WEB TRAFFIC AND LOG DATA ANALYSIS

[1]Abdul Hamid Qureshi, [2] Dr. Geetanjali Amarawat

[1] HOD Computer Science, [2] Associate Professor

[1]International Indian School, Jeddah , [2] Department of Computer Science and Engineering, Madhav University Abu Road Rajasthan

ABSTRACT

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely *pre-processing*, *pattern discovery*, and *pattern analysis*. This paper describes each of these phases in detail. Given its application potential, Web usage mining has seen a rapid increase in interest, from both the research and practice communities. This paper provides a detailed taxonomy of the work in this area, including research efforts as well as commercial offerings.

INTRODUCTION

The ease and speed with which business transactions can be carried out over the Web has been a key driving force in the rapid growth of electronic commerce. Specifically, ecommerce activity that involves the end user is undergoing a significant revolution. The ability to track users' browsing behaviour down to individual mouse clicks has brought the vendor and end customer closer than ever before. It is no possible for a vendor to personalize his product message for individual customers at a massive scale, a phenomenon that is being referred to as *mass customization*.

The scenario described above is one of many possible applications of *Web Usage mining*, which is *the process of applying data mining techniques to the discovery of usage patterns from Web data*, targeted towards various applications. Data mining efforts associated with the Web, called *Web mining*, can be broadly divided into three classes, i.e. content mining, usage mining, and structure mining.

This paper provides an up-to-date survey of Web Usage mining, including both academic and industrial research efforts, as well as commercial offerings.

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). 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. The principal kind of *inter-page* structure information is hyper-links connecting one page to another.
- 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. This includes registration data and customer profile information.

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, and single-site browsing behaviour 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 behaviour 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.

However, the site usage data recorded by server logs may not be entirely reliable due to the presence of various levels of caching within the Web environment. Cached page views are not recorded in a server log. In addition, arty important information passed through the POST method will not be available in a server log. Packet sniffing technology is an alternative method to collecting usage data through server logs. Packet sniffers monitor network traffic coming to a Web server and extract usage data directly from TCP/IP packets. The Web server can also store other kinds of usage information such as cookies and query data in separate logs. Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the site visitors. Tracking of individual users is not an easy task due to the stateless connection model of the HTTP protocol.

Cookies rely on implicit user cooperation and thus have raised growing concerns regarding user privacy, which will be discussed in Section 6. Query data is also typically generated by online visitors while searching for pages relevant to their information needs. Besides usage data, the server side also provides content data, structure information and Web page meta-information (such as the size of a file and its last modified time).

The Web server also relies on other utilities such as CGI scripts to handle data sent back from client browsers. Web servers implementing the CGI standard parse the URI 1 of the requested file to determine if it is an application program. The URI for CGI programs may contain additional parameter values to be passed to the CGI application. Once the CGI program has completed its execution, the Web server send the output of the CGI application back to the browser.

Client Level Collection

Uniform Resource Identifier (URI) is a more general definition that includes the commonly referred to Uniform Resource Locator (URL). Client-side data collection can be implemented by using a remote agent (such as JavaScript's 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 JavaScript's 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. JavaScript, 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 behaviour.

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 behaviour of a group of anonymous users, sharing a common proxy server.

Data Abstractions

The information provided by the data sources described above can all be used to construct/identify several data abstractions, notably *users, server sessions, episodes, clickstreams,* and *page views.* In order to provide some consistency in the way these terms are defined, the W3C Web Characterization Activity (WCA) has published a draft of Web term definitions relevant to analysing Web usage. A *user* is defined as a single individual that is accessing file from one or more Web servers through a browser. While this definition seems trivial, in practice it is very difficult to uniquely and repeatedly identify users. A user may access the Web through different machines, or use more than one agent on a single machine. A *page view* consists of every

file that contributes to the display on a user's browser at one time. Page views are usually associated with a single user action (such as a mouse-click) and can consist of several files such as frames, graphics, and scripts. When discussing and analysing user behaviours, it is really the aggregate page view that is of importance. The user does not explicitly ask for "n" frames and "m" graphics to be loaded into his or her browser, the user requests a "Web page." All of the information to determine which files constitute a page view is accessible from the Web server. A *click-stream* is a sequential series of page view requests. Again, the data available from the server side does not always provide enough information to reconstruct the full click-stream for a site. Any page view accessed through a client or proxy-level cache will not be "visible" from the server side. A *user session* is the click-stream of page views for a single user across the entire Web. Typically, only the portion of each user session that is accessing a specific site can be used for analysis, since access information is not publicly available from the vast majority of Web servers. The set of page-views in a user session for a particular Web site is referred to as a *server session* (also commonly referred to as a *visit)*. A set of server sessions is the necessary input for any Web Usage analysis or data mining tool. The end of a server session is defined as the point when the user's browsing session at that site has ended. Again, this is a simple concept that is very difficult to track reliably. Any semantically meaningful subset of a user or server session is referred to as an *episode* by the W3C WCA.

WEB USAGE MINING

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.

Pre-processing

Pre-processing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery.

Pattern Discovery

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. However, it is not the intent of this paper to describe all the available algorithms and techniques derived from these fields.

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.

TAXONOMY

Since 1996 there have been several research projects and commercial products that have analysed Web usage data for a number of different purposes. This section describes the dimensions and application areas that can be used to classify Web Usage Mining projects.

Taxonomy Dimensions

While the number of candidate dimensions that can be used to classify Web Usage Mining projects is many, there are five major dimensions that apply to every project - the data sources used to gather input, the types of input data, the number of users represented in each data set, the number of Web sites represented in each data set, and the application area focused on by the project. Usage data can either be gathered at the server level, proxy level, or client level. All projects analyse usage data and some also make use of content, structure, or profile data. The algorithms for a project can be designed to work on inputs representing one or many users and one or many Web sites. Single user projects are generally involved in the personalization application area. The projects that provide multi-site analysis use either client or proxy level input data in order to easily access usage data from more than one Web site. Most Web Usage Mining projects take single-site, multi-user, server-side usage data (Web server logs) as input.

PRIVACY ISSUES

Privacy is a sensitive topic which has been attracting a lot of attention recently due to rapid growth of e-commerce. It is further complicated by the global and self-regulatory nature of the Web. The issue of privacy revolves around the fact that most users want to maintain strict anonymity on the Web. They are extremely averse to the idea that someone is monitoring the Web sites they visit and the time they spend on those sites.

On the other hand, site administrators are interested in finding out the demographics of users as well as the usage statistics of different sections of their Web site. This information would allow them to improve the design of the Web site and would ensure that the content caters to the largest population of users visiting their site. The site administrators also want the ability to identify a user uniquely every time she visits the site, in order to personalize the Web site and improve the browsing experience.

The main challenge is to come up with guidelines and rules such that site administrators can perform various analyses on the usage data without compromising the identity of an individual user. Furthermore, there should be strict regulations to prevent the usage data from being exchanged/sold to other sites. The users should be made aware of the privacy policies followed by any given site, so that they can make an informed decision about revealing their personal data. The success of any such guidelines can only be guaranteed if they are backed up by a legal framework.

CONCLUSIONS

This paper has attempted to provide an up-to-date survey of the rapidly growing area of Web Usage mining. With the growth of Web-based applications, specifically electronic commerce, there is significant interest in analysing Web usage data to better understand Web usage, and apply the knowledge to better serve users. This has led to a number of commercial offerings for doing such analysis. However, Web Usage mining raises some hard scientific questions that must be answered before robust tools can be developed. This article has aimed at describing such challenges, and the hope is that the research community will take up the challenge of addressing them.

REFERENCES

1. Data mining: Crossing the chasm, 1999. Invited talk at the 5th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining(KDD99).
2. Charu C Aggarwal and Philip S Yu. On disk caching of web objects in proxy servers. In CIKM 97, pages 238-245, Las Vegas, Nevada, 1997.
3. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference, pages 487-499, Santiago, Chile, 1994.
4. Virgilio Almeida, Azer Bestavros, Mark Crovella, and Adriana de Oliveira. Characterizing reference locality in the www. Technical Report TR-96-11, Boston University, 1996.
5. Martin F Arlitt and Carey L Williamson. Internet web servers: Workload characterization and performance implications. 1EEE/A CM Transactions on Networking, 5(5):631-645, 1997.
6. M. Balabanovie and Y. Shoham. Learning information retrieval agents: Experiments with automated web browsing. In On-line Working Notes of the AAAI Spring Symposium Series on Information Gathering from Distributed, Heterogeneous Environments, 1995.
7. Alex Buchner and Maurice D Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. SIGMOD Record, 27(4):54-61, 1998.
8. L. Catledge and J. Pitkow. Characterizing browsing behaviors on the world wide web. Computer Networks and ISDN Systems, 27(6), 1995.
9. M.S. Chen, J. Hart, and P.S. Yu. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6):866- 883, 1996.
10. M.S. Chen, J.S. Park, and P.S. Yu. Data mining for path traversal patterns in a web environment. In 16th International Conference on Distributed Computing Systems, pages 385-392, 1996.
11. Roger Clarke. Internet privacy concerns conf the case for intervention. 42(2):60-67, 1999.
12. E. Cohen, B. Krishnamurthy, and J. Rexford. Improving end-to-end performance of the web using server volumes and proxy filters. In Proe. ACM SIGCOMM, pages 241-253, 1998.
13. Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Grouping web page references into transactions for mining world wide web browsing patterns. In Knowledge and Data Engineering Workshop, pages 2-9, Newport Beach, CA, 1997. IEEE.

14. Robert Codley, Bamshad Mobasher, and Jaideep Srivastava. Web mining: Information and pattern discovery on/th/e world wide web. In International Conference on Tools with Artificial Intelligence, pages 558- 567, Newport Beach, 1997. IEEE.
15. Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1(1), 1999.
16. Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. Discovery of interesting usage patterns from web data. Technical Report TR 99-022, University of Minnesota, 1999.
17. T. Fawcett and F. Provost. Activity monitoring: Noticing interesting changes in behavior. In Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 53-62, San Diego, CA, 1999. ACM.
18. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In Proc. ACM KDD, 1994.
19. David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. In Conference on Hypertext and Hypermedia. ACM, 1998.
20. Chi E. H., Pitkow J., Mackinlay J., Pirolli P., Gossweiler, and Card S. K. Visualizing the evolution of web ecologies. In CHI '98, Los Angeles, California, 1998.
21. Bernardo Huberman, Peter Pirolli, James Pitkow, and Rajan Kukose. Strong regularities in world wide web surfing. Technical report, Xerox PARC, 1998.
22. T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In The 15th International Conference on Artificial Intelligence, Nagoya, Japan, 1997.
23. Reagle Joseph and Cranor Lorrie Faith. The platform for privacy preferences. 42(2):48-55, 1999.
24. H. Lieberman. Letizia: An agent that assists web browsing. In Proe. of the 1995 International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995.
25. Stephen Lee Manley. An Analysis of Issues Facing World Wide Web Servers. Undergraduate, Harvard, 1997.
26. B. Masand and M. Spiliopoulou, editors. Workshop on Web Usage Analysis and User Profiling (WebKDD), 1999.
27. B. Mobasher, N. Jaln, E. Hart, and J. Srivastava. Web mining: Pattern discovery from world wide web transactions. (TR 96-050), 1996.
28. Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Creating adaptive web sites through usagebased clustering of urls. In Knowledge and Data Engineering Workshop, 1999.
29. Olfa Nasraoui, Raghu Krishnapuram, and Anupam Joshi. Mining web access logs using a fuzzy relational clustering algorithm based on a robust estimator. In Eighth International World Wide Web Conference, Toronto, Canada, 1999.