A DATA MINING HYPERTEXTBOOK: DESIGN, IMPLEMENTATION AND EXPERIENCE

Rakesh M. Verma
Computer Science Department
University of Houston
713 743-3348
rverma@uh.edu

Ping Chen
Computer and Math Science Dept.
University of Houston-Downtown
713 221-2764
chenp@uhd.edu

ABSTRACT

In addition to a richer learning experience integrating text and multimedia - such as audio commentary with slides and visualizations of key algorithms and concepts - hypertextbooks for computer science courses can provide immediate feedback to the reader. Due to the explosion of data and information in the Internet age, data mining is becoming a core area in the undergraduate computer science curriculum. In this paper, we discuss our experience in designing, teaching, and improving an undergraduate data mining course by designing and implementing a hypertextbook for the key concepts in the course. Course evaluations conducted show the effectiveness of our approach.

INTRODUCTION

Following [8], we define a hypertextbook as "a teaching and learning resource that uses all relevant Web technologies to provide readers with a rich, active-learning environment" and is delivered through a Web browser. It typically includes text, slides with sound tracks, images, self-check exercises and quizzes with immediate feedback, and most importantly, visualizations or animations of the key concepts and algorithms. It may also have multiple levels based on the user background, e.g., a beginner level and an advanced user level. Hypertextbooks were explored by a working group at ITiCSE 2006, which emphasized the deployment of educational visualizations [8, 9].

Visualizations of algorithms and data structures has been a long computer science tradition (e.g., see [7, 2] and references contained therein). However, as others have observed before us [8, 10], even though high quality visualizations have been developed for several concepts and algorithms, their usage has been quite limited. A survey performed by the ITiCSE 2002 Working Group "Improving the Educational Impact of Algorithm Visualization" reported that the two of the main reasons for this lack of use are the time required and the lack of integration with existing teaching materials [6].

That report has provided the impetus to approaches that provide tools and/or generators for producing content in a short amount of time. For the integration aspect the 2006 ITiCSE Working group had proposed hypertextbooks integrating text, images, audio, video and visualizations. Recently, some groups such as [8, 10] have started creating semi-automated environments for authoring hypertextbooks as well.

In this paper, we present our design, implementation and experience with a hypertextbook for an undergraduate data mining course. Due to the explosion of data and information, data mining has become a core area in computer science. Data mining techniques aim to clean, analyze, and extract interesting patterns and knowledge from large datasets, and a course aimed at the undergraduate, upper-level Computer Science major has become both feasible and desirable. Computing Curricula 2001 includes an elective data mining course in undergraduate Computer Science curriculum [1].

In 2003, the second author had developed an undergraduate Data Mining course including lecture notes, assignments, and projects. Based on this experience, the major

challenge in teaching and learning data mining at the undergraduate level is that due to high complexity of data structures and algorithms in data mining techniques, students often emerge with a shaky understanding and experience enormous difficulty when dealing with real-world problems [5]. In Spring 2009, we completely redesigned the data mining course with improved courseware including new presentation slides, guest lectures, real-world projects and a hypertextbook.

Below we discuss our experience of designing and integrating a data mining hypertextbook in the undergraduate data mining course offered in Fall 2009 and Spring 2010 semesters at the University of Houston and in Spring 2011 at University of Houston-Downtown. We believe that high-quality active-learning courseware is essential for the success of the data mining course. As has been observed earlier by others [8, 10], a complete hypertextbook for an undergraduate or graduate course requires a significant amount of effort. Hence we have selected the key concepts and algorithms that give students the most difficulty, in our experience, and included them in our hypertextbook.

This paper is organized as follows. In Section 2 we present the design and implementation of the hypertextbook that is used in our data mining course. Data collection on the hypertextbook is described in Section 3, evaluation results are presented in Section 4, related work in Section 5 and we conclude in Section 6.

DATA MINING HYPERTEXTBOOK

Studies have shown that accommodating a student's learning style can significantly increase his/her academic performance [4], which is especially important when introducing new and advanced topics into undergraduate Computer Science education. The new generation of students is exposed to interactive and multimedia technologies, such as video games, facebook, etc., from a very early age. Moreover, the current student population is increasingly attuned to the use of technology for learning. Thus this generation of students is even more turned-off by paper-pencil format exercises and passive presentations than previous generations. Keeping these factors and the huge effort involved in implementing a complete hypertextbook for the course in mind, we have implemented a hypertextbook for four key concepts of the data mining course: clustering, association rules, classification and visualization of data.

In addition to a Help Section and References, The UH Data Mining Hypertextbook, "Data Mining -The Hypertextbook[1]," has four chapters currently: Chapter 1 -Decision Trees, Chapter 2 -Association Analysis, Chapter 3 -Visualization, and Chapter 4 -Cluster Analysis. For each chapter, the data mining hypertextbook has two levels: beginner and advanced. The materials for the two levels are color coded: green for beginner level and blue for advanced level. For example, the Chapter on Decision Trees has four sections for the beginner and five sections for the advanced user.

For each concept and each level the hypertextbook includes the key definitions and algorithm(s) for each concept, slides with audio tracks, images, plus self-check exercises or quizzes. The difference in the two levels is that the beginner level has the introductory definitions and algorithms whereas

[1] http://www.hypertextbookshop.com/dataminingbook/working_version/

the advanced level has the more advanced algorithms and concepts. The concepts are presented in hypertext pages, powerpoint slides with and without audio, and in picture formats. The quizzes include multiple-choice questions, short-answer questions, and computational exercises. Answers to the quizzes can be checked immediately.

The Data Mining Hypertextbook also includes user-data-driven animations of two key algorithms so far: the decision tree classification algorithm and the association rule mining algorithm, Apriori. Descriptions of these algorithms can be found in popular data mining texts such as [11, 3]. We have selected the hypertextbook authoring environment proposed by [8] for our hypertextbook.

A screen shot of the hypertextbook is after the references. It shows images to help the user for the Decision Tree Builder Java Applet, which lets the user to run the decision tree algorithm on either data provided by us or on user-supplied data.

DATA COLLECTION

The Data Mining Hypertextbook was used in the Data Mining Course at the University of Houston during the Fall 2009 and Spring 2010 semesters and at University of Houston-Downtown during Spring 2011. The course was offered as a senior level elective class at both universities. At the University of Houston, during Fall 2009 there were five students in the course and during Spring 2010 enrollment increased to nine students even though the course was offered back to back and is not a required course. At University of Houston-Downtown, the course was offered online during Spring 2011 and therefore the data collection had to be done through email communication.

During the three semesters, all students were surveyed with an instrument designed by the first author on which there were 10 questions with a 5 point likert scale and one free response question. One survey was not returned by students at the University of Houston and only four surveys were returned at the University of Houston-Downtown, so the total number of surveys received is 17. Of the 10 questions on the survey, four compared the hypertextbook format with a traditional book, two asked about the animation applets, one asked about the technical aspects of using the hypertextbook, one was on the exercises included in the hypertextbook and two were prospective questions asking students whether they would like to switch to hypertextbooks for other courses. The applet questions are omitted below since the Fall 2009 version had two questions about one applet (only one had been implemented at the time), and the Spring 2010 version had one question each on two different applets. The question that asked about the technical aspects of using the hypertextbook had to be redesigned since it was misinterpreted by some students on the first version, so we only report the results for this question for the Spring 2010 and Spring 2011 versions of the course (n = 12). One student did not respond to the last two questions in the table.

Question	SA/A	N	SD/D
The materials were easy to study	15	1	1
I enjoyed the hypertextbook more than a traditional book	12	5	0
I felt more actively engaged in learning using the hypertextbook	13	2	2
Hypertextbook made it easier for me to visualize & understand concepts	16	1	0
The Uvaluate exercises were very helpful	10	4	3
I had very little trouble or no trouble in using it	10	1	1
I would be willing to use hypertextbooks for all subjects	13	1	2
I would pay up to $20 for a hypertextbook rather than pay for a traditional book	12	1	3

Feedback on Hypertextbook

Here are the student comments to the free response solicitation ("Additional written comments are appreciated, so please use the space below"):

• "Although it is not the same feel as a paper book, the pros outweigh the cons. As long as you have an internet connection, you can use it and that is why it is extremely good."

• "hypertextbooks is a great idea. It is easy to access and it is not expensive."

• "My only real complaint was the method for navigation. It was very unintuitive and disjointed."

• "I like this idea and think that we can improve more: ..."

• "The hypertextbook is good, but there are few flaws." Some of the exercises did not work well in Firefox Mozilla according to this student.

• "The tests from the hypertextbook need to be more punishing, similar to web-ct where it records time and attempts based on accounts."

• "This additional material has been very helpful for my project and for my final exam. Furthermore, I think other classes should use similar techniques to help enhance the material, which provides students with different approaches along the learning process."

Clearly, a majority of the students appreciate the hypertextbook format, have very little trouble with using it, and feel that they are able to visualize and understand course concepts better with this interactive format. We have more work to do on improving some of the exercises. Again, a majority of the students are also willing to consider hypertextbooks for all their subjects and are even willing to pay a moderate amount for a hypertextbook.

RELATED WORK
Hypertextbooks have been used before in a few disciplines such as biology and physics. Surprisingly there are not many hypertextbooks for computer science. For some research on the use of hypertextbooks in computer science, see [1, 8, 10].

CONCLUSION
To optimize student learning efficiency, the newly designed course is carefully presented through the interactive format of a hypertextbook. Evaluation shows that our approach is significantly appreciated by the students and course enrollment has increased

as a result even though the course is not a required one. In the future we plan to add animations for the two other chapters (clustering and visualization) to the hypertextbook and also improve the exercises by recording the number of attempts and/or the time taken by the student.

Acknowledgment

REFERENCES

[1] Christopher M. Boroni, Frances W. Goosey, Michael T. Grinder, and Rockford J. Ross. Engaging students with active learning resources: hypertextbooks for the web. *In SIGCSE*, pages 65–69, 2001.

[2] R. Cavalcante, T. Finley, and S.H. Rodger. A visual and interactive theory course with j flap 4.0. *In ACM SIGCSE '04*, pages 140–144, 2004.

[3] Jiawei Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc, 2005.

[4] C J Jackson, E V Hobman, N L Jimmieson, and R Martin. Comparing different approach and avoidance models of learning and personality in the prediction of work, university, and leadership outcomes. *British Journal of Psychology London England* 1953, 100(Pt 2):283–312, 2009.

[5] David R. Musicant. A data mining course for computer science: primary sources and implementations. *In SIGCSE,* pages 538–542, 2006.

[6] T. Naps, G. Rößling, et al. Exploring the role of visualization and engagement in computer science education. Report of the Working Group on improving the educational impact of algorithm visualization. *In ITICSE*, 2002.

[7] Thomas L. Naps and Guido R¨oßling. Jhavé -more visualizers (and visualizations) needed. *Electr. Notes Theor. Comput. Sci.,* 178:33–41, 2007.

[8] Rockford J. Ross. Hypertextbooks and a hypertextbook authoring environment. *In ITiCSE*, pages 133–137, 2008.

[9] Guido Rößling, Thomas L. Naps, Mark S. Hall, Ville Karavirta, Andreas Kerren, Charles̆ Leska, Andrés Moreno, Rainer Oechsle, Susan H. Rodger, Jaime Urquiza-Fuentes, and J. Angel Velázquez-Iturbide. Merging interactive visualizations with hypertextbooks and course management. *SIGCSE Bulletin*, 38(4):166–181, 2006.

[10] Guido Rößling and Teena Vellaramkalayil. First steps towards a visualization-based computer science hypertextbook as a moodle module. *Electr. Notes Theor. Comput. Sci.*, 224:47–56, 2009.

[11] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Education, 2006.

Decision Tree Builder

Figure 1. Main window of Decision Tree Builder