AC 2009-1644: IMPROVING DATABASE ENGINEERING CURRICULUM

Reza Sanati-Mehrizy, Utah Valley State College

Afsaneh Minaie, Utah Valley University

Improving Database Engineering Curriculum

Abstract:

Our university is a liberal art university with primarily undergraduate programs located in an area with many technology oriented business. In our Computer Science program, we offer a Database Engineering area of specialization which includes a number of database related courses but does not include any data mining related course.

A study has shown that some universities and colleges offer very few database related courses and do not offer any data mining course. On the other hand, many universities offer more than one database related courses and they also offer data mining course(s). But mostly these universities offer their data mining course as a graduate course. Therefore, the students who graduate from the universities like our university with no graduate computer science program will not have the opportunity to get Data Mining knowledge.

To improve our Database Engineering curriculum, we have decided to add a new Data Mining course to this curriculum. The paper first presents this curriculum and then elaborates the content detail of this Data Mining course.

Introduction:

In addition to associate degrees, the Computer Science and Pre-Engineering department offers a Bachelor's Degree in Computer Science with four areas of specialization: Computer Science (traditional), Computer Engineering, Computer Networking, and Database Engineering.

A study and curriculum comparison has shown that some universities and colleges offer only one database related course in their undergraduate curriculum while others, are offering more than one course in this field. The reason is that it is impossible to teach all the required subjects (theory, application, administration, etc.) in only one course. Also, experience indicates that the database job market expects our graduates to have enough expertise to be able to install and configure their Database Management System, write application programs, design the database, maintain and administer their database system. For this reason, we decided that it might be best to add a new area of specialization to our Computer Science program called Database Engineering.

Recently, we have been evaluating the content of our Database Engineering curriculum to make sure that we are covering enough materials in this track. In this process, we consulted with our advisory board members, industry experts and academic professionals in this field. This consultation has concluded that our Database Engineering curriculum does not address Data Mining / Data Warehousing areas. To remedy this issue, our Database Engineering curriculum has been extended by addition of a new data mining course. This course will be a core course for our Database Engineering area of specialization while it can be an elective for the other areas.

The following sections outline the Database Engineering curriculum and then the content of the new Data Mining course will be presented.

Database Engineering Curriculum

Currently, our Computer Science department is offering a Database Area of Specialization within computer science program. In order to graduate with a Computer Science degree, students must complete 123 semester hours of course work. These 123 credit hours include 36 credit hours general study, 45 credit hours of Computer Science core requirements, and 42 credit hours requirements for student graduating in Database Area of Specialization listed as follows:

General Study (36 Credit Hours)

All students graduating from XXX must complete the following 36 credit hours general study.

• ENGL 1010	Introduction to Writing	3.0
• ENGL 2020	Intermediate Writing	3.0
• PHIL 2050	Ethics & Value	3.0
• ECON or HIST or PLSC		3.0
• HLTH 1100 or PES 1097		2.0
• COMM 1020	Public Speaking	3.0
• COMM 2110	Interpersonal Communication	3.0
• PHYS 2210 & 2215		5.0
• PHYS 2220 & 2225		5.0
• FINE ART		3.0
• BIOLOGY		3.0

Computer Science Core Requirements (45 Credit Hours)

• CS 1400	Object-Oriented Programming I	3.0
• CS 1410	Object-Oriented Programming II	3.0
• CS 2300	Discrete Structure I	3.0
• CS 2420	Object-Oriented Data Structures	3.0
• CS 2600	Fundamentals of Data Communications	3.0
• CS 2810	Assembly Language & Computer Architecture	3.0
• CS 301R	Invited Speakers Series	1.0
• CS 3050	Computer Ethics	3.0
• CS 3060	Operating Systems Theory	3.0
• CS 3240	Introduction to Computational Theory	3.0
• CS 3690	Advanced Topics in Data Communications	3.0
• Math 1210	Calculus I	5.0
• Math 1220	Calculus II	5.0
• Math 2040	Principles of Statistics I & II	4.0

Database Engineering Requirements (42 Credit Hours)

All students graduating in database area of specialization must complete 42 credit hours which consist of 27 credit hours core and 15 credit hours electives:

Database Engineering Core Courses:

1. **CS 2450, Software Engineering (3.0 CR)**
   Presents concepts, methodology and best-practices necessary to develop large scale software projects. Includes step-wise software requirements analysis, design, implementation, testing and release. Discusses software generation, reuse, scheduling, verification, and maintenance. Emphasizes current "real world" industry best-practices and tools.
   Prerequisite: CS 2420.

2. **One of the followings:**
   CS 3220, Visual Basic Software Development
   CS 3250, Java Software Development
   CS 3260, C#.net Software Development
   CS 3370, Advanced C++ Software Development

3. **CS 3410, Human Factor in Software Engineering (3.0 CR)**
   Uses a systems approach in designing interactive multimedia products to create user experiences that enhance and extend the way people work and communicate. Stresses an iterative process of design and evaluation based on theory and good practice are needed to create usable products.
   Prerequisite(s): CS 3220 or 3250 or 3370 or 3550 or INFO 2200

4. **INFO 3410, Database Systems (3.0 CR)**
   Advanced SQL queries, DB Programming (Stored Procedures, Triggers, Control Structures), Physical design/physical environment (Data types, Physical Integrity, DBMS managed constraints, Primary key, Foreign key, Databases in Internet/Client Server environments, Front end/Back end connectivity (ADO.NET, OBDC, OLE DB, JBDC), Connections, Commands, Data sets, Introduction to Database Administration (Concurrency issues, Transaction Processing, Backup and recovery), Introduction to Data warehousing, Introduction to Distributed databases.
   Prerequisite: (INFO 1420 or CS 2810) and (INFO 2050 or CS 3520) and (INFO 2200 or CS 1410)

5. **CS 3520, Database Theory (3.0 CR)**
   Introduces theory, concepts, architecture, and use of Database Management Systems (DBMS). Presents the relational and object-oriented database models used in both centralized and client/server databases. Discusses basic constraints and Structured Query Language (SQL), database models,

History of database systems, database design, entity relationship and enhanced entity relationship, UML, primary key, foreign key, mapping ER/EER to schema, functional dependencies and normalization process, merging schemas, indexing, disk storage, basic file structures and hashing, relational algebra and relational calculus relating to database management system, join operations (inner join, outer join, semi join), etc.
Prerequisite: CS 2300, CS 2420.

6. **CS 4100, Database Management System Construction (3.0 CR)**
   This course looks at issues involved in actually implementing a DBMS. Students will implement from scratch a relational DBMS with added features such as indexing.
   Prerequisite: CS 3520 and (CS 3220 or 3250 or 3260 or 3370).

7. **INFO 4410, Database Administration (3.0 CR)**
   Installation and configuration, Setting up the operational environment, Setting up users and permissions, securing the database, distributed databases, clustering, replication, Performance monitoring and tuning, reconfiguring the database, developers vs. operational needs vs. management needs, Backup and recovery.
   Prerequisite: INFO 3410 or INFO 3440.

8. **CS 4500, Advanced topics in Database Systems (3.0 CR)**
   Transaction Processing, Concurrency Control Techniques, Database Recovery Techniques, Database Security and Authorization, Database Integrity, Distributed Databases and Client-Server Architectures, Load balancing. Data Warehousing, Data Mining, Database Machines, Mobile Database, Multimedia Database, GIS, Genome Data Management, Data Fragmentation, Data Encryption, Locking, Deadlock, etc.
   Prerequisite: CS 3520 or INFO 3410**.**

9. **CS 4620, Data Mining (3.0 CR)**
   An introduction to the process of knowledge discovery and the basic theory of automatic extracting models from data, validating those models, solving the problems of how to extract (mine) valid, useful, and previously unknown interesting patterns from a source (database or web) which contains an overwhelming amount of information. Explaining various models(decision trees, association rules, linear model, clustering, bayesian network, neural network) and how to apply them in practice. Algorithms applied include searching for patterns in the data, using machine learning, and applying artificial intelligence techniques. Students will study and implement several relevant algorithms, and use existing tools to mine real-world, business driven databases.
   Prerequisite: CS 3520 Database Theory

Database Engineering Electives (15 Credit Hours)

Students graduating in database area of specialization must complete an additional 15 credit hours from the following list of electives not already taken:

- CS 4470 Artificial Intelligence 3.0
- CS 3660 Web Server Administration Programming 3.0
- CS 3670 Network Programming 3.0
- CS 4410 Human Factors in Software Engineering 3.0
- CS 3220 Visual Basic Software Development or
  CS 3250 Java Software Development or
  CS 3260 C#.net Software Development or
  CS 3370 Advanced C++ Software Development 3.0
- CS 481R Internship Work Experience (max of 3) 3.0
- CS 3400 Software Engineering I 3.0
- CS 4400 Software Engineering II 3.0
- CS 4230 Software Testing and Quality Assurance 3.0
- CS 3550 Internet Software Development 3.0
- CS 4510 Operating System Design and Simulation 3.0
- CS 3540 Game Programming 3.0

More courses from the list of database electives can be transferred to the list of database core courses if someone desires to do so.

Rational for Undergraduate Data Mining Course

As the volume of data is increasing so quickly and businesses are becoming more data-rich but relatively knowledge-poor, data mining courses is becoming more popular in the field of computer science. Data mining courses are mostly offered at the graduate level. The reason could be because a data mining course may require the knowledge of statistics, neural network and machine learning. This will not provide an opportunity for some graduates from higher educations institutions with liberal art programs to gain the extremely important knowledge of data mining.

An undergraduate data mining course can introduce students to data mining concept by applying statistical, neural network, and machine learning algorithms necessary for knowledge discovery. In the Data Mining course, students will gain considerable knowledge and insight into the process of discovering patterns in large amounts of data for the purpose of solving problems, gaining knowledge, and making predictions. Teaching Data Mining course at the University of Minnesota, Morris[1] has shown that teaching data mining at the undergraduate level is appropriate and can be successful. Using data mining software packages such as Weka [2], will provide students the hands-on experience with data mining applications that they need in their career. Using these software packages, students apply the algorithms they have learned to solve the problems.

An undergraduate data mining course provides a strong opportunity for students to practice data structures, and enhance their understanding of algorithms [3]. Further more, it can be designed to integrate current research topics into the curriculum and promote the student's critical thinking and problem-solving skills [4]. Students taking this type of course will gain experience with applied research and engaged learning and will be more prepared for graduate research and study.

CS 4620: Data Mining Course Content

Data mining course can be custom built around the idea of using research-level papers as the primary reading material for the course, and implementing data mining algorithms for the assignments or it can be constructed based on an existing data mining textbook. A few team projects to implement data mining algorithms using data mining software packages such as Weka will provide students to gain hands-on experience. The following presents a sixteen weeks course content and some possible projects. (*) marks more advanced topics which can be skipped for a less advanced course. This course content is built based the data mining course outline form KDnuggets company [5].

Week1: Introduction: Machine Learning and Data Mining
- Data Flood
- Data Mining Application Examples
- Data Mining and Knowledge Discovery
- Data Mining Tasks

Week2: Machine Learning and Classification
- Machine Learning and Classification
- Examples
- *Learning as Search
- *Bias
- Weka

Week3: Input: Concepts, instances, attributes
- What is a concept?
- What is an example?
- What is an attribute?
- Preparing the data

Week4: Output: Knowledge Representation
- Decision tables
- Decision trees
- Decision rules
- Rules involving relations
- Instance-based representation

Week5: Classification - Basic methods
- OneR (1-rule)
- NaiveBayes

Week6: Classification: Decision Trees
- Top-Down Decision Trees

- Choosing the Splitting Attribute
- Information Gain and Gain ratio

Week7: Classification: CART

- CART (Classification and Regression Tree) Overview and Gymtutor Tutorial Example
- Splitting Criteria
- Handling Missing Values
- Pruning : Finding Optimal Tree

Week8: Classification: More Methods

- Handling Numeric Attributes
  Finding Best Split
- Dealing with Missing Values
- Pruning
  Pre-pruning, Post-Pruning, Estimating Error Rates
- From Trees to Rules

Week9: Classification: More Methods

- Rules
- Regression
- Instance-based (Nearest neighbor)

Week10: Evaluation and Credibility

- Introduction
- Classification with Train, Test, and Validation sets
  Handling Unbalanced Data; Parameter Tuning
- *Predicting Performance
- Evaluation on "small data": Cross-validation
- *Bootstrap
- Comparing Data Mining Schemes
- *Choosing a Loss Function

Week11: Evaluation - Lift and Costs

- Lift and Gains charts
- *ROC (Receiver Operation Characteristics)
- Cost-sensitive learning
- Evaluating numeric predictions

Week12: Data Preparation for Knowledge Discovery

- Data understanding
- Data cleaning
- Date transformation
- Discretization

Week 13: More on Data Preparation for Knowledge Discovery

- False "predictors" (information leakers)
- Feature reduction, leaker detection
- Randomization
- Learning with unbalanced data
- Principle Component Analysis

Week14: Clustering

- Introduction
- K-means

- Hierarchical

Week15: Associations
- Transactions
- Frequent itemsets
- Association rules
- Applications

Week16: Visualization
- Graphical excellence and lie factor
- Representing data in 1,2, and 3-D
- Representing data in 4+ dimensions
    - Parallel coordinates
    - Scatterplots
    - Stick figures
    - Course Summary

Additional Materials as Time Allows:

Applications: Targeted Marketing and Customer Modeling
Direct Marketing Review
Evaluation: Lift, Gains
Lift and Benefit estimation

Data Mining and Society;
Future Directions
Data Mining and Society: Ethics, Privacy, and Security issues
Future Directions for Data Mining
Web mining, text mining, multi-media data

Course Evaluation:

Midterm Exam
Final Exam
Team Projects:
Project #1: SQL Exercise
Project #2: Association Rule Mining
Project #3: Weka Exercise
Project #4: Hierarchical Clustering
Project #5: Logistic Regression

Conclusion

Students graduating in our computer science in the Database Engineering area of specialization will have a strong background in this field which makes them more marketable. Graduating students with this extensive database knowledge satisfies the area's industrial demand. Also, having such an area of specialization will have positive effects on our program enrolment.

Graduates from other programs may select any of these new database courses as core or elective. Higher education institutions abroad may benefit from implementing this area of specialization in their computer science program. This paper provided detailed descriptions of our Database Engineering track that one can use to build a similar area of specialization.

References

1. Lopez, Dian and Ludwig, Luke, "Data Mining at the Undergraduate Level", Proceedings of the 34th Annual Midwest Instruction and Computing Symposium, 2001.
2. Weka, http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html.
3. David R. Musicant, **"**A data mining course for computer science: primary sources and implementations", Proceedings of SIGCSE annual conference, 2006.
4. Terri L. Lenox and Carolyn Cuff, "Development of a Data Mining Course for Undergraduate Students", Proceedings of the 19th Annual Conference for Information Systems Educators, 2002.
5. http://www.kdnuggets.com/data_mining_course/course_outline.html, January, 2009.