Anomaly Detection in web logs Using Big Data Analytics

S.Sathya,
Ph.D Research Scholar
Alagappa University
Karaikkudi.

Dr.E.Ramaraj,
Proffessor& Head
Department of Computer Science
Alagappa University,
Karaikkudi

ABSTRACT

Web logs create and stored as record in a web server automatically. To get information about website use can analyze such web server logs. Weblog processing is a very challenging for various environments with lots of server. The information about user interest and behavior is stored in web log serve. In such an environment log data is large, coming at high speed in various formats. Along these to process such huge datasets we need an equal preparing(Hadoop). Hadoop runs the enormous information where a huge amount of data is prepared through group of ware equipment. In this paper we present the way for Anomaly detection by pre-processing the web log files, and to store the web logs using the architecture of Hadoop Framework. The objective of this paper is to analyze web log details for Anomaly Detection.

KEYWORDS : Anomaly,HADOOP,Analytics

INTRODUCTION

In present situation, so many service provider's organizations are quick to know whether they give the best service to clients. Log records contain all activities that happened at client side gets by the service provider website or web application. Each hit to the Website will be signed in a log record. These log records are put away in web servers. Each hit of the Website will be signed in a log record. These log records are put away in web servers. The web log detail is one line of content for each hit to the site and contains data about who visited the webpage, where they originated from, and what they did on the site. These log records convey a valuable data for service provider about website traffic patterns, user activity, customer interest etc.

PHASES IN WEB LOG MINING

Web log mining is of three types such as preprocessing of web data, Patterns discovery, and Patterns analysis. Preprocessing is an essential point in web log mining process. The main phases of web log mining are appeared in FIG 1.

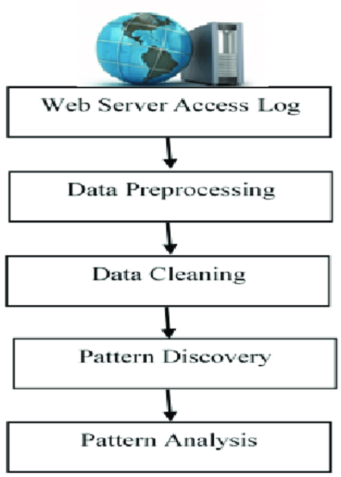

FIG1: PHASES IN WEB LOG MINING

Data Preprocessing: It is the initial period of web log mining. The web log information is a raw data and it is not legitimately utilized directly .In this stage, we applying strategy to transform raw data into an understandable format. Real world information is repeated, inadequate, unpredictable, and ailing in specific practices or inclines, and is probably going to contain numerous mistakes. Data preprocessing is a demonstrated strategy for settling such issues. Information pre handling plans crude information for additional preparing.

Pattern Discovery: The results from pre-handling will be utilized to discover frequent client access design. In design revelation will be utilize various information mining procedure like as affiliation rule, classification, clustering, and successive pattern system to discover significant data. The outcome that has been removed can be represented in many ways, for example, diagrams, outlines, table, and so on.

Pattern Analysis: The result of pattern discovery stage isn't straightforwardly utilized for analysis. In this stages will build up a tool that can assist experts with understanding the data has been separated. Tools or techniques that can be utilized in this stage like perception strategies, OLAP investigation and information Query component.

WEB LOG ANALYSIS

Web server logs click stream data which can be useful for mining purposes. Web log analysis is plain text (ASCII) file which contain information about Name of user IP Address, Access Request, Time Stamp, Error codes, URL that Referred, etc. There are following types of server logs: Transfer log, Agent log, Error log and Referrer log [8]. The transfer and the agent log are said to be standard whereas the error and referrer log are considered optional as they may not be turned on. Every log entry record the traversal from one page to another storing IP number and all the related information [7].

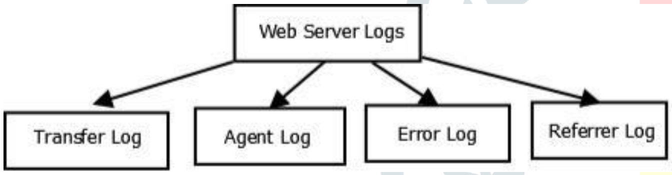

FIG 2. Taxonomy of Web Server Logs

The access log contains all information that provides to the client by the server. The error log files hold a list of any server error. These two log files very common and important to bring the required information in accessing the user behavior during suspected user quest. The Referrer and agent log file provide the information about user's browser, operating system and version of browser. The Referrer log file [7] is used to allow websites and web servers to identify.

PROPOSED METHODOLOGY

Weblog processing is a very challenging for various environments with lots of server. In such an environment log data is large, coming at very high speed and have various formats. The information about user interest and behavior is stored in web log server. Big data concept is essential to handle such large data sets.So many organizations such as e-commerce, healthcare, banking, media system has huge amount of data and stored in common storage place clouds. User interest and behavior is stored in web log server. Web logs are converted into event logs, where the user behavior is captured. The correlation among the sequence of events is created and proposes a set of queries to find user interest in visiting the website.

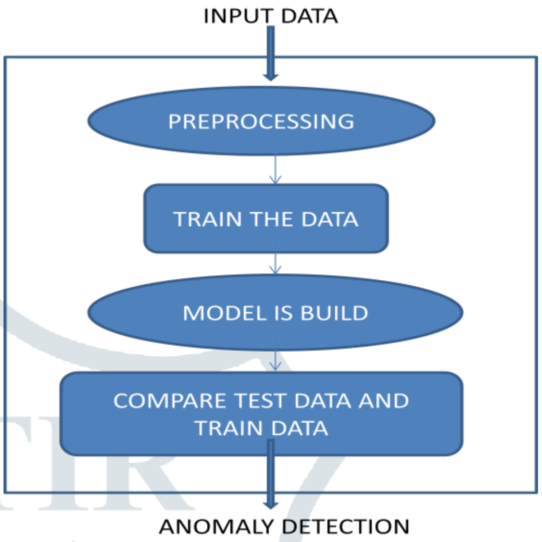

FIG 3: ANOMALY DETECTION IN WEB LOGS

Anomaly detection is used in all operations that done over the networks. Sometimes it can be as used as a preprocessing technique to remove anomalous data from the dataset. Mostly Machine learning techniques are used in anomaly detection. While the supervised learning is applied in removing the anomalous data from the dataset it results in accuracy.Unsupervised techniques can be used to uncover hidden structures, like find groups of photos with similar cars, but it's difficult to implement and is not used as supervised learning. Unsupervised techniques may be used as a first step before applying supervised learning.

When users and the number of applications get increases the web logs in server are difficult to frame correlation for the events. To overcome the limitations Big data is used. In proposed method, while extending the web server log data in various organizations is collected by the Big data tools .Then preprocessing is done to filter the noisy data. Data are transferred to HDFS. The Map Reduce method converts data from unstructured into structured data. Store the structured data in table using HIVE. Extract the required feature from the table with HQL (Hive Query Language).

Data can be split into testing and training data. In training model, Clustering Algorithm is used to separate normal and abnormal behavior of the user. Classification is done between test data and training data to find the anomaly behavior in log data.

ALGORITHM :

Step 1: Get the various input data from cloud servers stored in log server.
Step 2: Data is preprocessed. Here unwanted data are erased.
Step 3: Big data Analytics is first used for structuring the data.
Step 4: Transfer the log data to HDFS.
Step 5: Convert data into structured data and load in table using HIVE.
Step 6: Extract the required feature from the table with HQL(Hive Query Language).
Step 7: Data can be splited into testing and training data.
Step 8: With the Training data create a training model, which implements machine learning algorithm.
Step 9: Test data and trained data compared for ANOMALY detection.

WORKING MODEL USING HADOOP

The Hadoop Distributed File System (HDFS) and distributed file system has many similarities.. HDFS is heavy in fault-tolerant and is designed to be deployed on low cost when compared to rest hardware. HDFS provides high throughput for the application data and is suitable for which having very large data sets.

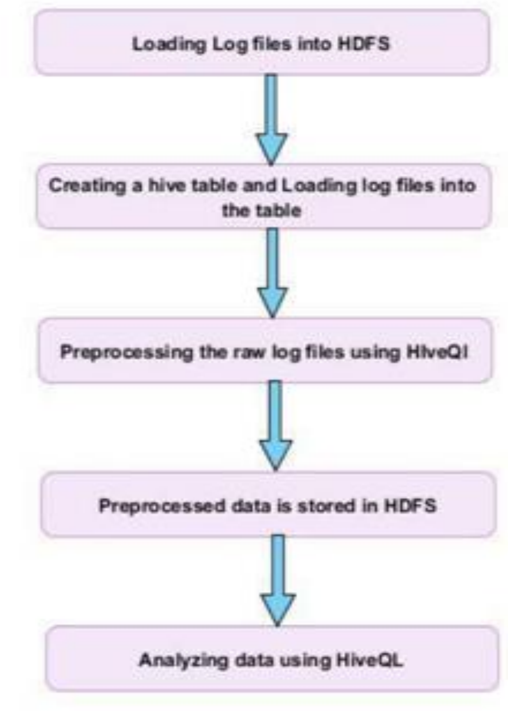
Loading Log files into HDFS
Creating a hive table and Loading log files into the table
Preprocessing the raw log files using HiveQl
Preprocessed data is stored in HDFS
Analyzing data using HiveQL

FIG 4: WEB LOGS IN HADOOP FRAMEWORK

The proposed work has been done on web logs using Hadoop is shown in FIG 4.The work has many phases, where the storage and processing is done in HDFS. Web server log files are first copied into Hadoop file system and then loaded to Hive table. Data cleaning, which is done using Hive query Language, is the first phase carried out as a pre-processing step. Log files generated from the web server is of very large volume of data that cannot be handled by a traditional database or additional programming languages for computation.

Web logs are files consist of number of records that correspond to automatic requests generated by web servers. The records usually be in a large volume some erroneous, and incomplete information. In our methodology first unwanted error information in web log files carrying requests from web servers, are removed in pre-processing with the entries that have a status of "error" or "failures.The identification of status code is the important task carried out in the data cleaning .Only the log lines with particular status code is consider as correct log. Therefore only the lines with the correct status code value are extracted and stored in Hive table for analysis. The next step is to identify unique user, unique fields of date, status code, and URL referred in each and every log files in log data These unique values are retrieved and used for further analysis.

Hive is an important tool in the Hadoop that provides a Structured Query Language called HiveQL to query the data stored in Hadoop Distributed File system. The log files that are stored in the HDFS are loaded in to a hive table and cleaning action is taken out. The cleaned web logs data are processed further for Anomaly detection using Machine learning Algorithms.

RESULTS

Here web logs were taken from https://www.kaggle.com/shawon10/web-log-dataset website for time period of 29/NOV/2017 to 28/FEB/2017 and the following results were obtained.

This dataset has 16008 rows and 4 columns. Columns that included here are IPAddress , Time, URL, Status[8].

SAMPLE DATA:

IP	Time	URL	Sta us
10.128.2.1	[29/Nov/2017:06:58:55	GET /login.php HTTP/1.1	200
10.128.2.1	[29/Nov/2017:06:59:02	POST /process.php HTTP/1.1	302
10.128.2.1	[29/Nov/2017:06:59:03	GET /home.php HTTP/1.1	200
10.131.2.1	[29/Nov/2017:06:59:04	GET /js/vendor/moment.min.js HTTP/1.1	200
10.130.2.1	[29/Nov/2017:06:59:06	GET /bootstrap-3.3.7/js/bootstrap.js HTTP/1.1	200
10.130.2.1	[29/Nov/2017:06:59:19	GET /profile.php?user=bala HTTP/1.1	200
10.128.2.1	[29/Nov/2017:06:59:19	GET /js/jquery.min.js HTTP/1.1	200
10.131.2.1	[29/Nov/2017:06:59:19	GET /js/chart.min.js HTTP/1.1	200
10.131.2.1	[29/Nov/2017:06:59:30	GET /edit.php?name=bala HTTP/1.1	200
10.131.0.1	[30/Nov/2017:07:07:57	GET /js/vendor/moment.min.js HTTP/1.1	200
10.131.0.1	[30/Nov/2017:07:08:06	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	302
10.128.2.1	[30/Nov/2017:07:08:06	GET /countdown.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200
10.130.2.1	[30/Nov/2017:07:24:34	GET /robots.txt HTTP/1.1	404
10.129.2.1	[30/Nov/2017:07:24:34	GET / HTTP/1.1	302

TABLE 1 : SAMPLE DATA

The entries in table 1 each field in a record are described below.

10.128.2.1

This is the IP address of the client (remote host) which made the request to the server. The address of the machine can be said as the IP address used by clent. If a proxy server is between the user and the server, so the address can be the address of the proxy, rather than the original machine.

[30/Nov/2017:07:24:34]

The time that the server finished processing the request. The format is:

[day/month/year:hour:minute:second]
day = 2*digit
month = 3*letter
year = 4*digit
hour = 2*digit
minute = 2*digit
second = 2*digit

GET /profile.php?user=bala HTTP/1.1

The above request is from the client side. The line gets the information of the user like user connection and user account information. First, the method used by the client is GET. Second, the client requested the resource profile.php?user=bala, and third, the client used the protocol HTTP/1.1. The client can send more than one request independently for each log.

Status

The status code is created and sent by the server to the client ,whenever server encountered the logs from client side..This information is useful for the error counting and it only says whether the request resulted in a successful response or not.

(codes beginning in 2), a redirection.
(codes beginning in 3), an error caused by the client.
(codes beginning in 4), or an error in the server.

Data cleaning process take place in the above data to reduce the size .The main advantage of data cleaning process is in producing results in quality with grate efficiency. The results after Pre-processing are gen in TABLE 2. The result shows the size reduction of data .Data cleaning convert the raw data into usable format.

CATOGORY	RAW DATA	AFTER CLEANING
FILE SIZE	1090kb	90.83kb
NO.OF ROWS	158472	16000

TABLE 2: AFTER PREPROCESSING

Here total number of visit in different IP address is counted in the given data .

S.NO	IP ADDRESS	TOTAL NO.OF VISIT
1	10.128.2.1	2941
2	10.129.2.1	1286
3	10.130.2.1	2895
4	10.131.0.1	2961
5	10.131.2.1	1247

TABLE 3: TOTAL NO.OF VISIT

Errors

In server, the log entries may be in variety and be with various errors. Error can be occurring when the user tries to access web pages. The various errors can be in server side or in client side as recorded during the time of accessing the website. The error is counted and listed as number of hits in TABLE 4 ,when the user accesses the website.

S.NO	STATUS	NO.OF.HITS
1	200	11382
2	300	4156
3	400	262

TABLE 4: NO. OF ERRORS IN WEB LOGS

CONCLUSIONS

Web sites are one of the important tool for making advertisements. In order to get usage details of a specific web site, we need to do log examination that helps enhance the business methodologies and also produce well reports. In this paper with the help of Hadoop framework web server log files are analyzed. Big Data tools are used to analysis will give reports about web pages, client‘s movement, in which part of the web site clients are interested. From these reports can verify what parts of the site more accessed, potential clients, what are the regions from which the site is getting more hits, and so on. This will help to detect anomaly activities. Log analysis can be done using many different techniques however what is important is response time. HDFS model provides parallel distributed processing and reliable data storage for huge volumes of web log files. Hadoop‘s ability of moving processing to data rather than moving data to processing helps enhance response

REFERENCES

[1] Barani Priyanga R,2Dr.K.Anitha Kumari, 3Dharani D,2018,A Survey on Anomaly Detection using Unsupervised Learning Techniques, IJCRT ,Volume 6, Issue 2.

[2]Yi Zhang, Weiwei Chen, and Jason Black. 2010. Anomaly Detection in Premise Energy Consumption Data. IEEE, 978-1-45771002-5/11.

[3] VarunChandola, Arindam Banerjee and Vipin Kumar. 2009. Anomaly Detection : A Survey To Appear in ACM Computing Surveys.09.

[4] S. E. Salama, M. I. Marie, L. M. E. Fangary, and Y. K. Helmy, “Web server logs preprocessing for web intrusion detection,” vol. 4, no. 4, 2011, pp. 123–133. [9] L. K. J. Grace, V. Maheswari, and D. Nagamalai, “Analysis of web logs and web user in web mining,” vol. abs/1101.5668, 2011.

[5] M. A. T. G. Castellano, A. M. Fanelli, “Log data preparation for mining web usage patterns,” vol. abs/1101.5668, 2007, pp. 371–378. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel.

[6]Agarwal B., Mittal N., Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques, Procedia Technology; 6; 2012; p. 996- 1003.

[7] B Shalem Raju, Dr. K. Venkata Ramana,”Analysis of Web Server Logs Using Apache Hive to Identify the User Behaviour”; Vol-3, Issue-1, 2017; ISSN: 2454-1362.

[8]Web page reviewed https://www.kaggle.com/shawon10/web-log-dataset

[9] Manoj Kumar, Mrs. Meenu,” Analysis of visitor’s behavior from Web Log using Web Log Expert Tool “.

[10]. Dhawan, Sanjeev, and Swati Goel. "Web Usage Mining: Finding Usage Patterns from Web Logs." American International Journal of Research in Science, Technology, Engineering & Mathematics (2013): 203-207.

[11] Shukla, Rajesh, Sanjay Silakari, and P. K. Chande. "Web Personalization Systems and Web Usage Mining: A Review." International Journal of Computer Applications 72, no. 21 (2013).