TRANSFORMING THE CURRICULUM WITH BIG DATA: THE NEED FOR DATA RESOURCES IN THE COMPUTER SCIENCE CURRICULUM

Brandon Olson
Department of Computer Information Systems
The College of St. Scholastica
Duluth, Minnesota 55811
bolson1@css.edu

Thomas Gibbons
Department of Computer Information Systems
The College of St. Scholastica
Duluth, Minnesota 55811
tgibbons@css.edu

Abstract

Instruction in the computer science classroom commonly relies on the existence of data to support input and output processes, database queries, and data analysis. As a result, data becomes essential to the instruction of topics across the curriculum. Unfortunately, this data is often lacking in depth, breadth, and applicability to the students' lived experiences. A need exists for large data sets spanning multiple domains, where the domain areas are familiar to the students. This paper reviews the current data sources available and highlights the need for such large data resources. The authors' experiences in obtaining such resources from both public sources and private companies are described. Additionally, a strategy for applying the large data set across the computer science curriculum is outlined to further demonstrate the need for these data resources.

1 Introduction

In the computer science curriculum there is significant attention given to the techniques and skills for developing software, while little attention is given to the data that drives this software. Based on the current computer science body of knowledge (ACM/IEEE, 2008), less than 4% of the 290 minimum core lecture hours are dedicated to information management topics. Although the purpose of this paper is not to argue for increased focus on information management topics, there is a need to draw more attention and resources to the data that powers these applications and supports the curriculum.

Without sufficient and quality data resources, the computer science curriculum is restricted to less engaging and more prescriptive methods of learning. A larger volume of data and data that is meaningful to the students is needed to place the student at the center of the learning environment and allow the student to learn through exploration and inquiry. This environment results in increased critical thinking skills.

2 Need for Big Data

The Computer Science / Computer Information Systems (CS/CIS) department at The College of St. Scholastica has recognized the importance of data, and in particular, the need for large volumes of data. The need for data is clear for such courses as introduction to databases, database modeling, and advanced database. However, needs also exist in traditional programming classes such as introduction to programming and advanced programming, as well as specialty courses like server-side web development, mobile applications, and artificial intelligence. Each of these courses requires the use of data either directly as a means to study the data or indirectly as a means to power the software applications. Unlike the tools used for software development, the data cannot be simply installed as part of a package solution, but must be created or discovered and customized to fit the needs of the classroom.

In the past, courses have used existing data sources such as the Microsoft Northwind database or other sample databases available on the Internet. However, in many cases, the database was developed by the instructor or the textbook author in order to fit the needs of the curriculum and assignments. Regardless of the source for the data, the size and complexity of the data remained small and simplistic. As a result, interactions with the data were restricted to more prescriptive exercises in order to ensure meaningful results with the limited data set. This controlled environment does not support the critical thinking and exploratory learning the department wanted to promote. The department wanted to pursue an inquiry-based learning environment.

The inquiry-based learning approach is a learning-centered form of instruction used to engage students in the investigation of real-world questions. This approach uses open-ended problems where students are able to explore, create, and communicate ideas rather than simply identifying facts and following a smooth path to a solution (What is IBL? - The Academy of Inquiry Based Learning, n.d.). The application of this approach places the student at the center of the learning, where they direct their own learning through exploration and experimentation. The inquiry-based learning method has been shown to significantly increase student learning (Magnussen, Ishida, &

Itano, 2000) and promote higher levels of critical thinking through interactive and self-directed learning (Hall, 2005). Through the use of exploration and experimentation, students are able to learn through exploration and discovery and are able to develop the critical thinking skills in not only responding to problems, but to develop new questions.

Using an instructional approach like the Modified More Method (Cohen, 1982), a series of inquiry-based learning modules could be incorporated into the curriculum based on a large-scale set of data. Each of these modules would require groups of students to study a topic, make discoveries using database and data analysis tools, and communicate the findings to the class. This inquiry-based learning approach enables the students to better understand the material through the exploration and communication of the questions and answers. Unfortunately, the current data environment does not include large volumes of authentic data. The current data resources are limited and do not support such exploration. A larger and deeper data set is needed to provide opportunities for students to apply the techniques covered in the curriculum, to explore patterns in the data, and to discover answers to questions and identify new questions.

In addition to the need for a larger and deeper data set, there exists a need for this data to be meaningful to the students. The act of exploration and discovery is predicated on the assumption that students are able to relate to the data. Access to relevant and real-world experiences promotes the perception of usefulness and importance resulting in increased student performance and retention (Davis & Fineli, 2007; Pintrich & Zusho, 2002; Wigfield & Eccles, 2000). Providing authentic real-world data related to subject domains within the lived experiences of the students will therefore increase performance and retention of the students. Therefore, a large data set containing relevant and meaningful data is needed to support the inquiry-based learning used to increase critical thinking, performance, and retention.

3 Past Solutions

The faculty members in the CS/CIS program at St. Scholastica have relied on a number of different options for database assignments. Commonly the textbook sample databases are used in the courses. Although these are easy to access for both students and teachers and often explained in the textbook, they lack a level of sophistication that inhibits inquiry-based learning.

Currently the database modeling course uses a textbook by Coronel, Morris and Rob (2010). This text provides a number of example databases, including the more advanced database for an aviation company for use in the chapter on advanced SQL. The primary problem with the database is its small size, both in the size of tables and the number of records in each table. For example, the Aircraft table, which holds records of every plan in the company, has only four records. Likewise, the Customer table contains records for only nine customers and the customer data is limited to seven attributes, three of which store the customer's name. A database of this size severely limits the types of questions students can investigate. With only seven customers and four planes, it is difficult to define inquiry-based learning problems which allow the student to learn advanced SQL.

An alternative to the sample databases provided in the textbooks are the sample databases that come with the database management software. The quality of these databases has varied over the years and they have generally been limited in scope. The Northwind sample database that comes with Microsoft Access 2010 has twenty different tables, but each table has a small number of records. The Customer table has only 29 customers and the Employee table has nine employees. Similar to the textbook sample databases, these databases generally limit the inquiry-based problems students can take on.

Over the years, our faculty members have also developed a number of their own databases for use in the different database courses. These are usually designed to fit the requirements of a particular assignment or learning outcome. But these also tend to be small in size and score. A typical instructor-designed database has three to five tables and each table has less than a dozen records. While instructor designed databases tend to address the narrow needs of a single classical learning outcome, they do not lend themselves to rich inquiry-based learning problems.

A number of websites provide access to large public data sets. Some of the primary sources for this data release copies of the data in a raw format. An example of this is the US Government, which provides releases of census data from the US Census Bureau (http://www.census.gov/main/www/cen2000.html). Most of the data from primary sources is not structured for relational databases and provides challenges for use in the classroom.

A number of organizations restructure this raw data from primary sources and provide it in a more usable format. Some of these organizations only provide web-based methods for exploring and analyzing the data. An example of this is Google's Public Census Data page (http://www.google.com/publicdata/explore?ds=kf7tgg1uo9ude_) which allows you to graph different components of the census data, but not download it. Some organizations do provide well-structured data sets for free download. An example of this is Infochips, which provides a variety of free and fee-based data sets. One set is the daily stock prices of all the companies on the NYSE from 1970-2010. The main problem with these large data sets is that they are generally one large table and don't have the relational structure needed in a database. More specifically, they do not have multiple tables connected by one-to-many relationships, which are the key to many learning activities in database courses.

After attempting to apply a variety of data sources, including the textbook sample databases, instructor-developed databases, and free public domain databases, it was clear these solutions did not address the data needs for the department's curriculum and a different approach to data resources was needed.

4 Seeking Partnerships

As part of an effort to build and deploy a new approach to data resources supporting the curriculum and to realize the benefits from an authentic and large data set, the authors sought partnerships with both industry and government organizations. The authors established an agreement with a local national retail sales organization to obtain legacy sales data for use in the CS/CIS curriculum. This data was to be incorporated into a large database repository and made

available to students through virtual desktop access to database tools. The authors also sought funding through the National Science Foundation (NSF) in the form of a Transforming Undergraduate Education in Science, Technology, Engineering, and Mathematics (TUES) grant to support the investments in the technology and curriculum development needed to take advantage of the data set. This effort was aimed at improving student learning through the use of inquiry-based modules and to increase access to the data for students enrolled at remote and online campuses through the use of the virtual desktop.

Although a partnership with the local retail sales firm was made, this partnership was dependent upon the receipt of the NSF TUES grant. Unfortunately, the department was not awarded the grant. The grant reviewers found value in the large data set and the partnership with industry, but the limited ability to share the data set with external institutions and the high reliance on grant funding to support the hardware infrastructure resulted in overriding concerns. The department was not able to sufficiently scale the virtual environment using minimal grant funding for the technology infrastructure. As a result, the grant request was not funded.

5 Next Steps

Without the support of grant funding, further options must be pursued. We are pursuing a number of options for accessing large databases without grant funding. First, we continue to search the web for good examples of large databases that cover problems domains relevant to our students and are in a format we can download to include in our courses. This includes investigating the data and tools available through Teradata University Network.

A second option, which we started using in an Applied Statistics course, is using data from sources internal to the college. Some learning assignments for the statistics course were designed around a table of 1000+ university courses, which are schedule each semester. The College provides access to a spreadsheet listing the course information including the course name, instructor, meeting times, and location. Unfortunately, this is just a single table and we are investigating options for using anonymized data from other related tables in our courses.

As a final approach, the department continues to discuss collaboration opportunities with our industry partners and works to find a mutually beneficial option for including authentic corporate data in the curriculum. As the use of data in the curriculum continues to expand, the need for a robust and authentic data source becomes more important.

6 Future Needs

The need for a robust set of authentic data continues to grow. The department is currently preparing to offer a new course in business intelligence in response to the increasing demand for more data analysis skills. This course will require the use of a large data set to support multiple forms of quantitative analysis and must include authentic data to provide meaningful results. Students in this course will develop hypotheses, design an analysis, conduct the analysis, and

draw conclusions and recommendations based on the findings. This course will provide an ideal platform to apply the inquiry-based learning environment.

In addition to the new business intelligence course a need exists to further expand the existing curriculum to incorporate data analysis techniques and skills into existing courses. Faculty have expressed concerns over the analytical skills of our students and, by offering data analysis opportunities, we hope to further develop these analytical skills and improve critical thinking. Similar to the needs of the database and programming courses, data is needed to facilitate an inquiry-based learning environment for the curriculum. The use of exploration, experimentation, and problem solving will foster the critical thinking skills needed to prepare students to succeed in the workforce and to support their continued growth. A large data set will support inquiry-based learning and can be applied to courses within the department's curriculum and may also be integrated into the curriculum of other academic departments.

References

The Academy of Inquiry-Based Learning (n.d.). What is IBL? - The Academy of Inquiry Based Learning. Retrieved from http://www.inquirybasedlearning.org/?page=What_is_IBL

ACM/IEEE Interim Review Task Force (2008). *Computer science curriculum 2008: An interim revision of CS 2001*. Retrieved from http://www.acm.org//education/curricula/ComputerScience2008.pdf

Cohen, D.W. (1982). A modified Moore method for teaching undergraduate mathematics. *The American Mathematical Monthly*, *89*(7), 473-490.

Coronel, C., Morris, S., Rob, P. (2010). Database Systems: Design, Implementation, and Management. Boston, MA: Course Technology.

Davis, C.G., & Finelli, C.J. (2007). Diversity and retention in engineering. *New Directions for Teaching and Learning,* (111), 63-71.

Hall, T.E. (2005). Critical thinking, self-direction and online learning: A practical inquiry perspective in higher education (Doctoral dissertation, University of South Dakota). Retrieved from http://dl.acm.org/citation.cfm?id=1104107.

Magnussen, L., Ishida, D., & Itano, J. (2000). The impact of the use of inquiry-based learning as a teaching methodology on the development of critical thinking. *Journal of Nursing Education, 39*(8), 360-364.

Pintrich, P., & Zusho, A. (2002). Student motivation and self-regulated learning in the college classroom. In J.C. Smart and W.G. Tierney (Eds.), *Higher education: Handbook of theory and research* (pp. 371-810). Boston, MA: Kluwer.

Wigfield, A., & Eccles, J.S. (2000). Expectancy-value theory of motivation. *Contemporary Educational Psychology*, *25*(1), 68-81.