zhang. "Research of Web Text Mining" computer and network. pp:505-506.
- [10] Lizhen Liu, Zhaowei Li, Aiqing Wang, Haijun He, "The Research of Decision Support Vector Machine in Web Information Classification". Computer Supported Cooperative Work in Design. 16-18 April 2008. pp:196 - 200.
- [11] Jia-qi Zou; Guo-long Chen; Wen-zhong Guo; "Chinese Web page classification using noise-tolerant support vector machines", Natural Language Processing and Knowledge Engineering, 2005. from 30 Oct to 1 Nov. 2005. pp: 785 - 790.