An Interactive Courseware for Learning Data Warehousing on the Web

Prof. Meiliu Lu, California State University at Sacramento

Meiliu Lu, a professor of Computer Science Department, College of Engineering and Computer Science, California State University at Sacramento since 1987. She received her Ph.D. in Computer Science, University of Illinois, 1987. Meiliu Lu has published in research areas such as web applications for user-paced learning, machine learning applications, bioinformatics, data mining and data warehousing, and distributed algorithms. She has taught both undergraduate and graduate level courses in areas such as Data Warehousing and Data mining, Computing Theory and Programming Languages, Introduction to Computer Programming, Machine Learning, Knowledge-based Systems, Artificial Intelligence, Web Database Applications, Data structures and Algorithms, and Algorithms and Paradigms.
An Interactive Courseware for Learning Data Warehousing on the Web

1. Introduction

Building on our own experiences, we address two issues related to global engineering education\textsuperscript{3, 4} in most universities today in this paper. These issues are, how can our students gain global competence without going abroad? and can we find an effective way under current resource limitations to help our students to possess some of most valuable attributes of global competence such as self-guided learning? In answer to the first question, we discuss issues and practices of the past decade in our institution which encouraged us to continue on the road of developing web-based courseware in engineering education. In answer to the second question, we present two interactive web-based courseware developed by international graduate students and then describe how these two learning tools helped both local and international students to learn key concepts of data analytics at their own pace. Both developers and users of these learning tools have participated in such unique experience of technical communication with their peers.

Data analytics is playing a significant role in science and engineering education in this digital information era. Data warehouses provide online analytical processing tools for the interactive analysis of multidimensional data of various granularities. The objective of this project is to develop a web-based interactive courseware to help students or beginning data warehouse designers in learning data warehousing. Developers of this project include Computer Science international graduate students from India. The targeted primary users are students of a computer science course called Data Warehousing and Data Mining. Other possible end users of the courseware can be anyone who has access to internet and have interests of learning data warehousing basics.

In this paper, we will introduce motivation, design and implementation of our data warehousing courseware. Of course, we will also share the lessons that we had learned from this project. The method used in this project can be generalized to other engineering fields which has globe educational needs on self-paced learning tools.

2. Motivation: Information Technology and Education Globalization

In the College of Engineering and Computer Science at California State University at Sacramento, over 90 percent of undergraduate students are from California, and around 70 – 80 percent of graduate students are international students from developing countries such as India, China, and other countries around the world. As indicated in a recent message from our Dean Macari, studies show that 85 percent of our Engineering and Computer Science graduates choose to remain in Sacramento to work locally after they complete their education. Even in today’s sluggish economy, our engineering graduates are employed at a higher rate than professionals in many other segments of the economy from Sacramento to Silicon Valley. Our University has a global education program, however, our Engineering and Computer Science students rarely goes abroad for their study. Most of them have part
time job to support themselves and sometimes their families. They have the needs for anytime anywhere learning to stay globally competitive. Education should meet the needs of variety of learners and therefore IT is important in meeting this need.

In our Computer Science MS graduate program, each MS student is required to complete a degree project. This degree project provides the student an excellent opportunity for technology study and application oriented research. Each MS student project is supervised by a faculty mentor and the project is reviewed by a committee of two computer science faculty members. As a result of a collective effort of MS degree projects, several computer science MS students contributed to this courseware project. The idea is very simple – to speed up future students learning by using information technology to develop a courseware on a set of topics of importance to many students. We choose data warehousing as the central theme for this project due to its significant role in data analytics.

The advantage of using information technology is that time-consuming work routines, such as giving an illustrative examples or making a demo, can increasingly be performed by means of this technology. Our courseware did exactly that. It saved the valuable time of instructor so that more time can be devoted to the processing of information and production of knowledge. This also creates an effective way of collaborative learning among international and local students.

In this project we have developed a set of courseware to reinforce learning of the key concepts of the data warehousing using a case study approach. The courseware now contains two working learning tools: (1) data warehousing courseware; and (2) ETL (Extraction, Transformation, and Load) courseware. We have been using this courseware in a Data Warehousing and Data Mining course at California State University in past two years. Students had very positive feedback on their learning experiences with the tool. We are now in the process of expanding to include one more sub-system, OLAP (On Line Analytical Processing) cube courseware.

The main advantages of using this web-based interactive courseware to supplement classroom instructions or textbook are as follows:

(a) The contents of the courseware can be updated by the instructor periodically to keep up with the rapid development of the field of data warehousing and data mining.

(b) The incremental developing courseware provides an ideal platform for the instructor to share latest research development results with students so that we can make the technology transfer a little faster to our students.

(c) This project provides a self-paced interactive learning tool not only to the students taking a course on data warehousing but also to any engineer in the world who have to build a data warehouse from scratch and needs to learn fundamental concepts quickly. From schema design to OLAP query execution, we explain concepts step by step and illustrate implementation techniques with examples from our case study. An end user can run various demo queries interactively as he/she is making progress in his/her own pace.

(d) Our digital courseware offers a low cost alternative to traditional expensive textbooks. Students usually do not want to buy additional reference books due to the high costs.
3. Courseware Design and Implementation

Data warehousing courseware

The data warehouse has been playing a critical role in data preprocessing and integration. It allows quick retrieval of input data for data mining or data analysis tools. The outcome of data reporting, data analysis and data mining can then be used for supporting decisions making on budget analysis, resource allocation, forecasting and prediction. To illuminate data warehousing basic concepts, design principle, and performance enhancement techniques, we developed this courseware. This web-based tool assists beginner data warehouse designers to reinforce their understanding of the basic design concepts of data warehousing via a case study. In this case study, the data sources include the student enrollment data from the California State University at Sacramento and, enrollment-related social and economic data of California. The main objective of this data warehouse is to prepare input data for an existing data mining system for student enrollment prediction. Using the case study, we demonstrate the procedure to build a data warehouse and reveal some common incorrect practices which should be avoided in the design process. This project provides a self-paced learning tool not only to the students taking a course on data warehousing but also to the beginner data warehouse designers who have to build a data warehouse quickly from scratch. Figure 1 shows the courseware tool’s introduction page.

Figure 1. Data warehousing courseware learning tool

This courseware tool is a 3-tier web application designed using PHP, HTML, CSS and JavaScript. It is supported by MySQL queries and stored procedures. The data from the data sources resides on a MySQL server. The courseware includes a set of 4 demonstrations covering following topics: (1) fundamentals of data warehouse, (2) data warehouse design principle, (3) building an enterprise data warehouse using an incremental approach, and (4)
aggregation. Each demonstration provides detailed description on building a data warehouse with text, diagram, and ready-to-go query runs. Furthermore, the theory behind each subject is outlined and a set of quiz problems are provided for self-evaluation.

One of our most important goals for developing this data warehousing courseware is to get students personally engaged in the use of the courseware to understand the fundamental concepts of data warehousing. We conducted a survey on this tool in our Data Warehousing and Data Mining class, CSC 177, in spring 2010 semester at California State University, Sacramento for the upper division undergraduate and graduate students in the Computer Science Department. The overall assessment from this student group on this courseware is extremely encouraging to us. The following is a brief summary of the positive feedbacks from the student group:

1. Very accessible
2. Helpful to understand the fundamentals of DW
3. Very helpful to learn from the examples
4. Complement the course lectures very well
5. Easy to follow steps and illustrations
6. Simplicity and natural progression of the website
7. Add on quizzes can be handy in review for tests

We also collected improvement suggestions from the students. One of the comments is that a data preprocessing component is missing. Therefore, we decided to develop an ETL courseware to meet the need of our students.

ETL courseware

Extract, Transform and Load (ETL) is a fundamental process used to populate a data warehouse and an important step towards data integration which is a key step for data preprocessing for many data mining projects. It involves extracting data from diverse sources, transforming the data according to business requirements, and loading them into target databases. In the transform phase a well-designed ETL system should also enforce data quality, consistency, and conformance so that data from different source systems can be integrated. Once the data is loaded into target systems in a presentation-ready format, the end users can run queries against them to generate reports which help effective decision making. Even though ETL process consumes roughly 70% of computing resources they are hardly visible to the end users of a data warehouse. The objective of this project is to create a simple web-based ETL tool which can be accessed by anyone with internet access. This ETL tool is also supported by a brief on-line tutorial on ETL process. Guests will have limited access to this tool and registered users can have complete access. Using this tool, data can be extracted from text files or MySQL tables, integrated and loaded into target MySQL tables. Before loading the data into target MySQL tables, various transformations can also be applied to them. The tool is developed using HTML, PHP, Korn shell scripts and MySQL.
The architecture of data warehouse consisted of operational data layer, data access layer, metadata layer and informational access layer. An ETL tool can be used in data access layer to extract data from source systems and load them into the data warehouse. Developers of data warehouse may build their ETL tool by hand-code or use an existing ETL tool. A commercial ETL tool has many advantages over hand-coded ETL code. It makes the development process of ETL simpler and faster. Technical people with broad business skills who are not professional programmers can use ETL tools effectively. Many ETL tools generate metadata automatically at every step of the process thus enforcing consistent metadata throughout the process. They also have integrated metadata repositories which can be synchronized with other source systems, target systems and other Business Intelligence tools. They deliver good performance with very large datasets by using parallelism concepts such as pipelining and parallelism. They come with built-in connectors for most of source and target systems. Most of the ETL tools these days have built-in schedulers to run the ETL code at scheduled times. An overview of ETL process is show in Figure 2.

![Figure 2. Overview of ETL process](image)

However, interested users who would like to learn more about ETL tools may not have access to them. This is because commercial ETL tools are expensive for small or medium size project developers. They also need expensive hardware to run on and specialists to configure them before a new user can start using them. Most of them are not open source and web-based. The web-based ETL tool created in this project helps overcome the above challenges. This web-based tool is accessible freely to anyone with internet access and very user-friendly. Beginning ETL developers can use this tool to get a feel of what an ETL tool does before they dive into understanding complex commercial ETL tools.

When users, guest or registered, open the homepage they have an option to either go to courseware, a brief tutorial on ETL as shown in Figure 3, or go to the ETL tool as shown in Figure 4. When they click on the tool link, first page that they see is the login page. When they don’t have a username and password they can click on the guest link to access. Once they are in the tool, as registered regular users of the tool they can continue use the tool in
the order of extract, transform and load pages; as registered administrators, they can also manage the users by deleting or adding operations.

Figure 3. ETL courseware tutorial page

Figure 4. ETL courseware ETL web tool page
The main objective of the ETL tool is to provide free access to interested audience to learn what ETL is all about and how an ETL tool works. The tool provides interactive graphical user interface and is user-friendly. It can extract from heterogeneous sources, land them in landing zone, apply various types of transformations, stage them in staging zone and finally load the transformed data into database tables. The heterogeneous sources can be flat files or MySQL tables. There are several transformations available to apply on the landed source data. It can serve as a good learning tool for students who are learning concepts of data preprocessing for data mining tasks. There are still some limitations in our current version of ETL tool which need future enhancement. The first limitation is limited number of source and target connectors. Currently the tool has only flat file and MySQL table connectors. Oracle database, MS SQL server database and XML file connectors could be added. The second enhancement would be to add more transformations to the ETL tool.

4. Summary

We have always felt that the future of the world economy and the future of global engineering education go hand-in-hand. That is why we should do everything we can to continue to prepare students to thrive on the leading edge of the innovation economy.

Using IT in the global education is an investment in the economic growth of the future. In this paper we have describe some of our effort and experience of using IT in an important topic in software engineering, data warehousing. Based on the demands of our students, we have designed and implemented a set of web-based interactive courseware using open source technology. These learning tools are not only available to our local students but also available to everyone else who is interested in learning data warehousing and has internet access. In addition to many online resources available to global engineering students today, such as lecture notes presented on YouTube and KHAN Academy, we are glad that we did our part of contribution to the global engineering learning community.

Bibliography


