Student Modeling for a Web-based Learning Environment: a Data Mining Approach

Tiffany Y. Tang and Gordon McCalla
Department of Computer Science
University of Saskatchewan
57 Campus Drive, Saskatoon, SK S7N 5A9 CANADA
{yat751@mail.usask.ca; mccalla@cs.usask.ca}

Introduction

This ongoing research focuses on how data mining techniques, if incorporated into web learning environments, can enhance the overall qualities of learning.

In a web-based learning environment, where both the tutors and learners are separated spatially and physically, student modeling is one of the biggest challenges. Traditional student modeling techniques are inapplicable in these systems when tutors are overwhelmed by the huge volumes of sequential data (Agrawal and Srikant 1995) generated as learners browse through the web pages. Web mining techniques, including clustering and association rules mining, could be applied to extract *hidden* and *interesting* knowledge to facilitate instructional planning and student diagnosis. Web mining in education is not new. It has been applied to mine aggregate paths for learners engaged in a distance education environment (Ha, Bae and Park 2000); to recommend relevant words to students based on text mining from their browsed documents (Ochi et al. 1998); to recommend e-articles for students based on key-word-driven text mining (Tang et al. 2000), and to analyze learners' learning behaviors (Zaiane and Luo 2001). The research proposed here will go beyond *usage* mining to consider the content of the pages that have been visited. In an e-learning system, both learners' browsing behaviors and course content are important to derive learners' learning levels, intentions, goals, interests, or abilities. Incorporating course content can aid in an understanding of learners' browsing habits. In particular, understanding the learners' browsing behaviors can facilitate, say, the personalization of course contents delivered.

Artificial intelligence in education (AIED) systems typically employ a knowledge base, a student model, and instructional plans. For a web-based AIED system, web mining becomes part of student modeling. Traditional usage data (Cooley 2000) keeps a lot of information that is not needed. But we do need the knowledge of content and complexity of each page. Finding and using such knowledge is *tractable* in our domain since we can annotate course web pages with metadata and the knowledge base, and instructional plan also give context for the properties of each page. The system can relate its mined knowledge of page contents and student navigation patterns to students' level of understanding (Tang et al. 2001) to decide upon appropriate feedback to them.

Data Clustering for Web Learning

Among mining techniques of particular interest in web-based learning environments is data clustering. It can, for example:

- Promote group-based collaborative learning

Traditionally, learning systems focus more on how to individualize course contents and delivery. However, in web-based learning environments where both the number of students and the size of the information can be huge, to reduce the cost and the computational burdens on the system, group-based learning will also be useful.

Data clustering is a powerful tool to find clusters of students with similar learning characteristics based on their path traversal patterns and the content of each page they have visited. The clusters of students can be used to promote effective group learning, e.g., assigning students from different clusters so as to form effective learning groups for collaboration. In addition, after we find a cluster of learners with similar browsing paths, we could extract course contents along the paths to create fragmented contents (group-based course content delivery). These fragmented course contents can also be selected for recommendation, and the clustered paths can be used to sequence the curriculum for other students in the future (group-based instructional planning).

- Provide incremental learner diagnosis

Incremental clustering can be performed to help diagnose learners as they browse through the system. This is consistent with the "just-in-time" modeling proposed by (McCalla, Vassileva and Bull 2000).

A Clustering Algorithm Based on Large Generalized Sequences

We briefly describe a clustering algorithm that should be useful in clustering data for web-based learning environments. This algorithm is based on the notion of *large* *generalized items*. Consider a collection of transactions $\{T_1, T_2,\ldots T_n\}$, where each transaction T_i is a set of sequences $\{t_1, t_2,\ldots t_p\}$. According to (Gaul and Schmidt-Thieme 2000), *generalized sequences* are defined as:

$$T^{gen} = \left\{ t \in \left(T \cup \{*\}\right)^* \mid \nexists i \in N \quad \text{such} \quad \text{that} \qquad t_i = t_{i+1} = * \right\}$$

where $*$ represents wildcard.

The idea behind this notion is that the navigation paths of different users might not be completely or even partially matched on a one-to-one mapping. Generalized sequences allow path deviations and retain path orders. However, among all the generalized sequences, we are only interested in those large ones. A *large* generalized sequence is the sequence whose frequency of occurrence is larger than a user specified *minimum support θ*. Formally, it is defined as:

$$T_l^{\;gen} = \left\{ T^{\;gen} \subseteq T \mid \text{Support} \left(T^{\;gen} \right) > \theta \right\}$$

In the context of our domain, adopting generalized sequences is more viable to derive traversal patterns than *maximal forward sequences* (Chen, Park and Yu 1998) or *longest repeated sub-sequences* (Pitkow and Pirolli 1999).

A general algorithm for finding all frequent generalized subsequences is proposed in (Gaul and Schmidt-Thieme 2000). The research proposed here will go a step further by first searching for all generalized subsequences among learners' traversal paths (i.e. path fragments). Then these path fragments will be further clustered based on course contents of each page so as to characterize learners' browsing behaviors. For example, we might find the following two large generalized sequences:

$B * D * H * I$ and

$B * DH * I$

They might belong to either two different clusters or one cluster, depending on what other pages lie in between page D and H for the first fragment and the contents of page D and H. If the pages in between D and H are those which can be regarded as additional readings for the topic covered in page D, then for learners who take the first path, we might infer that they are interested in more knowledge concerning these topics. Thus, course contents would be used to further cluster all these frequent path fragments and actions could then be taken accordingly such as recommending further readings for students, assigning collaborative work for them, etc.

Future Works

We are constructing a web-based learning environment as a test bed for our research. More data mining algorithms will be designed for our domain, since data mining algorithms are known to be dependent on application areas (Han and Kamber 2000). In the context of the proposed systems, we will also quantify the interestingness and usefulness of discovered knowledge.

Acknowledgments. This work is supported through a research assistantship from the Canadian Natural Sciences and Engineering Research Council.

References

Agrawal, R., and Srikant, R. 1995. Mining Sequential Patterns. In *Proc. of the Eleventh International Conference on Data Engineering (ICDE),* 3-14,Taiwan.

Chen, M.S.; Park, J.S.; and Yu, P.S. 1998. Efficient Data Mining for Path Traversal Patterns. *IEEE Trans. Knowledge and Data Engineering* 10(2): 209-221.

Cooley**,** R. 2000. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. Ph.D diss., Dept. of Computer Science, University of Minnesota.

Gaul, W., and Schmidt-Thieme, L. 2000. Mining Web Navigation Path Fragments. In *Proc. of 2000 Workshop on Web Mining for E-Commerce—Challengers and Opportunities*, Boston.

Ha, S.H.; Bae, S.M.; and Park, S.C. 2000. Web Mining for Distance Education. 2000. In *Proc. of IEEE International Conference on Management of Innovation and Technology (ICMIT)*, vol.2, 715-219.

Han, J.W., and Kamber, M. 2000. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

McCalla, G.; Vassileva, J.; and Bull, S. 2000. Active Learner Modeling. In *Proc. of the Fifth International Conference on Intelligent Tutoring Systems,* 53-62.

Ochi, Y.**;** Yano, Y.; Hayashi, T.; and Wakita, R. 1998. JUPITER: a Kanji Learning Environment Focusing on a Learner's Browsing. In *Proc. of the Third Asia Pacific Conference on Computer Human Interaction,* 446-451.

Pitkow, J., and Pirolli, P. 1999. Mining Longest Repeating Subsequences to Predict World Wide Web Surfing. In *Proc. of the USENIX Symposium on Internet Technologies and Systems*, 139-150.

Tang**,** C.; Lau, R.W.H.; Li, Q.; Yin, H.; Li, T.; and Kilis, D. 2000. Personalized Courseware Construction Based on Web Data Mining. In *Proc. of the First International Conference on Web Information Systems Engineering* (*WISE 2000*) vol.2, 204-211.

Tang, T.Y.; Chan, K.C.; Winoto, P.; and Wu, A. 2001. Forming Student Clusters Based on Their Browsing Behaviors. In *Proc. of the Ninth International Conference on Computers in Education*, vol. 3, 1229-1235.

Zaiane, O., and Luo, J. 2001. Towards Evaluating Learners' Behavior in a Web-based Distance Learning Environment. In *Proc. of IEEE International Conference on Advanced Learning Technologies*, 357-360, Madison, WI.