Web usage Mining: Web user Session Construction using Map-Reduce

Neha Sharma [α] & Pawan Makhija [σ]

Abstract- Web Usage Mining deals with the understanding of user behavior while interacting with the website by using various log files. The whole process of Web Usage Mining gets completed in three phases namely Data Preprocessing, Pattern Discovery and Pattern Analysis. Data Preprocessing is important because it takes 80% of the time of the whole process of Web Usage Mining. Data Preprocessing involves Data Cleaning, User Identification, and Session Identification. In Session Identification we find out the set of pages visited by a user within the duration of one particular visit to a website, also called as Sessionization.

In paper[1], we proposed a new method for session construction. As the size of log files are very large so there is a requirement of an approach for Session Identification by which processing time of our proposed method will be reduced to a great extent.

In this paper, we used Map-reduce method to calculate sessions in which we combine both time and user navigation method. This approach is faster than the existing approach because we have performed the whole process in distributed environment.

Keywords: web mining, web server logs, web usage mining (WUM), map reduce, session identification.

I. Introduction

Web Usage Mining deals with observing user behavior, while interacting with web site, by accessing various log files to extract knowledge from them. This knowledge can be applied for reorganizing the website contents by giving a personalization and recommendation that is more efficient as compared to previous one by improving the links and navigation which in turns increase the rate of advertisement. This will results the users to access the website in a comfortable manner which obviously generate more revenue to them.[2] This scheme comprises of three steps as data preprocessing, data mining and pattern analyzing. Data preprocessing contains three steps as data cleaning, user identification, session identification. Session identification is an crucial step in data processing of web log mining. A session is defined as multiple requests made by a user for a single navigation. A user may have a single or multiple sessions during a particular period. Basically sessions are identified either by Time based method or by Navigation based method. Here, we proposed a unique approach for user session identification by blending Time based method with Navigation based method to get better results.

To increase the pace of Sessionization, the process is performed on distributed systems using Map-reduce. Map- reduce [3] is a programming model and an associated implementation for processing and generating large data sets that supports fault tolerance, automatic parallelization, scalability, and data locality-based optimizations. Users define a Map function that will use this key/value pair for processing the data to generate a set of intermediate key/value pairs and a Reduce function will be called that concatenates all intermediate values related with the same intermediate key.

II. Motivation

Map Reduce is a programming model and an associated implementation for processing and generating large data sets. This process takes a set of input key and value pairs and generates an list of key and value pairs. The user of the Map-Reduce library classifies this calculation as two function as map and reduce functions.

The Map function takes a pair of input and generates a list of intermediate key and value pairs. The values grouped with the help of the Map-Reduce library is fed to the Reduce function.

The Reduce function accepts the output that was generated by the library as value and key pair, merge them to produce a small set of values e.g. zero or one value. The intermediate values that were produced during invocation are feded into the Reduce function with the help of an iterator. This will enable the user to handle large set of values so that it will be stored easily in the memory.

III. Proposed Approach

In order to enhance the performance of the proposed method in [1], we have used Map-Reduce method to lower the session generation time.

We have applied Map-Reduce on the time-based method, maximal forward sequence method and our proposed method[1]. The results that were generated during this approach has tremendously reduces the session generation time as it was fasten up by the Map and Reduce function.

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

The experiment is performed on the log data of www.smartsync.com on 8 Dec 2013.

IV. Testing and Results

The input data that was supplied during our proposed method are the access log files of the www.smartsync. com web server. Because data of log files are large, we have taken the log dataset of only one day (dated 8 Dec 2013) of size 1 GB, 2 GB, and 4 GB. Table-1 shows the time required for completing the process on a single system and multiple systems (ET=Execution Time):

Table-1: Comparison of Time Requirement by an Existing Methods and the Proposed Method on Single System and Multiple System

	Methods	Time in (Seconds)		
		1 GB dataset	2 GB dataset	4 GB dataset
1.	ET on Single System by Time based Method	4932	10221	19312
2.	ET on Single System by Maximal Forward Sequence Method	5014	9142	19514
3.	ET by Proposed Approach (Single System)	4821	8912	19455
4.	ET on Multiple Systems by Time based Method	2187	5346	10132
5.	ET on Multiple Systems by Maximal Forward Sequence Method	2034	5674	10314
6.	ET by Proposed Approach (Multiple Systems)	2512	5112	10248

Figure-1 shows the graphical representation of Table-1 for comprising the time requirement in completing the process by an existing method and the proposed method.

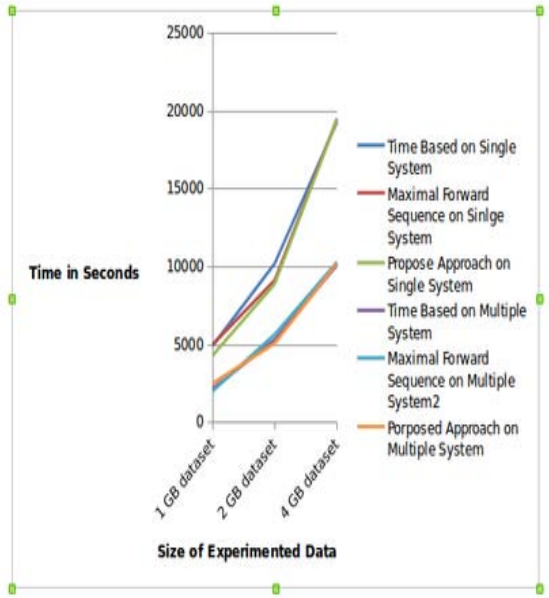

Figure-1: Graph for comparison of various time requirements

V. Conclusion

The information available on the web is increasing day by day in a fast manner. This lets the user to have a lot of data to access freely on the web. Our method have generated sessions that took less time comparable to the existing method. The experiment on 1GB, 2GB, and 4GB data shows that the new method proposed in [1] generates more sessions (3102) than the traditional Time Based Method (2875) and Maximal Forward Sequence Method (2742). As per the result shown in Table-1 with the proposed approach, this process takes less time in completion because of Map Reduce method.

References Références Referencias

1. Neha Sharma, Pawan Makhija, "Web Usage Mining:A novel approach for web user session construction" GJCST vol. 15 issue 3 version 1.0, 2015.
2. Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplifed Data Processing on Large Clusters" OSDI 2004.
3. Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod, "Web Usage Mining: A Review on Process", IEEE 2013.

4. 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.
5. He Xinhua, Wang Qiong, "Dynamic Timeout-Based A Session Identification Algorithm", IEEE 2011.
6. 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 .
7. 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.
8. Jozef Kapusta, Michal Munk, Martin Drlík, "Cut-off Time Calculation for User Session Identification by Reference Length" IEEE 2012.
9. 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.
10. G. Arumugam, S. Sugana, "Optimum algorithm for generation of user session sequences using server side web user logs", IEEE, 2009.
11. 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.