Web Usage Mining for a Better Web-Based Learning Environment

Osmar R. Zaïane
Department of Computing Science, University of Alberta
Edmonton, Alberta, Canada
zaiane@cs.ualberta.ca

Abstract

Web-based technology is often the technology of choice for distance education given the ease of use of the tools to browse the resources on the Web, the relative affordability of accessing the ubiquitous Web, and the simplicity of deploying and maintaining resources on the World-Wide Web. Many sophisticated web-based learning environments have been developed and are in use around the world. The same technology is being used for electronic commerce and has become extremely popular. However, while there are clever tools developed to understand on-line customer's behaviours in order to increase sales and profit, there is very little done to automatically discover access patterns to understand learners' behaviour on web-based distance learning. Educators, using on-line learning environments and tools, have very little support to evaluate learners' activities and discriminate between different learners' on-line behaviours. In this paper, we discuss some data mining and machine learning techniques that could be used to enhance web-based learning environments for the educator to better evaluate the leaning process, as well as for the learners to help them in their learning endeavour.

1. Introduction and background

With the rapid development of the World-Wide Web (WWW), the increased popularity and ease of use of its tools, the World-Wide Web is becoming the most important media for collecting, sharing and distributing information. Many organizations and corporations provide information and services on the Web such as automated customer support, on-line shopping, and a myriad of resources and applications. Web-based applications and environments for electronic commerce, distance education, on-line collaboration, news broadcasts, etc., are becoming common practice and widespread. The WWW is becoming ubiquitous and an ordinary tool for everyday activities of common people, from a child sharing music files with friends to a senior receiving photographs and messages from grandchildren across the world. It is typical to see web pages for courses in all fields taught at universities and colleges providing course notes and related resources even if these courses are delivered in traditional classrooms. It is not surprising that the Web is the means of choice to architect modern advanced distance education systems. Distance education is a field where web-based technology was very quickly adopted and used for course delivery and knowledge sharing. Typical web-based learning environments such as Virtual-U [5] and Web-CT [13] include course content delivery tools, synchronous and asynchronous conferencing systems, polling and quiz modules, virtual workspaces for sharing resources, white boards, grade reporting systems, logbooks, assignment submission components, etc. In a virtual classroom, educators provide resources such as text, multimedia and simulations, and moderate and animate discussions. Remote learners are encouraged to peruse the resources and participate in activities. However, it is very difficult and time consuming for educators to thoroughly track and assess all the activities performed by all learners on all these tools. Moreover, it is hard to evaluate the structure of the course content and its effectiveness on the learning process. Resource providers do their best to structure the content assuming its efficacy. Educators, using Web-based learning environments, are in desperate need for non-intrusive and automatic ways to get objective feedback from learners in order to better follow the learning process and appraise the on-line course structure effectiveness. On the learner's side, it would be very useful if the system could automatically guide the learner's activities and intelligently recommend on-line activities or resources that would favour and improve the learning. These tools do not exist yet and to the best of our knowledge there is no distance leaning system to date that provides such automated facilities either on the learner’s side or educator’s side. In the field of electronic commerce, however, given the lucrative prospects, a significant research effort has been made to devise elaborate methods to take advantage of customers’ accesses and purchase behaviours in order to enhance the purchasing experience and customer satisfaction by user profiling and smart recommendations, and thus increase profit. For example, systems for recommendation such as Amazon.com that suggests books to purchase related to a current purchase based on preference information and similar user purchases, or recommendation of movies with moviefinder.com, use collaborative filtering which predicts a person’s preferences as a linear weighted combination of other people’s preferences. Recently, researchers have used web access history to make web sites more adaptive and personalized and hence more attractive to visitors, which is critical to keep customers loyal. WUM [8] is a special web sequence analyser for improving web pages layout and structure based on the history of access sequences. Entire conferences and workshops have been dedicated to web usage analysis for the benefit of e-commerce [10,11,12]. While the analogy with e-commerce seems straight forward, it is certainly not as simple as it appears. It is true that in e-commerce the goal is to increase sales and profit and it is achieved by understanding customer access behaviour [2], and in e-learning the goal might be to improve the learning and it could be also achieved by understanding learners’ access patterns. However, many concepts involved are fundamentally different. For instance a purchase transaction, hence a session, which is a fundamental building block for most web usage mining algorithms, is somehow defined starting from a initial access to the web site to a purchasing or order operation, usually in a very short time frame (i.e. the same access session). In e-learning, a learning session can span many access sessions. To learn a concept or attain an exact result in a quiz, many access sessions, spread over many days and even weeks may be needed. Moreover, while the goal in e-commerce sites may be clear, for example encouraging the customers to buy more products and keeping them loyal, the goals in e-learning are vague, difficult to qualify/quantify and subjective.

Web-based course delivery systems, like any web site or web-based application, rely on web servers to provide access to resources and applications. Every single request that a Web server receives is recorded in an access log mainly registering the origin of the request, a time stamp and the resource requested, whether the request is for a web page containing an article from a course chapter, the answer to an on-line exam question, or a participation in an on-line conference discussion. The web log provides a raw trace of the learners' navigation and activities on the site. In order to process these log entries and extract valuable patterns that could be used to enhance the learning system or help in the learning evaluation, a significant cleaning and transformation phase needs to take place so as to prepare the information for data mining algorithms. The following section presents the issues related to web log cleaning and transformation. Section 3 enumerates some important data mining tasks that can be adopted in web usage mining. Section 4 illustrates with examples how web usage mining can be useful to enhance web-based learning environments. Finally, Section 5 presents some concluding remarks.

2. Web Log Cleansing

There is an assortment of web log analysis tools available [2]. Most of them, like NetTracker, webtrends, analog and SurfAid, etc., provide limited statistical analysis of web log data [16]. For example, a typical report has entries of the form: "during this time period t, there were *n* clicks occurring for this particular web page p". However, the results provided by these tools are limited in their abilities to help understand the implicit usage information and hidden trends. New products use more sophisticated and complex analytic means but are generic, require important manual intervention and often resort to sampling due to the huge size of web logs [2]. The most commonly used method to evaluate access to web resources or user interest in resources is by counting page accesses or “hits”. However, this is not sufficient and often not correct. Web server log files of current common web servers contain insufficient data upon which to base thorough analysis. However, they contain useful data from which a well-designed data mining system can discover beneficial information. Web server log files customarily contain: the domain name (or IP address) of the request; the user name of the user who generated the request (if applicable); the date and time of the request; the method of the request (GET or POST); the name of the file requested; the result of the request (success, failure, error, etc.); the size of the data sent back; the URL of the referring page; the identification of the client agent; and a cookie, a sting of data generated by an application and exchanged between the client and the server. A log entry is automatically added each time a request for a resource reaches the web server. While this may reflect the actual use of the resources on a site, it does not record reader behaviours like frequent backtracking or frequent reloading of the same resource when the resource is cached by the browser or a proxy. It is important to note that the entries of all users are mixed in the log, simply ordered chronologically even though one single page request from a user may generate multiple entries in the server log. One major problem in web log mining is to identify unique users and associate users with their access log entries. In e-learning applications, however, the problem is simplified since users are not anonymous but need to login to the system as registered learners. However, identifying sessions is a non-trivial task. The goal is to identify sequences of activities from the collection of mixed log entries as described above, and model them as sessions of learning activities to be presented to the educators for evaluation and interpretation, or forwarded to advanced data mining tools to further discover intrinsic useful patterns. The major steps for web log data transformation can be summarized as follows:

- Remove irrelevant entries
- Identify access sessions
- Map access log entries to learning activities
- Complete traversal paths
- Group access sessions by learner to identify learning sessions
- Integrate with other data about learners and groups of learners

Removing irrelevant entries is the simple task of weeding out requests for images, for example, or non-user requests such as web crawler requests etc. Identifying sessions is a demanding task. The aim is to recognize sequences of events such as A➔B➔C➔B➔D... where A, B, C, D, etc. are page or script requests. The challenge is to recognize the beginning and the end of sessions. The problem comes from the fact that HTTP, the protocol used for information exchange between web servers and browsers, is stateless and does not keep track of semantic sessions. In e-commerce applications, the end of sessions are usually the purchase of a product or the checkout of an e-cart, and idle times between requests that exceed 25 to 30 minutes are use to identify cuts between sessions. This heuristic is not necessarily true in the on-line learning context since learners can wander in other sites gathering relevant information while their access session at the e-learning site is still on hold. Moreover a learning session can span over days with different accesses. Many pages in e-learning applications are dynamically generated by script requests such as quiz pages, conference messages, etc. Mapping access log entries with actual learning activities consists of replacing script calls with their assigned parameter values with concrete activities. This is an arduous task that assumes thorough knowledge of the application scripts and their respective parameters and requires a mapping table provided by the application designers.

The result is a sequence of learners' relevant on-line activities of the form: Login➔ExerciseList➔SubmissionQuiz1➔ExerciseList➔ReadConferenceMessage… Completing the traversal paths consists of inferring cache hits and proxy meddling based on the structure of the web site and how pages and activities are effectively linked together. Finally, integrating the cleaned click streams with existing data about learners can be very valuable and beneficial. Such data could be the profiles of the learners, their quantitative and qualitative evaluations, etc. For instance combining the grades associated with completed activities with the sequences of events leading to these activities can help discover appropriate patterns that can help discriminate between sequence of activities that yield good results and sequence of events that are not as effective.

The web log cleaning and transformation phase often consumes 80% to 95% of the effort and resources needed for web usage mining [2]. The result of the pre-processing is a database of sets of pertinent activity sequences grouped by learner. This is usually modeled with sequences of tokens associated with user identification stored in flat files that current data mining algorithms can act upon. The information can also be stored in a data warehouse like in [16] allowing ad-hoc online analytical processing. The other two phases after data pre-processing are pattern discovery using intricate data mining algorithms, and pattern evaluation [9].

3. Useful Data Mining Tasks

What is needed are summarization trends and patterns that can be interpreted by educators delivering their courses on-line. Due to the importance of e-commerce and the lucrative opportunities behind understanding on-line customer purchasing behaviours, there is tremendous research effort in developing data mining algorithms and systems tailored for e-business related web usage data mining [4]. In addition to descriptive statistical analysis provided by most web access log analysis tools such as calculating hit frequency, average, median, etc., length and duration of sessions and other limited low-level statistical measures, there have been some data mining approaches adapted specifically for web usage mining. The most used methods are association rules mining, clustering, classification, sequential pattern analysis and dependency modeling [9], as well as prediction. These techniques are primarily used for personalization, system improvement such as web caching and network traffic improvements, site modification, and marketing intelligence [9]. None of these applications, however, was tailored to distance learning, but the methods are general enough that e-learning systems could benefit from them. Association rules generation is the discovery of relationships between items in transactions. It is typically used for market basket analysis to discover rules of the form “x% of customers who buy item A and B also buy item C.” Clustering is an unsupervised grouping of objects, while classification is a supervised grouping. In web mining, the objects could be users, events, sessions, pages, etc. Sequential pattern analysis is similar to association rules but takes into account the sequences of events. In other words, the fact that a page A is requested before another page B is captured in the patterns discovered. All these techniques were designed for knowledge discovery from very large databases of numerical data [6] and were adapted for web mining and applied in on-line business with relative success.

4. Enhancing Web-Based Learning Environments

WebSIFT [1] is a set of comprehensive web usage tools that is able to perform many data mining tasks and discover a variety of patterns from web logs. A versatile system, WebLogMiner [16], uses data warehousing technology for pattern discovery and trend summarization from web logs. However these wide-ranging tools are not integrated in e-learning systems and it is cumbersome for an educator who doesn’t have extensive knowledge in data mining to use these tools to improve the effectiveness of web-based learning environments. A new web usage mining system dedicated for e-learning is being developed to allow educators to assess on-line learning activities [15]. For an educator using a web-based course delivery environment, it could be beneficial to track the activities happening in the course web site and extract patterns and behaviours prompting needs to change, improve, or adapt the course contents. For example, one could identify the paths frequently and regularly visited, the paths never visited, the clusters of learners based on the paths they follow, etc. For a learner using a web-based course delivery environment, it could be beneficial to receive hints from the system on what subsequent activity to perform based on similar behaviour by other “successful” learners. For example, the system could suggest shortcuts to frequently visited pages based on previous user activities, or suggest activities that made similar learners more “successful”. It could also be beneficial if the system adapts the course content logical structure to the learner’s learning pace, interest, or previous behaviour. Web-based course content is not always presented and structured in an intuitive way. By analyzing common traversal paths of the course content web pages or frequent changes in individual traversal paths, the layout of the course can be reorganized or adapted to better fit the needs of a group or an individual. We see two types of data mining in the context of e-learning: off-line web usage mining and integrated web usage mining. Off-line web usage mining is the discovery of patterns with a standalone application. This pattern discovery process would allow educators to assess the access behaviours, validate the learning models used, evaluate the learning activities, compare learners and their access patterns, etc. We have designed and implemented a prototype of such an application as a tool for educators to apply association rules to discover relationships between learning activities that learners perform, sequential analysis to discover interesting patterns in the sequences of on-line activities, and clustering to group similar access behaviours [15]. While most data mining algorithms need specific parameters and threshold values to tune the discovery process, the users of web usage mining applications in the context of e-learning, namely educators and e-learning site designers, are not necessarily savvy in the intricate complexities of data mining algorithms. For this purpose we have tried to design new algorithms that need minimum input from the user and automatically adjust to the web log data at hand. In [3] we propose a totally non-parametric approach for clustering web sessions. Off-line web usage mining helps educators put in question and validate the learning models they use as well as the structure of the web site as it is perused by the learners. In contrast, integrated web usage mining is a process of discovering patterns that is incorporated with the e-learning application. This encompasses adaptive web sites, personalization of activities, and automatic recommenders that suggest activities to learners based on their preferences as well as their history of activities and the access patterns discovered from the communal accesses. We are currently designing a recommender based association rule mining similar to the text categorization we developed in [14]. The idea consists of discovering relevant associations between learning activities and generating association rules that are applied in real time when in a current session the activities of the antecedent of a rule are verified then the activities in the consequent of the rule are suggested to the learner as the recommended next step in the learning session. The algorithm for text categorization presented in [14] can also be used to automatically categorize learners' messages sent on an asynchronous conferencing system in order to help the educators better assess the information exchange in a course related forum.

5. Conclusions and Future Work

The Web is an excellent tool to deliver on-line courses in the context of distance education. However, counting only on web traffic statistical analysis does not take advantage in the potential of hidden patterns inside the web logs. Web usage mining is a non-trivial process of extracting useful implicit and previously unknown patterns from the usage of the Web. Significant research is invested to discover these useful patterns to increase profitability of e-commerce sites. However, the goals of these applications and methods, “turning visitors into purchasers”, are different from the goals in e-learning: “turning learners into effective better learners.” We have seen some examples where data mining techniques can enhance on-line education for the educators as well as the learners.

While some tools using data mining techniques to help educators and learners are being developed, the research is still in its infancy. In addition, with the awareness of the potential advantages of integrated web usage mining and the insufficient data recorded by web servers, there is a need for more specialized logs from the application side to enrich the information already logged by the web server. This added value by specific event recording on the e-learning side will give clicksteams and the patterns discovered a better meaning and interpretation.

References

[1] R. Cooley, B. Mobasher, J. Srivastava, *Web Mining: Information and Pattern Discovery on the World Wide Web*, Proceedings of the ninth IEEE international conference on Tools with AI, 1997.

[2] H. A. Edelstein, *Pan for Gold in the Clickstream*, Informationweek, March 2001, http://www.informationweek.com/828/mining.htm

[3] A. Foss, W. Wang, O. R. Zaïane, *A Non-Parametric Approach to Web Log Analysis*, Proc. Web Mining Workshop, in conjunction with the SIAM International Conference on Data Mining, Chicago, IL, USA, April 7, 2001

[4] M. N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim, *Data Mining and the Web: Past, Present and Future*, Proceedings of WIDM99, Kansas City, U.S.A., 1999.

[5] C. Groeneboer, D. Stockley, T. Calvert, *Virtual-U: A collaborative model for online learning environments*, Proceedings Second International Conference on Computer Support for Collaborative Learning, Toronto, Ontario, December, 1997.

[6] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann Publisher, 2001

[7] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, M.-C. Hsu, `` FreeSpan: Frequent Pattern-Projected Sequential Pattern Mining", Proc. 2000 Int. Conf. on Knowledge Discovery and Data Mining (KDD'00), Boston, MA, August 2000

[8] M. Spiliopoulou, L. C. Faulstich, K. Winkler, *A Data Miner analyzing the Navigational Behaviour of Web Users*, Proceedings of workshop on Machine Learning in User Modeling of the ACAI'99, Creta, Greece, July, 1999.

[9] J. Srivastava, R. Cooley, M. Deshpande, P. Tan, *Web Usage Mining: Discovery and Applications of Usage Patterns form Web Data*, SIGKDD Explorations, Vol.1, No.2, Jan. 2000.

[10] The International Workshop on Web Knowledge Discovery and Data Mining, Kyoto, Japan, April 18, 2000, http://www.ntu.edu.sg/home/awkng/wkddm2000.htm

[11] Third International Workshop on Advanced Issues of E-Commerce and Web-based Information Systems San Jose, CA, USA, June 21-22, 2001 http://www.chutneytech.com/wecwis2001.html

[12] Third WEBKDD workshop on data mining for web applications: Mining Log Data Across All Customer TouchPoints, San Francisco, CA, USA, August 26, 2001, http://robotics.Stanford.EDU/~ronnyk/WEBKDD2001/index.html

[13] WebCT: http://www.webct.com/

[14] O. R. Zaïane and Maria-Luiza Antonie, Automatic Text Categorization using Association Rule Mining, submitted to the Journal of Intelligent Information Systems, Special Issue on Automated Text Categorization, 2001

[15] O. R. Zaïane, J. Luo, *Towards Evaluating Learners' Behaviour in a Web-Based Distance Learning Environment*, Proc. IEEE International Conference on Advanced Learning Technologies (ICALT 2001), Madison, WI, USA, 6-8 August 2001

[16] O. R. Zaïane, M. Xin, J. Han, *Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs*, Proceedings from the ADL'98 - Advances in Digital Libraries, Santa Barbara, 1998.