Big Data Approach to Enhancing Quality of Web Application Contents

[1]Ms Swapna Sahu, [2]Mr Anuj kumar Pal

[1]M.Tech Scholar, [2]Assistant Professor
Department of CSE
Bansal Institute of Research and Technology
RGPV, Bhopal (M.P.), India

***Abstract*— In this paper we propose a technique for selecting big data approach since it enhances the quality of web contents. Big data is a technique for data sets that are so compound that conventional data processing application software is not enough to deal with them. Study of server log data can provide noteworthy and useful information. Information provided can help to find out user perception. This can improve the effectiveness of the Web sites by adapting the information structure to the users' behavior. The several web usage mining methods for extracting useful features is discussed and employ all these techniques to cluster the users of the domain to study their behaviours comprehensively. The goal of this project is to create a application that can able to give web log access information and store in a sophisticated way which is use to analyze user behaviour by mining enriched web access log data using BIG DATA HADOOP technology. The contributions of this research are a data enrichment that is content and source based and a tree-like visualization of frequent navigational sequences.**

***Index Terms*—** Big data, Web Mining, Web Usage, Hadoop

Introduction

Internet (WWW) is exceptionally famous and intelligent. It has turned into a critical wellspring of data and administrations. Web information has turned out to be more prominent and because of that web mining has pulled in part of consideration in late time [1]. Web content mining and web utilization mining. Cooley et al. [2, 3] presented the term web utilization mining in 1997 and concurring to their definition; it is the programmed disclosure of client gets to designs from web servers. Web utilization mining is an imperative innovation for understanding client's practices on the web what's more, is one of the most loved territory of numerous specialists in the late time. Gotten client get to examples can be utilized as a part of assortment of uses, for instance, one can monitor beforehand got to pages of a client. These pages can be utilized to distinguish the ordinary conduct of the client and to make forecast about sought pages [4]. Web webpage by making groups of clients with comparable get to designs and by including navigational connections.

Visit get to conduct for the clients can be utilized to recognize required connects to enhance the general execution of future accesses. Perfecting and reserving approaches can be made on the premise of much of the time got to pages to enhance inactivity time. In addition, utilization examples can be utilized for business knowledge in request to enhance deals and notice

Web structure mining is the way toward finding the association between site pages. Web content mining incorporates mining, extraction and combination of valuable information and learning of Web page content. Web Usage Mining is a strategy of extricating valuable data from the Web Log, e.g. the example in which a client experiences distinctive Web pages [2,3].

Web Log is by and large uproarious and equivocal. Web applications are expanding at a colossal speed and its clients, are expanding at exponential speed [4].

1.1 Web usage Mining

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the (Mining means extracting something useful or valuable from a baser substance, such as mining gold from the earth.) Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site[5,6].

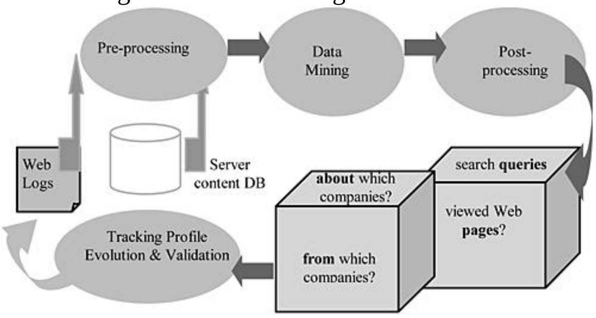


Figure 1 Web Usage Mining Approach

1.2 Big Data Approach

There is a lot of growth in web data. The data has been so large it wouldn't be handled by traditional mining methods. Big data is a terminology that express the large volume of data – both structured and unstructured. The 3Vs have been expanded to the characteristics of big data**.**

•Volume

–the size of data is now bigger than terabytes and petabytes. This huge dimension makes it difficult to analyze using conventional techniques

•Velocity

–big data should be used to mine large quantity of data within a pre-defined period of time. The long-established methods of mining may take huge time to mine such a volume of data.

•Variety

–big data comes from various sources. It is designed to handle structured, semi-structured as well as unstructured data.

Hadoop

Apache Hadoop is open source software which processes on large scale storage on commodity hardware. Hadoop has been designed such that its software framework can automatically manage and deal with hardware failure. Hadoop Map-Reduce and HDFS are designed with the help of Google Map-reduce and Google file system.

2.Related Work

2.1 Improving pattern quality in web usage mining by using semantic information

The impact of semantic data on example quality is assessed through a suggestion structure. Proposals are produced by either considering the continuous route designs regarding a solitary idea or considering blend of successive route designs regarding a few ideas. Test comes about demonstrate that coordinating semantic data gives deliberation that outcomes in impressive change of example quality. What's more, this approach handles new thing issue in proposal [15]. Both single idea and the consolidated affiliation rules have higher accuracy and scope values than the traditional Web utilization mining (without the utilization of semantic data). The change is higher for blend of affiliation tenets, henceforth, we can find that, when the measure of contributing semantic data expands, the example quality increments also.

The investigation on the single idea examples might be utilized for comprehension the client's aim. The one that has the most noteworthy accuracy and scope may mirror the client's plan for route. Another perception is that the expansion in window tally negligibly affects the accuracy and the scope, thus latest visit has all the earmarks of being the best one on the proposal. A fascinating outcome finished up from the tests is that, all together to accelerate the suggestion era, up to 30% of the guidelines can be disposed of with little decline in the quality.

2.2 Mining Generalized Associations of Semantic Relations from Textual Web Content

This paper has proposed a systematic approach for discovering knowledge from free-form textual Web content. Specifically, we present an automatic semantic relation extraction strategy to extract RDF metadata from textual Web content and an algorithm known as GP-Close for mining generalized patterns from RDF metadata. The experimental result shows that the GP-Close algorithm based on mining closed generalization closures can substantially reduce the pattern redundancy and perform much better than the original generalized association rule mining algorithm Cumulate in terms of time efficiency. The pattern analysis based on human validation shows that the proposed method is promising and useful**.**

2.3 Association Rules Mining from the Educational Data of ESOG Web-Based Application

In this paper, we show the KDD procedure [12] which incorporates the utilization of the Apriori calculation for the affiliation rules mining from the instructive information of ESOG Web-based application. In this paper we displayed the KDD stages for the affiliation rules mining the ESOG database which contains instructive information[17]. This procedure created 127 affiliation decides that could help and guide Greek Educators and School Managers to make instructive choices, plan learning exercises agreeing their understudy's advantages and productively deal with the classroom (isolate class into gatherings of understudies with comparative interests, adjust course's substance and so on). Amid the conduction of this work, many inquiries emerged that showed headings for future research.

2.4 Data Preparation for Mining World Wide Web browsing Patterns

This paper exhibits a few information planning procedures with a specific end goal to distinguish extraordinary clients and client sessions. A technique to isolate client sessions into semantically significant exchanges is characterized also, effectively tried against two different strategies. Exchanges recognized by the proposed strategies are utilized to find affiliation rules from genuine information utilizing the WEB MINER framework. This paper has exhibited the subtle elements of pre processing errands that are vital for performing Web Usage Mining, the utilization of information mining and learning disclosure systems to WWW server get to logs[13]. Future work will incorporate further tests to confirm the client perusing conduct show talked about web[18,19,20].

3.Problem identification

There has been tremendous increase in the amount of data generated by web nowadays. The data has been so large that it becomes complicated to analyze it with the help of our traditional mining methodology.

Big data is a general term for enormous quantity of data being collected from various sources, that are too large and raw in form. Big data deals with new challenge like complexity, security, risks ,to privacy. In this paper web logs are analyzed through big data and steps for effective implementation of log data has been suggested which help to increase their efficiency and reduce the complexity. This will help in diverse purposes.

4. Proposed Methodology

The proposed work will be divided in to three different parts:

1) Design and develop the front end for data collection and preparation.
2) Conversion of semi-structured data to structured data.
3) Analysing data using big data analysis approach.

4.1 Design and develop the front end for data collection and preparation.

As per the designing concern we had used HTML,CSS3 and JavaScript for front end. MySQl as a data source for our analysis and Tomcat apache as a web server to deploy our application in server side, we have used JSP as a server side programming. The snapshot of our application are given below in the figure 3. The skeleton of our program for collecting data is given below. In this web app different pages we have design like index.jsp, about.jsp, c.jsp, contact.jsp etc for collecting user information like ip address, browser info. Os, page, date, city and country etc. These data are further analysed using big data analysis.

In figure 4 the skeleton of our web application is shown. After accessing this web application by different client side through different geographic location the data stored in mysql data base the structure of our database is given as userId , visitDate, pageId, clientInfo, client_Ip, page_od, page_brw, page_country.

After Running web server (apache tomcat) we were fetch the url http://localhost:8080/WebAnalysis/index.jsp we got the output of our program in browser shown in figure 5. After Saving the data from the clients the table is shown in below figure 6. The above saved information is converted to CSV file using java program screenshot of file is given in figure 7.

After this using Hadoop programming we did our work. Here below some screen shot of hadoop cluster given below.

For our research we're going to be used check pattern statistics as mashable on-line news records to be had in mashable.Com. It is freely to be had for check and studies. For writing java program we're using notepad++ v6.Nine, Java improvement package version is JDK 1.7 for java environment and hadoop 2.3 for windows, and windows eight.1 operating system. Right here in this web website

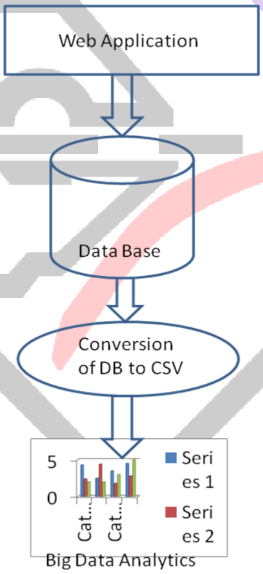
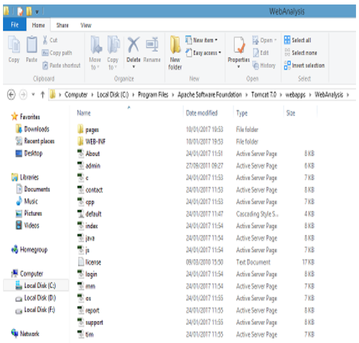


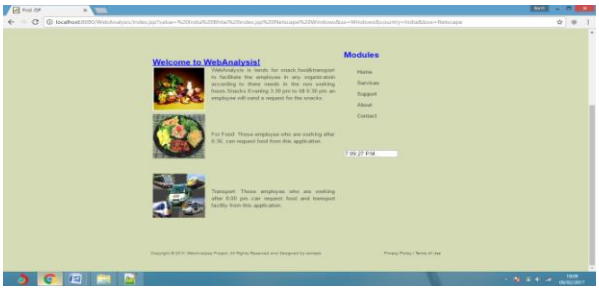
Figure 3 Proposed Methodologies



Computer › Local Disk (C:) › Program Files › Apache Software Foundation › Tomcat 7.0 › webapps › WebAnalysis ›

Name	Date modified	Type	Size
pages	10/01/2017 19:53	File folder	
WEB-INF	10/01/2017 19:53	File folder	
About	24/01/2017 11:51	Active Server Page	8 KB
admin	27/09/2011 09:27	Active Server Page	6 KB
c	24/01/2017 11:53	Active Server Page	7 KB
contact	24/01/2017 11:53	Active Server Page	8 KB
cpp	24/01/2017 11:53	Active Server Page	7 KB
default	24/01/2017 11:47	Cascading Style S...	4 KB
index	24/01/2017 11:54	Active Server Page	8 KB
java	24/01/2017 11:54	Active Server Page	8 KB
js	24/01/2017 11:54	Active Server Page	7 KB
license	09/03/2010 15:50	Text Document	17 KB
login	24/01/2017 11:54	Active Server Page	8 KB
mm	24/01/2017 11:54	Active Server Page	7 KB
os	24/01/2017 11:55	Active Server Page	7 KB
report	24/01/2017 11:55	Active Server Page	8 KB
support	24/01/2017 11:55	Active Server Page	8 KB
tim	24/01/2017 11:55	Active Server Page	7 KB

Figure 4 Snapshot of our web application



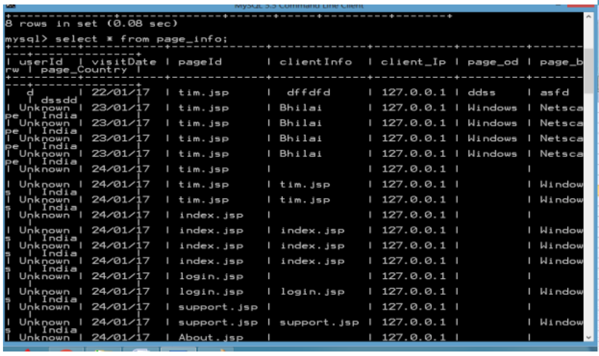
Welcome to WebAnalysis!

WebAnalysis is tends for snack,food&transport to facilitate the employee in any organization according to there needs in the non working hours. Snacks Evening 3:30 pm to till 5:30 pm an employee will send a request for the snacks.

For Food: These employee who are working after 6:30, can request food from this application.

Transport: These employee who are working after 8:00 pm can request food and transport facility from this application.

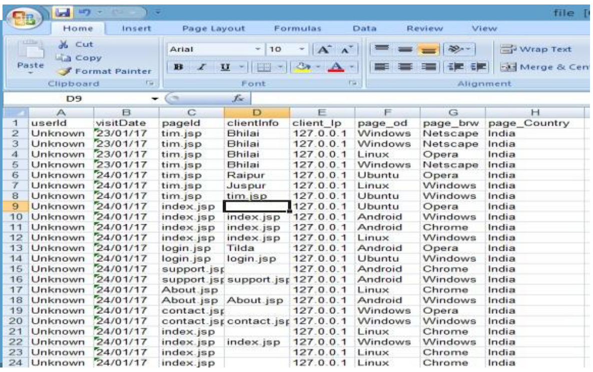
Figure 5. Output of index.jsp



```
+------------+------------+-----------+----------+------------+---------+
8 rows in set (0.08 sec)

mysql> select * from page_info;
+---------+------------+--------------+--------------+-----------+----------+---------
---+----------------+
| userId  | visitDate | pageId       | clientInfo   | client_Ip | page_od  | page_b
rw | page_Country |
+---------+------------+--------------+--------------+-----------+----------+---------
---+----------------+
|   d     | 22/01/17  | tim.jsp      |  dffdfd      | 127.0.0.1 | ddss     | asfd
   | dssdd        |
| Unknown | 23/01/17  | tim.jsp      | Bhilai       | 127.0.0.1 | Windows  | Netsca
pe | India        |
| Unknown | 23/01/17  | tim.jsp      | Bhilai       | 127.0.0.1 | Windows  | Netsca
pe | India        |
| Unknown | 23/01/17  | tim.jsp      | Bhilai       | 127.0.0.1 | Windows  | Netsca
pe | India        |
| Unknown | 23/01/17  | tim.jsp      | Bhilai       | 127.0.0.1 | Windows  | Netsca
pe | India        |
| Unknown | 24/01/17  | tim.jsp      |              | 127.0.0.1 |          |
   |              |
| Unknown | 24/01/17  | tim.jsp      | tim.jsp      | 127.0.0.1 |          | Window
s  | India        |
| Unknown | 24/01/17  | tim.jsp      | tim.jsp      | 127.0.0.1 |          | Window
s  | India        |
| Unknown | 24/01/17  | index.jsp    |              | 127.0.0.1 |          |
   |              |
| Unknown | 24/01/17  | index.jsp    | index.jsp    | 127.0.0.1 |          | Window
s  | India        |
| Unknown | 24/01/17  | index.jsp    | index.jsp    | 127.0.0.1 |          | Window
s  | India        |
| Unknown | 24/01/17  | index.jsp    | index.jsp    | 127.0.0.1 |          | Window
s  | India        |
| Unknown | 24/01/17  | login.jsp    |              | 127.0.0.1 |          |
   |              |
| Unknown | 24/01/17  | login.jsp    | login.jsp    | 127.0.0.1 |          | Window
s  | India        |
| Unknown | 24/01/17  | support.jsp  |              | 127.0.0.1 |          |
   |              |
| Unknown | 24/01/17  | support.jsp  | support.jsp  | 127.0.0.1 |          | Window
s  | India        |
| Unknown | 24/01/17  | About.jsp    |              | 127.0.0.1 |          |
```

Figure 6 Database with content



	A	B	C	D	E	F	G	H
1	userId	visitDate	pageId	clientInfo	client_Ip	page_od	page_brw	page_Country
2	Unknown	23/01/17	tim.jsp	Bhilai	127.0.0.1	Windows	Netscape	India
3	Unknown	23/01/17	tim.jsp	Bhilai	127.0.0.1	Windows	Netscape	India
4	Unknown	23/01/17	tim.jsp	Bhilai	127.0.0.1	Linux	Opera	India
5	Unknown	23/01/17	tim.jsp	Bhilai	127.0.0.1	Windows	Netscape	India
6	Unknown	24/01/17	tim.jsp	Raipur	127.0.0.1	Ubuntu	Opera	India
7	Unknown	24/01/17	tim.jsp	Juspur	127.0.0.1	Linux	Windows	India
8	Unknown	24/01/17	tim.jsp	tim.jsp	127.0.0.1	Ubuntu	Windows	India
9	Unknown	24/01/17	index.jsp		127.0.0.1	Ubuntu	Opera	India
10	Unknown	24/01/17	index.jsp	index.jsp	127.0.0.1	Android	Windows	India
11	Unknown	24/01/17	index.jsp	index.jsp	127.0.0.1	Android	Chrome	India
12	Unknown	24/01/17	index.jsp	index.jsp	127.0.0.1	Linux	Windows	India
13	Unknown	24/01/17	login.jsp	Tilda	127.0.0.1	Android	Opera	India
14	Unknown	24/01/17	login.jsp	login.jsp	127.0.0.1	Ubuntu	Windows	India
15	Unknown	24/01/17	support.jsp		127.0.0.1	Android	Chrome	India
16	Unknown	24/01/17	support.jsp	support.jsp	127.0.0.1	Android	Windows	India
17	Unknown	24/01/17	About.jsp		127.0.0.1	Linux	Chrome	India
18	Unknown	24/01/17	About.jsp	About.jsp	127.0.0.1	Android	Windows	India
19	Unknown	24/01/17	contact.jsp		127.0.0.1	Windows	Opera	India
20	Unknown	24/01/17	contact.jsp	contact.jsp	127.0.0.1	Windows	Windows	India
21	Unknown	24/01/17	index.jsp		127.0.0.1	Linux	Chrome	India
22	Unknown	24/01/17	index.jsp	index.jsp	127.0.0.1	Windows	Windows	India
23	Unknown	24/01/17	index.jsp		127.0.0.1	Linux	Chrome	India
24	Unknown	24/01/17	index.jsp		127.0.0.1	Linux	Chrome	India

Figure 7 Output of conversion Program

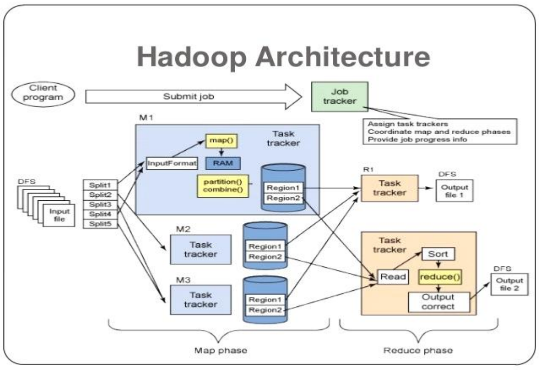
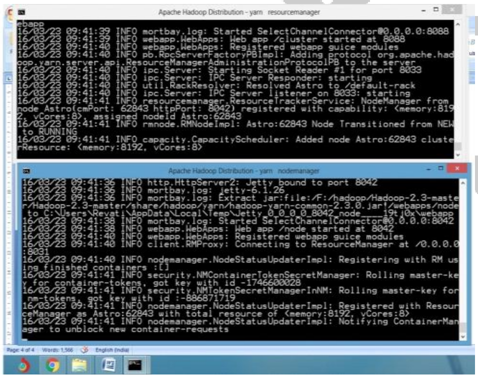


Figure 8 Hadoop Architecture

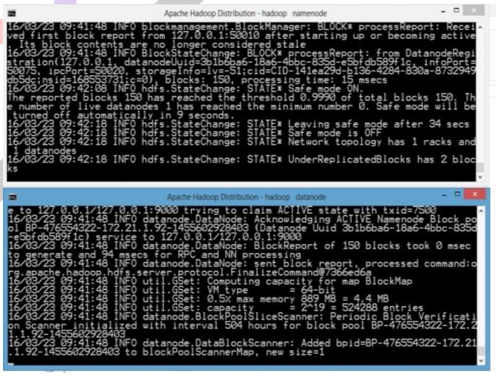
Open cmd prompt in admin mode and start hadoop demon using C:\Hadoop-2.3-master\sbin\start-yarn command snap shot is follows. After this start dfs by using C:\Hadoop-2.3-master\sbin\start-dfs command the snapshot in figure 10



```
ebapp
16/03/23 09:41:39 INFO mortbay.log: Started SelectChannelConnector@0.0.0.0:8088
16/03/23 09:41:39 INFO webapp.WebApps: Web app /cluster started at 8088
16/03/23 09:41:40 INFO webapp.WebApps: Registered webapp guice modules
16/03/23 09:41:40 INFO pb.RpcServerFactoryPBImpl: Adding protocol org.apache.had
oop.yarn.server.api.ResourceManagerAdministrationProtocolPB to the server
16/03/23 09:41:40 INFO ipc.Server: Starting Socket Reader #1 for port 8033
16/03/23 09:41:40 INFO ipc.Server: IPC Server Responder: starting
16/03/23 09:41:40 INFO util.RackResolver: Resolved Astro to /default-rack
16/03/23 09:41:40 INFO ipc.Server: IPC Server listener on 8033: starting
16/03/23 09:41:41 INFO resourcemanager.ResourceTrackerService: NodeManager from
node Astro(cmPort: 62843 httpPort: 8042) registered with capability: <memory:819
2, vCores:8>, assigned nodeId Astro:62843
16/03/23 09:41:41 INFO rmnode.RMNodeImpl: Astro:62843 Node Transitioned from NEW
 to RUNNING
16/03/23 09:41:41 INFO capacity.CapacityScheduler: Added node Astro:62843 cluste
rResource: <memory:8192, vCores:8>
```

```
16/03/23 09:41:36 INFO http.HttpServer2: Jetty bound to port 8042
16/03/23 09:41:36 INFO mortbay.log: jetty-6.1.26
16/03/23 09:41:36 INFO mortbay.log: Extract jar:file:/F:/hadoop/Hadoop-2.3-maste
r/Hadoop-2.3-master/share/hadoop/yarn/hadoop-yarn-common-2.3.0.jar!/webapps/node
 to C:\Users\Revati\AppData\Local\Temp\Jetty_0_0_0_0_8042_node____19tj0x\webapp
16/03/23 09:41:38 INFO mortbay.log: Started SelectChannelConnector@0.0.0.0:8042
16/03/23 09:41:38 INFO webapp.WebApps: Web app /node started at 8042
16/03/23 09:41:40 INFO webapp.WebApps: Registered webapp guice modules
16/03/23 09:41:40 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8031
16/03/23 09:41:40 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM us
ing finished containers :[]
16/03/23 09:41:41 INFO security.NMContainerTokenSecretManager: Rolling master-ke
y for container-tokens, got key with id -1746600028
16/03/23 09:41:41 INFO security.NMTokenSecretManagerInNM: Rolling master-key for
 nm-tokens, got key with id :-886871719
16/03/23 09:41:41 INFO nodemanager.NodeStatusUpdaterImpl: Registered with Resour
ceManager as Astro:62843 with total resource of <memory:8192, vCores:8>
16/03/23 09:41:41 INFO nodemanager.NodeStatusUpdaterImpl: Notifying ContainerMan
ager to unblock new container-requests
```

Figure 9 Yarn Hadoop



```
16/03/23 09:41:48 INFO blockmanagement.BlockManager: BLOCK* processReport: Recei
ved first block report from 127.0.0.1:50010 after starting up or becoming active
. Its block contents are no longer considered stale
16/03/23 09:41:48 INFO BlockStateChange: BLOCK* processReport: from DatanodeRegi
stration(127.0.0.1, datanodeUuid=3b1b6ba6-18a6-4bbc-835d-e5bfdb589f1c, infoPort=
50075, ipcPort=50020, storageInfo=lv=-51;cid=CID-141ea29d-b136-4284-830a-8732949
db5dc;nsid=168553731;c=0), blocks: 150, processing time: 15 msecs
16/03/23 09:42:08 INFO hdfs.StateChange: STATE* Safe mode ON.
The reported blocks 150 has reached the threshold 0.9990 of total blocks 150. Th
e number of live datanodes 1 has reached the minimum number 0. Safe mode will be
 turned off automatically in 9 seconds.
16/03/23 09:42:18 INFO hdfs.StateChange: STATE* Leaving safe mode after 34 secs
16/03/23 09:42:18 INFO hdfs.StateChange: STATE* Safe mode is OFF
16/03/23 09:42:18 INFO hdfs.StateChange: STATE* Network topology has 1 racks and
 1 datanodes
16/03/23 09:42:18 INFO hdfs.StateChange: STATE* UnderReplicatedBlocks has 2 bloc
ks
```

```
e to 127.0.0.1/127.0.0.1:9000 trying to claim ACTIVE state with txid=7500
16/03/23 09:41:48 INFO datanode.DataNode: Acknowledging ACTIVE Namenode Block po
ol BP-476554322-172.21.1.92-1455602928403 (Datanode Uuid 3b1b6ba6-18a6-4bbc-835d
-e5bfdb589f1c) service to 127.0.0.1/127.0.0.1:9000
16/03/23 09:41:48 INFO datanode.DataNode: BlockReport of 150 blocks took 0 msec
to generate and 94 msecs for RPC and NN processing
16/03/23 09:41:48 INFO datanode.DataNode: sent block report, processed command:o
rg.apache.hadoop.hdfs.server.protocol.FinalizeCommand@7366ed6a
16/03/23 09:41:48 INFO util.GSet: Computing capacity for map BlockMap
16/03/23 09:41:48 INFO util.GSet: VM type       = 64-bit
16/03/23 09:41:48 INFO util.GSet: 0.5% max memory 889 MB = 4.4 MB
16/03/23 09:41:48 INFO util.GSet: capacity      = 2^19 = 524288 entries
16/03/23 09:41:48 INFO datanode.BlockPoolSliceScanner: Periodic Block Verificati
on Scanner initialized with interval 504 hours for block pool BP-476554322-172.2
1.1.92-1455602928403
16/03/23 09:41:48 INFO datanode.DataBlockScanner: Added bpid=BP-476554322-172.21
.1.92-1455602928403 to blockPoolScannerMap, new size=1
```

Figure 10 DFS Of Yarn Hadoop

4.Result

We were store more than 100 records in our database for testing and analysis purpose. Here we have presented result one by one

4.1 Analysis using PageID

In our web application we have 15 jsp pages for accessing our clients. Pages are about.jsp, Admin.jsp, c.jsp, contact.jsp, cpp.jsp, default.jsp, index.jsp, java.jsp, js.jsp, login.jsp, mm.jsp, os.jsp, report.jsp, support, tim.jsp

Table 1

Page ID	tim.jsp	about.jsp	report.jsp	Admin.jsp	index.jsp	c.jsp	contact.jsp	cpp.jsp	Os.jsp	Default.jsp	Java.jsp	js.jsp	login.jsp	mm.jsp	Support
Hit	21	34	22	11	67	32	11	56	44	5	52	21	66	44	3

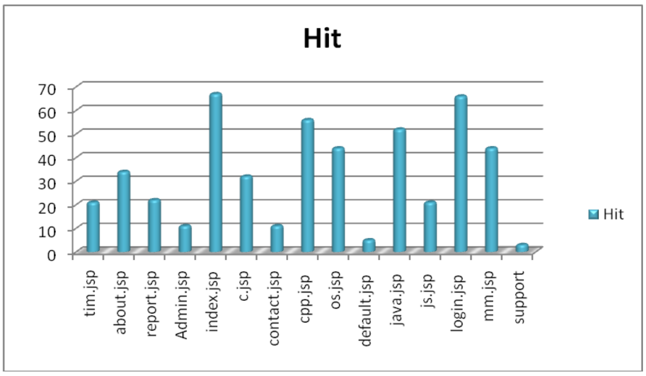


Figure 11 Analysis using Page id

In above graph and table we can conclude that index.jsp and login.jsp and cpp.jsp are most frequent pages that our client access.

4.2 Based on client info

We were listed city in below table refer Table 2

Table 2 city with Hits

City	Hit	City	Hit
Bhilai	23	Bilaspur	43
Durg	54	Champa	12
Raipur	33	Dongarhgadh	32
Raigarh	45	Tilda	21
Jagdalpur	24	Bhatapara	11

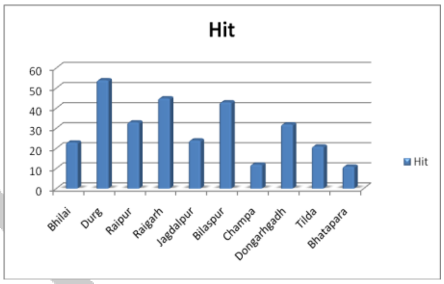


Figure 12 Analysis using different geographical region.

4.3 Analysis based on Operating System

Refer Table 3 for diiferent OS

Table 3 Os with Hit

OS	HIT
Window	70
Linux	3
Ubuntu	32
Sun	12
Macos	11

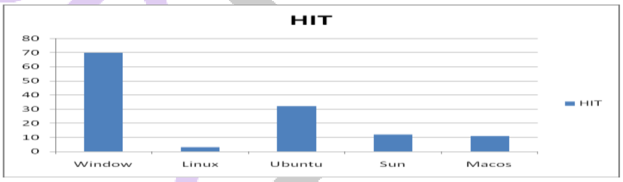


Figure 13 analysis using different OS

4.4 Analysis using Browsers Refer Table 4

Table 4 Different browser with its no of access.

Browser	Firefox	Chrome	IE	Safari	Opera	Edge	Chromium	Maxthon	SeaMonkey	Midori	SlimBrowser	Gnome	Rockmelt	Surf
Hit	54	56	23	44	32	2	43	0	5	21	7	1	4	1

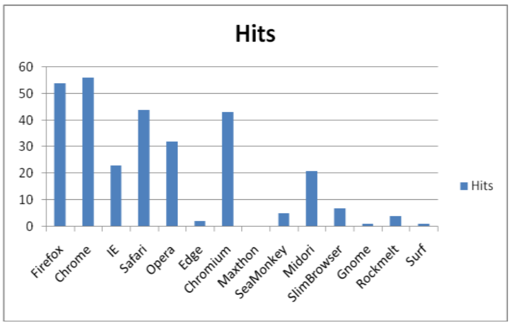


Figure 14 Analysis using Web Browser

6. Conclusion

The proposed system will give a privilege to managers and top level management employees to explore the hidden data and those relevant facts and data that will help to grow the business like online news, advertisement and marketing agency. Using this research we found that sampling / aggregation is simple, Reduce data movement and replication, it Bring the analytics as close as possible to the data, Optimize computation speed, User Behaviour, Frequent item set and Location centric analysis

References

[1] Cooley, R., Mobasher, B., and Srivastava, J, “Web mining: information and pattern discovery on the World Wide Web”, International Conference on Tools with Artificial Intelligence, Newport Beach, IEEE, 1997, pp. 558-567.

[2] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava,” Data preparation for mining World Wide Web browsing patterns”, Journal of Knowledge and Information System,1999,pp. 1-27.

[3] https://www.researchgate.net/publication/220802288_Recent_Developments_in_Web_Usage_Mining_Research

[4] Tan,P. N. and Kumar, V.: 2002, Discovery of Web Robot Sessions Based on their Navigational Patterns, Data Mining and Knowledge Discovery.

[5] Rapha¨el Nowak. Investigating the interactions between individuals and music technologies within contemporary modes of music consumption. First Monday, 19(10):Online, 2014.

[6] Association Rules Mining from the Educational Data of ESOG Web-Based Application, Stefanos Ougiaroglou1 and Giorgos Paschalis2, Dept. of Applied Informatics, University of Macedonia, Thessaloniki Greece Human-Computer Interaction Group, University of Patras, Patra, Greece stoug@uom.gr, gpasxali@upatras.gr, L. Iliadis et al. (Eds.): AIAI 2012 Workshops, IFIP AICT 382, pp. 105–114, 2012., Springer-Verlag Berlin Heidelberg 2012

[7] Andryw Marques, Nazareno Andrade, and Leandro Balby Marinho. Exploring the Relation Between Novelty Aspects and Preferences in Music Listening. In Proc. ISMIR, 2013.

[8] Joshua L. Moore, Shuo Chen, Thorsten Joachims, and Douglas Turnbull. Taste Over Time: The Temporal Dynamics of User Preferences. In Proc. ISMIR, 2013.

[9] C. Ding and J. Zhou, “Log Based Indexing to Improve Web Site Search,” *Proceedings of the ACM Symposium on Applied Computing*, Seoul, Korea, 2007, Mar 11-15, pp. 829 -833

[10] B. Mobasher, R. Cooley and J. Srivastava, “Automatic Personalization Based on Web Usage Mining,” *Communications of the ACM*, 2000, Vol. 43, pp. 142-151.

[11] A Framework for Web Usage Mining in Electronic Government Ping Zhou, ZhongjianLe School of Information Management, JiangXi University of Finance and Economic, NanChang ,China 330013 Zpjx@126.com, Zhou, P., Le, Z., 2007, in IFIP International Federation for Information Processing, Volume 252, Integration and Innovation Orient to E-Society Volume 2, eds. Wang, W., (Boston: Springer), pp. 487-496.

[12] Data Preparation for Mining World Wide Web browsing Patterns Robert Cooley*, Bamshad Mobasher, and J aideep Srivastava Department of Computer Science and Engineering University of Minnesota 4-192 EECS Bldg., 200 Union St. SE Minneapolis, MN 55455, USA

[13] Effectual Web Content Mining using Noise Removal from Web Pages. Sivakumar1 Published online: 24 April 2015 _ Springer Science+Business Media New York 2015, Wireless Pers Commun (2015) 84:99–121 DOI 10.1007/s11277-015-2596-7

[14] Improving pattern quality in web usage mining by using semantic information Pinar Senkul · Suleyman Salin, Knowl Inf Syst (2012) 30:527–541 DOI 10.1007/s10115-011-0386-4 , Received: 19 April 2010 / Revised: 5 January 2011 / Accepted: 6 February 2011 /Published online: 24 February 2011 © Springer-Verlag London Limited 2011

[15] Leung CW, Chan SC, Chung F (2006) A collaborative filtering framework based on fuzzy association rules and multiple-level similarity. Knowl Inf Syst 10(3):357–381

[16] Missaoui R, Valtchev P, Djeraba C, AddaM (2007) Toward recommendation based on ontology-powered web-usage mining. IEEE Internet Comput 11(4):45–52

[17] Mobasher B, Cooley R, Srivastava J (2000) Automatic personalization based on web usage mining. Commun ACM 43(8):142–151

[18] A.P. Sheth, C. Ramakrishnan, and C. Thomas, “Semantics for the Semantic Web: The Implicit, the Formal and the Powerful,” Int’l J. Semantic Web Information Systems, vol. 1, no. 1, pp. 1-18, 2005.

[19] Mabroukeh NR, Ezeife CI (2009) Using domain ontology for semantic web usage mining and next page prediction. In: Proceedings of conference on information and knowledge management (CIKM), pp 1677–1680

[20] Shahabi, C., Banaei-Kashani, F. and Faruque, J.: 2001,A Reliable,E ⁄cient,a nd Scalable System for Web Usage Data Acquisition,In : WebKDD’01Workshop in conjunction with the ACMSIGKDD 2001,Sa n Francisco, CA, August