An Approach of Big Data Analytics for quality enhancement of web application Content

Kavita Tandon
M.Tech Scholar, Computer Network
Bhilai Institute of Technology
Durg, India

Prof. D.P Mishra
Associate Professor, Computer Network
Bhilai Institute of Technology
Durg, India

Abstract—

In recent years, it has been observed that major publishing media is World Wide Web (www), which has attracted many interested people. Most of the web content is semi-structured and unstructured, so assembling those data into meaningful information is very tiresome. Web servers generated by people across the world is a huge amount of information in itself which is generated because of web users browsing actions. Numerous specialists have examined these supposed web get to log information to better comprehend and describe web clients. Information can be improved by adding the details of traversed pages and the starting point (e.g., geographic, hierarchical) of that particular solicitations. The objective of this venture is to make a web application that is ready to give web log get to data and store adroitly which is use to dissect client conduct by mining enhanced web get to log information utilizing BIG DATA HADOOP innovation. The few web use digging strategies for extricating valuable highlights is examined and utilize every one of these methods to bunch the clients of the area to contemplate their practices thoroughly. Main objective of this research is data enrichment which is substantial in nature, and source based and tree-like representation of regular navigational successions.

Keywords: Hadoop Big data, Web application log, Map-Reduce, Web Mining, Web Usage mining etc.

1. Introduction

Web (WWW) is exceptionally popular, which is basic wellspring of information and administrations. Web data has ended up being more unmistakable and in light of that web mining has pulled in part of thought in late time [1]. Web content mining and web use mining [2, 3] exhibited the term web usage mining in 1997 and agreeing to their definition. It is a customized disclosure, whose outlines client gets from web servers. Web use mining is a basic development for understanding client's practices of web is a standout amongst the most adored region of various experts in the late time. Clients get to illustrations can be used as a piece of collection of employments, for example, one can screen in advance got to pages of a client. These pages can be used to recognize the standard lead of the client and to make figure about looked for pages [4], web pages by influencing gatherings of clients with practically identical to get to plans and by including navigational associations.

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 path toward finding the association between site pages. Web content mining consolidates mining, drawing out and blending of significant data and learning of Web page content. Web Usage Mining is a methodology of removing significant information from the Web Log, e.g. the case in which a client encounters unmistakable Web pages[2,3].

Web Log is all things considered uproarious and obscure. Web applications are extending at an

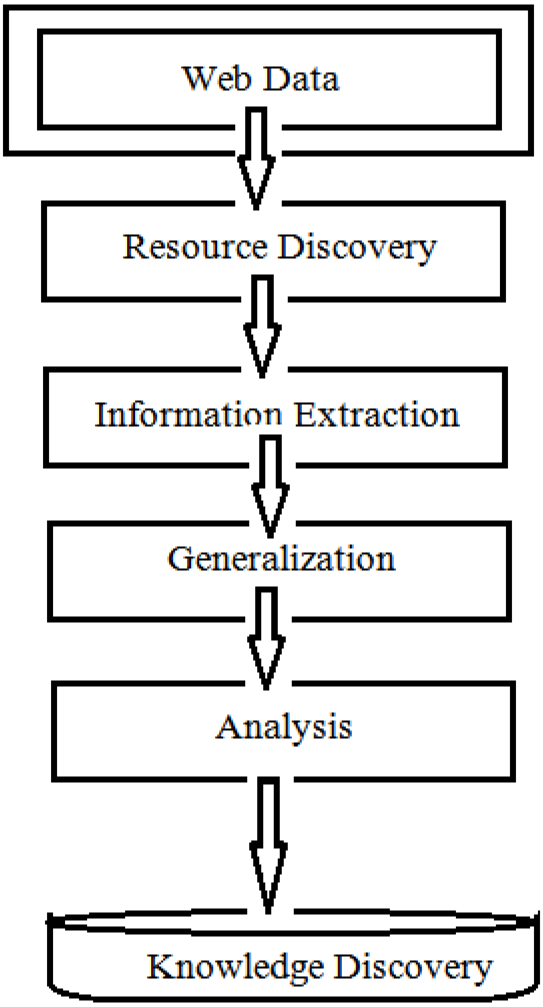


Figure 1 Quality Enhancement Methodology OF Web App

enormous speed and its clients, are growing simultaneously [4].

1.1 Big Data Analytic Approach

Big data is basically characterized by the volume of an informational collection. Huge informational indexes are by and large huge—measuring many terabytes—and once in a while crossing the limit of petabytes. The term huge information was gone before by extensive databases (VLDBs) which were overseen utilizing database administration frameworks (DBMS). Today, enormous information falls under three classifications of information sets—structured, unstructured and semi-organized.

Structured data sets include information which can be utilized as a part of its unique frame to infer comes about. Cases incorporate social information, for example, worker compensation records. Most present day PCs and applications are modified to produce organized information in preset configurations to make it less demanding to process.

Unstructured data sets include without appropriate arranging and arrangement. Cases incorporate human writings, Google query output yields, and so on. These irregular accumulations of informational indexes require all the more handling force and time for change into organized informational collections with the goal that they can help in determining unmistakable results.Semi-Structured informational collections are a blend of both organized and unstructured information. These informational indexes may have an appropriate structure but need characterizing components for arranging and handling. Cases incorporate RFID and XML information.

Semi-Structured informational indexes are a blend of both organized and unstructured information. These informational collections may have a legitimate structure but then need characterizing components for arranging and preparing. Illustrations incorporate RFID and XML information.

Huge information preparing requires a specific setup of physical and virtual machines to infer comes about. The preparing is done all the while to accomplish comes about as fast as could be expected under the circumstances. Nowadays enormous information preparing strategies additionally incorporate Cloud Computing and Artificial Intelligence. These advancements help in diminishing manual data sources and oversight via computerizing numerous procedures and assignments.

The developing idea of huge information has made it hard to give it a usually acknowledged definition. Informational collections are transferred the enormous information status in light of advancements and apparatuses required for their handling.

Cost reduction
BIG DATA ANALYTICS
Faster, better decision making
New products & services

Figure 2 Big Data Analytics

2. **Related Work**

2.1 **Use of semantic information to improve quality of patterns**

The effect of semantic information on illustration quality is evaluated through a recommendation structure. Recommendations are delivered by either considering the ceaseless course outlines with respect to a singular thought or considering mix of progressive course plans in regards to a couple of thoughts. Test happens show that organizing semantic information gives consultation, that results in great difference in illustration quality. Likewise, this approach handles new thing issue in proposition [15]. Both single thought and the combined connection rules have higher precision and degree esteems than the conventional Web use mining (without the usage of semantic information). The change is higher for mix of alliance fundamentals, from now on, we can find that, when the measure of contributing semantic information grows, the illustration quality additions moreover.

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 precise approach for finding learning from freestyle printed Web content. In particular, we exhibit a programmed semantic connection extraction procedure to separate RDF metadata from printed Web content and a calculation known as GP-Close to mine summed up designs from RDF metadata. The trial result demonstrates that the GP-Close calculation in light of mining shut speculation terminations can considerably lessen the example excess and perform much superior to anything the first summed up affiliation run mining calculation Cumulate as far as time effectiveness. The pattern analysis with respect to human validation shows that the proposed technique is promising and valuable.

2.3 **Association Rules Mining in ESOG application**

In this paper, author demonstrated the KDD strategy [12] which fuses the usage of the Apriori algorithm for the alliance rules mining from the educational data of ESOG Web-based application. This paper shows the KDD stages for the connection rules mining of ESOG database which contains educational information [17]. This method made 127 alliance concludes that could help and guide Greek Educators and School Managers to settle on enlightening decisions, design learning practices concurring their understudy's favourable circumstances and beneficially manage the classroom (detach class into social affairs of understudies with near premiums, modify course's substance et cetera). In the midst of the conduction of this work, numerous request rose that demonstrated headings for future research.

2.4 **Preparing Data for Mining WWW 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 displayed the unpretentious components of pre handling errands that are essential for performing Web Usage Mining, the use of data mining and learning revelation frameworks to WWW server get to logs[13].

This paper additionally displayed test comes about on manufactured information with the end goal of looking at exchange recognizable proof methodologies, and on certifiable modern information to outline a few of its applications[14,15,17]. The trades identified with the reference length approach performed dependably well on both the certifiable data and the made data. For the genuine information, just the reference length exchanges found decides that couldn't be sensibly derived from the structure of the Web locales. Since the traversed critical page is not generally the last one, the principal as it were exchanged related to the maximal forward reference approach did not perform well with genuine information having higher network level. The helper content exchanges prompted to an overwhelmingly huge arrangement of tenets, which confines the esteem of the information mining process. In future, it is expected to perform further tests to confirm that the client perusing conduct show talked about web[18,19,20].

2. Problem identification

Data collection and Sorting is big problem from the logs managers and records oriented dataset. Time and cost is much high in traditional data mining approach. Aggregation and summarization is big deal in web. Analysis of back end data is big task. Overall collection of relevant information is big deal all the way.

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 is given below in the figure 3. The skeleton of our program for collecting data is given below. In this web app, we have designed different pages like index.jsp, about.jsp, c.jsp, contact.jsp etc for collecting user information like ip address, browser information, 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 apache tomcat, we fetched the url http://localhost:8080/WebAnalysis/index.jsp and got the output of our program in program appeared in figure 5. After saving the information from the clients, the table is shown in beneath figure 6. The above saved data is changed over to CSV document utilizing java program screenshot of record 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 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. For this experimentation, 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.

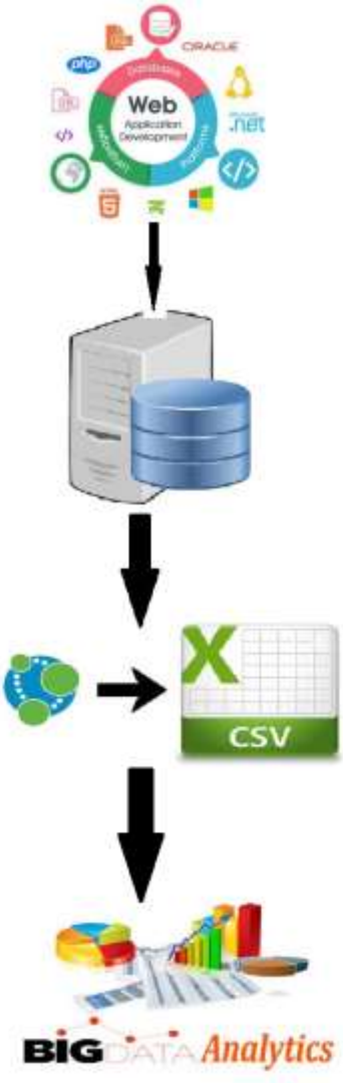


Figure 3 Proposed Methodologies

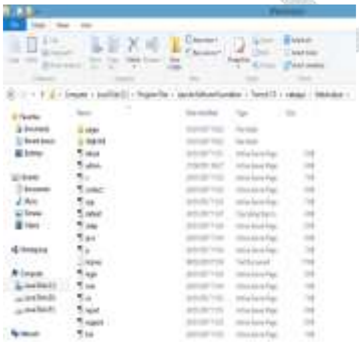
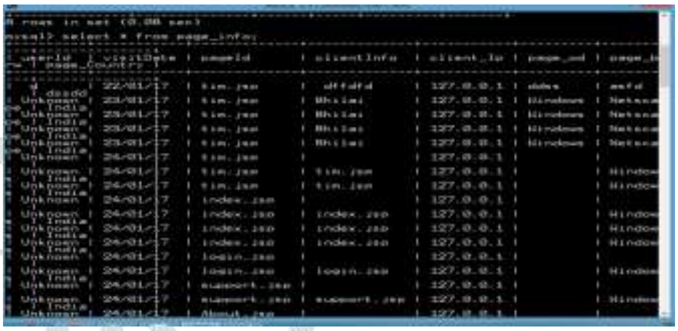


Figure 4 Snapshot of our web application

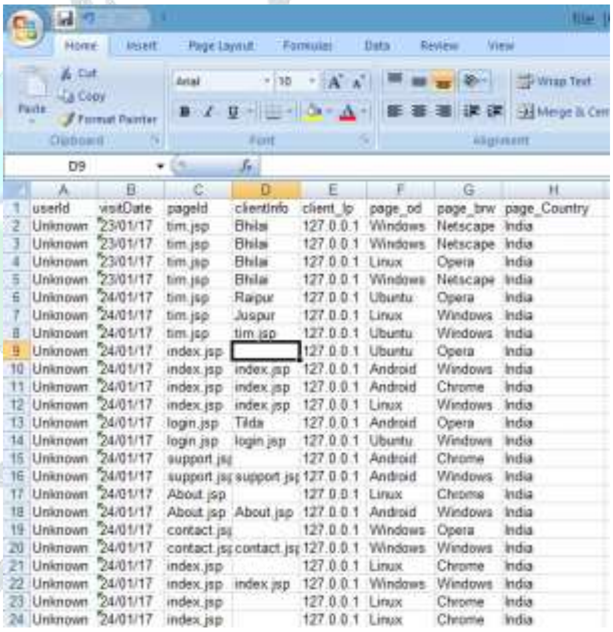
Figure 5. Output of index.jsp



```
8 rows in set (0.00 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 | ddds    | asfd
   | dsadd
| 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
  | India
| Unknown | 24/01/17 | tim.jsp      | tim.jsp     | 127.0.0.1 |         | Window
  | India
| Unknown | 24/01/17 | index.jsp    |             | 127.0.0.1 |         |
| Unknown | 24/01/17 | index.jsp    | index.jsp   | 127.0.0.1 |         | Window
  | India
| Unknown | 24/01/17 | index.jsp    | index.jsp   | 127.0.0.1 |         | Window
  | India
| Unknown | 24/01/17 | index.jsp    | index.jsp   | 127.0.0.1 |         | Window
  | India
| Unknown | 24/01/17 | login.jsp    |             | 127.0.0.1 |         |
| Unknown | 24/01/17 | login.jsp    | login.jsp   | 127.0.0.1 |         | Window
  | India
| Unknown | 24/01/17 | support.jsp  |             | 127.0.0.1 |         |
| Unknown | 24/01/17 | support.jsp  | support.jsp | 127.0.0.1 |         | Window
  | 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

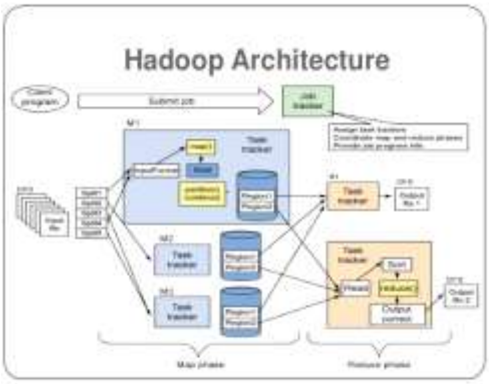


Figure 8 Hadoop Architecture

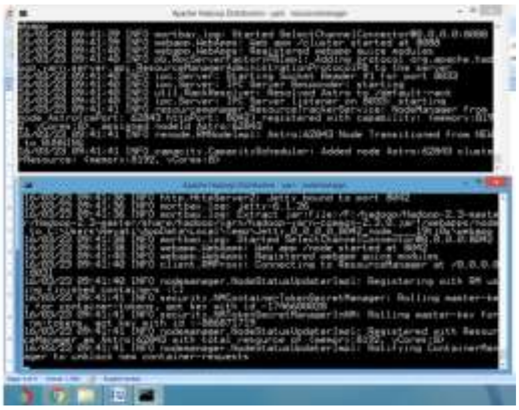


Figure 9 Yarn Hadoop

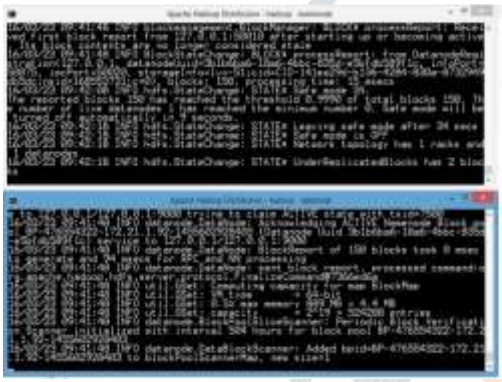


Figure 10 DFS Of Hadoop

3. Result

We have stored 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

PageID	Hit
tim.jsp	21
about.jsp	34
report.jsp	22
Admin.jsp	11
index.jsp	67
c.jsp	32
contact.jsp	11
cpp.jsp	56
os.jsp	44
default.jsp	5
java.jsp	52
js.jsp	21
login.jsp	66
mm.jsp	44
Support	3

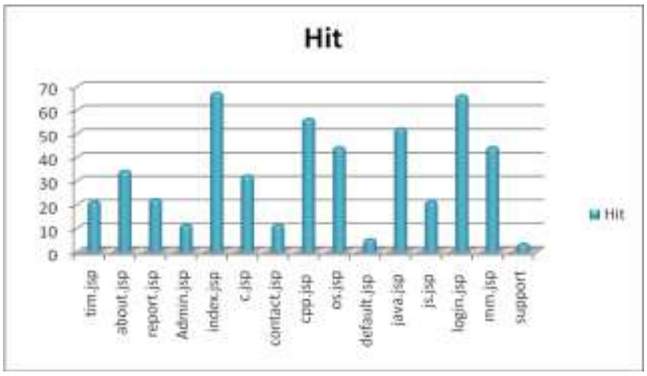


Figure 11 Analysis using Page id

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 have listed city in below table refer Table 2

Table 2 city with Hits

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

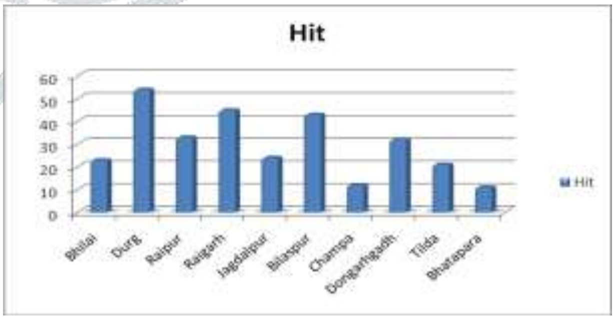


Figure 12 Analysis using Different geographical region.

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

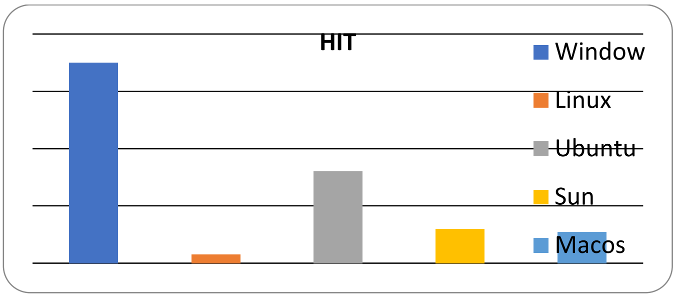


Figure 13 analysis using different OS

4.4 Analysis using Browsers Refer Table 4

Table 4 Different browser with its no of access

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

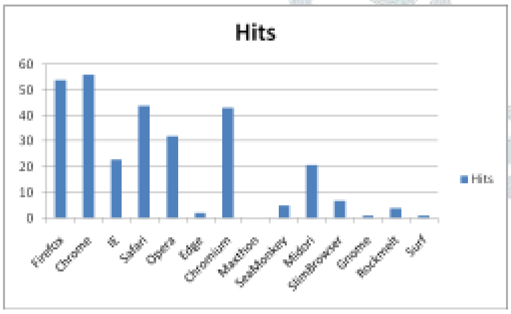


Figure 14 Analysis using Web Browser.

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. By reducing data movement and replication, it brings the analytics as close as possible to Data, Optimized 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