iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IX **Month of publication:** September 2020

DOI: https://doi.org/10.22214/ijraset.2020.31716

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

Finding User Navigational Patterns from Log Files using Hadoop Techniques: A Survey

Ruchi Patil[1], Ms. Pooja Trivedi[2]
[1]*Research Scholar, M. Tech CSE, CIIT Indore.*
[2]*Assistant Professor, CSE dptmnt, CIIT Indore.*

Abstract: ***In the World Wide Web the usage mining is used to find the user navigation patterns through extraction of knowledge from web usage logs. The effective mining produced by these log files and by the tools used for processing of these log files. In the course of these technologies extended log files can be created along with learning of user behavior. The learning of user behavior through user activities is analyzed while utilizing highly interactive systems. This paper presents the information of the methodology used, in which the focus is on learning the information-seeking process and on finding log errors and exceptions. The rest part of the paper discusses about the working and techniques used by web log analyzer.***
Keywords: Server Web Log files, Map Reduce, Hadoop, Web Mining

I. INTRODUCTION

Massive amount of information is carried by web pages that may not be concerned by the user. The key source of information is weblog data which concern the users visited links, time spent on page or link, browsing patterns.

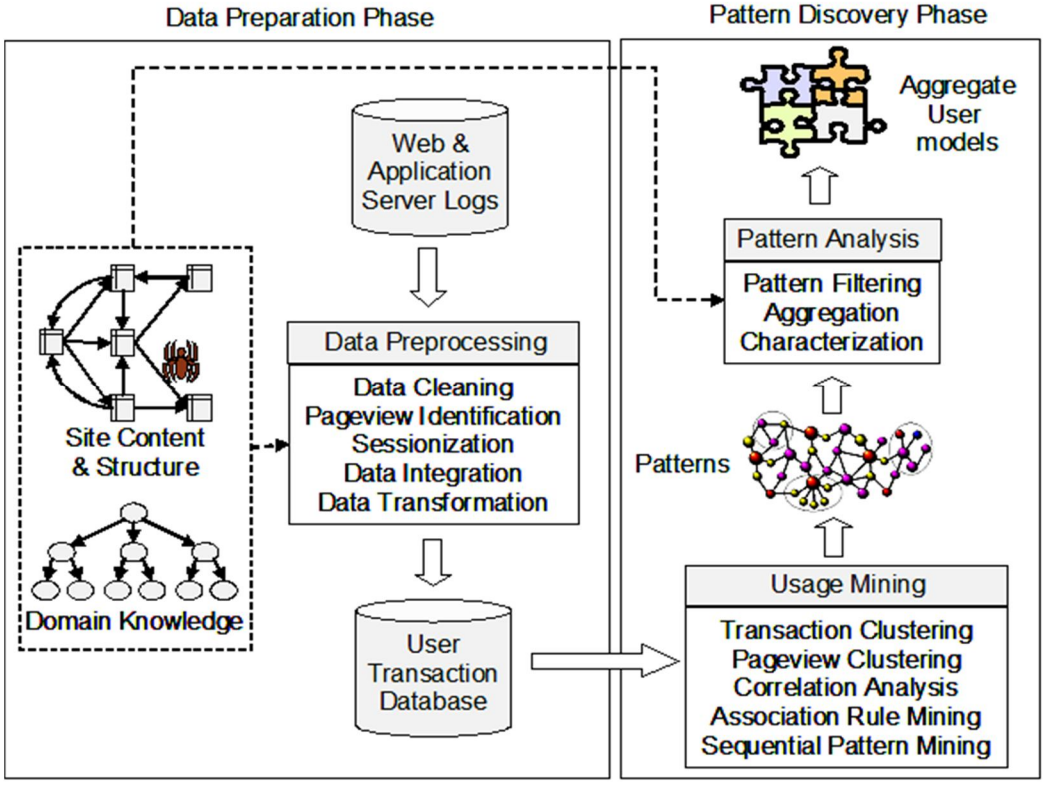


Figure 1 Process of Web Mining

This information is utilized for different applications like modified services, adaptive websites, , customer summary, generate attractive web sites etc. Web Usage Mining (WUM) is the chief application concerns mostly to the web data and estimate the user's visit behaviors and acquire their interests by examining the web log files. WUM preprocessing of log for cleaning data is the primary and essential step, to build it suitable for mining rationale [2].

A. Web Mining Consists of three types of Mining

1) Web Structure mining determines constructive knowledge from hyperlinks (or links), which correspond to the construction of the Web.

2) Web Usage mining identifies access patterns from Web usage logs, which documents each click made by each user.

3) Web Content mining: It extract useful information or knowledge from Web page contents.

Hadoop is an open source framework from Apache used to accumulate voluminous process and analyze data. MapReduce algorithm run by Hadoop, where the data is processed equivalent with others. In summary, for the computation of enormous and voluminous data hadoop platform is used.

B. Architecture of Hadoop has two Important layers as its Core

1) Processing/Computation layer (MapReduce),

2) Storage layer (Hadoop Distributed File System)

To counter the questions regarding security and compliance the server logs are analyzed by many IT organizations. IT organizations analyze server logs to answer questions about. A server log is a simple text file, used to record activities on the server. On the operations of network data capturing is taken by computer generated logs. Use full for managing network operations, especially for security and regulatory compliance. Various server log website proprietors are mainly attracted towards access logs which record hits and related information. These logs are in large amount thus resulting collection of large amount of data Big Data. Big data means really a big data; it is a collection of large datasets that cannot be processed using traditional computing techniques. It incorporates enormous volume, and variety with high velocity data. This data could be in any form like structured, semi structured or unstructured.

II. RELATED WORK

A. Samneet Singh et. al. this paper presents our cloud provider architecture that explores a search cluster for information indexing and query. The author proposed an integrated search head quarter cluster and semantic media wiki. This method accomplished through relaxation APIs to assist the investigation of cloud monitoring data. It facilitate person to delineate the admission and observe knowledge and prepare the processing results. This benefited from internet-based Media-Wiki interface

B. Joseph et.al. Explored different performance analysis for Hadoop Word Count workload utilizing different processors similar to Intel's and AMD's Bobcat E350. Analysis suggests that Hadoop Word Count is compute-sure workload in both map segment and scale down segment. It was observed that a higher efficiency/watt in comparison with AMD's Bobcat cluster could be achieved by Intel's ATOM cluster. Evaluating Intel's ATOM to Intel's Xeon X5690, the performance/buck for Xeon is better com- pared to the performance/buck for ATOM.

C. Yaxiong Zhao et al: suggested a cache framework for big-data functions based on knowledge-conscious. A novel cache description scheme and a cache request and reply protocol are designed. We enforce Dache by means of extending Hadoop. The obtained results illustrates that through Test bed Dache immensely progress on the finishing point time of MapReduce jobs.

D. Zhuoyao Zhang et al discovered a new framework for efficiency analysis. A keen examination is done on every map and concludes that knowledge processing stages and execution of every map duties contains typically and lessen. The methodology is concerned exclusively for MapReduce jobs and customer is delineate or the same. Customization is provided for services like scale back and handiest map.

E. NikzadBabaiiRizvandiet. al. presented an analytical system for modeling the dependency among configuration parameters and execution time of Map-diminish functions. The technique is accompanied in three stages named profiling, modeling, and prediction. In the first with some explicit elements of MapReduce configuration parameters an application can run on various occasions to outline the execution time of the pertaining on a given platform.

F. NikzadBabaii et.al. illustrated a provisioning method in context with MapReduce atmosphere for entire CPU usage in clock cycles of jobs.For a MapReduce job, a profile of complete CPU utilization in clock cycles is developed from the job prior executions with distinct values of two configuration parameters e.g., quantity of mappers, and quantity of reducers. For modeling the relation between configuration parameters and entire CPU utilization in clock cycles of the job a polynomial regression is utilized. We additionally in short learn they have an effect on of input information scaling on measured complete CPU usage in clock cycles. This derived mannequin together with the scaling outcome can then be used to provision the total CPU usage in clock cycles of the same jobs with one of a kind enter information dimension.

Table 2.1.Comparative Study of literature

S.No.	Paper Details	Work done	Limitation / Future Work
1	NehaGoel, Dr. C.K.Jha, "Preprocessing Web Logs: A Critical Phase In Web Usage Mining", 2015 (ICACEA),©2015 IEEE.	• An absolute tool for preprocessing phase. • Removes irrelevant and noisy data.	Restricted to few records and specific Website.
2	TanvirHabibSardar, Zahid Ansari, "Detection and Confirmation of Web Robot Requests for Cleaning the Voluminous Web Log Data", 2014 (IMPETUS), ©2014 IEEE.	• Used four methods for detection and confirmation of Web robots. • Robot.txt access check. • User agent check. • IP address check. • HTTP head request check.	Techniques represented are implemented in offline mode.Only one log format is used.
3	Shinil Kwon, Myeongjin Oh, Dukyun Kim, Junsup Lee, Young-Gab Kim, Sungdeok Cha, "Web Robot Detection based on Monotonous Behavior", ©Springer-Verlag Berlin Heidelberg 2012.	• Monitored the rate of behavioral changes in the user session known as "switching factor". • Presented switching factor for three features unassigned referrer field, file types and number of bytes from clients to the server.	• Real time detection is feasible. Real world demonstration is the major task for the future.

III. WEB LOG FILE

It is a sort of registry of pages based on web logs that can be accessed by various users at different moment of time. It can be maintained at both client and server side or at a proxy server. From all have its own pros and cons on searching the user navigation patterns and user relevant patterns. [5] Server Log: Stores data about requests carry out by client, therefore data consider commonly presently one source. Server Log details are given in Figure 2.

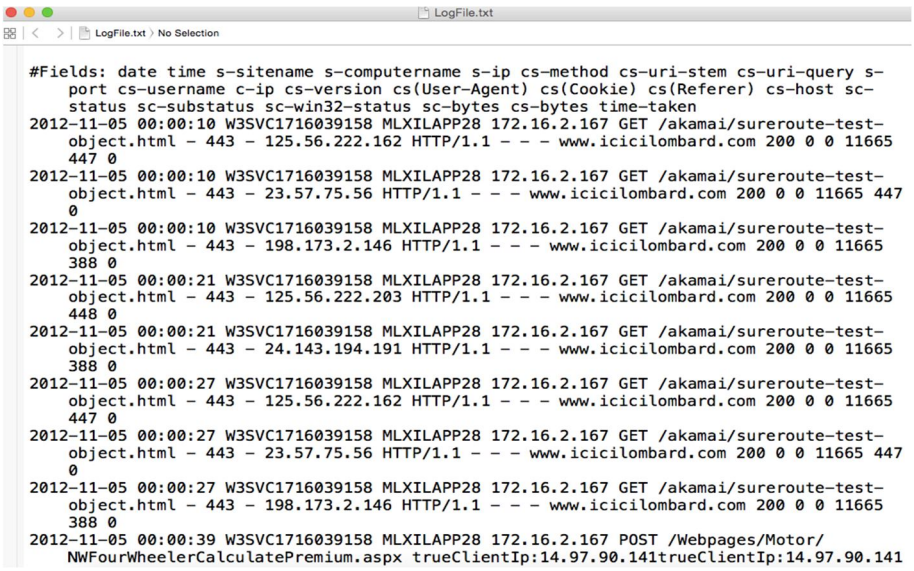


Figure 2 A Sample of Server Side Web Log

Client Log : Users behavior information are sent by the client itself to a repository. It can be executed by remote agents or via modification of source code of an accessible browser (such as Mosaic or Mozilla) to develop its data collection competencies.
Proxy Log: information is stocked up at the proxy side, as a result Web data look upon a number of Websites, but merely users whose Web clients clearthrough the proxy.

IV. MAP REDUCE

Map/Reduce first splits the input data set into independent chunk that are processed in a entirely equivalent manner. The Hadoop MapReduce framework sorts the outputs of the maps, which are then input tothe reduce tasks. Typically, both the input and the output of the job are stored in a file system.5-step parallel and distributed computation:

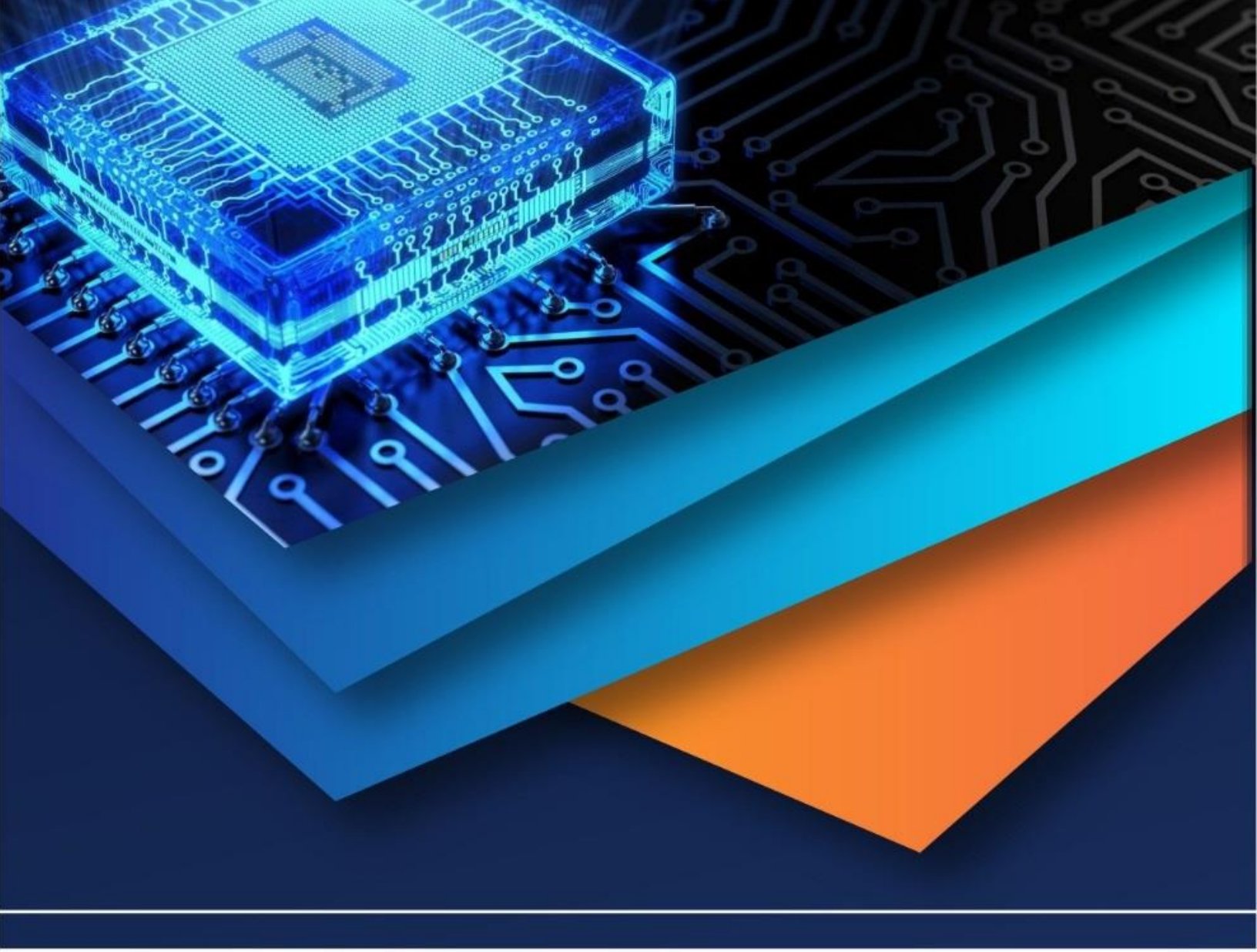
1) *Map() Input:* the "MapReduce system" allocates Map processors, give the K1 input key value each processor would work on, and offers the processor with every input data related with that key value.
2) *Map () Code:* Map () is run precisely one time for each K1 key value, producing output prepared by key values K2.
3) *"Shuffle":* The MapReduce system allocates Reduce processors, assigns the K2 key value each processor would work on, and provides that processor with all the Map-generated data associated with that key value.
4) *Reduce () Code:* Reduce () is run exactly once for each K2 key value produced by the Map step.
5) *Final Output:* The MapReduce system collects all the Reduce output, and sorts it by K2 to produce the final output.

V. CONCLUSION

During the analysis and taking the experiment results of the existing system discovers that data are more accurate than the classical process so that it gives better results from the old process. For this a wide range of existing methods, algorithms and architectures is studied for recognizing the issues detached and remains in web usage mining. In a while, this gives a brief categorization of various approaches, which has been suggested over the last few years on detection and removal of web robot request. The user navigation pattern recognition is important for prediction of users browsing behavior. This identification helps in reduction of access browsing time and eliminates the chances of visiting redundant pages. This will help in ease of network traffic. Therefore our proposed method focused on an improved algorithm that utilizes Hadoop environment with mapreduce to enhance the performance.

REFERENCES

[1] Samneet Singh and Yan Liu,"A Cloud Service Architecture for Analyzing Big Monitoring Data",ISSNll1007- 0214ll05/10llpp55-70 Volume 21, Number 1, February 2016
[2] JOSEPH A. ISSA, "Performance Evaluation and Estimation Model Using Regression Method for Hadoop WordCount", Received November 19, 2015, accepted December 12, 2015, date of publication December 18, 2015, date of current version December 29, 2015.
[3] Yaxiong Zhao, Jie Wu, and Cong Liu, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework",ISSNll10070214ll05/10llpp39-50 Volume 19, Number 1, February 2014
[4] Zhuoyao Zhang LudmilaCherkasova, "Benchmarking Approach for Designing a MapReduce Performance Model", ICPE'13, April 21-24, 2013
[5] NikzadBabaiiRizvandi, Albert Y. Zomaya, Ali JavadzadehBoloori, Javid Taheri1, "On Modeling Dependency between MapReduce Configuration Parameters and Total Execution Time", 2012
[6] NikzadBabaiiRizvandi, Javid Taheri1, Reza Moraveji, Albert Y. Zomaya, "On Modelling and Prediction of Total CPU Usage for Applications in MapReduce Enviornments", 2011
[7] Extracting WebLog of Siam University for Learning User Behavior onMapReduce -2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012) .
[8] Mining of Web Server Logs in a Distributed Cluster Using Big Data Technologies -(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 1, 2014.
[9] Tom White: Hadoop, "The Definitive Guide (1st edn.)", O'Reilly Media, Inc., United States of America, 2009.
[10] Hadoop MapReduce Change Log. Release0.22.1 – Unreleased.http://hadoop.apache.org/mapreduce/docs/r0.22.0/changes.h tml>, Accepted 02012012.
[11] Web Log Analysis for Security Compliance Using Big Data- Volume 5, Issue 3, March 2015 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering
[12] M. D. Kunder. World wide web size - daily estimated size of the world wide web. http://www.worldwidewebsize.com/, 2011. Last Visit: 2011 November.
[13] H. Liu and V. Ke_selj. Combined mining of web serverlogs and web contents for classifying user navigation patterns and predicting users' future requests. Data Knowl. Eng., 61:304{330, May 2007.
[14] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. E_ective personalization based on association rule discovery from web usage data. In Proceedings of the 3rd international workshop on Web information and data management, WIDM '01, pages 9{15, New York, NY, USA, 2001.
[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.
[16] Miki Nakagawa and BamshadMobasher, (2003)"Impact of site characteristics on Recommendation Models Based on Association Rules and Sequential Patterns", Proceedings of the IJCAI'03 Workshop on Intelligent Techniques for Web Personalization, Acapulco, Mexico, August 2003.
[17] F. Khalil, J. Li, and H. Wang. A framework for combining markov model with association rules for predicting web page accesses. In Proceedings of the Fifth Australasian Data Mining Conference (AusDM2006), pages 177–184,



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY