Global Journal of Computer Science and Technology: E
Network, Web & Security
Volume 15 Issue 3 Version 1.0 Year 2015
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Web usage Mining: A Novel Approach for Web user Session Construction

By Neha Sharma & Pawan Makhija
SGSITS University, India

Abstract- The growth of World Wide Web is incredible as it can be seen in present days. Web usage mining plays an important role in the personalization of Web services, adaptation of Web sites, and the improvement of Web server performance. It applies data mining techniques to discover Web access patterns from Web log data. In order to discover access patterns, Web log data should be reconstructed into sessions. This paper provides a novel approach for session identification.

Keywords: web mining, web server logs, web usage mining (wum), preprocessing, session identification.

GJCST-E Classification : H.3.5

Strictly as per the compliance and regulations of:

Web usage Mining: A Novel Approach for Web user Session Construction

Neha Sharma [α] & Pawan Makhija [σ]

Abstract- The growth of World Wide Web is incredible as it can be seen in present days. Web usage mining plays an important role in the personalization of Web services, adaptation of Web sites, and the improvement of Web server performance. It applies data mining techniques to discover Web access patterns from Web log data. In order to discover access patterns, Web log data should be reconstructed into sessions. This paper provides a novel approach for session identification.

Keywords: web mining, web server logs, web usage mining (wum), preprocessing, session identification.

I. Introduction

Web Usage Mining deals with understanding of user behavior, while interacting with web site, by using various log files to extract knowledge from them. This extracted knowledge can be applied for efficient reorganization of web site, better personalization and recommendation, improvement in links and navigation, attracting more advertisement. As a result more users attract towards web site hence will be able to generate more revenue out of it. [1]. Web Usage mining is made up with three procedures, as data preprocessing, data mining and pattern analyzing. Data preprocessing contains three steps as data cleaning, user identification, session identification. Session identification is an important step in data processing of web log mining. A session is defined as a group of requests made by a single user for a single navigation. A user may have a single or multiple sessions during a period of time. Presently sessions are identified either on Time based method or Navigation based method. Here, we proposed a novel approach for user session identification by combining both Time based method and Navigation based method.

II. Motivation

Web log mining is to discover the mode of users' accessing to web page through mining web logs. In the process, the designer's knowledge fields, the rate of his interesting and the users' visiting habit can be refined, which can optimize the site's structure, develop individual service and the control of the users that is useful strategies information for the designers and the managers. The most important and time- consuming link in mining web logs is the session identification in web log preprocessing. The users' session is a session aggregation covering more than one web services. The aim of session identification is to divide the users' page into an isolated identification.

III. Related Works

The focus of literature review is to study, compare and contrast the available session identification techniques. Traditional session identification algorithm is based on a uniform and fixed timeout. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes to 24 hours while 30 minutes is the default timeout by Cooley [2]. If the interval between two sequential requests exceeds the timeout, new session is determined.

Timeout algorithm uses a pre-fixed value of threshold for session identification in which if the interval between two sequential requests exceeds the threshold value, a new session is determined. According to He Xinhua and Wang Qiong [3], However, because of the uniform and fixed value, the algorithm cannot obtain efficient effect of session identification in several situations like (1) Different user results different reading speeds, (2) Even by the same user, different interest is shown on pages at different time, (3) Different page contains different contents. Therefore, the time taken is often different. They propose a session identification algorithm based on dynamic timeout, on the basis of traditional session identification algorithm. First, at beginning of the new session, the initial timeout is set for each page using the formula

$$\delta 0 = \alpha.t.(1 + \beta)$$

Where α denotes smooth coefficient ranging from 1.1 ~ 1.6 and β is an influence factor depend on link in and link out of the page.

Second, while requested page is put into the current session, the timeout will be recomputed selectively in order to make the timeout reflect the character of session using the formula

$$\delta' = \delta 0(tnew + t0) / 2t0$$

Where *t0* denotes primal timeout of the page, and *tnew* denotes the timeout of the page that put into

Author α σ : Department of Information Technology, SGSITS Indore (M.P.), India. e-mails: 1ne3haa@gmail.com, ne3haa@gmail.com, pawanmakhijaacro@gmail.com

current session and δ0 denotes the timeout by the adjustment last time.

In [7] Jozef Kapusta, Michal Munk and Martin Drlík, assume that the user goes over several navigation pages during her/his visit until she/he finds the content page with required information. The content page is a page where the user spends considerably more time in comparison with navigation pages. The content page is considered, the end of the session. The division of pages into content and navigation pages is based on the calculation of cut-off time C. When the cut-off time C is known the session can be created in such manner that we compare the time of particular web page visit with the cut-off time C. The session is then defined as a path through the navigation types of pages to the content page (the user spent there more time then C), they claim the content page is last page of session. The cut-off time C is calculated on the basis of exponential distribution of variable RLength(Time spent by user on individual page), here the assumption is that the variance of the times spent on the auxiliary pages is small then the content page.

In [8] Zhixiang Chen, Richard H. Fowler and Ada Wai-Chee Fu, designed two algorithms for finding maximal forward references (longest sequences of Web pages visited by a user without revisiting some previously visited page in the sequence) from very large Web logs. They consider two types of sessions as α-interval session and β -gap session,

where α-interval session insures the duration of a session may not exceed a threshold of (30 minute) and β-gap session insures the time between any two consecutively assessed pages may not exceed a threshold of β (20 minute). They define a URL node structure to store the URL and the access time of a user access record and a pointer to point to the next URL node and then the maximal forward reference session is calculated using both interval session and gap session.

In [9] G. Arumugam, S. Sugana,Suggested algorithm which does not require searching whole tree representing server pages. They employs concept of efficient use of data structure. Array List to represent web logs and user access list, hash table for storing server pages, two way hashed structures for Access History List, represents user accessed page sequences. Experiments reveals less time complexity and good accuracy of sessions generated as compare to results of maximal forward reference method and reference length method.

In [10] Dr.Antony and V.Chitraa, proposed a new technique for identifying sessions for extraction of user patterns. In the proposed method a matrix is constructed in which columns are the web pages and rows are users. Browsing time (BT) for a particular page is determined by finding the differences between the time fields of two consecutive entries of a same user and assumption is that the website Administrators fix the minimum time and maximum time (BTmin and BTmax) for all web pages as per the contents. Codification of pages are performed on the basis of BT, BTmin and BTmax and the sessions are calculated on the basis of this code. The result is shown in the form of matrix.

In [11] Peng Zhu, Ming-sheng Zhao proposed an improved algorithm based on average time threshold value. Experiments are conducted on the log files of Nanjing University Extra net user access logs. Because data of log files are very large, They selected the log test algorithm of only one day(March 15, 2008). Algorithm proposed in this paper, takes individual differences into account to define the threshold value of users' browsing pages, and identify long session page views, and divide the session less than the threshold into the next session. They proposed two algorithms from which first algorithm constructs session of individual user and the second algorithm disconnect the previous session into parts if there is no hyper link between two consecutive entries of logs.

IV. Proposed Approach

As we seen in the literature review the sessions are identifying either on the basis of time spend by user on particular web page or on the basis of user navigation in web site topology.

Time based method ignores the web site structure, the sessions generated by such type of methods are not generated right sessions as users reading speeds reflects the sessions. While in navigation method if particular user not moves back,it not generates the sessions.

In our approach we combine both method to generate more informative sessions. Initially sessions are generated by Maximal Forward reference method on these sessions the time based method have been applied with the threshold value of 10 minutes. The experiment is conducted on the log data of www.smartsync.com of dated 8 Dec 2013.

V. Testing and Results

The input data in this case are the access log files of the www.smartsync.com web server. Because data of log files are very large, we select the log test dataset of only one day (dated 8 Dec 2013) of size 1 GB, 2 GB and 4 GB. The Table-1 showing the number of session generated by existing approach and our approach.

Table1: Number of Sessions generated by various method

Name of Method	No. of Session generated 1 GB Dataset	2 GB Dataset	4 GB Dataset
1.Time based Method	983	1673	2875
2.Maximal Forward reference method	968	1640	2742
3.Proposed Approach	1120	1710	3102

Since the log data is very large in size it is not possible to count true sessions of whole data so we took 100 KB of data. In that data we manually found 53 true sessions and the number of session generated by the existing methods and the proposd method are compared. For finding the accuracy of proposed approach we have calc`ulated the ratio of generated session and true session. Table-2 showing the comparison of existing method and proposed method with true sessions. In the Table-2 S is number of sessions generated by methods and T is the true sessions counted manually.

Table 2 : Comparison of Session Generated by Existing Methods and Proposed Method with True Sessions

Methods	S	T	S/T %
Time based Method	42	53	79 %
Maximal Forward reference method	32	53	60 %
Proposed Approach	47	53	88 %

VI. Conclusion

The growth of the web has resulted in a huge amount of information that is now freely offered for user access. The several kinds of data have to be handled and organized in a manner that they can be accessed by several users effectively and efficiently. The experiment on 4 GB data shows that the new method proposed in this report generates more sessions (3102) than the traditional Time Based Method (2875) and Maximal Forward Sequence Method (2742). On comparing with the true sessions on 100 KB data, the accuracy of session is increased to 88%.

References Références Referencias

1. Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod, "Web Usage Mining: A Review on Process", IEEE 2013.
2. Robert.Cooley,Bamshed Mobasher, and Jaideep Srinivastava, " Web mining: Information and Pattern Discovery on the World Wide Web", *In International conference on Tools with Artificial Intelligence*, pages 558-567, Newport Beach, IEEE,1997.
3. He Xinhua, Wang Qiong, "Dynamic Timeout-Based A Session Identification Algorithm" , IEEE 2011.
4. J. Zhang, Ali A. Ghorbani, "The Reconstruction of user session from a server log using improved time oriented heuristic", *IInd Annual Confernnce on Communication Networks and Service Research*, IEEE, 2004.
5. Fang Yuankang and Huang Zhiqui, "A session identification algorithm based on frame page and page threshold", *Computer Science and Information Technology (ICCSIT)*, 3rd IEEE International Conference, 2010
6. R. F. Dell et al.,"Web user session reconstruction using integer programming", International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/ACM/WIC, 2008.
7. Jozef Kapusta, Michal Munk, Martin Drlík, "Cut-off Time Calculation for User Session Identification by Reference Length" IEEE 2012.
8. Zhixiang Chen, Richard H. Fowler and Ada Wai-Chee Fu," Linear Time Algorithms for Finding Maximal Forward References", *Intl Conf On Info Tech: Coding and Computing (ITCC03)*, Proc. of the 2003 IEEE.
9. G. Arumugam, S. Sugana, "Optimum algorithm for generation of user session sequences using server side web user logs", IEEE, 2009.
10. Dr.Antony Selvadoss Thanamani, V.Chitraa, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", *International Journal of Computer Applications*, Volume 34– No.9, November 2011.
11. Peng Zhu, Ming-sheng Zhao," Session Identification Algorithm for Web Log Mining", IEEE 2010.
12. Nirmala Huidrom, Neha Bagoria, "Clustering Techniques for the Identification of Web User Session", *International Journal of Scientific and Research Publications*, Volume 3, Issue 1, January 2013.

This page is intentionally left blank