Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Excerption of User Profile from Web Log Data using Hadoop Framework

Dipali Patil
RSCOE, IT Department,
Pune University, India.

Snehal Patil
RSCOE, IT Department,
Pune University, India.

Payal Balsetwar
RSCOE, IT Department,
Pune University, India.

Vishakha Sabale
RSCOE, IT Department,
Pune University, India.

Apurva Kulkarni
RSCOE, IT Department,
Pune University, India.

Abstract— ***With the high development of Internet, e-commerce websites now routinely have to work with log datasets which are up to a few terabytes in size. How to remove messy data timely with low cost and find out useful information is a problem we have to face. The mining process involves several steps from pre-processing the raw data to establishing the final models. To address the problem of extracting and maintaining a very large number of user profiles from large scale data, we first describe the different scalable implementations of the proposed framework. Then we will see the challenges they faced in the implementation. And at the end we will see how hadoop can be used as an efficient solution for the problem.***

Keywords– User Profile, Web Logs, WEB data mining; Hadoop Framework, MapReduce.

I. Introduction

With the rapid development of Internet, e-commerce websites have brought unprecedented huge records from users. Behavior information of the web users are concealed in the web log. Web log is a log automatically created and maintained by a web server. Every "hit" to the Web site, including each view of a HTML document, image or other object, is logged. The raw web log file format is essentially one line of text for each hit to the web site. This contains information about who was visiting the site, where they came from, and exactly what they were doing on the web site. Web mining is the application of data mining techniques to discover patterns from the web. The term Web Data Mining is technique used to crawl through various web resources to collect required information. The web log mining can find characteristics and rules of the users' visiting behavior to improve the service quality to users. A user profile is often used to classify a given user into predefined user segments (e.g., by demographics or tastes) or to capture the online behavior of the user including the user's private interests and preferences. A user profile is defined as an "abstract model that summarizes the relevance of each URL on a site relative to a group of users sharing a similar interest. The user profile will be characterized by previous user sessions. A user session is a set of web pages that a user examined within a specified time period. To organize the user sessions into profiles, the website administrator could examine each user session and manually place them into profiles. However, this is unrealistic for many reasons, for instance, many web sites have millions of users accessing it daily and it is impossible to find patterns by mere manual examination. Therefore, we can easily see that an unsupervised clustering algorithm is an attractive choice for clustering user sessions into profiles of several types of typical users since it relies on user access patterns and is capable of examining large amounts of data in a fairly reasonable amount of time. To address the problem Shaharam, Lisa, Bidyut (2006) proposed a new technique of Competitive Agglomeration for Relational Data (CARD) Algorithm.

II. Existing System

The Competitive Agglomeration for Relational Data (CARD) Algorithm is one such clustering algorithm that is designed to organize user sessions into profiles, where each profile would highlight a particular type of user. To prepare data for the CARD algorithm, clustering user sessions requires the use of relational data, which can be represented by a matrix of similarities between each session to all other sessions. Although the CARD algorithm has found useful user profiles in prior publications, there are several assumptions made during the implementation of the algorithm that are questionable.
Though, the CARD algorithm was in the use to extract user profile from large dataset, it does have limitations such as an extended execution time. It uses Porter Stemming algorithm for keyword parsing to generate similar data using URL syntactic similarity. This process will add too much execution time to the overall algorithm. In addition, the methods that prepare the input data for the CARD algorithm's use employs concepts which seem to be incomplete. One more reason behind extended execution time required by CARD algorithm is that it runs serially. Though, the parallel implementation of CARD algorithm was possible which made execution process relatively fast it was found that most of the time spent in execution is during early data preparation phase.

Processing activities of billions of users on a daily basis imposes many challenges such as :-

1) The first and foremost is, how to build user profiles in an efficient way while efficiently expanding these profiles with the new user daily activities ?
2) A second important challenge is, how to cope with the disk input/output overhead when processing billions of users through series of successive operations.
3) A third challenge consists of dealing with the problem of optimizing multiple nodes at the same time, thus, the need for forming positive and negative training instances for each of the nodes in an independent fashion and the possibility of repetitive processing of users exposed for different nodes.
4) After modification, parallel implementation of algorithm was possible but there was a lack of scalability.
5) The large web data is in an unstructured format so transforming such a large mixture of complex and unstructured data into a structured format for later analysis is a major challenge.
6) It uses SQL Databases. The SQL by design can handle the structured data so, the processing of unstructured data is difficult.
7) As the data is very large, failure can occur. Recovery of such a large data is also a big issue.
8) Workload is on single machine, so there are the chances of system failure or system crash.

Suggested improvements were to test parallel implementation on larger data sets. Another improvement would be to change the dataset from a computer science web site to a larger and more diverse website. Due to these limitations of existing system, a new system which can overcome these limitations was needed.

III. Related Work

A solution to above discussed problems is to adopt a "Hadoop Framework and MapReduce Distributed Programming technology" to mine User Profile from large web log data. MapReduce is a programming model for processing large data sets, MapReduce is typically used to do distributed computing on clusters of computers.

The underlying technology was invented by Google back in their earlier days. Google's innovations were incorporated into Nutch, an open source project, and Hadoop was later spun-off from that. Yahoo and Apache has played a key role developing Hadoop for enterprise applications. Hadoop runs jobs on hundreds of terabytes of data. Hadoop is also used at Facebook, Amazon, and Last.fm. Because of its high efficiency, high scalability, and high reliability, MapReduce framework is used in many fields, such as life science computing, text processing, web searching, graph processing, relational data processing, data mining, machine learning, and video analysis.

Here are some features of Hadoop that overcomes the problems of existing system :

1) It scales linearly. This means you can double the servers in the cluster to halve data processing time.
2) It's also true for storage capacity. The highly-fault tolerant Hadoop Distributed File System (HDFS) allows you to share the load of containing the source data amongst as many servers as you need.
3) Hadoop source data doesn't need to be constrained to columns and rows.
4) Hadoop provides data recovery as it stores multiple copies of data, so data stored on a server that goes offline or dies can be automatically replicated from a known good copy.
5) To handle an unstructured data SQL is implemented on top of hadoop framework as an execution engine.
6) Because it is intended to run on commodity hardware, Hadoop is architected with the assumption of frequent hardware malfunctions. It can gracefully handle most such failures.
7) Even if the system fails or crashes it doesn't affect the data, as it is divided and stored on different nodes in cluster.

Thus, to solve the problems of existing system we adopt a hadoop framework. In this way hadoop can be used to overcome the limitations and to improve a performance of system.

A system architecture is shown in Fig 1. which shows how hadoop is useful to extract user profiles from a large set of data. The proposed system solves the problem of performance and scalability. In our system data is analyzed using Map Reduce algorithm and Hadoop framework. It takes huge computation for analysis if the size of data is large. Therefore we are using hadoop cluster which divides the task into subtask and distribute this subtasks among the child nodes. Logs are given as an input to the system which are processed by MapReduce module. The result is stored in Oracle DB. The user can see graphical representation of analysis reports through web application.

IV. Conclusion

With the wide use of Internet in e-commerce it is difficult to process web log data and to extract user profiles from them. One of the system earlier used to extract user profiles from

mass data was based upon a CARD algorithm. A CARD algorithm found useful to user profiles but it has limitation such as extended execution time. Also the system was unable to support some features which are necessary while extracting data from mass data. So, to overcome the drawbacks of this system we proposed a scalable user profiling solution, which is implemented on top of Hadoop and MapReduce framework.

The proposed system not only solves the time constraint but also support many features which are required. Future work will include the extension of our framework with some more features.

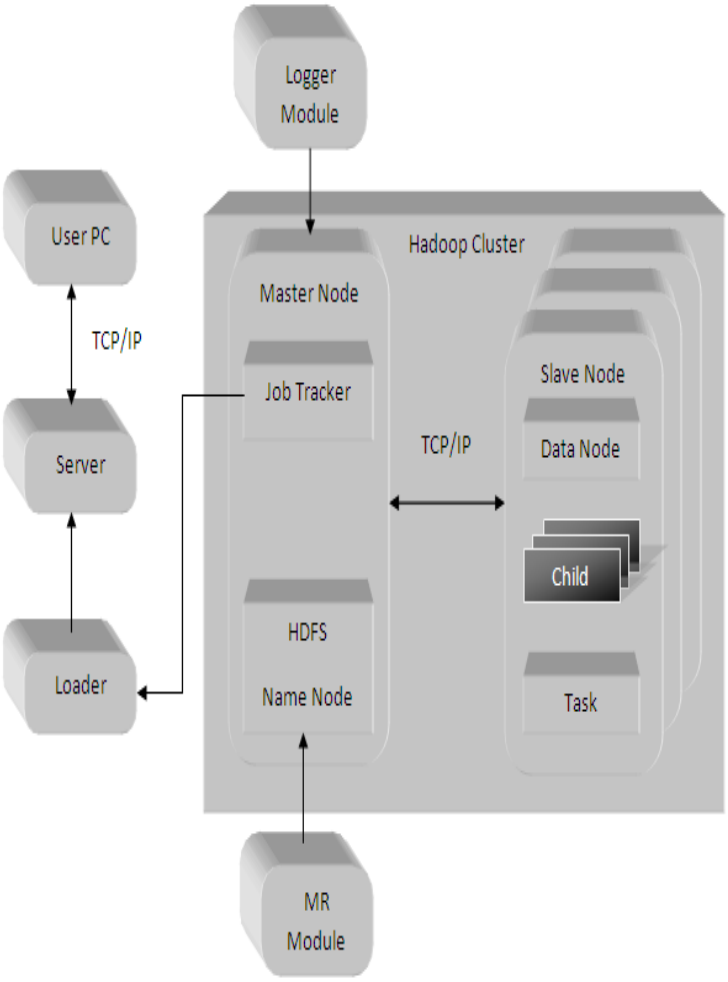

Fig 1. System Architecture

References

[1] Rahimi, Shahram; Gandy, Lisa; and Gupta, Bidyut, "Extracting Web User Profiles Using a Modified CARD Algorithm" (2006). http://opensiuc.lib.siu.edu/cs_pubs/21

[2] Huang Lan, Wang Xiao-Wei, Zhai, "Extraction of User Profile Based on the Hadoop Framework",(2009) College of Computer Science and Technology Jilin University Changchun, China.

[3] Michal Shmueli-Scheuer, Haggai Roitman, David Carmel, Yosi Mass, David, "Extracting User Profiles from Large Scale Data", (2009)

[4] Mohamed Aly, Andrew Hatch, Vanja Josifovski, Vijay K. Narayanan, "Web-Scale User Modeling for Targeting",(2009), Yahoo! Research, Santa Clara, CA 95051, USA.

[5] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[6] The Hadoop website. [Online]. Available: http://hadoop.apache.org

