Web crawlers Data Mining Techniques for Handling Big Data Analytics

Mr.V.NarsingRao[2]
Sphoorthy Engineering College,
Nadergul,R.R.District

Mr.K.Vijay Babu[1]
CMR Engineering College,
Medchal.

Abstract:

Web becomes an important part of organization. The huge usage data as a result of interaction of users and Web can be extracted to be knowledge applied in various application. Analysis of web site regularities and patterns in user navigation is getting more attention from business and research community a web browsing becomes an everyday task for more people around the world. This extremely large-scaled data called Big data are in terms of quantity, complexity, semantics, distribution, and processing costs in computer science, cognitive informatics, web-based computing, cloud computing, and computational intelligence. The size of the collected data about the Web and mobile device users is even greater. To provide the ability to make sense and maximize utilization of such vast amounts of web data for knowledge discovery and decision-making is crucial to scientific advancement; we need new tools for such a big web data mining. Visualization is a tool, which is shown to be effective for gleaning insight in big data. Apache Hadoop and other technologies are emerging to support back-end concerns such as storage and processing, visualization-based data discovery tools focus on the front end of big data on helping businesses explore the data more easily and understand it more fully.

Keywords: Web crawlers data mining, Visualization, Visual Web Mining and Hadoop.

Introduction

Web mining is the application of data mining techniques to extract useful knowledge from web data that includes web documents, hyperlinks between documents, usage logs of web sites, etc. This technique enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet. As there is large amount of data present in web pages, the World Wide Web Data Mining may include content mining, hyperlink structure mining, and usage mining. All of these approaches attempt to extract knowledge from the Web, produce some useful results from the knowledge extracted, and apply the results to certain real-world problems. As there are large amount of users of internet and the web, there is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest organization to fill in their stocks appropriately for each month depending on the predictions they have laid out through this analysis of buying trends. With the lots amount of Reviews and FAQ’s are present, buying trends shown through web data mining can help you to make forecast on your inventories as well. Web mining is used in Information Retrieval IR systems, such as search engines, trafficking measures, were traffic is traced and monitored.

Web data mining technology is opening avenues on not just gathering data but it is also raising a lot of concerns related to data security. There is loads of personal information available on the internet and web data mining had helped to keep the idea of the need to secure that information at the forefront. As data is a collection of facts from the grids of web servers along the web usually of unorganized form in the digital universe. Around 90% of data present in today’s world are generated in last two years [2]. This large amount of the data available in the internet is generated either by individuals, groups or by the organization over a particular period. The volume of data becomes larger day by day as the usage of Internet and the web makes an interdisciplinary part of human activities. Rise of these data leads to a new technology such as big data that acts as a tool to process, manipulate and manage very large dataset along with the storage required.

Data can come from a variety of sources and in a variety of types that includes not only structured traditional relational data, but also semi-structured and unstructured data. Machine generated data such as click stream logs and email logs of unstructured data are larger in magnitude when compared with human generated data that cannot fit into a traditional data warehouses for further analysis. Numerous research works are carried out in web log mining, Apache hadoop the application software is discussed and reviewed below. Proposed the smart miner framework that extracts the user behavior from web logs with the use of hadoop Map reduce.

Analysis of web site usage data involves two significant challenges that firstly the volume of data, arising from the growth of the web, and secondly, the structural complexity of web sites. In this paper we apply Web Data Mining with Information Visualization techniques to the web domain in order to benefit from the power of both human visual perception and computing can be term as Visual Web Mining. In response to the two challenges of such large amount of Big Data, we propose a generic framework, where we apply Data Mining techniques to large web data sets and use Information Visualization methods on the results. The goal is to correlate the outcomes of mining Web Usage Logs and the extracted Web Structure, by visually superimposing the results. We propose several new information visualization diagrams, analyze their utility, and elaborate on the architecture of a prototype implementation. The Visual Web Mining (VWM) is as the application of Information Visualization techniques on results of Web Mining in order to further amplify the perception of extracted patterns, rules and regularities, or to visually explore new ones in web domain . In this paper we are introducing the web data mining technique and its implementation for handling the big amount web data with VWM and Apache hadoop Map reduce framework to handle big data.

DataMining Process:

Data mining is the process of non-trivial discovery of useful knowledge from implied, previously unknown, and potentially useful information from data in large databases. Hence, it is called as a core element in knowledge discovery, often used synonymously. The data is integrated and cleaned so that the relevant data is retrieved. Data mining presents discovered data that is not just clear to data mining analysts but also for domain experts who may use it to derive actionable recommendations. Successful applications of data mining include the analysis of genetic patterns, graph mining in finance, expert system to get proper advice and consumer behavior in marketing. Traditional data mining uses structured data stored in relational tables, spreadsheets, or flat files in the tabular form. With the growth of the Web and text documents, Web mining and text mining are becoming increasingly important and popular.

Web Mining

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services. Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined that are web content mining, web structure mining and web usage mining.

1. Web Content Mining

Web content mining is the process of extracting useful information from the contents of web documents and pages. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables shortly said as multimedia data. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on some other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP) is also going on [6]. While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited.

2. Web Structure Mining

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. The analyzed web resources contain the actual web site, the hyperlinks connecting these sites and the path that online users take on the web to reach a particular site [6]. A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

3. Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage 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 behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered: The web usage mining the content of the raw data for web usage mining on the one hand, and the expected knowledge to be derived from it on the other, pose a special challenge [7].

CHARACTERISTICS OF WEB DATA

Following are some of the characteristics of web data that makes it hard to mine:

1. Information on the Web is mainly in heterogeneous form. Due to the diverse authorship of Web pages, multiple pages may present the same or similar information using completely different words and/or formats. This makes integration of information from multiple pages a more challenging task.

2. There is a significant amount of information present on the Web is linked. Hyperlinks are exist among Web pages within a site and across different sites. Within a site, hyperlinks serve as information organization mechanisms but when it is present across different sites, it represents implicit conveyance of authority to the target pages. That represent, those pages that are linked or pointed to by many other pages are usually high quality pages or authoritative pages simply because many people trust them.

3. The information on the Web is noisy. This noise is comes from two main sources. First, a typical Web page contains many pieces of information, e.g., the main content of the page, navigation links, advertisements, copyright notices, privacy policies, etc. But for a particular application, only part of the information is useful. The rest is considered as a noise. To perform fine-grain Web information analysis and data mining, the noise should have to be removed. Second, because of the Web does not have quality control of information, i.e., one can write almost anything that one likes, a large amount of information on the Web is of low quality, erroneous, or even misleading.

4. The Web is also provides services. Most commercial Web sites allow people to perform useful operations at their sites, e.g., to purchase products, to pay bills, and to fill in forms, by which important personal information is going from one place to another on the web.

5. The Web is dynamic, as the information on the Web changes constantly. Keeping up with the change and monitoring the change are important issues for many applications.

6. The Web is a virtual society. The Web is about not only data, information and services, but also about interactions among people, organizations and automated systems. One can communicate with people on anywhere and anytime basis all over the world easily and instantly, and express one's views on anything in Internet forums, blogs and review sites.

All these characteristics present both challenges and opportunities for mining and discovery of information and useful knowledge from the Web. In this paper, we are mainly focusing on web data mining techniques, for mining of images, videos and audios, [8,9] as this large amount of data leads to the big data. Its todays need to mine and handle such a large amount of big data. In further section we are proposing the Visualization based method and one software application i.e. apache hadoop trying to handle such a large amount of web data.

TECHNIQUES FOR HANDLING BIG WEB DATA

Visual Web Mining Architecture

The architecture of implementing the visual web mining is shown in below Figure1. We target one or a group of websites for analysis. Input of the system consists of web pages and web server log files. Access to web log is done by the local file system, or by downloading it from a remote web server. A web robot (webbot) is used to retrieve the pages of the website . In parallel, Web Server Log files are downloaded and processed through a sessionizer and a LOGML file is generated.

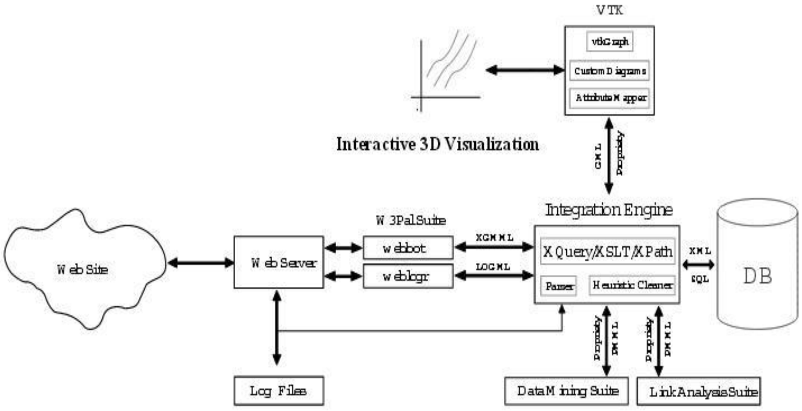

Figure 1: Sample implementation architecture of VWM

Figure 1: Sample implementation VWM architecture

- Data Mining Suite, Link Analysis Suite and User Pro filing/Modeling Suite need special data formats as input and produce output in propriety formats hence the Integration Engine is required to covert the data.

- The Integration Engine is a suite of programs for data preparation process like extracting, cleaning, transforming, integrating data. Finally, loading into database and later generating graphs in XGML.
- The engine uses XQuery, XSLT and Regular Expressions on both standard and propriety data formats.

- Much effort is put into enhancing performance of the transformation system in the Integration Engine and the database. By extracting user sessions from web logs, this yields results of roughly related to a specific user. User sessions are then converted into a special format for Sequence Mining using cSPADE.

- Outputs are frequent contiguous sequences with a given minimum support. These are imported into a database, and non-maximal frequent sequences are removed, i.e. we consider only the maximal frequent contiguous sequences. Later different queries are executed against this data according to some criterion, e.g. support of each pattern, length of patterns, etc.

Handling Big Web Data with Hadoop Map reduce

Big Data is an emerging growing dataset beyond the ability of a traditional database tool to handle. As the use of internet and the web is becoming a daily concern of many individuals, the growth of data is becoming

so high beyond the imagination of normal internet user. In addition, such a large amount of data leads to the big data problem. Hadoop rides the big data where the massive quantity of information is processed using cluster of commodity hardware. Web server logs are semi-structured files generated by the computer in large volume usually of flat text files. It is utilized efficiently by Map reduce as it process one line at a time. This paper performs the session identification in log files using Hadoop in a distributed cluster. Apache Hadoop Map reduce a data processing platform is used in pseudo distributed mode and in fully distributed mode. The framework effectively identifies the session utilized by the web surfer to recognize the unique users and pages accessed by the users. The identified session is analyzed in R to produce a statistical report based on total count of visit per day. The results are compared with non-hadoop approach a java environment, and it results in a better time efficiency, storage and processing speed of the proposed work .

Hadoop is a flexible infrastructure for large-scale computation and data processing on a network of commodity hardware. It allows applications to work with thousands of computational independent computers and petabytes of data. The main principle of hadoop is moving computations on the data rather the moving data for computation. Hadoop is used to breakdown the large number of input data into smaller chunks and each can be processed separately on different machines. To achieve parallel execution, Hadoop implements a

MapReduce programming model. Map Reduce is a java based distributed programming model consists of two phases: a massively parallel "Map" phase, followed by an aggregating "Reduce" phase [13]. Map Reduce is a programming model and an associated implementation for processing and generating large data sets. As it is work on java programming language which is best for to handling dynamic nature of the data and as most of the web data is in dynamic in nature, it suitable best for such a big amount of web data.

CONCLUSION:

In this paper, we have studied the web data mining technique, which is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. Now in todays advanced world, web becomes an important part of many of all organizations, businesspersons and daily individuals. As a web data is of very much different formats, we have studied the characteristics of web data. As it is very much important to mine particular data from web, we have studied two effective technique to mine this big data one with apache hadoop Map Reduce and second with visualization based technique called as Visual Web Mining (VWM).

REFERENCES

[1] Web Mining and Knowledge Discovery of Usage Patterns [Online] available:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.6743&rep=rep1&type=pdf.

[2]http://www.web-datamining.net/

[3]Murat Ali, Ismail Hakki Toroslu, "Smart Miner: A New Framework for mining Large Scale Web Usage
Data," WWW 2009, April 20-24. 2009 Madrid, Spain. ACM 978-1-60558-487-4/09/04.

[4]A. H. Youssefi, D. J. Duke, M. J. Zaki, and E. P. Glinert.

[5]Oren Etzioni. The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65-68, 1996.

[6]Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining - Concepts, Applications & Research Directions*", University of Minnesota, Minneapolis*, MN 55455, USA.

[7] Data Mining and Web Mining [Online] available: https://www.wiwi.hu-berlin.de/professuren-en/quantitativ/wi/forschung-en/dwm/standardseite-en.

[8]C. Djeraba, O. R. Zaiane, and S. Simoff. (eds.). *Mining Multimedia and Complex Data*. Springer, 2003.

[9]P. Perner. *Data Mining on Multimedia Data*. Springer, 2003.

[10]J. Punin and M. Krishnamoorthy. Wwwpal system- a system for analysis and synthesis of web pages. In Proc. WebNet, 1998.

[11]J. Punin, M. Krishnamoorthy, and M. J. Zaki. Logml: Log markup language for web usage mining. In WebKDD Workshop, ACM SIGKDD , pages 88–112, 2001.