Web Usages Mining and Knowledge Discovery

Prof.BadgujarMadhuri Nitin
Sheth N.K.T.T. College of Commerce,
Near C.K.P.P. Hall Behind,
Thane(W)-400-610
madhubadgujar25@gmail.com

Prof. Mane Arti Ashok
S.H. Muthacollege,
Aadharwadi,Kolivali
Kalyan(w)-421301
mane88arti@gmail.com

Abstract--Web mining means extracting useful information from large data source. There are 3 types of web mining first is web content mining, the second one is web structure mining and last is web usages mining. Web usages mining is used for automatic pattern discovery. Which helps to improve e-business by identifying exact user requirements.Web usages mining is maintaining log files of users which helps to identify user behavior about the site.
This paper represent the over view of important concepts of Web usage mining as well as different source of data. . This paper is an effort in analyzing the views and methodologies which are stated by various authors on various processes in mining the web.

Keywords--Web Mining,Data Pre-processing,Pattern discover, Pattern analysis, Personalization, Proxy Server.

INTRODUCTION

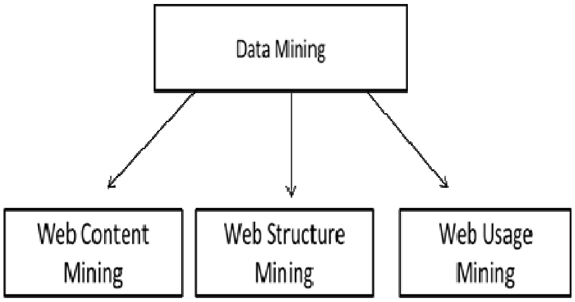
Data Mining is extraction knowledge from large amount of data, which is collected by several different sites. It is continuous improvement or evolutionary process of access, storeand generation of data in real time. Data mining is use in the business application community because it is supported by three technologies that are:

a) Massive data collection
b) Powerful multiprocessor computers
c) Data Mining Algorithm

Web mining –It is the application of data mining techniques which discover patterns from the Web. The goal of data mining is to extract knowledge from large amountsof data. Web mining is a technique to discover and analyze the useful information from the Web data.
Web involves three types of data:

1) **Data on the Web(content)**: The visible data in the Web pages. A major part of it includes text and graphics
2) **Web structure data:** Data which describes the organization of the website.
3) **Web log data (usage):** Data that describes the usage patterns of Web pages, such as IP addresses, page references, and the date and time of accesses and various other information depending on the log format.

Diagrammatically it can be represented as follows:

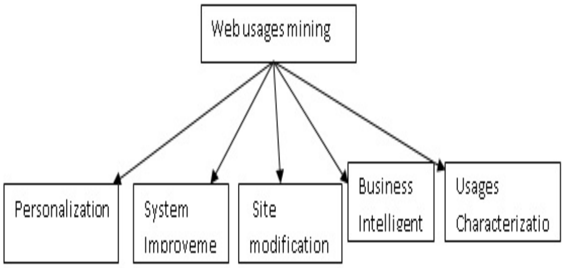


Web Usage Mining

Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web. Web Usage Mining is that part of Web Mining which deals with the extraction of knowledge from server log files. Source data mainly consist of the textual i.e. logs, that are collected when users access web servers and might be represented in standard formats.It involves the automatic discovery of user access patterns from one or more web server. Web usage mining tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web.
There are following process of web usage mining.

1) **Preprocessing**:
Conversion of the raw data into the data abstraction necessary for further applying the data mining algorithm.
2) **Pattern Discovery:**
It is the key component of web usages mining, which converges the algorithms and techniques from data mining, machine learning, statistics and pattern recognition etc. research categories.

3) **Pattern Analysis**:
   It is the validation and interpretation of the mined patterns.Major Application for Web Usages Mining



Research Methodology

The research methodology to be adopted for the study is proposed to be as follows:

1) **Sources of data:**
   a) **Number of hits:** This number usually signifies the number of times any resource is accessed in a Website.
   b) **Number of Visitors**: It is a human who navigates to your website and browses one or more pages on your site.
   c) **Visitor Referral Website:** The referral website gives the URL of the website which is being referred to by the particular website in consideration.
   d) **Time and Duration:** This information in the server logs give the time and duration for how long the Website was accessed by a particular user.
   e) **Path Analysis:** Path analysis gives the analysis of the path a particular user has followed in accessing contents of a Website.
   f) **Visitor IP address:** This information gives the Internet Protocol address of the visitors who visited the Website in consideration.
   g) **Browser Type:** This information gives the information of the type of browser that was used for accessing the Website.
   h) **Cookies**: A message given to a Web browser by a Web server.
   i) **Platform:** This information gives the type of Operating System that was used to access the Website.

Objectives

1) Make data in full structured format to overcome effort requires for preprocessing and parsing of data before extraction.

2) Elimination of irrelevant information such as image files.

3) To improve user navigation through perfecting and caching.

4) To improve the customer satisfaction, acquire new customers, retain customer, predict future visit patterns

5) To design of web personalizes system which aims development of robust and flexible system.

Hypothesis

1) **Hypothesis 1:**
   i. The data in the log files are unstructured, unorganized and irrelevant in nature due to which requirement of customer may not be understood by business resulting in reduction of growth of business.
2) **Hypothesis 2:**
   i. Web Usage Mining will help in creation of structured design which will reduce errors, saves server response time and satisfy customer preferences.

Plan of work

1) Focus is on business specific issues such ascustomer attraction, customer retention, cross sales. based on their browsing patterns customer behavior is identify.
2) Web servers log files and other sources of traffic data is used for analysis of visitors traffic information, errors that are generated and traffic faces while browsing. At the same time online behavior of user is also analyze.
3) After obtaining data, it is combined with relational databases, on which the data mining techniques are implemented.

2) **Collection of Data:**
   Collection of data from based upon,
   a) **Server side /web data:**
   It contains data generated by visits to a web site. Typical data includes IP address, page reference and access time.
   b) **Client Side:**

Collect data from client side by using client side java script, applet, etc.

c) **Business data**:
It is data in traditional systems generated by the respective business;

d) **Meta data**:
These data describing the web site itself (content and structure).

e) **Proxy Server:**
Used tocollect navigation data. And collection of data from intermediate server between browser and web server that is by using proxy server. It contain,

i) Characterize the browsing behavior of group of user by using proxy server.

ii) One can keep track of previously accessed pages of a user. These pages can be used to identify the typical behavior of the user and to make prediction about desired pages. Thu personalization for a user can be achieved through web usage mining.

iii) Frequent access behavior for the users can be used to identify needed links to improve the overall performance of future accesses. Perfecting and caching policies can be made on the basis of frequently accessed pages to improve latency time. Common access behaviors of the users can be used to improve the actual design of web pages.

3) Modifications to a Web site:
Usage patterns can be used for business intelligence in order to improve sales and advertisement by providing product recommendations.

4) Editing :
The collected data will be properly edited so as to eliminate irrelevant data.

5) Classification :
The edited data will be recorded and classified in the manner which can facilitate proper tabulation.

6) Scope of the study :
It is used to identify customer behavior, improving customer service and relationship, measuring the success of marketing efforts, and so on.

CONCLUSION

1) The Web usage mining method were successfully tested on log files and user session's identification.
2) Log file play important role in business e-commerce or e-business to identify number of users visits to site.
3) Access the web data has become huge in nature and lot of transactions taking place by the seconds so that data is not completely structured format and need lots of pre-processing and parsing for extraction of actual required information.

REFERENCES

[1] Abraham. A., Business Intelligence from Web Usage Mining, Journal of Information & Knowledge
Management (JIKM).

[2]K. R. Suneetha, Dr. R. Krishnamoorthi, Identifying User Behavior by Analyzing Web Server Access Log File, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009

[3] Ajith Abraham Natural Computation for Business Intelligence from Web Usage Mining

[4] SANJAY MADRIA, SOURAV S BHOWMICK, W. -K NG, E. P. LIM, Research Issues inWeb Data Mining.

[5] Federico Michele Facca and Pier Luca Lanzi Recent Developments in Web Usage Mining Research

[6]*Jaideep Srivastava, PrasannaDesikan, Vipin Kumar* Web Mining - Concepts,Applications& Research Directions

[7] VijayashriLosarwar, Dr.Madhuri Joshi Data Preprocessing in Web Usage Mining, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore

[8]Robert Cooly,BamshadMobasher and JaideepShrivastava Data Preparation For Mining World Wide Web Browsig Patterns, Knowledge and Information Systems1(1999)00 00

[9] BettinaBerendt, BamshadMobasher, Myra SpiliopoulouWeb Usage Mining for E-Business Applications,ECML/PKDD-2002 Tutorial, 19 August 2002

[10] Baoyao Zhou1, Siu Cheung Hui, and Alvis C. M. Fong, Web Usage Mining for Semantic Web Personalization

11]RajniPamnani, PramilaChawan,Web Usage Mining: A Research Area in Web Mining