International Journal of Advanced Science and Technology
Vol. 29, No. 8, (2020), pp. 6222-6149

Analysis of Web Logs In Pre-Processing and Log Access Frequency

S.Sathya,
Ph.D Research Scholar
Alagappa University
Karaikkudi.
EmailID :sathyaasundar@gmail.com
Phone: 9566277634

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

ABSTRACT

Nowadays, global huge web has grown to be a large repository or garage for retrieving, storing, sharing and additionally distribute the statistics or information. Net is a dominant platform from wherein the expertise may be determined to look at web user behavior. Each and each interplay of person with the web might be recorded or saved in a textual content file which is largely referred to as as internet Log file. Web logs create and saved as document in an internet server robotically. To get information about internet site use can analyze such internet server logs. Log processing is a completely tough for various environments with masses of server. The information about user interest and behavior is saved in web log server.Here on this work net mining is accomplished with web log documents for reading user interactions with net or person accesses. This is known as net utilization mining. This paper, specializes in web log preprocessing and mining techniques and their applicable limitations for web usage mining.

KEYWORDS : *Anomaly, knowledge discovered, Pre-processing, Pattern Discovery, Pattern Analysis, Web Log File.*

INTRODUCTION

In today's world as using internet has extended exceptionally, there is the want for knowledge every and every web consumer behaviour that allows you to enhance the enterprise and to satisfy consumer expectations. So the studies on web utilization Mining by means of the various researchers and knowledge and analysts is going on.Data cleaning includes disposing of unwanted fields of web log records, eliminate the logs with file names like .gif, jpeg, jpg, java script, css, robot.txt, and many others, and also removing the logs with failed HTTP ststus code. User identification involves identification of user by assuming each combination of IP address, Agent and Operating System as a single user.In session identification the session is described because the combination of pages accessed with time given. Like 'S' denotes a session. session 'S' shows the set of pages accessed by way of a selected user.

PHASES IN WEB LOG MINING

ISSN: 2005-4238 IJAST

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.

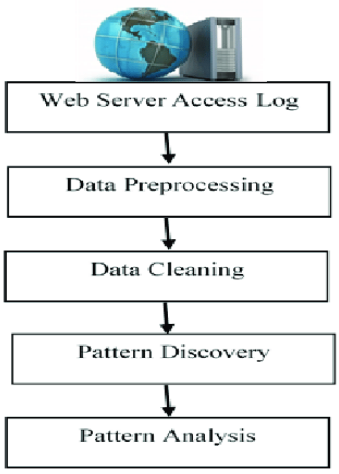


FIG1: PHASES IN WEB LOG MINING

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. Data collection is the foremost step in web usage mining where in the web log data is collected. 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.Each record line in this web log file represents a user interaction with the web server. This dataset has 16008 rows and 4 columns. Columns that included here are IPAddress , Time, URL, Status

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. Web patterns discovery step is performed to discover interesting patterns or knowledge to analyze web user behavior. Once the pre-processing step is completed the patterns or knowledge can be discovered from the pre-processed web log data in pattern discovery step. Different methods or techniques are used to discover association rules or frequent patterns like “statistical methods” and also data mining methods like “Path analysis”, “Association rule”, “Sequential patterns”, “Clustering” and “classification”. These are performed on web log files so as to detect interesting patterns to study web user behavior.

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 pattern analysis is the process in which uninterested patterns are removed out from patterns discovered in previous pattern discovery step. Here the patterns discovered are analysed by making use of some of OLAP tools or by SQL query mechanism.

ISSN: 2005-4238 IJAST

WEB LOG FILES

This collected web server log files will be containing huge amount of log records with some kind of unwanted data too. Kindly it's little difficult to deal with such huge amount of web log data. So this kind of unwanted or unnecessary data has to be removed out before proceeding to next step. This is done in pre-processing step by applying different pre- processing techniques. Some of the pre-processing are "Data cleaning", "User identification", "Session identification", "Data transformation", "Path completion".

Web log files are the text files which get generated whenever there is a interaction between user and the web. Each user interaction with web will be recorded as a single record in the web log file. Generally web log file records contains fields such as IP address, URL accessed, time stamp, number of bytes, method used for making request and protocol details. These web log files can be used to understand or study the web user behaviour. The data which is stored in web log files will be consisting of huge amount information with some kind of incomplete and unwanted data too. Data mining techniques can be applied to on web log files to remove out unnecessary data and then finding patterns out of pre- processed data for analysing the data to study web user behaviour.
A sample web log record is shown below,

123.46.7.79.8 - [12/Mar/2012:04:06:50 -0500]
—GET/HTTP/1.0‖ 200 3240

Where,

- 123.46.7.79.8- IP address
- "-"(hyphen) indicates Anonymous user id
- 12/Mar/2012:04:06:50- Web page access time
- -0500- The time zone
- GET/HTTP- HTTP request method
- 200- HTTP status code
- 3240- Number of bytes transmitted

IP	Time	URL	Staus
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%20 Testing%20Contest HTTP/1.1	302
10.128.2.1	[30/Nov/2017:07:08:06	GET /countdown.php? name=RUET%20OJ%20Server%20 Testing%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

ISSN: 2005-4238 IJAST

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]

GET /profile.php?user=ABC 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=ABC, and third, the client used the protocol HTTP/1.1. The client can send more than one request independently for each log.

Log records contain all activities that happened at user side gets by the server applicationThese log records convey a valuable data for service provider about website traffic patterns, user activity, customer interest etc.

WEB LOG MINING

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].

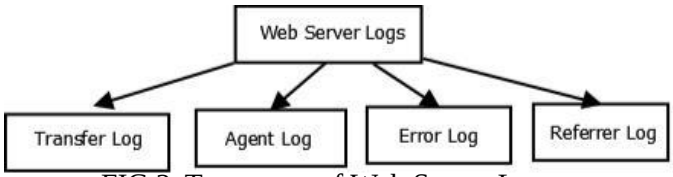


FIG 2. Taxonomy of Web Server Logs

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 is used to allow websites and web servers to identify.

Sequential pattern mining models are applied to discover the frequent web usage patterns between the page requests, session time and browsing history, etc. However, these sequential models have certain limitations such as:

- Need to maintain huge data structure in memory space throughout the execution due to the
- Lack of predicting a user's next access patterns based on historical data.

Web usage mining applications are used to find the web visitors' profiles and their behavior in terms of strengths and weaknesses of their web applications. The main issue focused by any web usage model is data increases per second with different server log file formats. Learning about the customer's behavior , predict their requirements in the future , monitoring the file structure and content of the web service according to their navigation behavior is necessary. Accurate web usage patterns could help to

ISSN: 2005-4238 IJAST

improve the new users, retain existing customers, optimize cross sales, customers' interest, etc. The usage decision patterns can improve the web server efficiency by using different caching techniques so as to minimize the server response time. The user's profile could be designed by integrating customer's page navigation paths with other attributes such as server response, session time, page duration, hyperlink and page content.

Applications of web usage mining include mining conceptual visiting user profile hierarchies and interesting patterns from the web log files for building the frequent web access structures using tree based Markov model or association models. Since web usage mining approaches consider only server logs due to security issue of information on the client side. The set of limitations of the server side are :

- IP addresses and sequence of page requests in the log file are not a reliable fields, because some pages are cached by the web server or browser and proxy.
- It is difficult to interpret the session duration in the server log file, as the same IP address can be used different users at different intervals (i.e. 30 minutes default time).
- Also server log files are difficult to predict without log preprocessing.
- Since server log files have different structures and formats, it is difficult to apply same preprocessing or knowledge based techniques.

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.

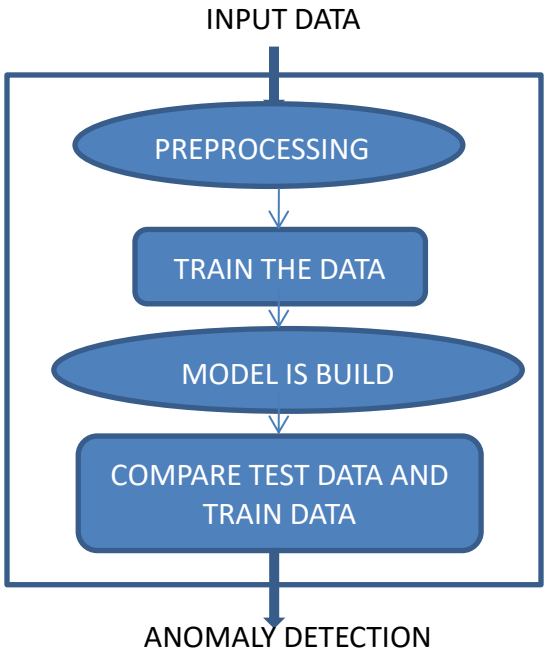


FIG 3: ANOMALY DETECTION IN WEB LOGS

ISSN: 2005-4238 IJAST

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.

ISSN: 2005-4238 IJAST

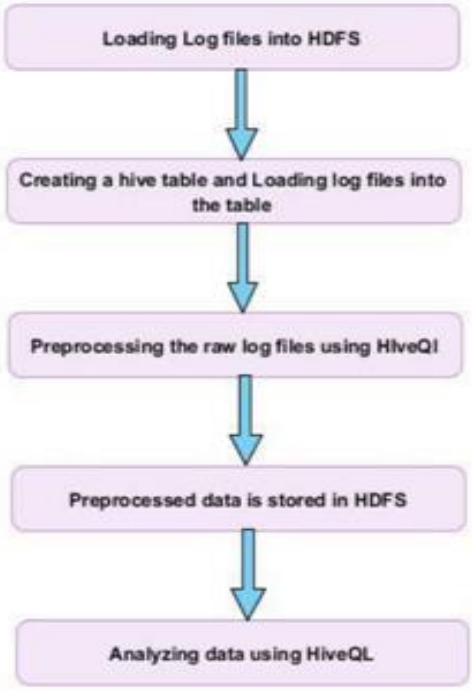


FIG 4: WEB LOGS IN HADOOP FRAMEWORK

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.

RELATED WORK IN DATA PREPROCESSING

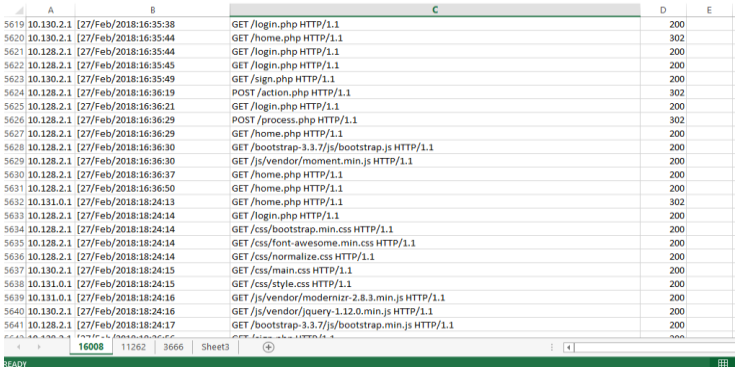
The information existing in the web is diverse and unstructured. Consequently, the preprocessing segment is a requirement for find out patterns. The objective of preprocessing is to change the raw click stream data into a set of user profiles. Data preprocessing presents a number of exceptional challenges

ISSN: 2005-4238 IJAST

which led to a diversity of algorithms and heuristic techniques for preprocessing step such as integration and cleaning, user and session identification etc. A variety of research works are approved in this preprocessing part for combination sessions and transactions, which is used to determine user behavior patterns.

In Data Collection process, the data are collected from the website https://www.kaggle.com/shawon10/web-log-dataset. for time period of 29/NOV/2017 to 28/FEB/2017 and the following results were obtained.Each record line in this web log file represents a user interaction with the web server. This dataset has 16008 rows and 4 columns. Columns that included here are IPAddress , Time, URL, Status.The raw data for mining purpose is collected as log data. It contains approximately 16008 records in Common log file format. The sample log file used for the task was in raw log format.

After the process completion the data copied in excel sheet and displayed as output



	A	B	C	D	E
5619	10.130.2.1	[27/Feb/2018:16:35:38	GET /login.php HTTP/1.1	200	
5620	10.130.2.1	[27/Feb/2018:16:35:44	GET /home.php HTTP/1.1	302	
5621	10.128.2.1	[27/Feb/2018:16:35:44	GET /login.php HTTP/1.1	200	
5622	10.128.2.1	[27/Feb/2018:16:35:45	GET /login.php HTTP/1.1	200	
5623	10.130.2.1	[27/Feb/2018:16:35:49	GET /sign.php HTTP/1.1	200	
5624	10.128.2.1	[27/Feb/2018:16:36:19	POST /action.php HTTP/1.1	302	
5625	10.128.2.1	[27/Feb/2018:16:36:21	GET /login.php HTTP/1.1	200	
5626	10.128.2.1	[27/Feb/2018:16:36:29	POST /process.php HTTP/1.1	302	
5627	10.128.2.1	[27/Feb/2018:16:36:29	GET /home.php HTTP/1.1	200	
5628	10.128.2.1	[27/Feb/2018:16:36:30	GET /bootstrap-3.3.7/js/bootstrap.js HTTP/1.1	200	
5629	10.128.2.1	[27/Feb/2018:16:36:30	GET /js/vendor/moment.min.js HTTP/1.1	200	
5630	10.128.2.1	[27/Feb/2018:16:36:37	GET /home.php HTTP/1.1	200	
5631	10.128.2.1	[27/Feb/2018:16:36:50	GET /home.php HTTP/1.1	200	
5632	10.131.0.1	[27/Feb/2018:18:24:13	GET /home.php HTTP/1.1	302	
5633	10.128.2.1	[27/Feb/2018:18:24:14	GET /login.php HTTP/1.1	200	
5634	10.128.2.1	[27/Feb/2018:18:24:14	GET /css/bootstrap.min.css HTTP/1.1	200	
5635	10.128.2.1	[27/Feb/2018:18:24:14	GET /css/font-awesome.min.css HTTP/1.1	200	
5636	10.128.2.1	[27/Feb/2018:18:24:14	GET /css/normalize.css HTTP/1.1	200	
5637	10.130.2.1	[27/Feb/2018:18:24:15	GET /css/main.css HTTP/1.1	200	
5638	10.131.0.1	[27/Feb/2018:18:24:15	GET /css/style.css HTTP/1.1	200	
5639	10.131.0.1	[27/Feb/2018:18:24:16	GET /js/vendor/modernizr-2.8.3.min.js HTTP/1.1	200	
5640	10.130.2.1	[27/Feb/2018:18:24:16	GET /js/vendor/jquery-1.12.0.min.js HTTP/1.1	200	
5641	10.128.2.1	[27/Feb/2018:18:24:17	GET /bootstrap-3.3.7/js/bootstrap.min.js HTTP/1.1	200	

16008 | 11262 | 3666 | Sheet3

FIG 5: CLEANED DATA IN SHEET1

A Weblogs includes series of transactions up to date frequently, while customers accessing the web sites. It incorporates of numerous entries like IP address, fame code and quantity of bytes transferred, categories and time stamp. The person hobby may be categorized based on categories and attributes and it's far helpful in identifying user conduct. The log query parser is to convert unstructured log to structured log based on person interest. The logs that is saved in internet log documents will be which includes massive amount of records with some form of incomplete and undesirable logs too. It is hard to deal with whole statistics that is large in size. So, unwanted or bored data can be eliminated by way of processing the logs. In popular the aim of web usage mining is to discover interesting styles or navigational patterns or knowledge about web usage.

Data cleaning Log data is stored in database for supplementary processing of data by way of queries and program .Data file acquired was very enormous and it obtains approximately 80% of total time to mine the data. In data cleaning process, the unnecessary information is removed from the log database.

The data cleaning obtains the following steps:

Step1: Elimination of the entries having image files, graphic or multimedia files. The records which are accessing file with extension gif,jpg, jpeg etc. are to be removed.After performing this step around 11262 records left.

Step 2: The elimination of entries with failed status code. A variety of status codes for HTTP 1.1 in this step the entries having status code of 200 will be retained, rest are removed.

ISSN: 2005-4238 IJAST

Step3: Removal of records with bytes transferred field zero. The records having entries zero in the byte transferred field specifies that the requested page is not opened, and is to be removed. After performing the above two steps the number of records left are 3666.

An algorithm for data cleaning

1) Start the process
2) Scan the Log Records one by one in log file
3) For every record in log file
4) Read all the fields
5) If status code = Success
6) Then check URL
7) If(URL = *.txt, *.mpg,*.gif , *.css, *.jpg)
8) Then Remove details from URL
9) Else Save records
10) End if
11) Move to next record
12) End if

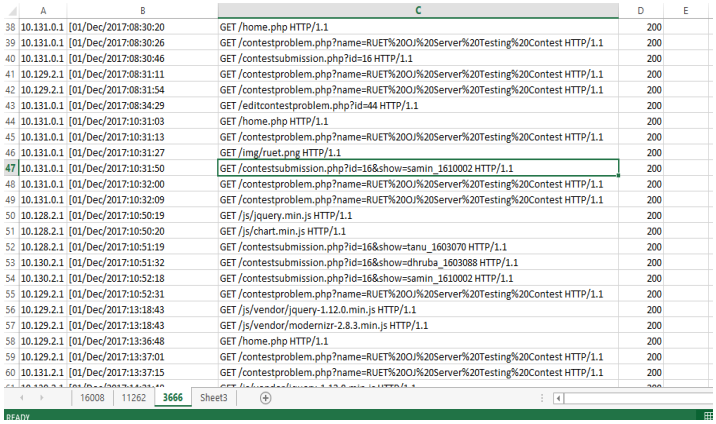
Stop

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..Data cleaning convert the raw data into structured data.

To clean the web log data, read the web log file and calculate all the record. The method is so as to, we read character by character from the file and evaluate the character from ASCII value of space and enter key and count up all the record from web log file.

Web log File Cleaning: In this action, the irrelevant log entries are deleted from the log file. This can be completed by examination in the request field of the log file, the suffix of the website URL requested by the user. These suffixes notify us the authentic format or extension of the web files requested by user. Contained by the log file, we will receive only those files which have extensions like .html, .asp, .aspx, .php. So we can also delete every log entries taking extensions like .gif,.jpeg, .flv, .mp3, .mp4, etc. We can also delete log entries with empty URL or having request methods other than GET and POST. We can also delete all those log entries with status code other than 200 .At the end the cleaned log file is organized for the next steps.

ISSN: 2005-4238 IJAST



	A	B	C	D	E
38	10.131.0.1	[01/Dec/2017:08:30:20	GET /home.php HTTP/1.1	200	
39	10.131.0.1	[01/Dec/2017:08:30:26	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200	
40	10.131.0.1	[01/Dec/2017:08:30:46	GET /contestsubmission.php?id=16 HTTP/1.1	200	
41	10.129.2.1	[01/Dec/2017:08:31:11	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200	
42	10.129.2.1	[01/Dec/2017:08:31:54	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200	
43	10.131.0.1	[01/Dec/2017:08:34:29	GET /editcontestproblem.php?id=44 HTTP/1.1	200	
44	10.131.0.1	[01/Dec/2017:10:31:03	GET /home.php HTTP/1.1	200	
45	10.131.0.1	[01/Dec/2017:10:31:13	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200	
46	10.131.0.1	[01/Dec/2017:10:31:27	GET /img/ruet.png HTTP/1.1	200	
47	10.131.0.1	[01/Dec/2017:10:31:50	GET /contestsubmission.php?id=16&show=samin_1610002 HTTP/1.1	200	
48	10.131.0.1	[01/Dec/2017:10:32:00	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200	
49	10.131.0.1	[01/Dec/2017:10:32:09	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200	
50	10.128.2.1	[01/Dec/2017:10:50:19	GET /js/jquery.min.js HTTP/1.1	200	
51	10.128.2.1	[01/Dec/2017:10:50:20	GET /js/chart.min.js HTTP/1.1	200	
52	10.128.2.1	[01/Dec/2017:10:51:19	GET /contestsubmission.php?id=16&show=tanu_1603070 HTTP/1.1	200	
53	10.130.2.1	[01/Dec/2017:10:51:32	GET /contestsubmission.php?id=16&show=dhruba_1603088 HTTP/1.1	200	
54	10.130.2.1	[01/Dec/2017:10:52:18	GET /contestsubmission.php?id=16&show=samin_1610002 HTTP/1.1	200	
55	10.129.2.1	[01/Dec/2017:10:52:31	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200	
56	10.129.2.1	[01/Dec/2017:13:18:43	GET /js/vendor/jquery-1.12.0.min.js HTTP/1.1	200	
57	10.129.2.1	[01/Dec/2017:13:18:43	GET /js/vendor/modernizr-2.8.3.min.js HTTP/1.1	200	
58	10.129.2.1	[01/Dec/2017:13:36:48	GET /home.php HTTP/1.1	200	
59	10.129.2.1	[01/Dec/2017:13:37:01	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200	
60	10.131.2.1	[01/Dec/2017:13:37:15	GET /contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest HTTP/1.1	200	

16008 | 11262 | 3666 | Sheet3

READY

FIG 6: CLEANED DATA COPIED IN SHEET3

Most popular page

In this section we can find out most popular Page. From this section firstly we can apply data preprocessing technique. Then we counted the frequency of the access page, the term Frequency of page means that the numbers of visibility of that page in web log file. This method we implement in java language which read the string line by line and checked it with other string. If the string gets matched again and again then we increment its counting by one each time and this counting only shows its frequency. This process is repeated until we reach the end of file.[4]

Algorithm to find frequency of log file

1. Start the execution.
2. Read the log data file.
3. If the file is not empty, can start the process.
4. Initialize a variable counter with value zero.
5. Read all the logs one by one from the log data.
6. Check whether the current log is valid or not.
7. If it is valid compare the text in log with other logs.
8. If it is equal then increment the counter variable by one.
9. Else, move on to the next log.
10. If the log is not valid move on to the next log.
11. The counter value is the output printed as frequency of current log,
12. Stop the execution.

ISSN: 2005-4238 IJAST

	A	B	C	D	E	F
1	IP	Time	URL	Staus	Frequency	
2	10.130.2.1	[28/Feb/2018:13:19:20	GET /contestproblem.php?name=RUET%20OJ%20TLE%20Testing%20Contest HTTP/1.1	200	525	
3	10.131.2.1	[17/Nov/2017:13:29:46	GET /contestshowcode.php?id=339&nm=shawon&cn=13 HTTP/1.1	200	5	
4	10.131.0.1	[30/Nov/2017:15:56:17	GET /contestshowcode.php?id=392&nm=dhruba_1603088&cn=16 HTTP/1.1	200	2	
5	10.131.2.1	[30/Nov/2017:17:57:36	GET /contestsubmit.php?id=45 HTTP/1.1	200	230	
6	10.131.0.1	[01/Dec/2017:15:32:02	GET /details.php?name=Factorial%20Factorization&cod=16 HTTP/1.1	200	97	
7	10.131.0.1	[30/Nov/2017:17:53:55	GET /details.php?name=Matrix%20For%20My%20valentine%20&cod=16 HTTP/1.1	200	4	
8	10.129.2.1	[25/Nov/2017:21:39:58	GET /details.php?name=Research%20Items&cod=16 HTTP/1.1	200	17	
9	10.129.2.1	[16/Nov/2017:16:17:31	GET /djquery/jquery-1.8.3.min.js HTTP/1.1	200	4	
10	10.128.2.1	[14/Dec/2017:00:00:28	GET /fonts/fontawesome-webfont.eot?v=4.6.3 HTTP/1.1	200	38	
11	10.130.2.1	[21/Feb/2018:05:05:39	GET /fonts/fontawesome-webfont.woff?v=4.6.3 HTTP/1.1	200	23	
12	10.128.2.1	[02/Mar/2018:15:45:42	GET /fonts/fontawesome-webfont.woff2?v=4.6.3 HTTP/1.1	200	224	
13	10.129.2.1	[25/Nov/2017:18:17:51	GET /fonts/glyphicons-halflings-regular.woff2 HTTP/1.1	200	5	
14	10.128.2.1	[02/Mar/2018:15:46:12	GET /js/jquery.min.js HTTP/1.1	200	771	
15	10.128.2.1	[02/Mar/2018:15:45:40	GET /js/vendor/modernizr-2.8.3.min.js HTTP/1.1	200	1659	
16	10.131.0.1	[02/Mar/2018:15:45:47	GET /js/vendor/moment.min.js HTTP/1.1	200	132	
17						
18						
19						
20						

FIG 7: FREQUENCY OF LOGS

RESULTS

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

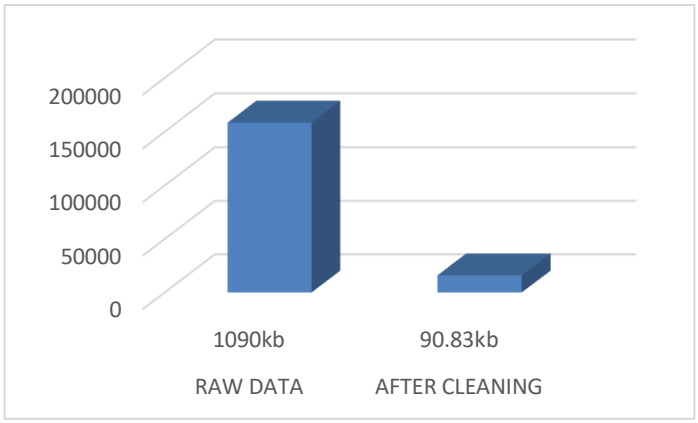


CHART 1 : PREPROCESSING CHART

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.

Errors

ISSN: 2005-4238 IJAST

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

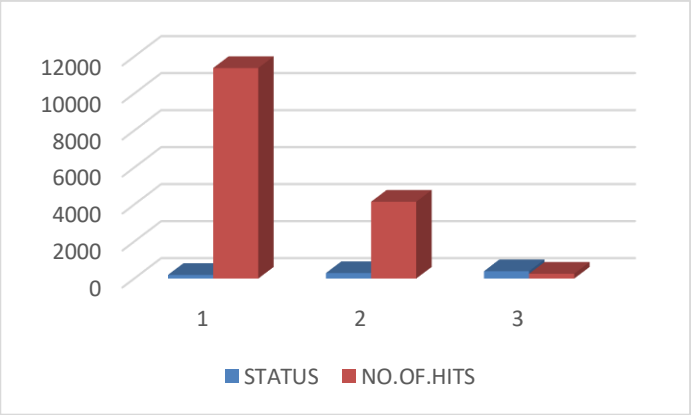


CHART 2 : ERRORS IN LOGS

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.

ISSN: 2005-4238 IJAST

[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] Manoj Kumar, Mrs. Meenu," Analysis of visitor's behavior from Web Log using Web Log Expert Tool ".

[9] 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).

[10]Jayanti Mehra," An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining" Volume 13, Number 2 (2018) pp. 1227-1232**,** International Journal of Applied Engineering Research ISSN 0973-4562.

[11]P.Nithya," Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots", 2012 National Conference on Computing and Communication Systems (NCCCS)

ISSN: 2005-4238 IJAST