KNOWLEDGE DISCOVERY ON THE INTERNET (WEB MINING TOOL AND TECHNIQUE)

MOHD SHAHID

Declaration

The Declaration of the author for publication of Research Paper in The Indian Journal of Research Anvikshiki ISSN 0973-9777 Bi-monthly International Journal of all Research: I, MOHD SHAHID the author of the research paper entitled **KNOWLEDGE DISCOVERY ON THE INTERNET (WEB MINING TOOL AND TECHNIQUE)** declare that, I take the responsibility of the content and material of my paper as I myself have written it and also have read the manuscript of my paper carefully. Also, I hereby give my consent to publish my paper in Anvikshiki journal, This research paper is my original work and no part of it or it's similar version is published or has been sent for publication anywhere else. I authorise the Editorial Board of the Journal to modify and edit the manuscript. I also give my consent to the Editor of Anvikshiki Journal to own the copyright of my research.

Abstract

With the explosive growth of information sources available on the World Wide Web today, Web has turned to be the largest information source available in this planet. This paper will focus on Web usage mining. Generally speaking, Web usage mining consists of three phases: Pre-processing, Pattern discovery and Pattern analysis. A detailed description will be given for each part of them, however, special attention will be paid to the user navigation patterns discovery and analysis.

Keywords: Knowledge Discovery, Web usage mining, web mining, weblog

1. INTRODUCTION

Web data mining is a process that discovers the intrinsic relationships among Web data, which are expressed in the forms of textual, linkage or usage information, via analyzing the features of the Web and web-based data using data mining techniques[2]. Particularly, we concentrate on discovering Web usage pattern via Web usage mining, and then utilize the discovered usage knowledge for presenting Web users with more personalized Web contents, i.e. Web recommendation[3].

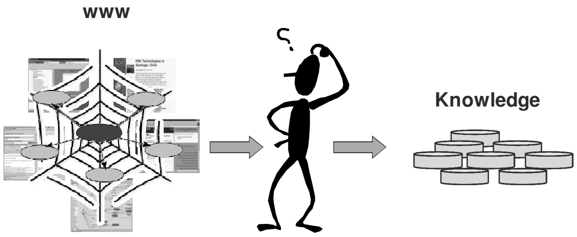


Figure 1: Web Mining

2. WEB DATA

One of the key steps in Knowledge Discovery in Databases is to create a suitable target data set for the data mining tasks. 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).

*Ph.D In C.S., Research Scholar, Singhania University,Pacheri Bari, Jhunjhunu, Rajasthan.

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.

Web Log

There are three types of log files, namely server logs, error logs, and cookie logs[4]. Server logs are either stored in the Common Log file Format or the more recent Extended Log file Format. Anatomy of A Log File is as follows:

1. Internet provider IP address: This can be either webminer.com or 204.58.155.58
2. Identification field: This usually appears as a dash, "-"
3. AuthUser: This is an ID or password for accessing a protected area
4. Date, time, and GMT (Greenwich Mean Time): Thu July 17 12:38:09 1999
5. Transaction: Usually "GET" filename such as /index.html/products.htm
6. Status or error code of transaction: Usually 200 (success)
7. Size in bytes of transaction (file size): 3234 Additional Fields in the Extended Log Format
8. Referrer: search engine and keyword used to find your Web site, such as http://search.yahoo.com/bin/search?p=data+miningý/index.html
9. Agent: browser used by your visitor, such as Mozilla/2.0 (Win95; I)
10. Cookie: .snap.com TRUE / FALSE 946684799 u_vid_0_0 00ed7085

Error logs store data of failed requests, such as missing links, authentication failures, or timeout problems. Apart from detecting erroneous links or server capacity problems — which, when satisfactorily corrected, can be seen as a compulsory form of customer satisfaction — the usage of error logs has so far proven rather limited for the discovery of actionable marketing intelligence. Cookies are tokens generated by the web server and held by the clients. The information stored in a cookie log helps to ameliorate the transaction less state of web server interactions, enabling servers to track client access across their hosted web pages. The logged cookie data is customizable, which goes hand in hand with the structure and content of the marketing data. A fourth data source that is typically generated on e-commerce sites is query data to a web server. For example, customers to an online store may search for products, or clients to a research database may search for publications. The logged query data must be linked to the access log through cookie data and / or registration information. There are currently no formal drafts for standards for handling query data, although new specification suggestions have reached draft *An Internet-enabled Knowledge Discovery Process* stage, for instance Resource Description Framework RDF. In order to make use of query data, it has to be grouped into logical, usually marketing-related clusters[6].

Data Sources

The usage data collected at the different sources will represent the navigation patterns of different segments of the overall Web traffic, ranging from single-user, single-site browsing behavior to multi-user, multi-site access patterns.

Server Level Collection

A Web server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the (possibly concurrent) access of a Web site by multiple users. These log files can be stored in various formats such as Common log or Extended log formats.

Client Level Collection

Client-side data collection can be implemented by using a remote agent (such as Javascripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities. The implementation of client-side data collection methods requires user cooperation, either in enabling the functionality of the Java scripts and Java applets, or to voluntarily use the modified browser. Client-side collection has an advantage over server-side collection because it ameliorates both the caching and session identification problems. However, Java applets perform no better than server logs in terms of determining the actual view time of a page. In fact, it may incur some additional overhead especially when the Java applet is loaded for the first time. Java scripts, on the other hand, consume little interpretation time but cannot capture all user clicks (such as reload or back buttons). These methods will collect only single-user, single-site browsing behavior. A modified browser is much more versatile and will allow data collection about a single user over multiple Web sites. The most difficult part of using this method is convincing the users to use the browser for their daily browsing activities. This can be done by offering incentives to users who are willing to use the browser, similar to the incentive programs offered by companies such as NetZero and All Advantage that reward users for clicking on banner advertisements while surfing the Web.

Proxy Level Collection

A Web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy caching can be used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides. The performance of proxy caches depends on their ability to predict future page requests correctly. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterizing the browsing behavior of a group of anonymous users sharing a common proxy server.

3. WEB USAGE MINING

Web usage mining is the automatic discovery of user access patterns from Web servers. Organizations collect large volumes of data in their daily operations, generated automatically by Web servers and collected in server access logs[7]. Other sources of user information include referrer logs which contain information about the referring pages for each page reference, and user registration or survey data gathered via CGI scripts. Analyzing such data can help organizations determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among other things. It can also provide information on how to restructure a Web site to create a more effective organizational presence, and shed light on more effective management of workgroup communication and organizational infrastructure[8]. For selling advertisements on the World Wide Web, analyzing user access patterns helps in targeting ads to specific groups of users.

As shown in following Figure 2, there are three main tasks for performing Web Usage Mining or Web Usage Analysis. This section presents an overview of the tasks for each step and discusses the challenges involved[6].

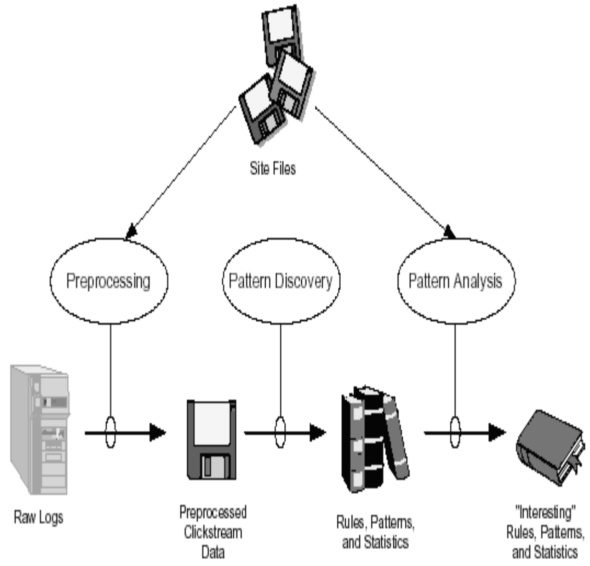

Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Figure-2: High Level web usages mining process

Preprocessing

The inputs to the preprocessing phase are the server logs, site files, and optionally usage statistics from a previous analysis. The outputs are the user session file, transaction file, site topology, and page classifications, one of the major impediments to creating a reliable user session file is browser and proxy server caching. Current methods to collect information about cached references include the use of cookies and cache busting. Cache busting is the practice of preventing browsers from using stored local versions of a page, forcing a new down-load of a page from the server every time it is viewed. none of these methods are without serious drawbacks. Cookies can be deleted by the user and cache busting defeats the speed advantage that caching was created to provide, and is likely to be disabled by the user. Another method to identify users is user registration; Registration has the advantage of being able to collect additional demographic information beyond what is automatically collected in the server log, as well as simplifying the identification of user sessions. However, again due to privacy concerns, many users choose not to browse sites that require registration and logins, or provide false information. The preprocessing methods used in the WEBMINER system are all designed to function with only the information supplied by the *Common Log Format* specified as part of the HTTP protocol by CERN and NCSA.

Data Cleaning

Techniques to clean a server log to eliminate irrelevant items are of importance for any type of Web log analysis, not just data mining. The discovered associations or reported statistics are only useful if the data represented in the server log gives an accurate picture of the user accesses to the Web site. The HTTP protocol requires a separate connection for every file that is requested from the Web server. Therefore, a user's request to view a particular page often results in several log entries since graphics and scripts are down-loaded in addition to the HTML file. In most cases, only the log entry oft he HTML file request is relevant and should be kept for the user session file. This is because, in general, a user does not explicitly request all of the graphics that are on a Web page, they are automatically down-loaded due to the

HTML tags. Since the main intent of Web Usage Mining is to get a picture of the user's behavior, it does not make sense to include file requests that the user did not explicitly request. Elimination of the items deemed irrelevant can be reasonably accomplished by checking the suffix oft he URL name. For instance, all log entries with filename suffixes such as, gif, jpeg, GIF, JPEG, jpg, JPG, and map can be removed. In addition, common scripts such as "count.cgi" can also be removed. The WEBMINER system uses a default list of suffixes to remove files. However, the list can be modified depending on the type of site being analyzed. For instance, for a Web site that contains a graphical archive, an analyst would probably not want to automatically remove all of the GIF or JPEG files from the server log. In this case, log entries of graphics files may very well represent explicit user actions, and should be retained for analysis. A list of actual file names to remove or retain can be used instead of just file suffixes in order to distinguish between relevant and irrelevant log entries.

User Identification

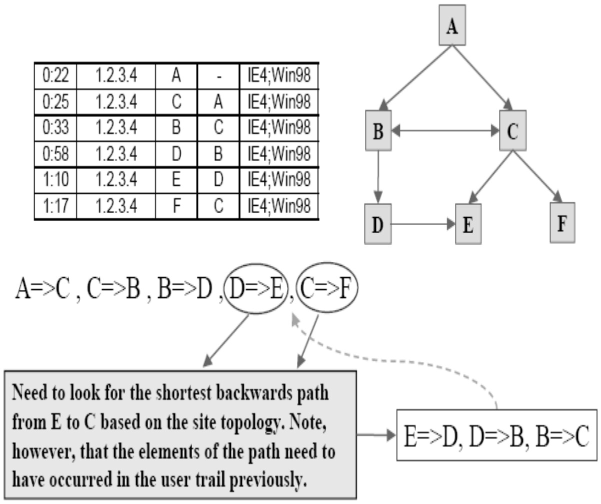
Next, unique users must be identified. As mentioned previously, this task is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. The Web Usage Mining methods that rely on user cooperation are the easiest ways to deal with this problem. However, even for the log/site based methods, there are heuristics that can be used to help identify unique users, even if the IP address is the same, if the agent log shows a change in browser software or operating system, a reasonable assumption to make is that each different agent type for an IP address represents a different user. The next heuristic for user identification is to use the access log in conjunction with the referrer log and site topology to construct browsing paths for each user. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, again, the heuristic assumes that there is another user with the same IP address.

Session Identification

For logs that span long periods of time, it is very likely that users will visit the Web site more than once. The goal of session identification is to divide the page accesses of each user into individual sessions. The simplest method of achieving this is through a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout, and establish timeout of 2 5.5 minutes based on empirical data.

Path Completion

Another problem in reliably identifying unique user sessions is determining if there are important accesses that are not recorded in the access log. This problem is referred to as *path completion*. Methods similar to those used for user identification can be used for path completion. If a page request is made that is not directly linked to the last page a user requested, the referrer log can be checked to see what page the request came from. If the page is in the user's recent request history, the assumption is that the user backtracked with the "back" button available on most browsers, calling up cached versions of the pages until a new page was requested. If the referrer log is not clear, the site topology can be used to the same effect. If more than one page in the user's history contains a link to the requested page, it is assumed that the page closest to the previously requested page is the source of the new request. Missing page references that are inferred through this method are added to the user session file. An algorithm is then required to estimate the time of each added page reference. A simple method of picking a time-stamp is to assume that any visit to a page already seen will be effectively treated as an auxiliary page. The average reference length for auxiliary pages for the site can be used to estimate the access time for the missing pages. A path completion example is illustrated below:



0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:17	1.2.3.4	F	C	IE4;Win98

A=>C , C=>B , B=>D , D=>E , C=>F

Need to look for the shortest backwards path from E to C based on the site topology. Note, however, that the elements of the path need to have occurred in the user trail previously.

E=>D, D=>B, B=>C

Figure 3: Path completion Example

Pattern Discovery

This is the key component of the web mining. Pattern discovery covers the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. It has separate subsections as follows.

Statistical Analysis: Statistical analysts may perform different kinds of descriptive statistical analyses based on different variables when analyzing the session file. By analyzing the statistical information contained in the periodic web system report, the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitation the site modification task, and providing support for marketing decisions

***Association Rules*:** In the web domain, the pages, which are most often referenced together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions.

Clustering: Clustering analysis is a technique to group together users or data items (pages) with the similar characteristics. Clustering of user information or pages can facilitate the development and execution of future marketing strategies.

Classification: Classification is the technique to map a data item into one of several predefined classes. The classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifier, Support Vector Machines etc.

Sequential Pattern: This technique intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions or episodes. Sequential patterns also include some other types of temporal analysis such as trend analysis, change point detection, or similarity analysis.

Dependency Modeling: The goal of this technique is to establish a model that is able to represent significant dependencies among the various variables in the web domain. The modeling technique provides a theoretical framework for analyzing the behavior of users, and is potentially useful for predicting future web resource consumption.

Pattern Analysis: Pattern Analysis is a final stage of the whole web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transform to a format can be assimilate easily. There are two most common approaches for the patter analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations. All these methods assume the output of the previous phase has been structured.

4. WEB USAGE MINING ARCHITECTURE

We have developed a general architecture for Web usage mining The WEBMINER is a system that implements parts of this general architecture. The architecture divides the Web usage mining process into two main parts. The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes preprocessing, transaction identication, and data integration components. The second part includes the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine. The overall architecture for the Web
Mining process is depicted in Figure 4 Data cleaning is the first step performed in the Web usage mining process. Some low level data integration tasks may also be performed at this stage, such as combining multiple logs, incorporating referrer logs, etc. After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. The goal of transaction identification is to create meaningful clusters of references for each user. The task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. The input and output transaction formats match so that any number of modules to be combined in any order, as the data analyst sees. Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data mining task. For instance, the format of the data for the association rule discovery task may be different than the format necessary for mining sequential patterns. Finally, a query mechanism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints.

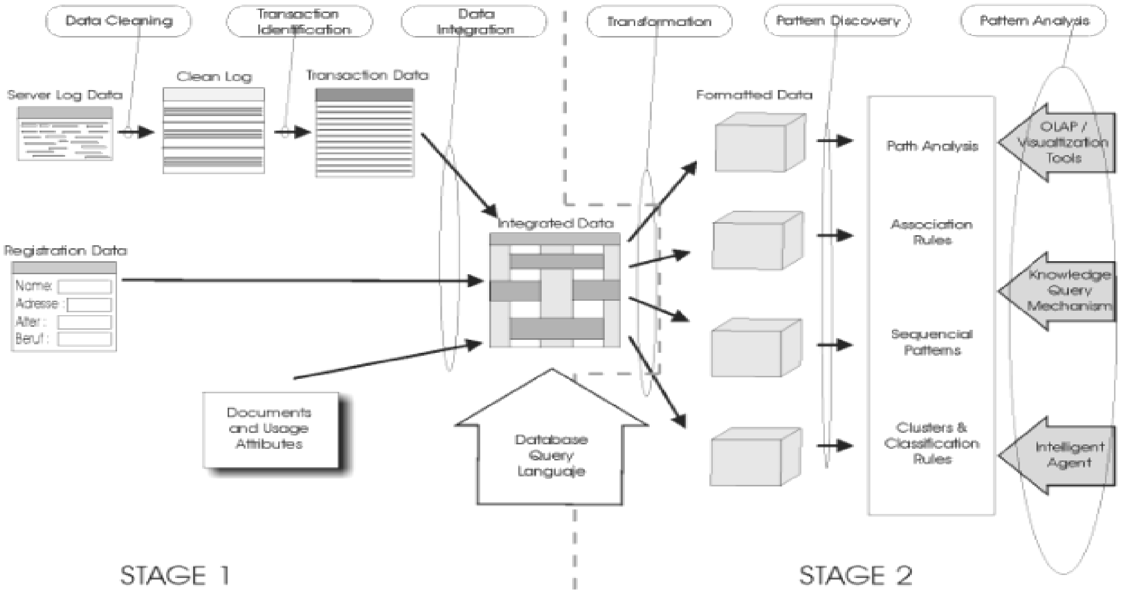


Figure 4: General Architecture of web usages mining

5. *CONCLUSION*

The term Web mining has been used to refer to techniques that encompass a broad range of issues. However while meaningful and attractive, this very broadness has caused Web mining to mean different things to different people and there is a need to develop a common vocabulary. Towards this goal we proposed a definition of Web mining and developed taxonomy of the various ongoing efforts related to it. We provided a general architecture of a system to do Web usage mining and identified the issues and problems in this area that require further research and development

References

[1]. RAYMOND KOSALA, HENDRIK BLOCKEEL, Web Mining Research: A Survey, Sigkdd Expirations, Acm Sigkdd, July 2000.

[2]. M. KOSHER. ALIKE - Archie-Like Indexing In The Web. In Proc. 1st International Conference On The World Wide Web, Pages 91--100, May 1994.

[3]. R. COOLEY, B. MOBASHER, AND J. SRIVASTAVA. Web Mining: Information And Pattern Discovery On The World Wide Web. In Proceedings Of The 9th Ieee International Conference On Tools With Artificial Intelligence (Ictai'97), 1997

[4]. R. KOSALA, H. BLOCKEEL. Web Mining Research: A Survey Data & Knowledge Engineering, Volume 53, Issue 3, June 2005, Pages 225-241

[5]. NASRAOUI, O. ET AL. , A Web Usage Mining Framework For Mining Evolving User Profiles In Dynamic Web Sites, Ieee Transactions On Knowledge And Data Engineering, Volume: 20 Issue:2 On Page(S): 202 – 215, 2008.

[6]. F. MASSEGLIA, ET AL. Web Usage Mining: Extracting Unexpected Periods From Web Logs, Data Mining And Knowledge Discovery Volume 16, Number 1, 39-65, 2007.

[7]. NAVEENA DEVI ET AL. Design And Implementation Of Web Usage Mining Intelligent System In The Field Of E-Commerce, Procedia Engineering Volume 30, 2012, Elsevier , Pp 20–27

[8]. MALIK, S.K. ET AL., Information Extraction Using Web Usage Mining, Web Scrapping And Semantic Annotation, In Procd. Of Ieee Cicn, 2011 Pp-465 - 469