APPLICATIONS AND TECHNIQUES OF WEB USAGE MINING

Kanwarpal Singh
Assistant Professor
PG Department of Computer Science
BBKDAV College, Amritsar

ABSTRACT
Web mining refers to the application of data mining techniques to the World Wide Web. Web usage mining is the process of extracting useful information from web server logs based on the browsing and access patterns of the users. The information is especially valuable for business sites in order to achieve improved customer satisfaction. Based on the user's needs, Web Usage mining discovers interesting usage patterns from web data in order to understand and better serve the needs of the web based application. In this paper, we present application areas and techniques of Web Usage Mining such as Education, Health, Human-computer interaction, and Social media.

Keywords: Web mining; clustering; classification;

1. INTRODUCTION

Web usage mining, a subset of Data Mining, is basically the extraction of various types of interesting data that is readily available and accessible in the ocean of huge web pages, Internet- or formally known as World Wide Web (WWW). Being one of the applications of data mining technique, it has helped to analyze user activities on different web pages and track them over a period of time. Basically, Web Usage Mining can be divided into two major subcategories based on web usage data [1].

Web Content Data are HTML, web pages, images audio-video, etc. Though it may differ from browser to browser the common basic layout/structure would be the same everywhere. XML and dynamic server pages like JSP, PHP, etc. are also various forms of web content data.

Web Structure Data is content arranged according to HTML tags (which are known as intrapage structure information). The web pages usually have hyperlinks that connect the main webpage to the sub-web pages. This is called Inter-page structure information. So basically relationship/links describing the connection between webpage's is web structure data.

Web Usage Data involves log data which is collected by the main above two mentioned sources. Log files are created when a user/customer interacts with a web page. The data in this type can be mainly categorized into three types based on the source it comes from: Server-side, Client-side, Proxy side.

There are other additional data sources also which include cookies, demographics, etc.

2. LITERATURE SURVEY

With the rapid growth of the World Wide Web, the study of knowledge discovery in web, modeling and predicting the user's access on a web site has become very important.

From the administration, business and application point of view, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e-Business, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, Digital Libraries, and so on. Web is becoming the necessity of the businesses and organizations because of its demand from the clients. Since the web technology largely feeds on ideas and knowledge rather than being dependent on fixed assets, it gave birth to new companies such as Yahoo, Google, Netscape, e-Bay, e-Trade, Expedia, Amazon and so on. With the large number of companies using the Internet to distribute and collect information, knowledge discovery on the web has become an important research area. With the explosive growth of information sources available on the World Wide Web, it has become necessary for organizations to discover the usage patterns and analyze the discovered patterns to gain an edge over competitors.

[2] Jespersen et al proposed a hybrid approach for analyzing the visitor click stream sequences. A combination of hypertext probabilistic grammar and click fact table approach is used to mine Web logs, which could be also used for general sequence mining tasks. Mobasher et al proposed the web personalization system, which consists of offline tasks related to the mining if usage data and online process of automatic Web page customization based on the knowledge discovered. LOGSOM (LOGSOM, a system that utilizes Kohonen's self-organizing map (SOM) to organize web pages into a two-dimensional map) proposed by Smith et al, utilizes a self-organizing map based solely on the users' navigation behavior, rather than the content of the web pages. LumberJack proposed by Chi et al builds up user profiles by combining both clustering of user sessions and traditional statistical traffic analysis using k–means algorithm. Joshi et al used relational online analytical processing approach

for creating a Web log warehouse using access logs and mined logs. Global Internet Usage Average Usage shows the current usage around the globe and in United States.

3. APPLICATIONS OF WEB USAGE MINING

Each of the applications can benefit from patterns that are ranked by subjective interesting. The general goal of Web Usage Mining is to gather interesting information about users navigation patterns (i.e., to characterize web users). This information can be exploited later to improve the web site from the users' viewpoint. The results produced by the mining of web logs can used for various purposes: (i) to personalize the delivery of web content; (ii) to improve user navigation through prefetching and caching; (iii) to improve web design; or in e-commerce sites (iv) to improve the customer satisfaction [2][3].

Web usage mining is used in the following areas:

1. Personalization of Web Content: 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 behavior by comparing the current navigation pattern with typical patterns which were extracted from past web log. Web personalization may be defined as catering to the user's need-based upon its navigational behavior tracking and their interests. Web Personalization includes recommender systems, check-box customization, etc. Recommender systems are popular and are used by many companies.

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

3. Prefetching and Catching: The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. Prefetching basically means loading of data before it is required to decrease the time waiting for that data hence the term 'prefetch'. All the results which we get from web usage mining can be used to produce prefetching and caching strategies which in turn can highly reduce the server response time.

4. Education: In education, data preparation will transform Web log files and profiles into data with the appropriate format. Pattern discovery will use a data mining technique, such as clustering, sequential pattern and association rule mining. Finally, recommendation will use the discovered patterns to provide personalized links or contents.

5. Health informatics: Internet can serve as the backbone for implementing supply chain solutions to add value to health care providers, their suppliers, and their patients. Health care information systems in some hospitals and clinics have been linked together with a local area network or a wide area network, network based health care systems have not been popular until the advent of the Internet. The three primary Internet applications that the healthcare industry uses, to varying degrees, are the Internet, intranets, and extranets. Doctors can use the Internet to do more than download information and communicate with other providers; it can also be used to send complex medical files across the Web.

6. Human-computer interaction: Web usage mining is a kind of web mining, which exploits data mining techniques to discover valuable information from navigation behavior of World Wide Web users. Web usage mining (WUM) is a new research area which can be defined as a process
of applying data mining techniques to discover interesting patterns from web usage data. Web usage mining provides information for better understanding of server needs and web domain design requirements of web-based applications.

7. Social media: Web usage mining also plays an important role in social networks analysis. Social network analysis is finding the communities embedded in the social network datasets, and moreover, analyzing the evolutions of the communities in dynamic networks. The evolution pattern as one kind of temporal analysis aspect sometimes could provide us an interesting insight from the perspective of social behavior. Recently, a considerable amount researches have been done on this topic. In the field of social network, Web community is also used to mean a set of users having similar interests. In Social networking products are flourishing. Sites such as twitter, Facebook, and Instagram attract millions of visitors a day, approaching the traffic of Web search sites[10].

4. TECHNIQUES IN WEB USAGE MINING

Web usage mining is the application of data mining techniques to discover usage pattern from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Data preprocessing consists of data filtering, user identification, session/transaction identification, and topology extraction [4][5].

Following are the techniques in Web Usage Mining:

1. Association Rules: In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. Association rule mining has been well studied in Data Mining, especially for basket transaction data analysis.

Association Rules help in the reconstruction of websites using the access logs. Access logs generally contain information about requests which are approaching the webserver. The major drawback of this technique is that having so many sets of rules produced together may result in some of the rules being completely inconsequential. They may not be used for future use too.

2. Classification: Classification is mainly to map a particular record to multiple predefined classes. The main target here in web usage mining is to develop that kind of profile of users/customers that are associated with a particular class/category. For this exact thing, one requires to extract the best features that will be best suitable for the associated class. Classification can be implemented by various algorithms – some of them include- Support vector machines, K-Nearest Neighbors, Logistic Regression, Decision Trees, etc. For example, having a track record of data of customers regarding their purchase history in the last 6 months the customer can be classified into frequent and non-frequent classes/categories. There can be multiclass also in other cases too [6].

3. Clustering: Clustering is a technique to group together a set of things having similar features/traits. There are mainly two types of clusters- the first one is the usage cluster and the second one is the page cluster. The clustering of pages can be readily performed based on the usage data. In usage-based clustering, items that are commonly accessed /purchased together can be automatically organized into groups. The clustering of users tends to establish groups of users exhibiting similar browsing patterns. In page clustering, the basic concept is to get information quickly over the web pages.

4. Sequential Patterns: The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. A new algorithm MiDAS (Mining Internet data for Associative Sequences) for discovering sequential patterns from web log files has been proposed that provides behavioral marketing intelligence for e-commerce scenarios [7].

5. Dependency modeling: Dependency modeling is another useful pattern discovery task in web mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the web domain. As an example, one may be interested to build a model representing the different stages a visitor undergoes while shopping in an online store based on the actions chosen (ie, from a casual visitor to a serious potential buyer.

6. Deviation/Outlier Detection: It contains techniques aimed at detecting unusual changes in the data relatively to the expected values. Such techniques are useful, for example, in fraud detection, where the inconsistent use of credit cards can identify situations where a card is stolen. The inconsistent use of credit card could be noted if there were transactions performed in different geographic locations within a given time window.

7. Pattern analysis: Pattern analysis is the last step in the overall Web Usage mining process. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL [8][9]. Another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

5. CONCLUSION

This paper has attempted to provide an up-to-date survey of the rapidly growing area of Web usage mining, which is the demand of current technology. In this paper a general overview of Web usage mining Applications and techniques, Web usage mining is used in many areas such as Education, E-Commerce, Health Informatics and so on. The major classes of recommendation services are based on the discovery of navigational patterns of users. The main techniques for pattern discovery are association rules, Classification, Clustering, pattern analysis etc.

REFERENCES

[1] Ajit Abhraham, Vitorino Ramos, Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming, to appear in CEC´03 - Congress on Evolutionary Computation, IEEE Press, Canberra, Australia, 8-12 Dec. 2003.

[2] Mitharam, M.D.: Preprocessing in Web Usage mining. International Journal of Scientific & Engineering Research 3(2), 1 (2012) ISSN 2229-5518

[3] Sharma, A.: Web Usage Mining: Data Preprocessing, Pattern Discovery and Pattern Analysis on the RIT Web Data

[4] Chaudhary, K., Gupta, S.K.: Web Usage Mining Tools & Techniques: A Survey. International Journal of Scientific & Engineering Research 4(6), 1762 (2013) ISSN 2229-5518

[5] F. Johnson and S. K. Gupta, "Web Content Mining Techniques: A Survey", *International journal of computer applications (0975-888)*, vol. 47, no. 11, June 2012.

[6] Priyanka Bharti and Sona Malhotra, "A Review Paper on Web Usage Mining and future request prediction", *Special Issue (ICFTEM-2014)*, vol. 33, pp. 33-37, May 2014, ISSN 0973-4414.

[7] Mitali Srivastava, Rakhi Garg and P. K. Mishra, "Preprocessing Techniques in Web Usage Mining: A survey", *International Journal of Computer Applications (0975 – 8887)*, vol. 97, no. 18, July 2014.

[8] Smith K.A. and Ng A., Web page clustering using a self-organizing map of user navigation patterns, Decision Support Systems, Volume 35 , Issue 2 (May 2003) Special issue: Web data mining, Pages: 245 – 256.

[9] I. V. Cadez, D. Heckerman, C. Meek, P. Smyth, S. White, Model-based clustering and visualization of navigation patterns on a Web Site, *Journal of Data Mining and Knowledge Discovery*, 7(4), 2003. (extended version of ACM SIGKDD 2000 conference paper).

[10] M. Rao, M. Kumari, and K. Raju, Understanding User Behavior using Web Usage Mining', international Journal of Computer Applications, 1(7), 2010, 55–61.