SESSION IDENTIFICATION BASED ON TIME INTERVAL IN WEB LOG MINING

Zhuang Like, Kou Zhongbao and Zhang Changshui
State Key Laboratory of Intelligent Technology and Systems,Department of Automation, Tsinghua University, Beijing 100084, P.R.China

Abstract: In this paper, we calculate the time intervals of page views, and analyze the time intervals to obtain a certain threshold, which is then used to break the web logs into sessions. Based on the time intervals, frequencies for each interval are counted and frequency vectors are obtained for each IP. Some IPs with special features of frequency distributions can be deemed as single users. For these IPs, we can define threshold for each individual IP, and separate sessions at the points of long access time intervals.

Key words: Web log mining, session identification, time interval

1. INTRODUCTION

The World Wide Web continues to grow at an astonishing rate in both the volume of traffic and the size and complexity of Web sites. The technique of Web mining just tries to acquire useful information and knowledge from the huge amount of information in WWW. Web mining can be classified into three categories: content mining, structure mining, and usage mining [1]. Web usage mining focuses on providing efficient techniques to analyze and predict behaviors of users while users access the Web. The data used for usage mining is the web logs. Through usage mining, behaviors of users, such as access frequency and regularly visited web page etc., are analyzed and the corresponding rules are obtained. Furthermore, the rules are used to improve the design of web site. Some software tools for web log analysis have been designed and used for real-world applications. [2-4]

The first step for web log mining is the pre-processing of the raw log data, which is the most important step for its results will impact the subsequent steps of pattern discovery and pattern analysis directly [5].

Two typical methods for web log mining are proposed by Chen et.al. [6] and Han et.al.[7]. Chen et.al. [6] introduce the concept of using the maximal forward references in order to break down user sessions into transactions for the mining of traversal patterns. Han et.al. [7] have loaded Web server logs into a data cube structure in order to perform data mining as well as traditional On-Line Analytical Processing (OLAP). Both the methods include the tasks of user identification and session identification in pre-processing step.

Unique users must be identified for further session identification and analyses. While the existence of local caches, corporate firewalls and proxy servers make the task greatly complicated [8]. Many assistant techniques such as cookies and CGI scripts etc. are used to detect users. However if the users don't collaborate, the only thing to do is to deem the one individual IP as a user. In [9], if the agent log shows a change in browser software or operating system, each different agent type for an IP address are deemed as a different user. But there are still many users using the same browser and operating system visit the web site from the same IP address.

To do further pattern discovery and analyses, web logs need to be converted into user sessions. A session is a serial of page views of a single user when accessing the entire Web. The purpose of session identification is to separate the web logs of one user into individual sessions for further analysis. A simple method to identify sessions is to determine a threshold. If the time between page accessing exceeds the threshold, we assumed that the user starts a new session. On current research on this area, this threshold is determined from fifteen minutes to one hour based on experiences. Obviously, some users may browse a site hastily and others may spend more time when they begin a page view. Thus using the uniform threshold for all users is not proper and bring out imprecision.

In this paper, we calculate the time intervals of page views for each IP address, and then count the frequency for each interval and obtain the frequency vector, which is used as the feature of one IP. Some IPs with special features of frequency distributions can be deemed as single users. For these IPs, we can define threshold for each individual IP, and divide sessions at the point of long access time interval.

The reminder of this paper is organized as follows: In Section 2, we establish the frequency vectors of IPs. Section 3 details and discusses several typical frequency distributions. We propose a novel method to analyze the feature of access frequency and estimate the threshold of session in Section 4. The last section is our conclusion.

2. FREQUENCY VECTORS OF IPS

The website of China National Tourism Administration (CNTA) [10] is a famous HTTP sever of China, which have large amounts access volume. The web logs from Oct 1st, 2001 to Oct 9th, 2001 of CNTA are adopted in our experiments. We delete some redundant information of the web logs, such as some assistant files with suffixes ``.gif", ``.swf", ``.css", ``.jpg", ``.cgi" etc.[5]. After preprocessing the raw data, we get 547,279 access records, all of which visited the sever from 21,438 IP addresses.

In Table 1, we show some access records of the data. For every record, it includes starting time of the access, URL, size, and so on. The starting time of a page view is key factor in our analysis of the session threshold. We define the time interval of page views as:

$$Diff\_T_k = T_{k+1} - T_k, \qquad k = 1, 2, \cdots, N-1, \tag{1}$$

where N is the amount of page views of one IP, T_k is the starting time of the k page view.

Table 1. A sample of server log

Date	Time	IP	Web page
10/02/2001	09:06:38	203.204.71.242	/26-zsyz/2j/dfly-12.asp
10/02/2001	09:07:41	203.204.71.242	/26-zsyz/2j/dfly-13.asp
10/02/2001	09:08:53	203.204.71.242	/26-zsyz/2j/dfly-14.asp
10/02/2001	09:09:58	203.204.71.242	/26-zsyz/2j/dfly-15.asp
10/02/2001	09:47:08	203.204.71.242	/12-gwcy/shopping/sichou.htm
10/02/2001	09:48:10	203.204.71.242	/12-gwcy/shopping/cixiu.htm
10/02/2001	09:49:38	203.204.71.242	/12-gwcy/shopping/china.htm
10/02/2001	09:49:57	203.204.71.242	/12-gwcy/shopping/tu/cixiou.jpg
10/02/2001	09:51:01	203.204.71.242	/12-gwcy/shopping/qiqi.htm

Figure 1 shows the time intervals of a serial of page views. In Figure 1, most of the time intervals converge from about 20 seconds to 140 seconds. We believe that in a normal session, the interval time of page views distribute in a range of value, the appearance of an extremely larger value may mean the beginning of another session. Then the 68th page views can be deemed as the beginning of a new session for the time interval between the 67th and the 68th page views is 2230 seconds, which is much larger than average value. The 68th page view just corresponds to the fifth access record showed in Table 1. For a special IP, the threshold to divide sessions can be obtained by the statistical analysis of the time intervals. This statistic is the frequency vector of the IP.

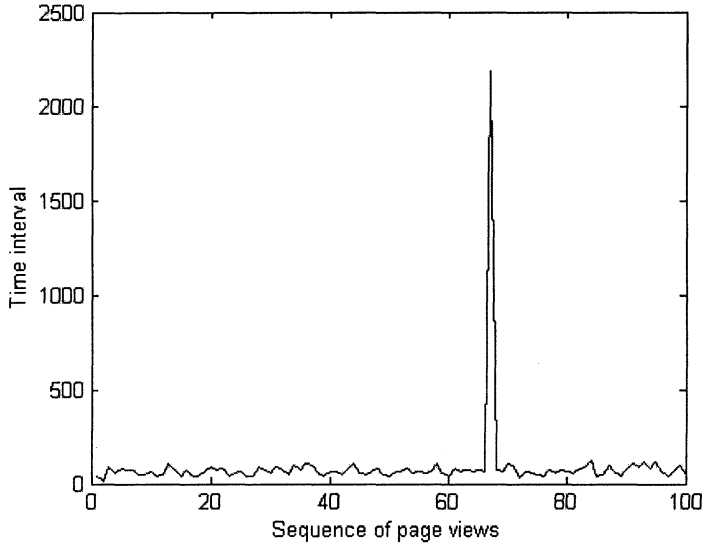

Figure 1. Time interval of a serial of page views

We define the frequency vector of an IP as

$$X^{(i)} = \{x_k^{(i)}, k = 1, 2, \cdots, M\}, \tag{2}$$

where i denotes the IP, k is the time interval between two adjacent page views, the unit is second. The value $x_k^{(i)}$ is the times that time interval k appears in the logs from IP i. M is the maximal time interval.

Since we think the very large value of time interval means the end of a session and starting of another one, these values don't express users' access behavior in one session. When take off 0.5% largest data in all access time intervals, the remaining data are with the maximal time interval of 897, which is just the M value we adopted. Thus, the frequency vectors of all IPs are established.

3. FEATURES OF FREQUENCY VECTORS

Figure 2 shows the cumulated frequency distribution for all the data, the frequency S_n is acquired by summing up the frequency vectors of all IP addresses.

$$S_n = \sum_i^n X^{(i)} \tag{3}$$

The value n is the number of IPs in the data. The distribution of S_n exhibits a straight line in logarithm histograms, which shows power-law distribution. Power-law characteristics are shared by many nature and society system, such as computer systems and Internet. That is, the frequencies of events to their sizes are often exhibiting power-law distributions. [11,12]

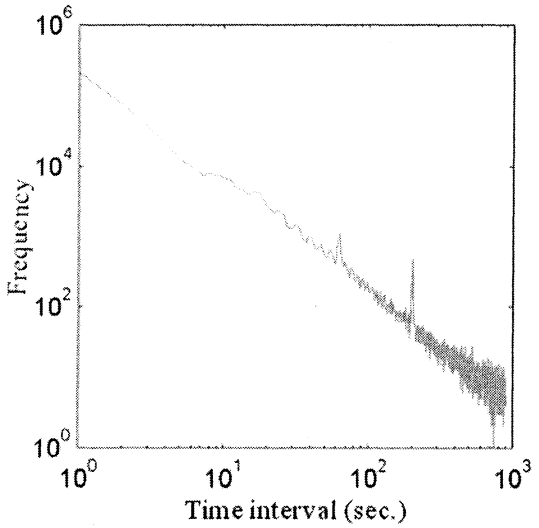

Figure 2. Frequency-time interval distribution for all IP addresses

For individual IPs, the frequency distribution can also be acquired. We observed these distributions of 25 IPs and get some interesting features. Two typical frequency distributions are showed in Figure 3.

The frequency distribution in Figure 3(a) also exhibits Power-law characteristics, just as the distribution of S_n showed in Figure 2. There are 101386 requests from this IP in 9 days, and from the figure, most of the accesses have very small time intervals, some requests are even occurred simultaneously. It seems that many users are accessing the web site according to the same IP, so it is probably a proxy server. We examined the IP and validated our supposition. For this kind of IP, session identification is extremely difficult, for they mixed with many users with different access features. Other techniques, such as using cookies or combining

content/structure information can be used to deal with these kinds of complexity.

However, distribution in Figure 3(b) exhibits Gaussian Characteristic. We think it shows the browsing behavior of one unique user. For a frequenter of a web site, the feature of his access web pages reflects his/her psychologies, interests, reading style and reading speed etc., while has little thing to do with the content of the web pages. Then we can identify unique users according to the frequency distribution.

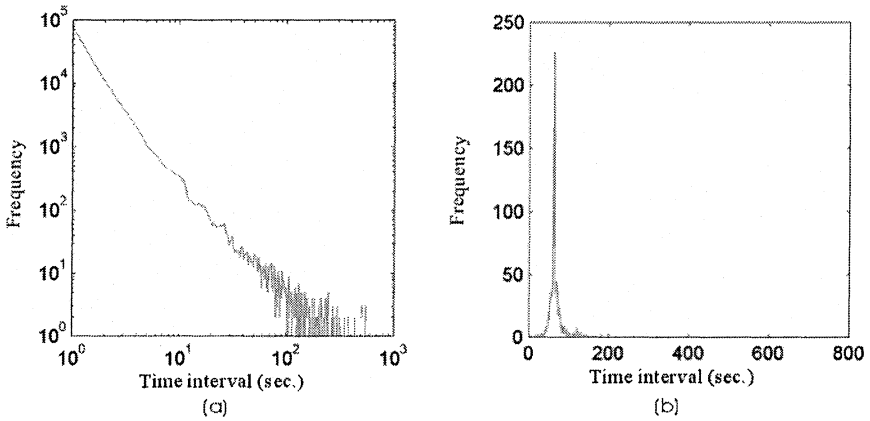

Figure 3. Two typical frequency distributions

4. THRESHOLD OF SESSIONS FOR INDIVIDUAL IPS

From Figure 3(a), we can derive that there are many accesses with small value of time interval when IP serve as a proxy server, for concurrent accesses to the web site split the time interval to pieces. On the other hand, if there are not accesses with small value of time interval, we can deem the IP as a unique user. In Figure 3(b), stable behavior of unique user is acquired. For these cases, sessions can be divided at the threshold of time interval.

If we have determined that an IP is used by one special user, we can calculate the mean μ and standard variance σ according to the frequency distribution. Then the threshold to divide sessions can be defined as $\mu + 3\sigma$. The threshold of two IP whose frequency distribution are showed in Figure 3(b) are calculated as $453.8(65.9+3\times 129.3)$.

Take the IP which shows the frequency distribution in Figure 3(b) as a example. Three are 1586 access records from this IP address. We do the

session identification step to these logs using our proposed method together with traditional uniform threshold method.

For our time interval based methods, we separate sessions when time interval is larger than the threshold 453.8 seconds that calculated above. Then the access serial can be divided into 27 sessions. The duration of sessions varies from 60 seconds to 12361 seconds. In Table 2, we list part of the accesses in a span of time, the column of "#" is added by the author for convenience. Using our method, the session can be separated between record 2^{nd} and 3^{rd},and between 112^{th} and 113^{th}, so this segment of logs can be separated into 3 sessions. It is proper in our intuition.

If the traditional uniform threshold are adopted, a fixed threshold are used to divide session, things may be different. For example, truncate sessions when 30 minutes arrived, all the logs from the IP are divided into 69 sessions. And the logs in Table 2 are divided into 6 sessions, the breakpoint are record 3^{rd}, 30^{th}, 61^{st}, 86^{th} and 113^{th} respectively. Obviously, this separation is reasonless, one serial of access are interrupt into pieces.

Table 2. Sample of logs for session identification

#	Date	Time	IP	Web page
1	10/04/2001	12:05:11	66.77.74.212	/12-gwcy/index.asp
2	10/04/2001	12:06:47	66.77.74.212	/23-dfly/index.asp
3	10/04/2001	18:00:42	66.77.74.212	/ziliao/ztjj/lyuxcs-7.asp
4	10/04/2001	18:02:47	66.77.74.212	/23-dfly/2j/gd.asp
	...	...	...	...
29	10/04/2001	18:30:33	66.77.74.212	/22-zcfg/dfly.asp
30	10/04/2001	18:31:33	66.77.74.212	/21-wxzw/2j/zxq-1.asp
	...	...	...	...
60	10/04/2001	19:00:36	66.77.74.212	/ziliao/zlk/2000gn100q.asp
61	10/04/2001	19:01:35	66.77.74.212	/30-bkzz/jianjie.asp
	...	...	...	...
85	10/04/2001	19:30:54	66.77.74.212	/HTML/point/whyc.htm
86	10/04/2001	19:31:54	66.77.74.212	/23-dfly/2j/sh.asp
	...	...	...	...
111	10/04/2001	19:58:56	66.77.74.212	/ziliao/lyjjyj/50.asp
112	10/04/2001	19:59:59	66.77.74.212	/31-lysd/fs.asp
113	10/05/2001	2:16:17	66.77.74.212	/12-gwcy/shopping/m_hebei.htm
114	10/05/2001	2:17:36	66.77.74.212	/ziliao/zlk/lxcywnj7.asp

5. CONCLUSION

In this paper, we propose a new method based on analysis of time intervals to do session identification. We separate sessions when the time interval is larger than a given threshold, which can be calculated by

analyzing the features of accesses of individual IP addresses. It is different from traditional method that defines uniform threshold based on experiences.

We first calculate the time intervals of page views for each IP, and then count frequencies for each interval and obtained frequency vector for each IP. Some IPs show stable features in frequency distributions of time intervals, and they can be deemed as single users. On this condition, we can calculate threshold for each individual IP, and separate sessions at special points, where the access time interval is larger than the threshold.

For IPs used by many users, a large amount of accesses are often appeared simultaneously. The time intervals are spited into pieces. For these kinds of IP, assistant technologies are needed to identify users. The method proposed in this paper for session identification is not proper.

REFERENCES

1. R.Kosala and H.Blockeel. Web mining research: a survey, *ACM, SIGKDD*, 2000.
2. SoftwareInc. Webtrends. http://www.webtrends.com,1995.
3. OpenMarketInc. OpenmarketWebreporter. http://www.openmarket.com,1996.
4. NetGenesisCorp. Netanalysisdesktop. http://www.netgen.com,1996.
5. J.Srivastava, R.Cooley, M.Dehpande, and P.N.Tan. Web usage mining: Discovery and applications of usage pattern from web data. *SIGKDD Explorations*, 1(2), 2000.
6. M.S.Chen, J.S.Park, P.S.Yu. Data mining For Path traversal patterns in a Web environment. *In:Pro cof the16th Int'l Confon Distributed Computing Systems,* HongKong,1996.
7. O.R.Zaiane, M.Xin, and J.Han. Discovering Web access patterns and trends by applying OLAP and datamining technology on Weblogs. *In:ProcofAdvances in Digital Libraries Conf. Santa Barbara*, CA,19-29,1998.
8. R.Cooley, B.Mobasher, and J.Srivastava, Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1), 1999.
9. P.Pirolli, J.Pitkow, and R.Rao. Silk from a sow's ear: Extracting usable structures from the Web. *In Proc. Of 1996 conference on Human Factors in Computing Systems (CHI-96)*, Vancouver, British Columbia, Canada, 1996.
10. http://www.cnta.com
11. Q.Chen, H.Chang, R.Govindan, etc. The Origin of Power Laws in Internet Topologies Revisited, *Proc. of IEEE Infocom*, 2002.
12. B.A.Huberman and L.A.Adamic, The nature of markets in the World Wide Web, *Quarterly Journal of Economic Commerce*, 1():5-12, 2000