ISSN:2231-0711
www.ijcset.net

Extraction of Pattern Sequence and Geographical Location of User using Web Analysis Approach

Monika Verma
RCET, Bhilai

***Abstract*- In the field of Web Mining more specifically Web Usage Mining an extraction of pattern sequence of navigations and session tracking of any user using Web Analysis Approach is difficult and along with the pattern sequence if we have to identify other details of any particular user like geographical location, IP address, name of web browser and operating system being used it become more tedious task. This paper introduces a new concept in which we would like to design an algorithm that will give detail about the pattern sequence of any user hitting a particular Web Applications along with other detail about that user like geographical location, IP address, name of web browser and operating system being used. This research work can be used in Market Analysis for the Web Application and its result will be helpful for different kind of advertisement for different segments of users for their different browser in Web Applications and Web Portals.**

***Keywords*- Web Mining, Web Usage Mining, Web Analysis Approach, Pattern Sequence, Session Tracking, Geographical Location, IP Address, Browser, Operating System, Web Application, Web Portal.**

I. INTRODUCTION

1. Data Mining

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful and a new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

1.2 Data Mining Process

Data Mining is a process of discovering various models, summaries, and derived values from a given collection of data. It is important to realize that the problem of discovering or estimating dependencies from data or discovering totally new data is only one part of the general experimental procedure used by scientists, engineers, and others who apply standard steps to draw conclusions from the data, refer figure 1. As we enter into the age of digital information, the problem of data overload looms ominously ahead. Our ability to analyze and understand massive data sets, as we call large data, is far behind our ability to gather and store the data. Large databases of digital information are ubiquitous.

Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls, and many more applications generate streams of digital records archived in huge business databases. Scientists are at the higher end of today's data-collection machinery, using data from different sources from remote-sensing platforms to microscope probing of cell details. Scientific instruments can easily generate terabytes of data in a short period of time and store them in the computer.

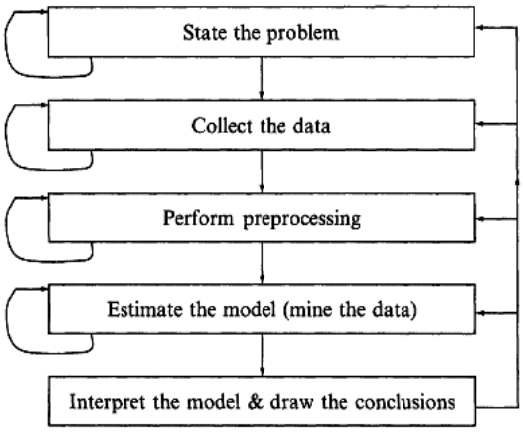

Figure 1: The data mining Process

1.3 Web Mining

Web mining [1] the application of machine learning (data mining) techniques to web-based data for the purpose of learning or extracting knowledge. Web mining encompasses a wide variety technique, including soft computing [PTMOO]. Web mining methodologies can generally be classified into one of three distinct categories: web usage mining, web structure mining, and web content mining, a survey of techniques used in these areas. In *web* usage mining the goal is to examine web page usage patterns in order to learn about a web system's users or the relationships between the documents. For

example, the tool presented by asseglia, association rules from web access logs, which store the identity of pages accessed by users along with other information such as when the pages were accessed and by whom; these logs are the focus of the data mining effort, rather than the actual web pages themselves. Rules created by their method could include, for example, "70% of the users that visited page A also visited page B." Similarly, the method of Nasraoui et *al.* examines web access logs. The method employed in that paper is to perform a hierarchical clustering in order to determine the usage patterns of different groups of users. Beeferman and Berger [BBOO] described a process they developed which determines topics related to a user query using click-through logs and agglomerative clustering of bipartite graphs.

1.3.1 Web Usage Mining

With the continued growth and proliferation of e-commerce[2], Web services, and Web-based information systems, the volumes of **clickstream** and **user data** collected by Web-based organizations in their daily operations has reached astronomical proportions. Analyzing such data can help these organizations determine the life-time value of clients, design cross-marketing strategies across products and services, evaluate the effectiveness of pro-motional campaigns, optimize the functionality of Web-based applications, provide more personalized content to visitors, and find the most effective logical structure for their Web space.

Web usage mining refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests.

1.3.2 Taxonomy of Web Mining

Taxonomy of Web mining along its two primary dimensions, namely Web content mining and Web usage mining. We also describe and categorize some of the recent work and the related tools or techniques in each area. This taxonomy is depicted in Figure 2.

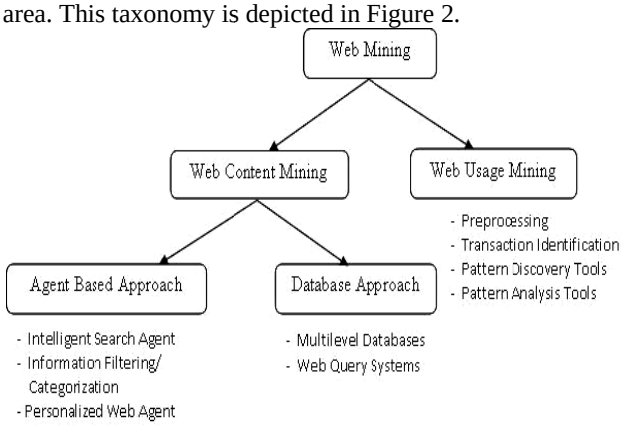

Figure 2: Taxonomy of Web Mining

2 PROPOSED METHODOLOGY

2.1 The following scripts make use of the MaxMind GeoIP Javascript, which is freely usable as long as you adhere to the Terms of Use below.

These scripts may be used to control access to your web site based upon where the user is located. This process has obvious limitations: in order for it to work, the user must have a javascript-enabled browser. That's not to say this is not a reasonably good control mechanism for your web site; but this is by no means a 100% solution to filtering traffic.

2.1.1 GeoIP Get Directions from Google by Current Location

Now here's an interesting concept for businesses (for example) who want their web site visitors to be able to get directions to their office easily: a script that detects the visitor's whereabouts, and then generates a Google Map with step-by-step directions to the destination (presumably, the place of business). In this example, a user will see a button that says "Get Directions". If clicked, a map is generated by Google Maps which gives the user step-by-step driving directions to the destination you specified in the configuration section--the CN Tower in Toronto in this example. The user never need go through the standard "get directions" procedure; GeoIP is used to determine where the user is driving from.

2.1.2 GeoIP JavaScript Web Service

GeoIP JavaScript is a service offered by MaxMind to return the Country, Region, City, Latitude, and Longitude for your web visitors [17]. It uses JavaScript and is very easy to program, works on both static and dynamically served web pages. If you only need to display the country, use GeoIP Country JavaScript service. In order to use this JavaScript on our website, a link back to the www.maxmind.com website should be provided, or a JavaScript attribution-free license can be purchased for $250/year.

Example:

Country Code: IN
Country Name: India
Region:
Region Name:
City:
Postal Code:
Latitude: 20.0000
Longitude: 77.0000

2.1.3 Source code

For example to find the country code the snap of the source code for this is

```
<script language="JavaScript" src=
"http://j.maxmind.com/app/geoip.js">
</script>
<br>Country Code:
<script language= "JavaScript">
document.write(geoip_country_code());
</script>
```

2.1.4 MaxMind JavaScript Web Service

MaxMind JavaScript is a service offered by MaxMind to return the country of your web visitors. It uses JavaScript and is very easy to program, works on both static and dynamically served web pages [18].

In order to use this JavaScript on your website, a link back to the www.maxmind.com website should be provided, or a JavaScript attribution-free license can be purchased for $250/year. The uptime for the JavaScript service is about 99.95%.

Example:

Country Code: IN

Country Name: India

2.1.5 GeoIP Geo-location Products

MaxMind GeoIP products deliver information on the geographic location and connection type of Internet visitors. APIs and associated documentation for databases can be found in the GeoIP Support Center. We can purchase multiple databases with one order by adding them to the shopping cart before checking out.

2.1.6 IP Address Locator

MaxMind GeoIP City/ISP/Organization Edition Results Table 1:

Table 1: MaxMind Geo IP

Hostname
Country Code
Country Name
Region
Region Name
City
Postal Code
ISP
Organization
Metro Code
Area Code

These results were generated with the Perl API and the commercial GeoIP City, GeoIP ISP, and GeoIP Organization databases.

3 EXPERIMENTAL RESULTS:

My research work is divided into 3 parts that is:

1. To find out the different areas concerning to the different users like
   - I. Date: The date of visit.
   - II. Page: The page that is visited.
   - III. Client Info: The information of the client.
   - IV. Client IP: IP address of the client.
   - V. Platform: OS that is used by the Users.
   - VI. Web Browser: The type of Browser used by the user.
2. To find out the numbers of hit for the particular page separately.
3. To find out the browsing patterns of different users of the web sites.

4 CONCLUSION:

In this paper we learned about web usage mining and the use of MaxMind GeoIP Javascript in a web application in which the first part of this study is to find out the date of the page accesses, the browser being used by the users, the operating system in which the browse is used, the page that is accessed and at last the total number of hits for the particular web page inside the web application can also be counted and display separately, is accomplished successfully using the java server pages technology.

REFERENCES:

[1] http://www.worldscibooks.com/etextbook/5832/5832_chap1.pdf
[2] http://maya.cs.depaul.edu/~mobasher/papers/12-web-usage-mining.pdf
[3] Olfa Nasraoui, Maha Soliman, Esin Saka, Antonio Badia and Richard Germain "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 2, February 2008.
[4] M.A. Maloof and R.S. Michalski, "Learning Evolving Concepts Using Partial Memory Approach," Working Notes AAAI Fall Symp. Active Learning 1995, pp. 70-73, 1995.
[5] M.A. Maloof and R.S. Michalski, "Selecting Examples for Partial Memory Learning," Machine Learning, vol. 41, no. 11, 2000.
[7] B. Mobasher, H. Dai, T. Luo, Y. Sung, and J. Zhu, "Integrating Web Usage and Content Mining for More Effective Personalization,"Proc. Int'l Conf. e-Commerce and Web Technologies (ECWeb '00), Sept. 2000.
[8] USER PROFILING: WEB USAGE MINING, *Miha Grcar*.
[9] Netcraft. Net craft Web Server Survey. http://www.netcraft.com/survey/archive.html
[10] http://www.w3.org/TR/WDlogfile.html
[11] P. Baldi, P. Frasconi, P. Smyth. "Modelling the Internet and the Web". pp. 171–209. ISBN: 0-470-84906-1. 2003.
[12] Maximilian Viermetz , Carsten Stolz K , Vassil Gedov, Michal Skubacz "Relevance and Impact of Tabbed Browsing Behavior on Web Usage Mining", Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence,2006
[13] H. Hannah Inbarani, K. Thangavel, A. Pethalakshmi, "Rough set based Feature Selection for Web Usage Mining "International Conference on Computational Intelligence and Multimedia Applications 2007.
[14] Z. Pawlak. "Rough Sets". International Journal of Computer and Information Sciences, 11(5): 341-356, 1982.
[15] Z. Pawlak. "Rough Sets: Theoritical Aspects and Reasoning about Data". Kluwer Academic Publishers, 1991.
[16] V. Estivill-Castro and J. Yang." Fast and robust general purpose clustering algorithms". In Pacific Rim International Conference on Artificial Intelligence, pages 208–218, 2000
[17] http://www.maxmind.com/app/javascript_city
[18] http://www.maxmind.com/app/javascript

Monika Verma received her Bachelor of Engineering in Computer Science & Engineering with honors from MP Christian College of Engineering and Technology, Bhilai (Chhattisgarh State) in year 2010. She has secured eighth position in university topper ranking over all engineering colleges in the state under Chhattisgarh Swami Vivekanand Technical University, Bhilai (Chhattisgarh State). She is pursuing M.Tech. 4th semester with specialization in CSE at department of CSE from Rungta College of Engineering and Technology, Bhilai (Chhattisgarh State) India.