Asian Journal of Science and Applied Technology (AJSAT)
Vol.3.No.1 2015 pp.70-73
available at: www.goniv.com
Paper Received :08-03-2015
Paper Accepted:20-03-2015
Paper Reviewed by: 1. R. Venkatakrishnan 2. R. Marimuthu
Editor : Prof. P.Muthukumar

WEB MINING USING RESEARCH ISSUES OF DATA MINING TECHNIQUES

T.Balasubramanian
Sri Vidya Mandir Arts And Science College
Uthangarai, Krishnagiri (Dt).
Tamilnadu, India
balaeswar123@gmail.com

ABSTRACT

Web mining is the application of the data miningwhich is useful to extract the knowledge. Web mining has been explored to different techniques have been proposed for the variety of the application.The World Wide Web is huge, unstructured, universal and heterogeneous. In recent years the growth of the World Wide Web exceeded all expectations.Web usage mining is one of thetechnique of web mining is very useful to discover knowledge from secondary data obtained from the interaction from users with the web.The paper discusses about web usage mining involves the automatic discovery of user access patterns from one or more Web servers. This paper provides a survey and analysis of current Web usage mining systems and technologies.Today
lot of businesses are happening on World Wide Web (WWW), it is very important for the website owner to provide a better platform to attract more customers for their site.

The purpose of Web mining is to develop methods and systems for discovering models of objects and processes on the World Wide Web and for web-based systems that show adaptive performance. Web Mining integrates three parent areas: Data Mining, Internet technology and World Wide Web, and for the more recent Semantic Web**.**

Keywords: Issues of data mining, Web mining, Web mining types

1. INTRODUCTION

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. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.Web is a collection of billion of documents. The web is very enormous, diverse, flexible, and dynamic.

Web mining is an important area in data mining where we extract the interesting patterns from the contents.

“Web Mining is the application of data mining techniques to the content, structure and usage of Web resources.Generally three kinds of information are handled in web site namely 1. Content 2.Structure 3.Log data.

- ✓ Content Mining - Analyses the content of Web resources. Mainly based on text mining techniques, but extensions to multimedia content is beginning to emerge in the research.
- ✓ Structure Mining - Analyses the hyperlink structure between Web pages.

- ✓ Usage Mining - Analyses the user’s clicks from Web server logs.

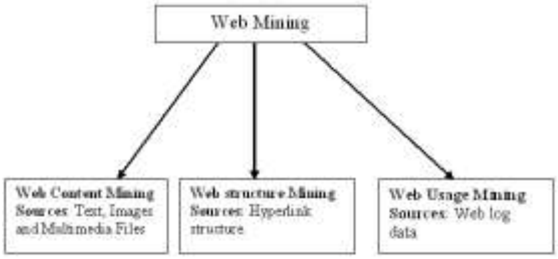
Web Mining
Web Content Mining Sources: Text, Images and Multimedia Files
Web structure Mining Sources: Hyperlink structure
Web Usage Mining Sources: Web log data

Fig.1. The types and sources of Web mining

1.1. Web Usage Mining

Web usage mining is the type of Web mining activitythat involves the automatic discovery of user access patterns from one or more Web servers. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by Web servers and collected in server access logs.Web usage mining has become a necessity task in order to provide web administrators with meaningful information about users and usage patterns for improving quality of web information and service performance. Successful websites may be those that are customized to meet user preferences both in the presentation of information and in relevance of the content that best fits the user.

1.2. Web Structure Mining

Web structure mining can be used by search engines to rank the relevancy between websites classifying them according to their similarity and relationship between them.

2. ORIGIN OF WEB MINING

Web mining techniques are the result of long process of research and product development. This evolution began when the amount of data kept in computer files and databases is growing at a phenomenal rate. At the same time users of these data are expecting more sophisticated information from them .A marketing manager is no longer satisfied with the simple listing of marketing contacts but wants detailed information about customers‟ past purchases as well as prediction of future purchases. Simple structured / query language queries are not adequate to support increased demands for information. Data mining steps is to solve these needs. Data mining is defined as finding hidden information in a database alternatively it has been called exploratory data analysis, data driven discovery, and deductive learning.

3. WEB PERSONALIZATION

Personalization includes using technology to accommodate the differences between individuals. Once restricted mainly to the Web, it is becoming a factor in education, health care (i.e. personalized medicine), television, and in both "business to business" and "business to consumer" settings. Social Network websites use personal data to provide relevant advertisements for their users. Websites like Google and Facebook are using account information to give better services. There are three categories of personalization:

- ✓ Profile / Group based
- ✓ Behavior based (also known as Wisdom of the Crowds)
- ✓ Collaboration based

Web personalization models include rules-based filtering, based on "if this, then that" rules processing, and collaborative filtering, which serves relevant material to customers by combining their own personal preferences with the preferences of like-minded others. Collaborative filtering works well for books, music, video, etc. However, it does not work well for a number of categories such as apparel, jewelry, cosmetics, etc. Recently, another method, "Prediction Based on Benefit", has been proposed for products with complex attributes such as apparel.

The Web Personalization process divides in to four distinct phases.

Collection of Web data–In this, implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules.

Explicit data usually comes from registration forms and rating questionnaires. In some cases, Web content, structure, and application data can be added as additional sources of data, to shed more light on the next stages.

Preprocessing of Web data–In this, Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step.

Analysis of Web Data–Also known as Web Usage Mining, this step applies machine learning or data mining techniques to discover interesting usage pattern and statistical correlation between web pages and user groups. This step frequently results in automatic user profiling, and is typically applied offline, so that it does not add a burden on the web server.

4. ADVANTAGES OF WEB MINING

Web mining is attractive for companiesbecause of several advantages. In the most general sense it can contribute to the increase of profits, be it by actually selling more products or services, or by minimizing the costs. In order to do this, marketing intelligence is required. This intelligence can focus on marketing strategies and competitive analyses or on the relationship with the customers.

From Sharon's point of view we could saythat she was pleased by the fact that the web site of her favorite bookstore displayed an interesting banner and she was not aware of any missed offers from this bookstore.

5. APPROACH OF WEB USAGE MINING

1. Requirement Analysis

Web access logs are the files that record the users' browsing information on the server. Many kinds of formats are available for web log files.

a.) Common web log format

A web server log file contains requests made to the web server, recorded in chronological order. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. A common log format file is created by the web server to keep track of the requests that occur on a web site. A standard log file has the following format as shown in figure Format of Standard Log File

- Remotehost is the remote hostname or its IP address,
- Logname is the remote logname of the user,
- Username is the username as which the user has authenticated himself,
- Date is the date and time of the request,
- Request is the exact request line as it came from the client,
- Status is the HTTP status code returned to the client, and
- Bytes is the content-length of the document transferred.
-

b.) Extended log file format

An extended common log format file is a variant of the common log format file simply adding two additional fields to the end of the line, the referrer Universal Resource Locator (URL) and the user agent fields:

- ✓ Referrer URL is the page the visitor was on when they clicked to come to this page.
- ✓ User Agent is whatever software the visitor used to access this site. It's usually a browser, but it could equally be a web robot, a link checker, a File Transfer Protocol (FTP) client or an offline browser.
- ✓

6. CONCLUSION

In this paper a general overview of Web usage mining is presented. Web usage mining is used in many areas such as e-Business, e-CRM, e-Services, e-Education, e-Newspapers, e- Government, advertising, Digital Libraries, marketing, bioinformatics and so on.Web Usage Mining (WUM) systems are specifically designed to carry out this task by analyzing the data representing usage data about a particular Web Site.More research work need to be done on the web mining domain as it will rule the web in the near future. Web mining along with semantic web known as semantic web mining is to be concentrated that is evolving which helps us to overcome the cons of web mining.

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it.

REFERENCES

1. Mohamad H. Hassoum "Fundamentals of artificial neural networks", PHI.

2. Laurene Fausett "Fundamentals of neural networks architectures, algorithms and applications", pearsoneducation.

3. Arun K Pujari "Data Mining Techniques", Universities Press.

4.J. Srivastva, P. Desikan, and V. Kumar, *Web mining – Concepts, Application and Research direction*, pp. 51, 2009.

5.O. Etzioni, "The World-Wide Web, Quagmire or Gold Mine?" *Communications of the ACM*, vol. 39, no. 11, pp. 65–68,1996.

6. R. Cooley, J. Srivastava, and B.Mobasher, "Web mining: Information and pattern discovery on the World Wide Web". In *Proc. of the 9th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'97)*, 1997.

7.Research on Personalized Recommendation Based on Web Usage Mining using Collaborative Filtering Technique, Taowei Wang, Yibo RenBrijendra Singh1, Hemant Kumar Singh2,"WEB DATA MINING RESEARCH: A SURVEY**"**, 978-1-4244-5967-4/10/$26.00 ©2010 IEEE.

8. Rajni Pamnani, Pramila Chawan 1 Qingtian Han, Xiaoyan Gao, "Web Usage Mining: A Research Area In Web Mining"**.**
Margaret H. Dunham, "Data Mining Introductory & Advanced Topics", Pearson Education.

9. Qingyu Zhang and Richard s. Segall," Web mining: a survey of current research,Techniques, and software", in the International Journal of Information Technology & Decision Making Vol. 7, No. 4 (2008) 683–720.

10. Q. Yang and X. Wu, 10 challenging problems in datamining research, Int. J Inform.Technol. Decision Making5(4) (2006) 597–604