WEB USAGE MINING: A COMPREHENSIVE STUDY ON INFORMATION EXTRACTION FROM DATA SOURCES, APPLICATIONS AND TECHNIQUES ON WEB

Prof. Mrunali Metri
Asst. Professor, NCRD's Sterling Institute of Management Studies, Nerul, Navi Mumbai.
Email: mrunalimetri@gmail.com

Prof. Rahul Wantmure
Asst. Professor, NCRD's Sterling Institute of Management Studies, Nerul, Navi Mumbai.
Email: rahul_wan2003@yahoo.co.in

Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by web servers. Web Mining enables to find out the relevant results from the web and is used to extract meaningful information from the discovery patterns kept back in the servers. Web Usage Mining is a type of Web Mining which mines the information of access routes/manners of users visiting the web sites. In this paper, we revisit, explore and discuss about the information extraction from Data Sources, Applications and techniques on web like Web Usage Mining.

Key words: *Web Usage Mining, Web Mining, Data Sources, Pattern Discovery, Web Log.*

INTRODUCTION:

The comfort and speed with which business transactions can be carried out over the web has been a key driving force in the explosive growth of electronic commerce. Specifically, e-commerce activity that involves the end user is undergoing a significant revolution. The ability to track users browsing behaviour down to individual mouse clicks has brought the vendor and end customer closer than ever before.

More precisely, Web Content Mining is that part of Web Mining which focuses on the raw information available in web pages; typical applications are content-based categorization and content-based ranking of web pages. Web Structure Mining is that part of Web Mining which focuses on the structure of web sites whereas, Web Usage Mining comprises of Web Mining which works with the extraction of knowledge from server log files.

Source data mainly consist of the (textual) logs that are collected when users access web servers, typical applications are those based on user modelling techniques, such as web

personalization, adaptive web sites, and user modelling. The recent years have seen the flourishing of research in the area of Web Mining and specifically of Web Usage Mining.

The scenario described above is one of many possible applications of Web Usage mining. It is the process of using data mining techniques to discover usage patterns with several sources of data from web data, targeted towards various applications. This paper provides a comprehensive study of information extraction from Data Sources on web in Section 1. Section 2 and 3 describes Applications and Techniques on web including different resources available to extract data from server such as server level data, client level data and proxy server data. Section 4 concludes the paper.

Web Usage Mining helps companies to create productive information relative to the future of their business growth and structuring their business. Some of this information can be derived from the cumulative information of lifetime user value, cross marketing strategies pertaining to various products and promotions of effective campaigning of the products. The accumulated usage data provides the firms with the ability to generate intended results effectively thereby having increase in their sales. By having web usage, the marketing skills of the companies can be developed and recreated thus can be at par with the competitors. Thus they can promote their activities, services and/or products at a higher pace.

WEB DATA:

Usage Mining is considerable to businesses using online marketing and those e-businesses that are based on the traffic provided through various search engines. The use of this type of Web Mining helps to collect the important information from the web sites where customers visit repeatedly. This helps to get complete analysis of a company's workflow. E-businesses rely on this information to assist the company to find an effective web server on which promotional activities, of its products and services, can be conducted.

In Web Mining, data can be gathered from different sources like, server-side, client-side, proxy servers, or obtained from an organization's database. One of the key steps in knowledge discovery in databases [2] is to create a suitable target data set for the Data Mining tasks. Whatever data collected together differ from each other in terms of the location, availability of data, the segment of population from which the data was gathered and its method of implementation.

There are many kinds of data that can be used in Web Mining. This paper classifies such data into the following types:

Content: The real data in the web pages, means, Web Content Mining is the extraction and integration of important data, information and knowledge from web page content.

Structure: It is the format that describes the organization of the content. According to the type of web structural data, web structure mining can be divided into two kinds:

a) Extracting patterns from hyperlinks in the web: (a hyperlink is a structural component that connects the web page to a different location).

b) Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.[3]

WORKING OF WEB USAGE MINING?

Web usage mining is accomplished by reporting visitors' traffic information where users are visited based on web server log files and other source of traffic data. Using these web server log files one can find the interest of user in particular area, i.e. how much traffic they are encountering, which type of data are they looking for, how many requests fail, and what kind of errors are being generated, etc. Thus, this information is handled by webmasters and system administrators, for finding various information about the traffic on web related to a particular site. After the Web traffic data is obtained, it may be combined with other relational databases, over which the data mining techniques are implemented. Through some data mining techniques such as association rules, path analysis, sequential analysis, clustering and classification, visitors' behavior patterns are found and interpreted.

However, web server log files can also record and trace the visitors' on-line behaviors. For example, after some basic traffic analysis, the log files can help us answer questions such as - from which search engine are visitors coming? What pages are the most and least popular? Which browsers and operating systems are most commonly used by visitors? etc.

DATA SOURCES

Web Usage Mining applications are based on data collected from three main sources [4]:

1. web servers,
2. proxy servers, and
3. web clients.

The usage data collected at the different sources will represent the navigation patterns of different segments of the overall web traffic, ranging from single-user, and single-site web browsing behaviour to multi-user and multi-site access patterns.

The Server Side: Web servers are surely the richest and the most common source of data. They can collect large amounts of information from the log files and in the log files of the databases they use. These logs usually contain basic information e.g.: name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented in standard format. When exploiting log information from web servers, the major issue is the identification of users' sessions.

Apart from web logs, users' behaviour can also be tracked down on the server side by means of TCP/IP packet sniffers. Even in this case the identification of users' sessions is still an issue, but the use of packet sniffers provides some advantages. In fact:

- Data are collected in real time
- Information coming from different web servers can be easily merged together into a unique log
- The use of special buttons can be detected so to collect information usually unavailable in log files

Packet sniffers are rarely used in practice because of scalability issue on web servers with high traffic, and the impossibility to access encrypted packets. Probably, the best approach for tracking web usage consists of directly accessing the server application layer.

The Proxy Side: Many internet service providers (ISPs) give to their customer proxy server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers. The filtered data can be collected from proxy side more effectively and promptly, rather than from direct server available for the particular URL.

For Example: Suppose we have to declare the results of various departments, under a website from a direct server, where we may declare multiple links for multiple departments. As we have various links on the same server, it may fetch more clicks (requests from clients) that may cause the server to deliver the result with a time delay. But by using the proxy server we may reduce the work process. The requests can be redirected to a dedicated proxy where we can get the results segregated without overloading any of the servers with much faster rate. Thus, the filtered results can be collected from proxy side more effectively and promptly. Thus, it becomes useful to get the data for data mining.

The Client Side: Usage data can be tracked also on the client side by using JavaScript, java applets, or even modified browsers. These techniques avoid the problems of users' sessions' identification and the problems caused by caching (like the use of the back button). In addition, they provide detailed information about actual user behaviours. However, these approaches rely heavily on the users' cooperation and rise many issues concerning the privacy laws, which are quite strict. There are many ways to perform web mining at the client side - one of the ways is the use of cookies.

The cookies are the data stored by the server at the client side such as user Id and preferences at the client end. With cookies, web servers can store short strings of information at the client. Though, the cookies help in storing the data while maintaining the state of the session, they cannot provide the functionality of cross browsing on the same machine. This becomes a challenge for web mining.

TECHNIQUES

Most of the commercial applications of Web Usage Mining exploit consolidated statistical analysis techniques. In contrast, research in this area is mainly focused on the development of knowledge discovery techniques specifically designed for the analysis of web usage data. Most of this research effort focuses on three main paradigms: association rules, sequential patterns, and clustering.

Association Rules are probably the most elementary data mining technique and, at the same time, the most used technique in Web Usage Mining. When applied to Web Usage Mining, association rules are used to find associations among web pages that frequently appear together in users' sessions.

Suppose we consider the example of a shopping cart, where three web pages have to be visited sequentially, and if a visitor visits the first and the second page then it is considered that user has obviously visited the third page as well. This proposes and evaluates some interesting measures to evaluate the association rules mined from web usage data.

Sequential Patterns are used to discover frequent sub-sequences among large amount of sequential data. In web usage mining, sequential patterns are exploited to find sequential navigation patterns that appear in users' sessions frequently.

The distinctive quality of sequential pattern has the form: Considering, the most of the users who first visited first page of our previous example and then visited second page, in the same session, have also accessed the third page. Sequential patterns might appear syntactically similar to association rules; in fact algorithms to extract association rules can also be used for sequential pattern mining. There can be two approaches of algorithms that are used to derive sequential patterns: one based on association rule mining and the other based on the use of methods such as tree-structures, data projection techniques, etc.

Clustering techniques look for groups of similar items among large amount of data based on a general idea of distance function which computes the similarity between groups. Clustering has been widely used in Web Usage Mining to group together similar sessions. The focus of web usage mining should be shifted from single user sessions to group of user sessions. It also proposes similarity graph in conjunction with the time spent on web pages to estimate group similarity in concept-based clustering. Through the various feedbacks acquired we can improve the results of Clustering using Genetic Algorithms.

APPLICATIONS

1. Navigations: The intention of Web Usage Mining is to gather some fascinating information about user's navigation patterns. This information can be exploited later to improve the web site from the users' point of view. The results produced by the mining of web logs can be used for various purposes viz,

- to personalize the delivery of web content
- to improve user navigation through caching and prefetching
- to improve design of commercial web sites
- to improve the customer satisfaction
- to have dynamic content on the website which can fetch the data from instant data mining

2. Personalized web user experience: Web Usage Mining techniques can be used to provide personalized web user experience. For instance, it is possible to anticipate, in real time, the user behaviour by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common

application; their aim is to recommend interesting links to products which could be interesting to users. Personalized Site Maps are an example of recommendation system for links. The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. Typically, Web Usage Mining can be used to develop proper prefetching and caching strategies so as to reduce the server response time.

3. Support to the Design the Web site: Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications and suggests proper modifications to web sites. The content and the structure of the web site can be dynamically reorganized according to the data mined from the users' behaviour.

4. E-commerce: Mining business intelligence from web usage data is artistically important for e-commerce web-based companies.

5. Customer Relationship Management (CRM): can have an effective advantage from the use of Web Usage Mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales and customer departure.

CONCLUSION:

Recently, Web Usage Mining has been gaining a lot of attention because of its potential commercial benefits. Web usage mining is the third category in web mining. We have discussed about the information extraction from Data Sources, Applications and Techniques**,** for web usage mining which allow companies to produce productive information pertaining to the future of their business function ability.

Web Usage Mining deals with the extraction of interesting knowledge from users logging information available on web servers which mines the information of access routes/manners of users visiting the web sites. Web Content Mining focuses on the raw information available in web pages. Some of this information can be derived from the collective information of lifetime user value, product cross marketing strategies and promotional campaign effectiveness.

Mining business intelligence from web usage data is artistically important for e-commerce web-based companies as well as for Customer Relationship Management. Therefore, it is

easily determined that Web Usage Mining has valuable uses for the marketing of businesses and has a direct impact to the success of their promotional strategies and internet traffic.

REFERENCE:

[1] http://www.web-datamining.net/usage/

[2] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, In Proc. ACM KDD, 1994, "data mining to knowledge discovery: An overview".

[3] http://en.wikipedia.org/wiki/Web_mining

[4] www.researchgate.net/publication

[5] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan (2000). Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations.

[6] Kurt Fenstermacher and Mark Ginsburg, MIS Department, University of Arizona, Eller College of Management-" Client-Side Monitoring for Web Mining".

[7] Damiani, E., Oliboni, B., Quintarelli, E., & Tanca, L. (2001). "Modeling Users' Navigation History" Paper presented at the International Joint Conference on Artificial Intelligence (IJCAI), Seattle, WA.

[8] Ginsburg, M., Therani Madhusudan.(2001) - "Pattern Acquisition to Improve Organizational Knowledge Management"- Paper presented at the AMCIS2001, Boston, MA.

[9] Federico Michele Facca and Pier Luca Lanzi – "Recent Developments in Web Usage Mining Research".