Information Systems Department
University of Maryland Baltimore County
Baltimore Maryland 21250

IS 733 Data Warehousing and Mining
Spring 2008

Instructor: Dr. Vandana Janeja
Office: ITE 429
e-mail vjaneja@umbc.edu : please put IS 733 in subject line
Course Delivery Site http://blackboard.umbc.edu
Office Hours: Thursday 2 – 4 (other times by appointment)

Meeting Times: Th 4:30pm - 7:00pm in ITE 469

Textbook:

Main text book : Jiawei Han and Micheline Kamber , Data Mining: Concepts and Techniques, 2nd ed.

Reference book: Ian H. Witten, Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)

Course Description: The purpose of this course is to provide a comprehensive discussion on using organizational data sources to enable decision support through warehousing and mining of data. The course will provide an in-depth understanding of the conceptual and application issues in the area of data mining and data warehousing. The major focus this semester will be on Data Mining algorithms and applications with hands on look at implementation issues.

Instructional Methods: Discussion, Lectures and Demonstrations

Attendance and Participation: Regular and punctual attendance is expected of all students. In the case of absence due to emergency (illness, death in the family, accident), religious holiday, or participation in official College functions, it is the student's responsibility to confer with the instructor about the absence and missed course work.

Class Preparation: All of the reading and homework assignments should be completed before the class in which the material is to be discussed. The students are expected to review the chapter before each class.

Course Requirements:

Regular Punctual Attendance
Tests
Class Assignments & Homework
Programming Projects

Grading:

IS instructors are expected to have exams and evaluations consisting of a mix of class work, test, homework and programming projects, which result in a reasonable distribution of grades. The break up for this class is as follows:

Readings / In Class Presentation: 15 points
Assignment 1 : 15 Points
Assignment 2 : 15 Points
Midterm (In class – closed book): 25 points
Implementation + Term Paper : 15+15

Grading criteria: With respect to final letter grades, the University's Undergraduate Catalogue states that, "A, indicates superior achievement; B, good performance; C, adequate performance; D, minimal performance; F, failure" There is specifically no mention of any numerical scores associated with these letter grades. Final letter grades in this course conform to the University's officially published definitions of the respective letter grades. In accordance with the published University grading policy, it is important to understand that final letter grades reflect academic achievement and not effort. While mistakes in the arithmetic computation of grades and grade recording errors will always be corrected, it is important to understand that in all other situations final letter grades are not negotiable and challenges to final letter grades are not entertained. For this course it is anticipated that "A" grades may be in the 90-99 range, "B" grades may be from 80-89 and "C" grades range from 70-79. All points from each Learning Unit are additive. Each student starts at zero points which is an "F", any other grade must be earned. This grading criteria is a guideline and is subject to change based on the discretion of the instructor.

In Class Presentation
The presentation is based on a paper (based on a chapter topic) that you have selected from a reading list or any other paper that you have selected after consulting with me. This is also an individual presentation for about 15 minutes using power point slides. In this presentation you can discuss the problem area, the methodology, any implementation or experimental results discussed in the paper. This is an overview presentation and you are not expected to go into all the scientific details, although the presentation should be clear and present the overall approach in an understandable manner.

Term paper
The term paper will be on a topic you have selected for your final implementation project. It is strongly recommended that you select the in class presentation paper related to the term paper topic. It should be at least 20-25 pages in length (double spaced, 12 point font size with 1" margin all round). It should be formatted like a journal paper (please refer to the IEEE format). The first page should have the title, your name, and an abstract not exceeding 500 words. It should be divided into the following sections with rough length of each shown within bracket: Introduction (2 pages), Background (7-10 pages), Methodology (5 pages), Implementation or Experimental results (5 pages), References (2-3 pages single spaced). Please make an appointment with me to discuss the term paper topic after giving it some thought. If you are unable to find a topic I can help you with it. The topic should be finalized by week 8 so that you have enough time to work on it. Some examples of term paper topic are: Outlier detection, privacy preserving data mining, time series/sequence analysis, mining biological data, mining financial data, healthcare/clinical applications, business applications etc. The methodology should include details of either the application of data mining/warehousing techniques in an area, or the development of new techniques that are validated with experimental results.

Academic Integrity: By enrolling in this course, each student assumes the responsibilities of an active participant in UMBC's scholarly community in which everyone's academic work and behavior are held to the highest standards of honesty. Cheating, fabricating, plagiarism, and helping others to commit these acts are all forms of academic dishonesty and they are wrong. Academic misconduct could result in disciplinary action that may include, but is not limited to, suspension or dismissal. Full policies on academic integrity should be available in the UMBC Student Handbook, Faculty Handbook, or the UMBC Directory.

Cheating in any form will not be tolerated in this class. *You may not copy other students' work or copy programs from the Internet. You will receive an F for any assignment found to be copied for the first time and any subsequent violations will result in immediate failure of the course.* **Any form of cheating will be reported and will stay on student's record for the rest of their term at UMBC with possible note on their transcripts.**

COURSE SCHEDULE
(Schedule subject to change)

Dates	Material Covered	Lab Work / work due
Week 1 – 1/31/08	Chapter 1. Introduction	

Week 2 – 2/7/08	Chapter 2. Data Preprocessing	
Week 3- 2/14/08	Chapter 3. Data Warehouse and OLAP Technology: An Overview	Lab excercises
Week 4 – 2/21/08	Chapter 4. Data Cube Computation and Data Generalization	Lab excercises - Assignment 1 handed out
Week 5 – 2/28/08	Chapter 5. Mining Frequent Patterns, Associations and Correlations	Assignment 1 Due
Week 6 – 3/6/08	Chapter 5 continued and Chapter 6. Classification and Prediction	Lab Excercises – Finalize selection of paper for in class presentation
Week 7 – 3/13/08	Mid Term	
Week 8 – 3/20/08	Spring Break March 16-23	Final Project one page writeup due via email
Week 9 – 3/27/08	Chapter 6 continued	Lab excercises
Week 10 - 4/3/08	Chapter 7. Cluster Analysis	
Week 11 – 4/10/08	Chapter 7 continued and Chapter 8. Mining Stream, Time-Series and Sequence Data	Assignment 2 handed out
Week 12 – 4/17/08	Chapter 8 continued	Assignment 2 Due
Week 13 – 4/24/08	Chapter 9. Graph Mining, Social Network Analysis and Multi-Relational Data Mining	
Week 14 – 5/1/08	Student Presentations	Student Presentations
Week 15 – 5/8/08	Chapter 10. Mining Object, Spatial, Multimedia, Text and Web Data Chapter - Discussion on Spatial Data Mining	
Week 16 – 5/15/08	Final Project/ term paper due via email	

Sample seminal papers for class presentation:

- R. Agrawal, T. Imielinski, A. Swami. "*Mining association rules between sets of items in a large database*." Proc. ACM SIGMOD Intl. Conference on Management of Data, pp. 207--216, 1993. 13, http://citeseer.ist.psu.edu/agrawal93mining.html > 900 citations
- Edwin M. Knorr and Raymond T. Ng. *Algorithms for mining distance-based outliers in large datasets*. In Proceedings of 24rd International Conference on Very Large Data Bases (VLDB 1998), pages 392--403. Morgan Kaufmann, San Francisco, CA, 1998, http://citeseer.ist.psu.edu/knorr98algorithm.html > 60 citations (For presentation you can focus on one of the algorithms only in the paper)
- T. Zhang, R. Ramakrishnan, and M. Livny. *BIRCH: An efficient data clustering method for very large databases*. In ACM SIGMOD Conference, 1996, http://citeseer.ist.psu.edu/zhang96birch.html > 200 citations
- Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. 1998. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international Conference on Management of Data* (Seattle, Washington, United States, June 01 - 04, 1998), http://www.cs.cornell.edu/johannes/papers/1998/sigmod1998-clique.pdf > 475 citations
- Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan, "*Web usage mining: Discovery and applications of usage patterns from web data*," SIGKDD Explorations, vol. 1, no. 2, pp. 1--12, Jan 2000, http://citeseer.ist.psu.edu/srivastava00web.html > 85 citations
- R. T. Ng and J. Han. Efficient and Effective Clustering Method for Spatial Data Mining. In Proceeding of the 20th VLDB Conference, pages 144--155, 1994, http://citeseer.ist.psu.edu/ng94efficient.html, > 240 citations

Other sources :

Data Mining Journals, Magazines, and Newsletters

- SIGKDD Explorations, the magazine of the professional association of data miners
- ACM TKDD: Transactions on Knowledge Discovery from Data
- Data Mining and Knowledge Discovery journal (now published by Springer).
- Other Journals and Magazines related to Data Mining
- Periodicals: Newsletters and Mailing Lists

Other Publications

- Text Mining, Text Analytics publications.
- Data Mining Overviews
- Business-oriented articles
- Technical articles
- DBLP Bibliography Server, index to database, logic programming, KDD, and other related publications.
- Web Mining publications

Useful Resources

1. **http://www-faculty.cs.uiuc.edu/~hanj/bk2/index.html**

2. Some important overview papers (Linked from KDNuggets) http://www.kdnuggets.com/publications/overviews.html)

- From Data Mining to Knowledge Discovery in Databases (PDF), Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, AI Magazine 17(3): Fall 1996, 37-54
- How to Choose a Data Mining Suite, by Robert Nisbet, DM Review, March 2004.
- Data Mining Tools: Which One is Best For CRM?, by Robert Nisbet, DM Review, January 2006.
- A Survey of Data Mining software Tools, by Michael Goebel and Le Gruenwald, SIGKDD Explorations, June 1999. Volume 1, Issue 1
- John F. Elder IV and Dean W. Abbott, A Comparison of Leading Data Mining Tools, tutorial presented at KDD-98, New York, New York, August 28, 1998.
- Crossing the Chasm: From Academic Machine Learning to Commercial Data Mining, Ronny Kohavi, invited talk at ICML-98, July 1998 See also software for data mining

Inclement Weather: Any work or test due on a class date that has been canceled due to inclement weather will be due the next class meeting. (If the semester's last exam is postponed, it will be given during the time period assigned during the University's official Final Exam week.)