ANALYSIS OF WEBSITE USAGE WITH USER DETAILS USING DATA MINING PATTERN RECOGNITION

K.Vinodkumar [1], Kathiresan.V[2], Divya.K [3]

[1] MPhil scholar, RVS College of Arts and Science, Coimbatore, India.
[2] HOD, Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore, India.
[3] MPhil scholar, PSG College of Arts & Science, Coimbatore, India.

Abstract

Web usage mining is a main research area in Web mining focused on learning about Web users and their interactions with Web sites to know the usage level of the website. The motive of mining is to find users' access models automatically and quickly from the vast Web log data, such as frequent access paths, frequent access page groups and user clustering. Through web usage mining, the server log, registration information and other relative information left by user access can be mined with the user access mode which will provide foundation for decision making of organizations. The main feature of this research is, this is a complete framework and findings in Web usage patterns from Web log files of a real Web site that has all the difficult aspects of real-life Web usage, including developing user profiles and external data describing database of the Web content.

Keywords: web log , server log ,user profile.

1. Introduction

The frame work "**WEB USAGE MINING**" is used to mine the web usage. This frame work uses the data mining concept "pattern recognition". Through this research we can determine the usage level of one particular website. From the details of the various user's behaviour we can predict the usage level of one particular web page.

Here the web page usage details will be automatically stored and maintained by the "Web log file". Here the registration process allows the user to register on the particular web site. User's personal details will be stored in the data base. Then the user can log in to the site. When one registered user logged into the particular page means the web log file stores the details about the user's various access on the page.

Until the user's log out the log file automatically stores these details. The web log file also maintain the details about the number of hits for one particular page (or) link, with in the particular time duration who all are visited one particular page, the log in and log out time details of one particular user and their various access details on the web pages.

Through the web log file details we can determine that which page (or) link is mostly visited by the users. From these details we can produce a report for the web site usage. Through the usage level we can provide an effective and user friendly solutions. This will help us to use the web as more profitable one for a particular business strategy.

2. Related Work

In Web Mining, data can be collected at the server side, client-side, proxy servers, or obtained from an organization's database (which contains business data or consolidated Web data).

Each type of data collection differs not only in terms of the location of the data source, but also the kinds of data available, the segment of population from which the data was collected, 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, i.e. the data the Web page was designed to convey to the users. This usually consists of, but is not limited to, text and graphics.

Structure: Data which describes the organization of the content. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. This can be represented as a tree structure, where the <html>tag becomes the root of the tree.

Usage: Data that describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of accesses.

User Profile: Data that provides demographic information about users of the Web site. Includes registration data and customer profile information.

3. Proposed System

This framework covers the various aspects related to the "WEB USAGE MINING". The proposed system uses the data mining concept "PATTERN DISCOVERY AND PATTERN ANALYSIS". The proposed system is developed to ease the process of

mine the usage level of one particular website. The uses of this application are,

- It used to identify the usage level of one particular web site.
- The various user's behaviour on one particular website will be monitored by this application.
- Through this application we can answer for the following six questions:
    1. How are people using this site?
    2. Which pages are being accessed most frequently
    3. To know which page (or) hyperlink got maximum number of hits.
    4. Users details will be easily maintained.
    5. With the help of the web log file details the reports will be produced.
    6. With the results of the reports we can use web as a more profitable one for particular business strategy.

3.1 Pattern Discovery

Pattern discovery techniques involve algorithms to discover interesting patterns from web data. Once user transactions or sessions have been identified, there are several kinds of access pattern mining that can be performed depending on the needs of the analyst. Some of these discovery techniques are discussed below.

3.1.1 Path Analysis

Graph models are most commonly used for Path Analysis. In the graph models, a graph represents some relation defined on Web pages and each tree of the graph represents a web site. Each node in the tree represents a web page and edges between trees represent the links between web sites and the edges

between nodes inside a same tree represent links between documents at a web site. When path analysis is used on the site as a whole, this information can offer valuable insights about navigational problems. Most graphs are involved in determining frequent traversal patterns and more frequently visited paths in a web site.

3.1.2 Association Rules

Predict the association and correlation among set of items where the presence of one set of items in a transaction implies with a certain degree of confidence the presence of other items. That is, it can discover the correlations between pages that are most often referenced together in a single server session/user session. It can provide the information: What are the set of pages frequently accessed together by web users? What page will be fetched next? What are paths frequently accessed by web users?. Implement association rules to on-line shopper can generally find out his/her spending habits on some related products.

3.1.3 Sequential Patterns

Sequential patterns discovery is to find the inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. Web log files can record a set of transactions in time sequence

3.1.4 Decision Trees

A decision tree is essentially a flow chart of questions or data points that ultimately leads to a decision.

3.1.5 Clustering

Clustering identifies visitors who share common characteristics. After you get the customers'/visitors' profiles, you can specify how many clusters to identify within a group of profiles, and then try to find the set of clusters that best represents the most profiles. Besides information from Web log files, customer profiles often need to be obtained from an on-line survey form when the transaction occurs.

3.1.6 Grouping

Users usually can draw higher-level conclusions by grouping similar information. how many visitors came from a Yahoo server. For example: http://search.yahoo.com/bin/search?p=Web+Miners

3.1.7 Filtering

Simple reporting needs require only simple analysis systems. However, as the company's Web becomes more integrated with the other functionality of the company, for example, customer service, human resources, marketing activity, analysis need to rapidly expand. For example, the company launches a marketing campaign. Print and television ads now are designed to drive consumers to a Web site, rather than to call an 800 number or to visit a store. Consequently, tracking online marketing campaign results is no longer a minor issue but a major marketing concern. Often it's difficult to predict which variables are critical until considerable information has been captured and analysed. Consequently, a Web traffic analysis system should allow precise filtering and grouping information even after the data has been collected. Systems that force a company to predict which variables are important before capturing the data can lead to poor decisions because the data will be skewed toward the expected outcome. Filtering information allows a manager to answer specific questions about the site. For example, filters can be used to calculate how many visitors a site received this week from Microsoft. In this example, a filter is set for "this week", and for visitors that have the word "Microsoft" in their domain name e.g.proxy12.microsoft.com.

3.1.8 Dynamic Site Analysis

Traditional Web sites were usually static HTML pages, often hand-crafted by Webmasters. Today, a number of companies, including Microsoft, make systems that allow an HTML file to be dynamically created around a database.

This offers advantages like, included centralized storage, flexibility, and version control.

3.3 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. 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 colours 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.

4. Conclusion

We presented a framework for mining, tracking, and validating evolving multifaceted user profiles on Web sites that have all the challenging aspects of real-life Web usage mining, include evolving user profiles and access patterns, dynamic Web pages, and external data describing ontology of the Web content. A multifaceted user profile summarizes a group of users with similar access activities and consists of their viewed pages, search engine queries, and inquiring and inquired companies. The choice of the period length for analysis depends on the application

or can be set, depending on the cross-period validation results. Even though we did not focus on scalability, the latter can be addressed by following an approach similar to, where Web click streams are considered as an evolving data stream, or by mapping some new sessions to persistent profiles and updating these profiles, hence eliminating most sessions from further analysis and focusing the mining on truly new sessions.

5. References

[1] "Web Analytics An hour a day", Avinash kaushik, Wiley India Publication,2007.

[2] "Web Analytics 2.0", Avinash kaushik, Wiley India Public ation,2010.

[3] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/ The MIT Press, 1996.

[4] J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

[5] Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

[6] Sholom M. Weiss and Nitin Indurkhya, "Predictive Data Mining: A Practical Guide", Morgan Kaufmann Publishers, 1998.

[7] Alex Freitas and Simon Lavington, "Mining Very Large Databases with Parallel Processing", Kluwer Academic Publishers, 1998.

[8] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 1988.

[9] V. Cherkassky and F. Mulier, "Learning From Data", John Wiley & Sons, 1998.

6. AUTHOR PROFILE:

K.Vinodkumar is an M.Phil scholar, Department of Computer Applications (MCA) in RVS College of Arts and Science, Coimbatore. He received his B.Sc., in 2010 from Bharathiar University and MCA in 2013 from Anna University, Chennai. He is pursuing his M.Phil. in the area of Data mining from Bharathiar University, Coimbatore. He got university rank in MCA.

Dr.V.Kathiresan is an HOD, Department of Computer Applications (MCA) in Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore. He received his B.Sc., in 2003 and MCA in 2006 from Bharathiar University, Coimbatore. He obtained his M.Phil. in the area of Data mining from Periyar University, Salem in 2007. He obtained his Phd in the area of Data mining. His research interest lies in the area of Data mining. He got Faculty Excellence Award from RVS College of Arts & Science for the Academic years 2007-08, 2008-09,2009-10, 2010-11, 2011-12 , 2012-13 and 2013-14 consecutively.

K.Divya is an M.Phil scholar, Department of Computer Science in PSG College of Arts and Science, Coimbatore. He received his BCA in 2013 from Bharathiar University and Msc(CS) in 2015 from PSG College of Arts and Science, Coimbatore. .