Research of Analysis of User Behavior Based on Web Log

Jia Li
Computing center
Anshan Normal University
Anshan, China
e-mail: assy_lj@163.com

***Abstract*—This paper first introduces the theory and knowledge related to Web logs, and then introduces a Web log mining process. As to this article is based on Web log processing, we draw valid user data according to a specific website users' access data through the preprocessing, and then research and analyze the users behaviors. The work can provide a theoretical basis for the management and optimization of the site for site managers. Experiments have showed that the work is effective.**

Keywords- data mining; Web log; user behavior analysis

I. INTRODUCTION

The purpose of Web mining is to find useful information and knowledge from the network hyperlink structure, page content and data used [1]. Although Web mining uses many data mining technology, due to the nature of heterogeneity and semi-structured or unstructured of Web data, traditional data mining techniques cannot be used for Web mining directly and completely. In the past decade, a lot of new mining techniques and algorithms have been invented and used to solve the problem. Based on the data types used in the mining process, Web mining tasks can be divided into three types: Web structure mining, Web content mining and Web usage mining [2].

Web usage mining is related to the user's access mode found by Web log, it records every click of each user [3]. The key issues in Web use is the pretreatment of click-stream using log, in order to obtain the correct data for mining [4]. The process of Web mining and data mining are similar, usually data collections are different. In traditional data mining, data has been collected and stored in data warehouse. For Web mining, the workload of data collection is a huge, especially for Web structure mining and content mining, which includes a large number of landing pages crawled over. When the data is collected, the next three processes are the same [5]: data preprocessing, Web data mining and post-processing. This paper studies the part of preprocessing in Web usage mining, and analyzes user behavior on data processed.

II. OVERVIEW OF WEB LOG

Web log records all the information of user's activities from sending request to the server. The original data source is from the network access log. Originally, log files are produced for debugging, log files can be found from three different places: 1. network server, 2. network proxy server, 3. client browser.

A. classification of Web log

a) server-side log

Generally, these logs are able to provide the most complete and accurate usage data, but having two shortcomings:

1) These logs include sensitive personal information, so the server owner is not announced.

2) These logs cannot record the access to cache page, the cache page is returned from local browser or proxy server directly, rather than Web server.

b) agent-side log

Proxy server gets Hypertext Transfer Protocol (HTTP) from the user, sends it to Web server, and Web server returns the results to the user through proxy server. There are some problems:

1) Building proxy server is a difficult task, such as TCP / IP and other advanced network programming is necessary.

2) The intercept of request is limited, which do not cover most of the requests.

3)If proxy server joins recorder, the performance of Web log system declines, , because requests of each page need to be processed by agent simulator.

c) Client log

Remote test of network is by downloading the designated software to record usage situation of Web or modify the source code of existing browser. HTTP caching can be used to achieve this purpose, these is a small part of information generated by the Web server, which is stored in user's computer, available for later access. The disadvantages of such a log are:

1) Design team must deploy specified software in advance and make end-users install on.

2) The technology makes it difficult to achieve compatibility for a series of operating systems and Web browser.

B. Structural Analysis

Web server logs are general text (ASCAII) file, which is independent of the server platform. Server softwares are different, but usually there are four types of server logs: transmission log, agent log, error log and reference log.

The first two types of logs are standard. Reference log and agents log may be open or not in the server, or be added to the transmission log and forming a extensional log form.

978-0-7695-5004-6/13 $26.00 © 2013 IEEE
DOI 10.1109/ICCIS.2013.165

Web log file is server recording information of user's requests for resources to a specific site each time. Most logs use common log format. The following is a log fragment taken from a web server:

65.52.109.26--
[18/Feb/2012:23:59:41+0800]"GET/index.php?do=rck&indus_id=6&slt_page_size=10&join=2&price=&style=&slt_order=2&p=Liaoning HTTP/1.1" 200 12311"-"
"Mozilla/5.0
(compatible; bingbot/2.0;
+http://www.bing.com/bingbot.htm)"

It shows the information as shown below:

Remote IP address or domain name: IP address is a 32-bit host address defined by the Internet Protocol; domain name is used to determine unique network address for any host online, an IP usually corresponds to a domain name.

Authorized user: username and password used when server requires user to be verified.

Date and time of login and logout

Mode requested: GET, POST, or HEAD method of CGI (Common Gateway Interface).

Status: HTTP status code returned to the user, such as 200 is "OK", 400 "not found".

Byte: length of text content transferred.

Remote log and agency-side log

Remote URL (Uniform / Universal Resource Locator)

"request": completely from request line of client

URL requested

C. HTTP Overview

Hypertext transfer protocol is used on the World Wide Web in 1990, is an application layer protocol. The first version of HTTP is HTTP/0.9, is a simple protocol used in the whole Internet, for raw data transfer. HTTP/1.0 is defined by RFC 1945. Table 1 is frequently-used status code of data transmission error and success by Hypertext Transfer Protocol.

Table 1 state of Hypertext Transfer Protocol

Status code	Significance
101	Switching Protocols
200	OK
202	Accepted
305	Use Proxy
400	Bad Request
401	Unauthorized
403	Forbidden
404	Not Found
408	Request Time-Out
500	Server Error
502	Bad Gateway

III. USER BEHAVIOR ANALYSIS BASED ON WEB LOG

The three main steps of Web usage mining are data preprocessing, pattern recognition and pattern analysis [6]. Data preprocessing includes removing unwanted data, in pattern discovery phase, extracting usage pattern from Web data by data mining techniques. Pattern discovery is the key part of Web mining, it covers a number of research areas, such as data mining, machine learning, statistics and pattern recognition. Statistical analysis, association rules, clustering, classification, sequence pattern and dependence modeling techniques are all used to discover the rules and patterns. The knowledge can be found in the following aspects, performance form of rules, tables, and icons, or performance form of other features, comparisons and predictions, as well as data classified from Web access logs. The purpose of this process is to extract interesting rules or patterns form output of pattern discovery process by eliminating irrelevant rules or patterns.

A. Data mining methods

The main work of this paper is to analyze Web log files, as well as to analyze the behavior of users, so contents of access log needs to be pretreated. Data preprocessing is to remove the irrelevant part from the log file. Figure 1 is an overview of data pretreatment, which comprises data cleansing and user identification.

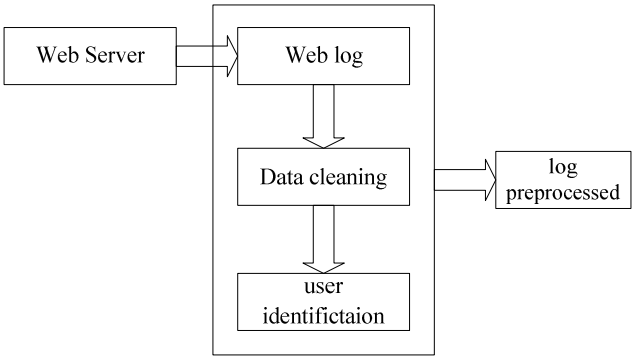

Figure 1. Overview of data preprocessing

Data cleansing is the first stage of data mining process. Some low-level merger work is performed at this stage, in order to combine multiple integrated reference logs. Entries which are irrelevant with data analysis and data mining will be removed. In data cleaning process, the error and failure entries must be removed, and some access records generated by automatic search engine will also be distinguished and removed from the access log. This process should also remove the access requests having nothing to do with the analysis, such as images, multimedia files and page style file, such as requests of graphic page content (*. Jpg & *. Gif)and file requests of any form initiated by robots or web spiders. By filtering out irrelevant data, the size of log file will be reduced and occupy less storage space, and which will facilitate the conduct of next task. Example, by request of filtering off the picture, after cleaning, the size of log file of Web server is reduced to less than 50% of the original document. Log entries must be divided into logical clusters through one or a series of transaction identification modules.

B. data analysis of user log

User identification is to identify each specific user by user's IP address. In Figure 2, in order to identify the user,

this paper proposes a provision: If the a new IP address appears, it represents a new user, if the IP address has not been changed, but the operating system have been changed or browser is not the same, which also indicates the presence of a new user, another reasonable assumption is that the same IP address using different proxies also indicates the new user.

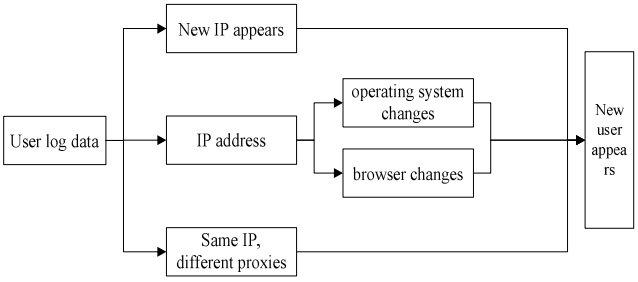

Figure 2. simple user log analysis

IV. EXPERIMENTAL DATA AND RESULTS ANALYSIS

In the process of this article, we have analyzed 564MB log file of server, do different analyses to identify the user behaviors.

By Table 2, we can get the detailed information of user access throughout every day of the period, in which the administrator can find the third day should be the most active day for users, at least the user clicks the most on this day, while administrators can analyze the customer's basic situation through long-term data analysis, if it needs to shut down the server, they can select the day when less visitors access. Figure 3 shows the overall situation of logged in users, Figure 4 shows the number of all clicks except login failures, and Figure 5 shows the range of independent visitors.

Table 2 information of user record

Serial number (day)	Log record	Number of IP address	independent visitor	click rate	Number of access failure
1	64576	7597	8767	27845	2812
2	60298	6640	7978	21173	2145
3	97643	17503	9985	28463	1762
4	67434	8985	10367	24675	2334
5	89703	18439	12447	33432	997
6	69454	15872	8743	22374	1792
7	87255	26643	11343	29874	1469

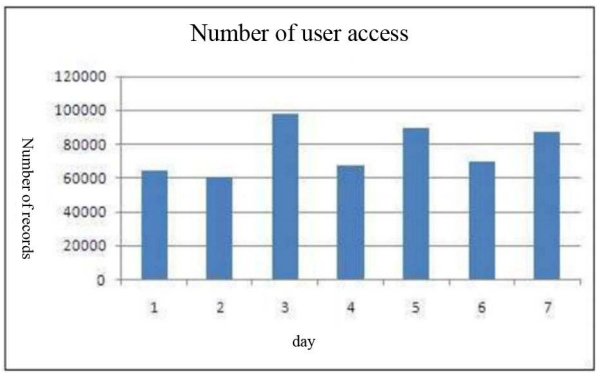

Figure 3. overall situation of logged user

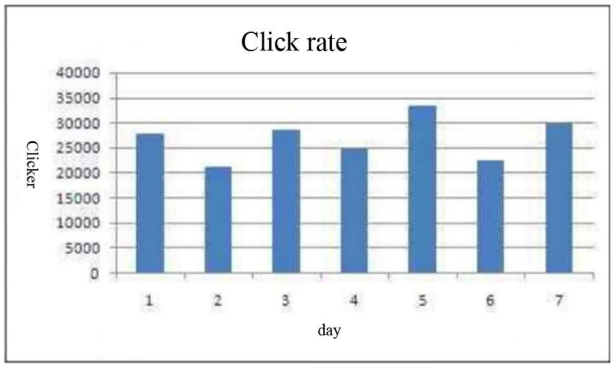

Figure 4. actual condition of click rate

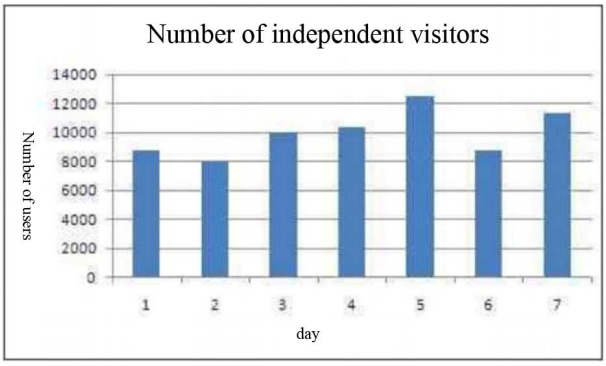

Figure 5. overall situation of independent visitor

V. CONCLUSION

Based on theory of Web log mining, this paper makes presentation and comparison for sources of a variety of logs, combining with the actual Web server log, analyzes contents of the log record in detail , finally, combining with web server logs, does multifaceted analysis for series of user access situation. The work can give specific guidance for the operators to optimize front page, improve user experience, as well as optimize structure, for example, dead chain in the website must be removed, unreasonable page jumps and substandard logic must be fine-tuned in detail , so as to retain the old users and attract new users further.

VI. REFERENCES

[1] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong et al.A Framework for Mining High Utility Web Access Sequences [J]. IETE technical review, 2011, 28 (1) :3-16

[2] Han JW, KAMBER M. data mining concepts and technology [M] Fan, X. Meng translation, Beijing: Mechanical Industry Press, 2001.

[3] Chen Baoshu, Jing Qimin. Data Preprocessing of Web Data Mining, [J] Computer Engineering, 2002, 28 (7):125-127.

[4] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang Ning Tan.Web usage mining: Discovery and Applications of usage patterns from web data [J]. SIGKDD Explorations, 2000,1 (2): 12-33.481-483.

[5] Song Jiangchun, Shen Junyi. Efficient and Pluripotent Mining Algorithm of Web Logs [J] Computer Research and Development, 2001, 38 (3):328-333.

[6] Zheng Mingchao. Comparison and Analysis of Classification Algorithm in Data Mining Technology [D] Lanzhou: Lanzhou Commercial College, 2007.