Web Mining Functions in an Academic Search Application

Jeyalatha SIVARAMAKRISHNAN, Vijayakumar BALAKRISHNAN
Faculty of Computer Science and Engineering,
BITS – PILANI, Dubai, U.A.E
jeylatha@yahoo.com, bv_uma@yahoo.com

This paper deals with Web mining and the different categories of Web mining like content, structure and usage mining. The application of Web mining in an academic search application has been discussed. The paper concludes with open problems related to Web mining. The present work can be a useful input to Web users, Web Administrators in a university environment.
***Keywords**: Database, HITS, IR, NLP, Web mining.*

1 Introduction

With the explosive growth of data available on the World Wide Web, discovery and analysis of useful information from the World Wide Web becomes a practical necessity. It has become increasingly necessary for users to utilize automated tools in finding the desired information resources and to track and analyze their usage patterns. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge.

Web mining can be broadly defined as the automated discovery and analysis of useful information from the web documents and services using data mining techniques. It is a large, interdisciplinary and dynamic scientific area, converging from several research communities such as database, information retrieval and artificial intelligence (especially from machine learning and natural language processing).

There are several important issues, unique to the Web paradigm that comes into play if sophisticated types of analyses are to be done on server side data collections. These include integrating various data sources such as server access logs, user registration or profile information; resolving difficulties in the identification of users due to the missing unique key attributes in collected data and the importance of identifying user sessions or transactions from usage data, site topologies and models of user behavior.

Web mining methodologies can generally be classified into one of three distinct categories: Web usage mining, Web structure mining and Web content mining. Web usage mining is also known as Web log mining, is the process of extracting interesting patterns in Web access logs. It analyzes navigational activities of web users. Web structure mining is the process of inferring knowledge from the World Wide Web. Web content mining is the process of extracting knowledge from the content of documents or their description, available on the World Wide Web.

2 Objectives and Motivation

The present work is intended to meet the following objectives:

1. To specify the functions for Web content, structure and usage mining.
2. To identify the Academic Search related functions where Web mining can be applied effectively.

The World Wide Web acts as a large repository of data. It is very much necessary to analyze the navigational activities of users and also extract meaningful information from online internet sites. The present research work is motivated by the increasing importance of Web mining in the contemporary internet applications and the main area of interest will be Academic Search Application. An Academic Search application involves retrieving voluminous information from the web, analyzing the web usage patterns of different users, document structure and hyperlinks.

3 Related Work

The World Wide Web (WWW) continues to grow at an astounding rate in both the sheer volume of traffic and the size and complexity of web sites. The complexity of tasks such as Web site design, Web server design and of simply navigating through a web site have increased along with this growth [1]. An important input to these design tasks is the analysis of how a web site is being used. The World Wide Web provides every internet user with access to an abundance of information, but it becomes increasingly difficult to identify the relevant pieces of information. Research in web mining tries to address this problem by applying techniques from data mining and machine learning to Web data and documents. When a user visits a website, behind the scenes the user leaves his impressions, usage patterns and also access patterns in the web servers log file [2]. The amount of Web information is growing rapidly and improving the efficiency and accuracy of Web information retrieval are very much essential. There are two fundamental issues regarding the effectiveness of Web information gathering: information mismatch and overload [3]. Web mining extends the ideas of Data Mining for handling web based data.

There exist several definitions with regard to Web mining in the literature. Web mining is the application of data mining techniques to discover patterns and analysis of useful information from the Web [4]. Web mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web [5]. Web mining can be defined as the automated discovery and analysis of useful information from the web documents and services using data mining techniques [6]. It discovers potentially useful and previously unknown information or knowledge from web data.

Data mining can be regarded as “database mining” and “knowledge discovery in databases” or KDD. Data mining involves extracting valid and interesting relationships from large collections of data using intelligent methods and statistical techniques [7].

The World Wide Web (WWW) is a popular and interactive medium to disseminate information today. The Web is huge, diverse and dynamic and thus raises the scalability, multimedia data and temporal issues respectively. This results in information overload. To be able to cope with the abundance of available information, users of the Web need assistance of intelligent software agents for finding, sorting and filtering the available information [8].

The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce and many other information services [9]. The web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information, providing rich sources for data mining [10]. However based on the following observations, the Web also poses great challenges for effective resource and knowledge discovery.

1. The Web seems to be too huge for effective data warehousing and data mining.
2. The complexity of Web pages is far greater than that of any traditional text document collection.
3. The Web is a highly dynamic information source.
4. The Web serves a broad diversity of user communities.
5. Only a small portion of the information on the Web is truly relevant or useful.

There are many index-based Web search engines that search the Web, index Web pages and build and store huge keyword-based indices that help locate sets of Web pages containing certain keywords.

The present work can act as input to the Web Administrators to analyze access patterns of Web content by different types of users.

4 Web Mining Categories

Web mining is an emerging area in Web Intelligence. Currently, a Web mining system can be viewed as the use of data mining

techniques to automatically retrieve, extract, generalize, and analyst information on the Web. Web mining is said to have three operations of interests – clustering (e.g. finding natural groupings of users, pages, etc), associations (e.g. which URLs tend to be requested together), and sequential analysis (e.g. the order in which URLs tend to be accessed). Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

Web mining can be decomposed into these subtasks, namely:

1. *Resource finding*: the task of retrieving intended Web documents.
2. *Information selection and pre-processing*: automatically selecting and pre-processing specific information from retrieved Web resources.
3. *Generalization*: automatically discovers general patterns at individual Web sites as well as across the multiple sites.
4. *Analysis*: validation and / or interpretation of the mined patterns.

By resource finding we mean the process of retrieving the data that is either online or offline from the web sources like text, relational data and semi structural data like XML. The information selection and pre-processing step is any kind of transformation processes of the original data retrieved in the IR process. Machine learning or data mining techniques are used for generalization. Humans play an important role in the information or *knowledge discovery process* on the Web since the web is an interactive medium. This is especially important for validation and/ or interpretation.

Google Scholar is a freely accessible Web search engine that indexes the full text of scholarly literature across an array of publishing formats and disciplines. Google Scholar Index includes most peer-reviewed online journals of the world's largest scientific publishers. Google Scholar allows users to search for digital or physical copies of articles, whether they be online or in libraries. It is a subset of the larger Google search index, consisting of full-text journal articles, technical reports, preprints, theses, books, and other documents, including selected Web pages that are deemed to be "scholarly" [11]. It retrieves document or page matches based on the keywords searched and then organizes the results. Web mining can be divided into three areas of interest based on which part of the Web to mine. They are Web content mining, Web structure mining and Web usage mining.

4.1 Web Content Mining

Web content mining [8] is the process to discover useful information from the content of a web page. The technologies that are normally used in web content mining are NLP (Natural Language Processing) and IR (Information Retrieval). Web content mining describes the discovery of useful information from the Web contents/ data/ documents. Web contents encompass a very broad range of data. Basically, the Web content consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks. The present work considers unstructured data such as free texts, semi-structured data such as HTML documents, and a more structured data such as data in the tables or database generated HTML pages. Web content mining can be viewed from two different points of view: IR (Information Retrieval) and DB (Database) views. The goal of Web content mining from the IR view is mainly to assist or to improve the information finding or filtering the information to the users usually based on either inferred or solicited user profiles, while the goal of Web content mining from the DB view mainly tries to model the data on the Web and to integrate them so that more sophisticated queries other than the keywords based search could be performed.

4.2 Web Structure Mining

Web Structure Mining [8] is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges

connecting related pages. Web Structure mining is the process of using graph theory to analyze the node and connection structure of a web site. It is used to discover structure information from the web and it can be divided into two kinds based on the kind of structure information used. They are *Hyperlinks* and *Document Structure* [12].

The first kind of web structure mining is extracting patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location. The other kind of the web structure mining is mining the document structure. It is using the tree-like structure to analyze and describe the HTML (Hyper Text Markup Language) or XML (eXtensible Markup Language) tags within the web page.

Web structure mining tries to discover the model underlying the link structures of the Web. The model is based on the topology of the hyperlinks with or without the description of the links. This model can be used to categorize Web pages and is useful to generate information such as the similarity and relationship between different Web sites. Web structure mining could be used to discover authority sites for the subjects (authorities) and overview sites for the subjects that point to many authorities (hubs). Web topology has been modeled using algorithms such as HITS (Hyperlink Induced Topic Search), Page Rank and CLEVER. These models are mainly applied as a method to calculate the quality rank or relevancy of each web page. Some applications of web structure mining include measuring the completeness of web sites by measuring the frequency of local links that reside on the same server, measuring the replication of web documents across the web warehouse (which helps in identifying for example mirrored sites), and discovering the nature of the links hierarchy in the web sites of a particular domain to study how the flow of information affects their design.

4.3 Web Usage Mining

Web usage mining [8] is also known as Web log mining, is the process of extracting interesting patterns in web access logs. Web Usage Mining is the application of data mining techniques to usage logs of large web data repositories in order to produce results than can be used in the design tasks.

In web usage mining, an application uses data mining to analyze and discover interesting patterns of user's usage data on the web. The usage data records the user's behavior when the user browses or makes transactions on the web site. It is an activity that involves the automatic discovery of patterns from one or more Web servers. Organizations often generate and collect large volumes of data; most of this information is usually generated automatically by Web servers and collected in server log. Analyzing such data can help these organizations to determine the value of particular customers, cross marketing strategies across products and the effectiveness of promotional campaigns, etc. The first web analysis tools simply provided mechanisms to report user activity as recorded in the servers. Using such tools, it was possible to determine such information as the number of accesses to the servers, the times or time intervals of visits and the domain names and the URLs of users of the Web server. However, in general these tools provide little or no analysis of data relationships among the accessed fields and directories within the Web space. Now more sophisticated techniques for discovery and analysis of patterns are emerging. These tools fall into to two main categories: Pattern Discovery Tools and Pattern Analysis Tools [13].

Web usage mining tries to analyze the sessions and access patterns of web users. While the Web content and structure mining utilize the real or primary data on the Web, Web usage mining mines the secondary data derived from the interactions of the users while interacting with the Web. The Web usage data includes the data from Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, book mark data, mouse clicks and scrolls, and any

other data as the results of interactions.
Web usage mining facilitates the discovery of Web access patterns. With Web usage mining, the user log can be analyzed. Some patterns about user behavior can be obtained from usage logs and then these patterns can be turned into a user profile. The user profiles are then utilized to filter incoming articles for the individual. The profiles can be constructed using a variety of learning techniques including the vector space model, genetic algorithm and the probabilistic model or clustering. A System Administrator can extract and analyze data recorded in web server log files by means of a script / program. There are a lot of commercial tools and programs intended to extract and analyze data recorded in Web server log files.
Web usage mining can be classified depending on the kind of usage data [14]. They are:

1. Web Server Data
   User logs are collected by the web server and typically include IP address, page reference and access time.
2. Application Server Data
   Commercial application servers such as Weblogic, StoryServer have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
3. Application Level Data
   New kinds of events can be defined in an application and logging can be turned on for them- generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above categories [15].

The main steps in Web usage mining comprises of data preparation, pattern discovery, pattern analysis and visualization [16, 17]. While extracting simple information from web logs is easy, mining complex structural information is very challenging. Data cleaning and preparation constitute a very significant effort before mining can even be applied [18]. The relevant data challenges include: elimination of classified into two main categories;

- Learning a user profile, or user modeling in adaptive interfaces (personalized).
- Learning user navigation patterns (impersonalized).

Web based learning environments are now extensively used as integral components of course delivery in tertiary education. To provide an effective learning environment, it is important that educators understand how these environments are used by their students [19].
A Web server access log contains a complete history of file accesses by clients. Most WWW access logs follow the Common Log Format specified as part of the HTTP protocol [20, 21]. A log entry, following this standard, contains the client IP address, user id, access time, request method, and the URL of the page accessed, the protocol used for data transmission, an error code, and the number of bytes transmitted.

5 Academic Search Application: Main Functions

A University consists of heterogeneous users such as students, faculty, System Administrator and Librarian. Each category of user has specific requirements while searching information on the World Wide Web. Web mining helps greatly in achieving the user's preferences.
Web mining facilitates in identification of workgroups and Special Interest Groups in various universities across the world. This will help faculty and students in confining the search to a particular group like Database Systems, Network Systems, Power Electronics and so on and get quick results, relating to technical papers and articles. Web mining helps in building an updated information base for easy access by students and faculty belonging to a specific research group. The information base contains listings of and links to conferences and upcoming events in a particular area such as Software Engineering, Database Systems and so on.

Web mining helps students for extracting information on higher education such as programs, courses, specialization, fee structure and stipend offered in various universities across the world and facilitates easy access from the updated information base. This will help students considerably in terms of saving time and effort in the search process.

Web Log files record useful information about Web usage in a University's Web Server. The Log file can be analyzed over a time period. The time period can be specified on hourly, daily, weekly and monthly basis. The following information can be gathered from the log files of an University's website:

- General summary on number of visitors and accessibility.
- Total Hits on number of pages/files accessed or attempted to be accessed.
- The source websites of the visitors.
- The browser used by the visitor to access the website.
- Error reports for identifying the problems and solving them.
- Report on file size, file type and directory/subdirectory visited.

6 Open Problems

With the fast increase in Web activities, Web data mining has recently become an important research topic and is receiving a significant amount of interest from both academic and industrial environments. While existing methods are efficient for the mining of frequent path traversal patterns from the access information contained in a log file, these approaches involve considerable overhead in evaluating associations in a given set of data [22].

The search engines have to greater extent been successful for practical applications. However, there exist some problems relating to retrieval accuracy. Web server logs provide a rich source of information about how users access a site. When users access a sequence of web pages, they usually have a particular information seeking task in mind. The degree of accuracy is decreased when the user follows more links [23, 24].

In the present days, the main problems encountered in web search are Information mismatch and voluminous results. The results often frustrate and consume precious time of the users. Most of the existing search engines perform searches that are keyword based. They get the search query from the users, search for the presence of keywords in the web documents, rank them and display the results to the user. However, it is also essential to consider the semantics of the user query and precise requirements of the user to yield reasonably good results.

We either browse or use the search service when we want to find specific information on the web. We usually specify a simple keyword query and the response from a web search engine is a list of pages, ranked based on their similarity to the query. However, today's search tools have the following problems:

1. Low precision: This is due to the irrelevance of many of the search results. We may get many pages of information which are not really relevant to our query.
2. Low recall: This is due to the inability to index all the information available on the web. Because some of the relevant pages are not properly indexed, we may not get those pages through any of the search engines.

Some of the problems encountered by the information users when interacting with the Web are finding relevant information, creating new knowledge out of the information available on the Web and Personalization of the information. The above problem can be termed as a query triggered process (retrieval oriented). Web mining techniques address the above mentioned issues.

7 Conclusion

This paper analyses the functions pertaining to Web Usage mining, Web Structure mining and Web Content mining. The application of web mining techniques for academic search application has been highlighted. Web mining continues to remain as a potential research area in the present scenario. Web Administrators and web users in an academic

environment are required to familiarize with Web mining techniques, so that they can analyze large volumes of web data more effectively.

References

[1] R. Cooley, B. Mobasher and J. Srivastava, "Data preparation for Mining World Wide Web Browsing Patterns," *Journal of Knowledge and Information* Systems, Vol. 1, 1999, pp. 5-32.

[2] P. Madiraju, Y. Q. Zhang, S. Owen and Z. Y. Sunderraman, "Graphical Web Mining Agent for Class Teaching Enhancement," *International Journal for Infonomics*, Vol. 3, 2006, pp. 243-249.

[3] X. Zhou, Y. Li, P. Bruza, S. T. Wu and Y. Xu, "Using Information Filtering in Web Data Mining Process," *Proceedings of the ACM International Conference on Web Intelligence*, Silicon Valley, California, USA, 2007, Nov. 2-5, pp. 163-169.

[4] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining : Information and Pattern Discovery on the World Wide Web," *ACM SIGKDD Explorations Newsletter*, 2000, Vol. 1, pp. 12-23.

[5] P. Galeas, *Web Mining.* [serial on the Internet]. [cited 2005 September 18]. Available at: http://www.galeas.de/webmining.html

[6] K. Markellos, P. Markellou, M. Rigou and S. Sirmakessis, "Web Mining: Past, Present and Future," *Proceedings of the NEMIS Launch Conference*, Patras, Greece, 2003, Apr. 5, pp. 26-36.

[7] B. Mobasher, N. Jain, E. H. Han and J. Srivastava, *Web Mining: Pattern Discovery from World Wide Web Transactions*. Minneapolis (MS): University of Minnesota, Department of Computer Science; 1996 Feb. Report No.: TR 96-050.

[8] R. Kosala, H. Blockeel, "Web Mining Research: A Survey," *ACM SIGKDD Explorations*, 2000, Vol. 2, pp. 1-15.

[9] J. Han and M. Kamber, *Data Mining Concepts and Techniques.* Morgan Kaufmann, San Fransisco, USA., 2001, pp. 435-441.

[10] M. M. Gore and A. K. Mishra, "Algorithm for Data Mining [CD-ROM]," *Proceedings of Winter School on Data Mining*, Allahabad, India, 2001.

[11] S. Sullivan, *Google Scholar offers Access to Academic Information.* [serial on the Internet]. [cited 2004 November 18]. Available at: http://searchenginewatch.com

[12] A. K. Pujari, *Data Mining Techniques.* Universities Press India, Private Limited, Hyderabad, India, 2001, pp. 231- 239.

[13] L. Borzemski, "The Usage of Data Mining to predict Web performance," *Cybernetics & Systems*, 2006, Vol. 37, pp. 587-608.

[14] S. Choenni, "Design and Implementation of a Genetic based Algorithm for Data Mining," *Proceedings of the 26th International Conference on Very Large Data Bases*, Cairo, Egypt, 2000, Sep. 10-14, pp. 33-42.

[15] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *ACM SIGKDD Explorations*, 2000, Vol. 2, pp. 12-23.

[16] J. R. F. Boullossa and G. Xexeo, *An Architecture for Web Usage Mining* [serial on the Internet]. [cited 2002 July 16]. Available at: http://www.boullosa.org/artigos/Arch-web-mining-2002.pdf

[17] J. R. Punin, M. S. Krishnamoorthy and M. J. Zaki, *Web Usage Mining: Languages and Algorithms*. Troy (NY): Rensselaer Polytechnic Institute, Department of Computer Science, 2001 Jan. Report, pp. 99-100.

[18] J. Ceddia, J. Sheard and G. Tibbey, "WAT – A Tool for Classifying Learning Activities from a Log File," *Proceedings of the Ninth Australasian*

Computing Education Conference (ACE2007), Ballarat, Australia, 2007, Jan. 30 – Feb. 2, pp. 11-18.

[19] B. Mobasher, R. Cooley and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," *Communications of the ACM*, 2000, Vol. 43, pp. 142-151.

[20] S. Sendhilkumar and T. V. Geetha, "Personalized ontology for Web Searach Personalization," *Proceedings of the 1st Bangalore Annual Computer Conference*, Bangalore, India, 2008, Jan. 18-20, pp. 1-7.

[21] J. C. Ou, C. H. Lee and M. S. Chen, "Efficient algorithms for incremental Web log mining with dynamic thresholds," *The VLDB Journal – The International Journal on Very Large Data Bases*, 2008, Vol. 17, pp. 827-845.

[22] C. Ding and J. Zhou, "Log Based Indexing to Improve Web Site Search," *Proceedings of the ACM Symposium on Applied Computing*, Seoul, Korea, 2007, Mar 11-15, pp. 829 -833.

[23] J. Zhou, C. Ding and D. Androutsos, *Improving Web site search using Web Server Logs*. 2006 Oct [cited 2006 Nov] Available at: http://delivery.acm.org/10.1145/1190000/1188996/html

[24] H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, *Data Mining: Next Generation Challenges and Future Directions*. Prentice Hall of India Private Limited, New Delhi, India, 2005, pp. 405-408.

Jeyalatha SIVARAMAKRISHNAN is a PhD student in Computer Science at BITS-Pilani, Dubai and Faculty of Computer Science and Engineering. Her area of research includes Web Mining Algorithms and she is actively associated with Microprocessors and Advanced Computing Laboratory.

Vijayakumar BALAKRISHNAN holds a PhD in Computer Science from BITS-Pilani, India from 2001 and has 18 years of teaching experience in Computer Science and Engineering and 6 years of Computer Industry. Currently he is Associate Professor, Computer Science & Engineering, BITS-Pilani, Dubai. He is a member of Professional bodies such as ISTE, LINUX Users Group in Dubai and organizing and judging committee member in UAE for technical events such as National Programming Contest and Technofest during the last 6 years.