Big Data Approach to Enhancing Quality of Web Application Contents

Ms Swapna Sahu, Mr Anuj Kumar Pal

1M.Tech Scholar, 2Assistant 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 behaviors 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 get 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](image)

![Figure 4 Snapshot of our web application](image)

![Figure 5. Output of index.jsp](image)
Open cmd prompt in admin mode and start hadoop demon using `C:\Hadoop-2.3-master\sbin\start-yarn` command snapshot is follows. After this start dfs by using `C:\Hadoop-2.3-master\sbin\start-dfs` command the snapshot in figure 10.

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>
<thead>
<tr>
<th>Page ID</th>
<th>tim.jsp</th>
<th>about.jsp</th>
<th>report.jsp</th>
<th>Admin.jsp</th>
<th>index.jsp</th>
<th>c.jsp</th>
<th>contact.jsp</th>
<th>cpp.jsp</th>
<th>Os.jsp</th>
<th>Default.jsp</th>
<th>Java.jsp</th>
<th>js.jsp</th>
<th>login.jsp</th>
<th>mm.jsp</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit</td>
<td>21</td>
<td>34</td>
<td>22</td>
<td>11</td>
<td>67</td>
<td>32</td>
<td>11</td>
<td>56</td>
<td>44</td>
<td>5</td>
<td>52</td>
<td>21</td>
<td>66</td>
<td>44</td>
<td>3</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>City</th>
<th>Hit</th>
<th>City</th>
<th>Hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhilai</td>
<td>23</td>
<td>Bilaspur</td>
<td>43</td>
</tr>
<tr>
<td>Durg</td>
<td>54</td>
<td>Champa</td>
<td>12</td>
</tr>
<tr>
<td>Raipur</td>
<td>33</td>
<td>Dongargadh</td>
<td>32</td>
</tr>
<tr>
<td>Raigarh</td>
<td>45</td>
<td>Tilda</td>
<td>21</td>
</tr>
<tr>
<td>Jagdalpur</td>
<td>24</td>
<td>Bhatapara</td>
<td>11</td>
</tr>
</tbody>
</table>

4.3 Analysis based on Operating System
Refer Table 3 for different OS

Table 3 Os with Hit

<table>
<thead>
<tr>
<th>OS</th>
<th>HIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window</td>
<td>70</td>
</tr>
<tr>
<td>Linux</td>
<td>3</td>
</tr>
<tr>
<td>Ubuntu</td>
<td>32</td>
</tr>
<tr>
<td>Sun</td>
<td>12</td>
</tr>
<tr>
<td>Macos</td>
<td>11</td>
</tr>
</tbody>
</table>

4.4 Analysis using Browsers Refer Table 4
Table 4 Different browser with its no of access.

<table>
<thead>
<tr>
<th>Browser</th>
<th>Firefox</th>
<th>Chrome</th>
<th>IE</th>
<th>Safari</th>
<th>Opera</th>
<th>Edge</th>
<th>Chromium</th>
<th>Maxthon</th>
<th>SeaMonky</th>
<th>Midori</th>
<th>SlimBrowser</th>
<th>Gnome</th>
<th>Rockmelt</th>
<th>Surf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit</td>
<td>54</td>
<td>56</td>
<td>23</td>
<td>44</td>
<td>32</td>
<td>2</td>
<td>43</td>
<td>0</td>
<td>5</td>
<td>21</td>
<td>7</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>
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 behavior, frequent item set and location-centric analysis

References
[6] Association Rules Mining from the Educational Data of ESOG Web-Based Application, Stefanos Ougiaroglou1 and Giorgos Paschalidis2, Dept. of Applied Informatics, University of Macedonia, Thessaloniki Greece Human-Computer Interaction Group, University of Patras, Patra, Greece stoug@uom.gr, gpasiali@upatras.gr, L. Iliadis et al. (Eds.): AIAI
[12] Data Preparation for Mining World Wide Web browsing Patterns Robert Cooley*, Bamshad Mobasher, and J aiddeep Srivastava Department of Computer Science and Engineering University of Minnesota 4-192 EECS Bldg., 200 Union St. SE Minneapolis, MN 55455, USA