ISSN:2231-0711

Girish Mahajan et al | IJCSET(www.ijcset.net) | September 2015 | Vol 5, Issue 9,341-343

User Sessionization using DFS: Skillful Way to Obtain User Sessions

[1]Girish Mahajan, [2]Makrand Samvatsar

[1] *Student M. Tech. (SS), Patel College of Science and Technology, Indore (MP), India*

[2] *Asstt. Prof., CS Deptt., Patel College of Science and Technology, Indore (MP), India*

Abstract **- In present scenario, every business need input regarding user's metadata related to behavioral and situational while accessing information through World Wide Web and also they just wanted to utilize this information in taking crucial and bold decision as quick as possible.**

Specially segment like online shopping, airline reservations, online gaming, online food chains are few examples from lot many businesses, where this kind of information play a vital role and affect profitability up to a great extent. When customer or user visit any web site over WWW, web servers and application servers tracks the audit information in logs.

The growth of web is tremendous as approximately one million pages are added daily. Because of it the web log files are growing at a faster rate and the size is becoming huge. These logs can be used to identify the patterns which can be used by any business. Patterns discovery [5] and analysis process required user sessions as input.

This paper suggests, user session identification process can be improved by combining right available techniques to get more effective and accurate results and using distributed file processing system like Hadoop, the overall processing can be speedup to a great extent.

Keywords **- Web Mining, Data Preprocessing, Pattern Analysis, Hadoop, Distributed File System.**

I. INTRODUCTION

As we can see in the world of internet, web sites are playing huge role in providing services to end user. This sites may vary from general purpose static content sites to very crucial, sensitive and high performant sites like online shopping, online food chains, and online insurance and banking. These web sites are useful source of information in day-to-day activities. This is the only reason that there is a rapid development of the World Wide Web in its volume of traffic, size and complexity of web sites.

Web Mining is also categorized in Web Content Mining, Web Structure Mining and Web Usage Mining. Web Mining traces user visiting behaviors and extracts their interests using patterns. Because of its direct application in e-commerce, web analytics, e-learning and information retrieval, Web Usage Mining has become one of the sensitive and crucial areas in Internet world.

Analyzing such data can help these organizations determine the life-time value of clients, design cross-marketing strategies across products and services, evaluate the effectiveness of pro-motional campaigns, optimize the functionality of Web-based applications, provide more personalized content to visitors, and find the most effective logical structure for their Web space.

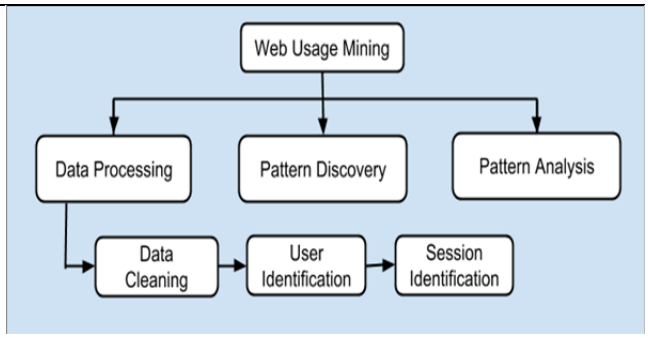


Figure 1: Web Usage Mining

There are three steps in pre-processing of log data [1].

- Data Cleaning
- User Identification
- Session Identification

Data preprocessing is vital phase of Web usage mining because of the complex nature of the Web architecture. The log file are processed to get reliable sessions for efficient mining. It requires data cleaning, user identification and finally user session identification.

Data cleaning involves removing irrelevant data from the input logs. User identification is the process of drives on URLs referencing with same IP address with different users. Identification of sessions is actually breaking references into user sessions [6] based on different methods. There are mainly two techniques available for session identification, Times based method and Navigational based method.

This paper focuses on effective way if using combination of best of both technique to get maximum number of user session as compared to using individual method for the same.

II. DATA PREPROCESSING

a. *Data Cleaning*: The primary objective of data cleaning is to eliminate irrelevant from log files. Raw log file may contain lots of unwanted data like metadata for graphics, videos, failed requests, local/global noise etc. which need to remove before using it.

b. *User Identification*: Ideally the user Identification is nothing but to identify the corresponding user from each log record. It is very difficult task to extract accurately distinguish users due to the existence of local cache, proxy servers and firewalls.

The fields which are useful to find unique users and sessions are [4]:

- IP address
- User agent
- Referrer URL

c. *Session Identification*: It is a process to discover different user sessions from the different types of logs specially web access log. Typically user session is a series of web pages browse in a single access. The goal of session identification is to divide the page accesses of each user into individual sessions. This process is also known as sessionization.

III. Literature Survey

a. Chitraa et al. [3] enhances Data Cleaning to remove irrelevant records from log file. They conduct an experiment in the proposed technique to obtain record consists of 2000 records in the log file in effective way.

b. Shaily et al. [7] suggested that in order to take full advantage of web usage mining and it's all techniques, it is important to carry out preprocessing stage efficiently and effectively.

c. Sanjay et al. [2] tried to deliver areas of preprocessing including data cleansing, session identification, user identification, etc. Once preprocessing stage is well-performed, we can apply data mining techniques using clustering, association, and classification for applications of web usage mining such as business intelligence, e-commerce, e-learning, personalization etc. Map Reduce provides best platform to process such files with huge set of data in MBs or GBs etc. Hadoop [8] provides perfect facility for such scenarios.

IV. Proposed System

Proposed system we have suggested to use the combination of Time based and Navigational methods sequentially for get the maximum number of sessions out of files with huge size of logs in it. We can write MapReduce jobs for each of the steps from Data Cleaning, User Identification and Sessionization.

A very powerful way to handle huge amount of data is by using HDFS, Hadoop Distributed File System, which provides way to distribute data among several machines connected in a network called cluster. Map-Reduce provides creation of such queries which run on all nodes trough mapper and collect the individual result to form as a whole in reducer.

Using the proposed system tool can handle any size of log file and there would be no limitation of the system. Also entire operation would be fail safe due to underlying mechanism provided by Hadoop and MapReduce.

This research suggest implementation of each sessionization [6] process using Hadoop Map-Reduce to improve processing performance. User session identification process can be improved by combining right available techniques to get more effective and accurate results and using distributed file processing system like Hadoop, the overall processing can be speedup to a great extent.

V. Result Analysis

An experiment is an orderly procedure carried out with the goal of verifying or establishing the validity of proposed system. Results of experimentations show the comparison of existing and proposed system.

The input data in this case are the access log files of web server. Because data of log files are huge in size, we select the log test dataset from different size varying from size 300 MB, 500 MB, 700MB and 1 GB.

Once cleaning the data, experiment is performed by Reference length based Method, Maximal Forward Sequence based Method and by Proposed Method on single system and multiple systems and the results are shown below.

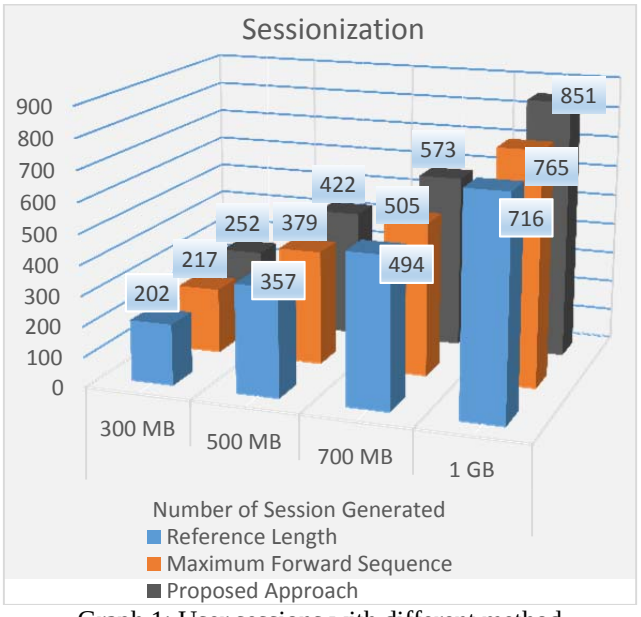
#	Name of Methods	Number of Session Generated			
		300 MB	500 MB	700 MB	1 GB
1	Reference Length	202	357	494	716
2	Maximum Forward Sequence	217	379	505	765
3	Proposed Approach	252	422	573	851

Table 1: User session comparison

VI. Conclusion

Our program generates the User Session Sequences in Distributed Environment. The experiment on 1 GB data shows that the new method proposed in this report generates more sessions (851) than the traditional Reference Length Based Method (716) and Maximal Forward Sequence Method (765).

On comparing with the true sessions on average size of data, the accuracy of session is increased drastically. This process takes lesser time in completion because of Map Reduce.



Graph 1: User sessions with different method

REFERENCES

[1] Dilip Singh Sisodia, Shirish Verma "Web Usage Pattern Analysis through Web Logs: A Review" IEEE 2012.

[2] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters" OSDI 2004

[3] V. Chitraa, Dr. Antony, Selvdoss Thanamani, "Web Log Data Cleaning For Enhancing Mining Process" Volume 01 – No.11, Issue: 03 December 2012, IJCCTS.

[4] Sheetal A. Raiyani , Shailendra Jain, "Efficient Preprocessing technique using Web log mining" International Journal of Advancements in Research & Technology, Volume 1, Issue6, November-2012.

[5] S. Baron, M. Spiliopoulou, and O. Gunther. "Efficient monitoring of patterns in data mining environments," Proceedings of 7th East European Conf. on Advances in Databases and Inf. Sys (ADBIS 03), LNCS, Springer, Sept. 2003, pp. 253-265.

[6] Yan Li and Boqin FENG "The Construction of Transactions for Web Usage Mining". International Conference on Computational Intelligence and Natural Computing, IEEE, 2009.

[7] L. Shaily, B. Mehul and M. Darshak. "Pre-processing: Procedure on Web Log File for Web Usage Mining". International Journal of Emerging Technologies and Advance Engineering (IJETAE), Dec 2012, Vol 2, Issue 12, Pg 419, 2012.

[8] Apache Foundation Web site, "HDFS Architecture". Apache Hadoop 2.4.1, Version: 2.4.1, 2014.

[9] K. Jozef, M. Michal and D. Martin. "User Session Identification Using Reference Length". DIVAI 2012 - 9th International Scientific Conference on Distance Learning in Applied Informatics, Pg. 175-184, 2012.

-