Internet Mining and its Phases

Manisha[1], Joni Birla[2], Gurpreet[3]
[1,2,3]Department of Computer Science & Engineering,
Ganga Institute of Technology and Management,
Kablana, Jhajjar, Haryana, India

***Abstract*— In this paper, we describe the data warehousing and data mining. Data Warehousing is the process of storing the data on large scale and Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.**

As massive amount of data is continuously being collected and stored, many industries are becoming interested in mining some patterns (association rules, correlations, clusters etc) from their database. Association rule mining is one of the important tasks that are used to find out the frequent itemset from customer transactional database. Each transaction consists of items purchased by a customer in a visit.

Internet mining is the application of data mining techniques to discover patterns from the Internet. Internet Usage Mining (IUM) is the process of application of data mining techniques over web data. The data sources are mainly the web server logs, proxy server logs and cookies stored in the user's computer. IUM is composed of three phases namely, preprocessing, pattern discovery and pattern analysis. This paper describes these phases in detail. A necessary introduction to Internet Mining is also provided for the purpose of background knowledge.

Keywords— ***Data warehousing and its architectures, Data Mining, Techniques of Data Mining, Internet mining.***

I. INTRODUCTION

Data warehousing helps us to store the data. Data warehouse architecture is primarily based on the business processes of a business enterprise taking into consideration the data consolidation across the business enterprise with adequate security, data modeling and organization, extent of query requirements, meta data management and application, warehouse staging area planning for optimum bandwidth utilization and full technology implementation.

The Data Warehouse Architecture includes many facets. Some of these are listed as follows:

Process architecture
Date Model architecture
Technology architecture
Information architecture

Resource architecture

PROCESS ARCHITECTURE

Describes the number of stages and how data is processed to convert raw / transactional data into information for end user usage. The data staging process includes three main areas of concerns or sub- processes for planning data warehouse architecture namely "Extract", "Transform" and "Load".

These interrelated sub-processes are sometimes referred to as an "ETL" process.

1) Extract- Since data for the data warehouse can come from different sources and may be of different types, the plan to extract the data along with appropriate compression and encryption techniques is an important requirement for consideration.

2) Transform- Transformation of data with appropriate conversion, aggregation and cleaning besides de-normalization and surrogate key management is also an important process to be planned for building a data warehouse.

3) Load- Steps to be considered to load data with optimization by considering the multiple areas where the data is targeted to be loaded and retrieved is also an important part of the data warehouse architecture plan.

DATA MODEL ARCHITECTURE

In Data Model Architecture (also known as Dimensional Data Model), there are 3 main data modeling styles for enterprise warehouses:

3rd Normal Form - Top Down Architecture, Top Down Implementation

Federated Star Schemas - Bottom Up Architecture, Bottom Up Implementation

Data Vault - Top Down Architecture, Bottom Up Implementation

Technology Architecture

Scalability and flexibility is required in all facets. The extent of these features is largely depending upon organizational size, business requirements, nature of business etc.

Technology or Technical architecture primary evolved from derivations from the process architecture, meta data management requirements based on business rules and security levels implementations and technology tool specific evaluation.

Besides these, the Technology architecture also looks into the various technology implementation standards in database management, database connectivity protocols (ODBC, JDBC, OLE DB etc), Middleware (based on ORB, RMI, COM/DOM etc.), Network protocols (DNS, LDAP etc) and other related technologies.

Information Architecture

It is the process of translating the information from one form to another in a step by step sequence so as to manage the storage, retrieval, modification and deletion of the data in the data warehouse.

Resource Architecture

Resource architecture is related to software architecture in that many resources come from software resources. Resources are important because they help determine performance. Workload is the other part of the equation. If you have enough resources to complete the workload in the right amount of time, then performance will be high. If there are not enough resources for the workload, then performance will be low.

II. DATA MINING

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

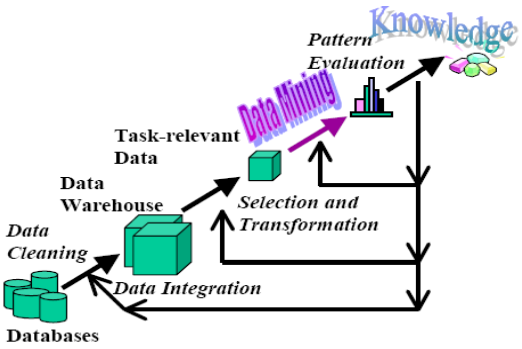

Fig:1 Data Mining is the core of Knowledge Discovery process

Data mining has its own tools and techniques to mine interesting information. When these tools and techniques are applied to the World Wide Web [as is or with some modifications and adaptations for the www environment], it can be called as Internet Mining.

So, Internet mining refers to discovery and analysis of useful information over the World Wide Web. Internet mining can be broadly classified into three categories:

- Content Mining
- Structure Mining
- Usage Mining

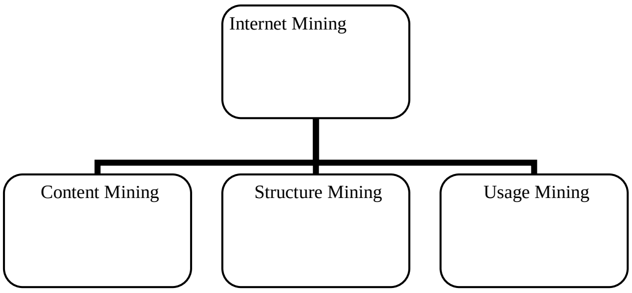

Fig:2 Types of Internet Mining

Content Mining:

Content Mining refers to mining of desired content over World Wide Web. Various search engines exists for the content mining, such as altavista, Lycos, WebCrawlar, MetaCrawlar etc.

Structure Mining:

Structure mining tries to discover the link structure of the hyperlinks at the inter-document level to generate structural summary about the Website and Web page.

Usage Mining:

Usage Mining refers to automatic knowledge mining of user access patterns from web servers. It includes,

Preprocessing
Pattern Discovery Tools
Pattern Analysis Tools

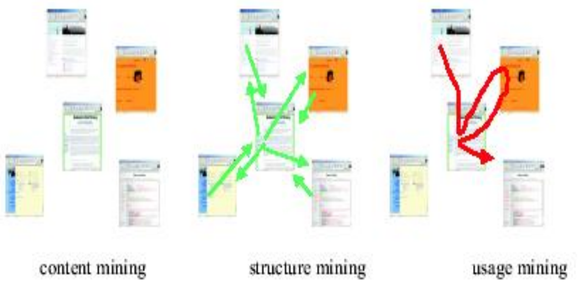

Figure 3: Types of Internet Mining

III. THE INTERNET USAGE MINING

Internet Usage Mining refers to automatic knowledge mining of user access patterns from different web servers.. It is the application of various techniques used in Data Mining to discover and analyze the usage patterns of web data.

Why Internet Usage Mining?

Internet has been growing at explosive rate since last decades. Lots of information is available on the internet. Millions of Websites exists and more are uploaded daily containing a lot of information. Billions of users browse on internet for different reasons, each searching for some interesting information. By Interesting Information, we refer to the information for which the user is browsing on internet, rest all information doesn’t seems to be interesting to him. How interesting the information is to a particular user, is identified by interestingness measures. Interestingness measures are used based on data mining techniques such as clustering, classification and association. These users needs tool and techniques [e.g. browsers], so that they can find needed information in a less time with more accurate results.

Another perspective is from the engineers, developers, web designers, and such professionals who strive to create more and more structured information, on structured websites. They are responsible for managing the structure of websites and providing interesting information in an interesting manner. They design tools and techniques for this and use them to manage websites by their content, and structure.

A very different perspective is from the companies who have invested millions into the web and web technologies. These are the organizations which are mostly based on E-Commerce, selling their products and services over the World Wide Web. For these organizations, it is very essential to keep the patterns of user visits, their profiles and their interestingness measures. This gives requirement for the development of client and server side intelligent systems that can mine knowledge across web.

So, it is essential to have some techniques and tools for satisfying the above said requirements. All these requirements give rise to “INTERNET MINING”. The term INTERNET MINING is very broad in its sense. But a special kind of internet mining called “INTERNET USAGE MINING” is the focus of the work presented here.

A number of organizations has invested highly on web technologies and carrying out business there. For example Amazon.com, ebay.com, buy.com etc. A lot of people access their websites across the world and does business with them. Analyzing this data can provide these organizations with the value of the customers. It helps the organizations to identify the “Good”, “Valued” and “Bad” customers based on their access patterns. This data also helps them for cross marketing strategies, their campaigns and others. Organizations can identify the effectiveness of their websites and also the effectiveness of their advertisements on different websites. Web Usage Mining helps them to identify the market segment and target interesting customers.

From where the data comes:

All the data, regarding the users is stored in their server access logs. Other sources include referrer logs which contains the information about referring pages from which the user has been referred to a particular page. User forms, survey results are also used as input. In Internet Usage Mining, data is collected at Web Servers, proxy servers, and organization’s own database. Various methods such as cookies, CGI Script, Java Script, forms, session tracking, query data, click streams and page views are frequently used in web usage mining.

The data that is required to perform includes web server logs, cookies, proxy server logs, surveys, registration forms filled by users, access patterns of users (click stream) etc. The data sources can be classified into three categories:

Collection of Data from Server:

These data sources include logs from web server. Web server logs are important because they provide major user access patterns. All the works that user performs on a website are recorded in logs in the web server. Web servers are the computers having special software installed on them which are used to fulfill the user requests. A web server software may be Apache Tomcat, BEA WebLogic, IBM’s WebSphere, Sun Microsystem’s J2EE Application server etc. Logs that are maintained can be in different formats.

So, care should be taken when data is collected from more than one web server. A web usage mining tool must be capable of processing logs of more than one web server software.

However, the logs stored in web servers cannot be called the complete input, as there are different levels of caching in the internet architecture. Often, clients are first directed to cache and then web servers. Moreover there are different data that are not logged in the web servers such as information passed through POST method. Other sources includes cookies. Cookies are special files that are generated by web servers to collect information about individual clients. For creating cookies, user must authorize web server to created cookies, as cookies concern with privacy. Various scripting languages such as CGI Script, Java Script, VB Script and Perl Script are also used to handle the data that is sent back to the web server from client browsers.

Collection of Data from Clients:

Client side collection requires user cooperation. The technologies includes Java Applets, and various scripts which requires users to enable them. Data from clients can also be collected by using modified browsers. But user must be made willing to use that browser. Different companies like NetZoro[9], YouMint[10] and AllAdvantage[11] offers users incentives for using modified browsers and clicking on the advertisements on them.

Collection of Data from Proxy Servers:

Data collection only from web servers is not efficient to perform web usage mining. This is because, not all the requests reach the web servers each and every time. To speed up the browsing of internet, proxy servers are also used thus reducing the load on a web server. So, proxy servers also acts as servers and also contain user access logs. These logs should also be analyzed to perform web usage mining.

IV. PROCESS OF INTERNET USAGE MINING

The process of internet usage mining is composed of three steps. As given in the figure,

1. Pre-processing
2. Pattern Discovery
3. Pattern Analysis

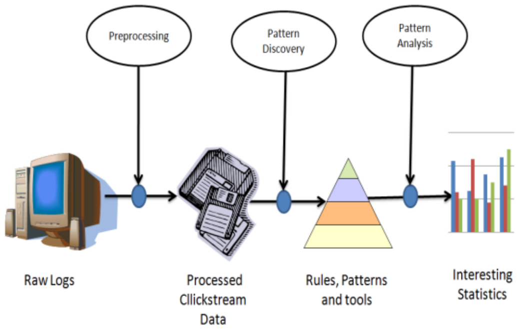

Fig: Web Usage Mining Process

Figure 4: Web usage mining process

Pre-processing:

Pre-processing is the process of preparing data received through server logs, proxy server logs and other data ready for pattern discovery and analysis task. The pre-processing task includes many processes. These are:

a. Data Cleaning: Involves removal of those log entries, which does not contribute to the data mining task. These unnecessary entries may be called noise.

b. Identification of users: Involves identification of users. It associates a page reference with a particular user. User identification is not an easy task because (i) a single IP address can be used by multiple users, (ii) Different IP addresses can be used by a single user

c. Identification of session: involves identification of session over a web server. It associates a group's web page references into user/server session. It also involves some issues: (i) a single IP address can have multiple server sessions, such as in case of proxy servers. (ii) Multiple IP address can have a single server session.

d. Path Completion: Due to proxy servers, and caching, it is not always possible to get complete data from web servers. The access paths shown in web server are incomplete if some page is referenced through proxy servers or cache. Path completion is the process of completing those incomplete paths.

Pattern Discovery:

Once the necessary transactions have been identified, the next step is the discovery of patterns. Pattern discovery phase extensively uses data mining algorithms. Various pattern discovery methods are:

Statistical Analysis: Statistical Analysis techniques are most commonly used techniques. These include frequency distribution, Mean, Mode, Median etc upon the web server logs. These techniques provide the basis for the IUM process. It provides the statistical data, and thus provides support for making market decisions.

Clustering: Clustering is division of data into groups of similar objects. A cluster represents objects that are similar between themselves. From machine learning perspective clusters corresponds to hidden patterns. Many clustering algorithms have been devised. Some major algorithms includes: Hierarchical Methods, K-means method, Grid based Clustering etc. In IUM, two type of clusters needs to be discovered: Usage Clusters and Page Clusters. Usage clusters helps to identify groups of users having similar browsing patterns. Page clusters helps to identify groups of pages with similar content. A dynamic clustering based model based on Markov Analysis is presented in [15]

Classification: Classification is a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items (referred to as traits, variables, characters, etc) and based on a training set of previously labeled items. Formally, the problem can be stated as follows: given training data {(x1, y1),….,(xn,yn)} produce a classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ which maps any object $\mathbf{x} \in \mathcal{X}$to its true classification label $y \in \mathcal{Y}$defined by some unknown mapping $g : \mathcal{X} \rightarrow \mathcal{Y}$(ground truth). For example, if the problem is filtering spam, then $\mathbf{x}_{\mathbf{i}}$is some representation of an email and y is either "Spam" or "Non-Spam". Statistical classification algorithms are typically used in pattern recognition systems. In WUM, we are interested in profiling users from same class. Classification algorithms includes: K-Nearest-Neighbor (KNN) Algorithm, Naïve Bayesian (NB) Algorithm, Concept Vector based Algorithms etc.

Association: Association Algorithms find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis. For example, Microsoft association Algorithm. In IUM, association algorithms are used to relate web pages which were referenced by a user in a single session.. Algorithms like Apriori can be used for association rule mining.

Sequential patterns: Sequential patterns tend to find inter-transaction patterns in such a way that one pattern is followed by another in a time sequential manner. Web logs are periodically recorded in Web Servers. These log entries also includes time-stamps associated with each user visit on the link. These sequential patterns can help organizations to predict the future visit time of the user over their website. It can also help to establish the relation that which file/page was visited most during which user session/day/time/week/month.

Pattern Analysis:

Pattern Analysis is the last step in our IUM process. This helps to analyze organizations that how customers are accessing their website, and which are the pages they mostly visits. The purpose of pattern analysis is to filter out uninteresting rules and analyze the interesting rules which were found during the pattern discovery process. The major techniques included in this phase include:

SQL Queries

Visualization Techniques

OLAP Techniques and

Usability analysis.

V. CONCLUSIONS

The Internet Usage Mining is special case of Data Mining where the usage patterns of web pages are analyzed. Web pages can be on one or more servers, and also can be in different formats. Internet Usage Mining is very useful tool for organizations who wants to keep their customer base. We provided a detailed survey of research in this area. Various softwares and tools are available in market for IUM. We also provided the demonstration of WebLogAnalyzer® by Nihuo™. Though, the survey is short as the area is not very well established. There is immense scope of research in this area for identifying new methods and tools to discover pattern and analyze them.

REFERENCES

[1] J-Han M.Kamber "Data mining: concepts and techniques"2nd edition ,Morgan Kaufman publication, August

[2] Bart Goethals" survey on frequent pattern mining".

[3] The World Wide Web Consortium Web Usage Characterization Activity (WCA). http://w3.org/WCA

[4] Software Inc. Webtrends. http://www.webtrends.com

[5] NetGenesis netAnalysis desktop, http://www.netgen.com

[6] J.Shrivastava, R.Cooley, M.Deshpandey, Pang-Ning Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data”, Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 USA. http://cs.umn.edu

[7] B.Mobasher, R.Cooley, J.Shrivastava, “Web Mining: Information and Pattern discovery on the World Wide Web”, Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA http://cs.umn.edu7