Towards Evaluating Learners' Behaviour in a Web-Based Distance Learning Environment

Osmar R. Zaïane and Jun Luo
Department of Computing Science, University of Alberta
{zaiane, jun}@cs.ualberta.ca

Abstract

The accessibility of the World-Wide Web and the ease of use of the tools to browse the resources on the Web have made this technology extremely popular and the means of choice for distance education. Many sophisticated web-based learning environments have been developed and are in use around the world. Educators, using these environments and tools, however, have very little support to evaluate learners' activities and discriminate between different learner's on-line behaviours. In this paper, we exploit the existence of web access logs and advanced data mining techniques to extract useful patterns that can help educators and web masters evaluate and interpret on-line course activities in order to assess the learning process, track students actions and measure web course structure effectiveness.

1. Introduction

The World-Wide Web is becoming the most important media for collecting, sharing and distributing information. Web-based applications and environments for electronic commerce, distance education, on-line collaboration, news broadcasts, etc., are becoming common practice and widespread. 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 [1] 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.

Web-based course delivery systems 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. While web logs are relatively information poor, present mixed accesses of different users, contain erroneous and irrelevant entries, and are extremely large, there are techniques for web log cleansing and transformation as well as advanced approaches for discovery of hidden and useful patterns from these access logs. Web usage mining refers to non-trivial extraction of potentially useful patterns and trends from large web access logs. In the context of web-based learning environments, the discovery of patterns from navigation history by web usage mining can shed light on learners' navigation behaviour and the efficiency of the models used in the on-line learning process. The patterns discovered can be used to evaluate learners' activities, but can also be used in adapting and customizing resource delivery, providing automatic recommenders for activities, etc. These patterns, however, cannot be extracted with simple statistical analysis.

Currently there is a variety of web log analysis tools available. Most of them, like NetTracker, webtrends, analog and SurfAid, etc., provide limited statistical analysis of web log data [8]. For example, a typical report has entries of the form: "during this time period t, there where *n* clicks occurring for this particular web page p".

However, the results provided by these tools are limited in theirs abilities to help understand the implicit usage information and hidden trends. What is needed is summarization of these trends that can be interpreted by educators delivering their courses on-line.

There are more sophisticated tools that use data mining techniques and go beyond these rudimentary statistical analyses. 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]. For example, WebSIFT [3] is a comprehensive web usage tools that is able to perform many data mining tasks. WUM [6] is special web sequence analyser for improving web pages layout and structure. A versatile system, WebLogMiner [8], uses data warehousing technology for pattern discovery and trend summarization from web logs.

Although these web usage-mining tools have been successfully applied to some degree in e-commerce applications, few of them are flexible enough to adapt to an on-line learning environment. Moreover, while the nature of the patterns to be discovered can be the same in both domain applications, the identification of users, hits and sessions as well as the interpretation of activities, and thus the needs of the application are significantly different. We suggest a flexible framework for web usage mining in the context of on-line learning systems where the users can express constraints at the data gathering and transformation stage, as well as at the patterns discovery and analysis steps. This way the users (i.e. educators) can tailor the data mining process to their needs and tasks at hand. The dilemma is that educators are already overwhelmed with complicated tasks pertaining to delivering courses on-line and should not be burdened with additional intricate data mining tasks, yet they need to iteratively interact with the data mining system in order to extract meaningful and useful patterns form learners' activity history. We have designed our system taking this into consideration. The complex algorithms are transparent to the users, but the needs can be simply expressed by constraining the system at different levels using plain filters and a straightforward query language to sift through the patterns discovered.

In the next section we describe the three-tier architecture for our open-structure and interactive web usage mining system. We briefly present in the third section some of our algorithms used and portray in the subsequent section, some of the experiments we conducted on real log files.

2. System framework

Data mining from web access logs is a process consisting of three consecutive steps: **data gathering and pre-processing** for filtering and formatting the log entries, **pattern discovery** which consists of the use of a variety of algorithms such as association rule mining, sequential pattern analysis, clustering and classification on the transformed data in order to discover relevant and potentially useful patterns, and finally **pattern analysis** during which the user retrieves and interprets the patterns discovered [7]. The pre-processing stage is arguably the most important step and certainly the most time consuming. Web usage data often contains irrelevant and misleading entries that need to be eliminated. Moreover, since hits of all users are combined and in impractical format in the web log, it is necessary to transform the entries into a format viable for data mining algorithms after identifying individual users, sessions and transactions.

Our web usage mining system also adopts this three-tier architecture, although we added the possibility to express specific constraints at the different levels of the system. In the context of an e-learning environment with a data-mining-based evaluation system, the users are often educators who are not necessarily savvy in data mining techniques. The constraint-based approach we suggest allows the user (i.e. educator) to simply express needs by specifying restrictions and filters during the pre-processing phase, the patterns discovery phase, or the patterns evaluation phase. Indeed, two educators using the same web server for their courses may have different requirements for learner behaviour evaluation. Even the same user evaluating different course activities at different times for different learners can have diverse requirements with regard to the data sources, the relevant attributes or the types of patterns sought for. Furthermore, defining filters during the pre-processing phase considerably reduces the search space, pushing constraints during the mining not only accelerates the process but also controls the patterns discovered, and expressing constraints at the evaluation phase helps sifting through the large set of patterns extracted. The ability to add limitation and control at all stages allows interactive data mining with ad-hoc constraint specification leading to the discovery of relevant and restrained patterns, pertinent to the evaluation task at hand. For instance, in our implementation, the user can pick filters in the pre-processing phase to select desired student or student group, the desired time period and/or the relevant subset of web pages in order to zero-in the learning tasks and activities to evaluate. In addition,

educators can define their interpretation of "session" and sequence of student's clicks, concepts important in the web log data transformation. For example, a session can be defined as the sequence of clicks of one student, which happen each time from "log in" and "log out" the web environment. Also, educators can define a session as a series of clicks of one student happening in the specified period after the certain specified action. Most data mining algorithms, thereafter, use these sessions as the basic units for searching patterns.

For the pattern discovery, several algorithms, including association rule mining, inter-session frequent pattern mining, intra-session frequent pattern mining, etc., have been chosen to discover the strong trends and relationships from web usage data. The constraints that are provided to the educators to state are mainly related to these algorithms. These constraints can be used to conduct the knowledge discovery process and limit the search space. Stating the constraints is the only (optional) interaction with the data mining modules. Knowledge of the intricate algorithms is not necessary. Another noteworthy point is that the architecture of the system allows a plug-and-play of new data mining modules without significant change in the system, allowing addition of new pattern discovery functionalities.

In the last stage of pattern analysis, the objectives are to make the discovered patterns easy to interpret for the decision makers. We have implemented intuitive graphic charts and tables for pattern visualization and understanding. We intend to add an ad-hoc query language that would allow the weeding-out of irrelevant patterns and the focus on knowledge discovered to use for the evaluation of learners' on-line.

3. Algorithms

The modular design of our system allows us to add as many new data mining algorithms as necessary without compromising the effectiveness of the pattern discovery and evaluation process. We have implemented a variety of algorithms with intuitive interfaces. For example we put into practice association rule mining for discovering correlations between on-line learning activities, two variants of sequential pattern mining for studying the sequences of on-line activities within a learning session or between sessions, and clustering to group learners with similar access behaviours. In this paper, we take the association rule discovery as an example.

Association rule discovery is a classical data mining problem [2]. It shows the correlations among items within transactions. Given an item set I (in our case a set of pages or URLs), and a transaction data set T where each transaction $t \subset I$, X, Y are two different item-sets, $X \subset I$, $Y \subset I$, $X \cap Y = \phi$, then, $X \Rightarrow Y$ is an association rule with two measures: support and confidence, where support is the percentage of the transactions containing $X \cup Y$ and confidence is the percentage of transactions containing Y on the condition of containing X. For example, an association rule looks like: 30.5% of the students who successfully finished Exercise 3 also accessed Section 4 of Chapter 2. Depending upon the support and confidence thresholds, a large number of rules can be discovered and sifting through them can be tedious. A constraint-based association rule can be more useful and interesting for the educators trying to evaluate with different requirements.

For the algorithms-related constraints, the educators can set the requirements like strong support threshold, strong confidence threshold as in other association rule discovery applications. However, in addition to those two constraints, the educators can also specify constraints on the item-sets X and Y. For instance, the educators can direct the algorithms to search for the rules that answers "How often the students check out on-line resources when they read the Section 1 of Chapter 1 in one transaction."

4. Experiment

Currently we are experimenting our system on web logs from two systems: an in-house built system at the Technical University of British Columbia (TechBC), a university that delivers most its courses on-line, and Virtual-U, a web-based learning environment built in the context of the TeleLearning Canadian Centres of Excellence. The example we use for illustration in this paper comes from a TechBC web log with records of 100 students' on-line activities in two courses, TECH 142 and TECH150 from September 14th, 1999 to December 17th, 1999. There are 200,433 entries in this web log file of a size of 109 Megabytes.

One typical entry in the original log file looks as follows:
1,1999-09-14 22:02:13,200, "/TECH150.1/Unit.2/Presentation.1/FAQ/index.html","-"
This entry shows that the user with ID "1" successfully visited at "1999-09-14 22:02:13" the web page "/TECH150.1/Unit.2/Presentation.1/FAQ/index.html".
Since the URL syntax of this web site encodes the structure of the site, when pre-processing the web log, we provide a way to generalize the log entries. For example, if a student visits these following web pages successively:
"/TECH142.1/Unit.1/LearningPath.1/ActivitySequence1/External1.html",
"/TECH142.1/Unit.1/LearningPath.1/ActivitySequence1/favicon.ico",

"/TECH142.1/Unit.1/LearningPath.1/ActivitySequence1/index.html",
"/TECH142.1/Unit.1/LearningPath.1/ActivitySequence1/tooltip.htc",
we can generalize these four clicks into one action, say, "Tech142.1, unit 1, Learning Path 1, and Activity Sequence 1". We might even generalize it in a higher-level like "tech142.1, unit 1". These drill-down and roll-up functionalities are provided to the decision makers to manipulate the data set and impose a concept level during the constraint-based mining process.

We assume that the educators use the "log-in" and "log-out" as the starting point and ending point of each transaction. We are taking 15 minutes as the upper limit of the time interval between two successive inter-transaction clicks to break the sequence of one student's click stream into the transactions.

Two experiments are presented in this paper to demonstrate the advantages of using constraint-based web usage mining in the context of e-learning. The first experiment is to find associations between visited pages using the whole web log and without use of interactive constraint specification. The second experiment takes advantage of constraint specification in particular at the data pre-processing phase. Both experiments aim at finding association rules of the same significant level with support=0.3 (supported by at least 30% of the sessions) and confidence=0.4 (the rule discovered is at least 40% confident). In the second experiment, we were interested at the students 1 to 12 and the web pages relevant to the course "TECH142". This could be the case were the educator would want to understand the on-line behaviour of students 1 to 12 who outperformed other students.

Although the second experiment deals with a subset of the filtered web log, it still finds 193 association rules with 17 frequently visited web pages, compared to 23 association rules with 4 frequent web pages found in the first experiment due to the support being 30% of the whole dataset mined. Moreover, rather than only showing the correlations among the 4 entry pages in the first experiment, the second experiment gives the educators a better idea about relationships with respect to "TECH142" web pages. For example, the educator can discover that 83% of the students who worked on "TECH142.3 Unit.1 LearningPath.1 ActivitySequence1" also visited the "PriorKnowledgeAssessment" of same Learning Path. The educator could act on this by either recommending activities or pages to students to improved their learning accomplishments, or change the structure of the on-line course towards a structure that helps the learners perform as sought for by the educator.

In summary, our system provides a powerful mechanism that makes it much easier for the educators to find the interesting rules that could be used for student access behaviour evaluation.

5. Conclusion and future work

Web usage mining has proven very useful in many e-Commerce web log analysis applications. However, the current web usage mining systems are limited in their ways to support interactive data mining and therefore they are limited in their ways to be applied in the field of web-based learning evaluation. We have implemented a system that takes advantage of the latest data mining techniques and pushes constraint specification at all stages of the web usage mining to help the educators control and guide the knowledge discovery, and effectively and efficiently understand the students' behaviours in e-learning sites.

We are in the process of enhancing the user interface of our system with the help of practitioners and educators using web-based learning environments in order to develop a more intuitive interface for constraint-based data mining and pattern visualization for the specific purpose of evaluating on-line learning.

7. References

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

[2] R. Agrawal, G. Srikant, Fast algorithms for mining association rules, Proceedings of the 20th VLDB conference, pp. 478-499, Santiago, Chile, 1994.

[3] 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.

[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] 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.

[7] 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.

[8] 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.