COMP9318 Course Introduction

Units of credit	6
Parallel teaching	no

1. Why this course

Job opportunities in Data Mining.

E.g., see the recent job ad (at http://www.cs.wisc.edu/dbworld/messages/2007-03/1173318394.html):

```
Intern Position Description:



Yahoo! SDS Overview
SDS' mission is to create value to consumers and marketers by
delivering a consumer-centric data platform and insights services that
maximize user engagement and enable innovative marketing solutions.

Data Driven Applications (DDA) Group Overview
Data Driven Applications Group is a part of SDS and is chartered to
rapidly spread the power of data and data mining to business centers
to help generate brand and top-line growth by being able to target
users with personalized advertisements and content.

Introduction:

The intern gets the opportunity to work on data driven projects which
create revenue of multi million dollars for Yahoo business units via
personalization and targeting. Of course, you also get the chance to
enjoy technical challenges worth multi million dollars.

Essential Job Functions:

Work with DDA team to:
* Process user data
* Study user behavior
* Build targeting models
* Analyze quality of services

Requirements:

* Exceptional software development skills in C, C++, Perl and Php.
* Strong technical problem solving skills
* Knowledge of data mining technologies, modeling, tuning and testing.
* Self motivated team player
* Good verbal and written communication skills
* Flexible in a dynamic product development environment

Length: 3 months

Location: Sunnyvale, CA

Contact:

If you feel excited about this opportunity, please send your resume
to:

Shu-Yao Chien
csy2007intern@yahoo.com
```

2. Pre-requisites

The formal pre-requisite for this course is COMP2011 Data Organisation and COMP3311/9311 Database Systems.

The knowledge that we assume from COMP2011 is:

- experience with procedural programming, and an understanding of a range of in-memory searching data structures (e.g., binary search trees, 2-4 trees, and hash-tables).

The knowledge that we assume from COMP3311/9311 is:

- experience with relational data model and SQL query language, and an understanding of data structures and algorithms to enable efficient and scalable management of massive amount of data (e.g., B+-tree indexes).

3. Course philosophy and teaching strategies

The learning foci in this course are primarily lectures (theoretical knowledge) and projects (practical knowledge). The course will have an emphasis on problem solving for real applications.

Students will learn the main contents of the course through lectures. Tutorials are available to assist students to obtain in-depth understanding of course materials and develop problem solving skills by working on tutorial questions.

4. Course aims

This course aims to introduce the foundation of data warehousing, the theories of various data mining techniques and explore the practice of developing data mining applications. This course is one of the advanced database course series. Other advanced database courses include: COMP9315 (DBMS Implementation), COMP9314 (Next Generation Database Systems), COMP9317 (XML and Databases), COMP9321 (E-Commerce Systems Implementation Infrastructure), etc.

The course is designed to be practical. As such, real-life examples of data mining issues and applications will also be used throughout the course.

5. Learning outcomes

Students successfully completing this course will be able to:

- understand the whole process of data mining and knowledge discovery from databases
- understand the data models and query processing mechanisms used in Data Warehouses
- understand various data mining techniques and their variants
- develop solutions for real problems using existing data warehousing and data mining technologies.
- appreciate the past, present and future of data warehousing and data mining technology

The learning outcomes are closely related to UNSW graduate attributes 1 -- 6. For example, bonus questions will be given in the assignment in order to encourage independent learning and critical thinking.

6. Administrative Components

See the course hompage (http://www.cse.unsw.edu.au/~cs9318) for (up-to-date) information regarding **Course Staff**, **Course Schedule**, and **Course Resource List**.

As the course is continuously developing together with the data warehousing and data mining fields, the course schedule is subject to change too. Please read the introduction slides (lecture notes of the first week) for the course schedule in the current offering.

7. Assessment

```
q      = average mark for quizzes
ass1   = mark for written assignment 1
proj1  = mark for programming project 1
exam   = mark for final exam

t      = (q + ass1 + proj1) / 3
grade  = (exam * t) / (0.5 * exam + 0.5 * t)
```

Both written and programming assignments are helpful to achieve deep understanding of the course materials and develop problem solving abilities.

Relationship to the Learning Outcomes:

- All assignments and the final examination are used to assess whether you have achieved the objectives of the subject.
- The programming assignment, in particular, gives you the opportunities to engage in real problems, and helps you to develop in-depth understanding of the course materials and develop problem-solving skills.

Grading Criteria: Grading criteria for each assessment will be detailed in the specification.

Late submission: Assignments/projects submitted late are subject to late penalties, which are specified in the assignment/project specifications. “Soft” Late penalties are normally used in this course, which only reduces the maiximum mark obtainable. Thus if the assignment is marked out of 10, and students A and B hand in assignments worth 9 and 7, both receiving 20% penalty, then the maximum mark obtainable is 8, so A gets min(9, 8) = 8 and B gets min(7,8)

= 7.

Assignment submission: Assignment submission procedure is described in the assignment specification document, which will be linked to this page when the assignment specification becomes available. Generally assignments are submitted electronically using the give program running on the School's computer systems (in labs, and on servers). Details are in the assignment specifications.

8. Reading e-mail

You should check your school e-mail frequently in case of announcements relating to this course. We assume that you read e-mail sent to your CSE account by the next working day during teaching sessions.

- How to forward your CSE E-mail to another e-mail address
- CSE E-mail guidelines

9. Academic honesty and plagiarism

Copying assignments is unacceptable. Assignments will be checked. The penalties for copying range from receiving no marks for the assignment, through receiving a mark of 00 FL for the course, to expulsion from UNSW (for repeat offenders). Allowing someone to copy your work counts as plagiarism, even if you can prove that it is your work.

Further details of the School plagiarism policy can be found here. (You acknowledged receipt of these rules when you obtained your CSE computer account, and the link above is for your convenience so that you can review the rules now.)

We are aware that a lot of learning takes place in student conversations, and don't wish to discourage those. However, it is important, for both those helping others and those being helped, not to provide/accept any programming language code in writing, as this is apt to be used exactly as is, and lead to plagiarism penalties for both the supplier and the copier of the codes. Write something on a piece of paper, by all means, but tear it up/take it away when the discussion is over.

If you are new to studying in Australia, be aware that attitudes to plagiarism at UNSW may be different from those in your home country. Make sure you are clear about the rules here at UNSW. In brief, and for the purposes of COMP9318, plagiarism includes copying or obtaining all, or a substantial part, of the material for your assignment, whether programming language code, or written or graphical report material, without written acknowledgement in your assignment from:

- location on the internet;
- a book, article or other written document (whether published or unpublished) whether electronic or on paper or other medium;
- another student, whether in your class or another class;
- a non-student (e.g. from someone who writes assignments for money)

Note that if you copy code or other material from another student or non-student with acknowledgement, you will not be penalised for plagiarism, but you are unlikely to get any marks for the copied material. If you use code found in a publication (on the internet or otherwise) then the marks you get for this will be at the marker's discretion, and will reflect the marker's perception of the amount of work you put into finding and/or adapting the code, and the degree to which you understand the code.

Note also that there is a big difference between being able to understand someone else's code, and writing that code yourself from scratch. A computer programmer has to be able to write code from scratch. The assignments provide opportunities for you to develop the skills necessary to write your own code. Use these opportunities!

10. Course Material

- All lecture notes, tutorial and lab notes will be posted on the course web site. These notes summarises the major contents and help you to understand the materials when you read the textbook later. You definitely need to read the corresponding chapters in the textbook to gain a full understanding of the course materials.
- [HK00] is the textbook for this course. We got three copies on **standard loan** in the library, two on **two hour loan**.

```
[HK00] Data Mining: Concepts and Techniques, Jiawei Han and
Micheline Kamber. Kaufmann Publishers, August 2000. ISBN:
1-55860-489-8.
```

- Reference books for this course are:

```
[WF00] Data Mining : Practical Machine Learning Tools and
Techniques with Java Implementations, Ian H. Witten, Eibe
Frank. Morgan Kaufmann, 2000. ISBN: 1558605525.
```

- Useful Resources

- KD Nuggets
- The OLAP Report

11. Further information

- Students enrolled in COMP9318 are expected to attend all classes
- The use of School of Computer Science and Engineering computing laboratories is subject to rules described in the [[http://www.cse.unsw.edu.au/people/studentoffice/policies/yellowform.html][Yellow Form]], which you acknowledge (electronic) receipt of when you receive your computing account. The Yellow Form also outlines what to do in case illness or misadventure that affects your assessment, and supplementary examinations procedures within the School of Computer Science and Engineering.
- Information on UNSW Occupational Health and Safety policies and expectations
- Equity and Diversity issues

12. Continual Course Improvement

Each year feedback is sought from students and other stakeholders about the courses offered in the School and continual improvements are made based on this feedback. UNSW's Course and Teaching Evaluation and Improvement (CATEI) Process (http://www.unsw.edu.au/learning/pve/catei.html) is one of the ways in which student evaluative feedback is gathered. Significant changes to courses and programs within the School are communicated to subsequent cohorts of students.

There were no serious issues noted in the previous evaluation. We are making incremental refinements to preserve the already high standard of the course.

Wei Wang, Feburary 2009