

CSE392 Data Mining, FALL 2006
Professor Anita Wasilewska

Meets Tuesday Thursday 2:20 pm - 3:40 pm

Place ESSCI 069

Professor Anita Wasilewska

e-mail address: anita@cs.sunysb.edu

Office phone number: 632 8458

Office location: Computer Science Department building, office 1428.

Professor Office Hours Tuesday, Thursday 11:30 - 1:00 pm and by appointments.

Textbook (on reserve in CS Library)

DATA MINING Concepts and Techniques

Jiawei Han, Micheline Kamber

Morgan Kaufman Publishers, 2003

LECTURES NOTES posted on the Web

Course Description Data Mining, called also Knowledge Discovery in Databases (KDD) is a new multidisciplinary field, It brings together research and ideas from database technology, machine learning, neural networks, statistics, pattern recognition, knowledge based systems, information retrieval, high-performance computing, and data visualization. Its main focus is the automated extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories.

The course will closely follow the LECTURES NOTES and some chapters of the book and is designed to give a broad, yet in-depth overview of the Data Mining field and examine the most recognized techniques in a more rigorous detail. It also will explore the newest trends and developments of the field in form of students talks based on newest applications.

Grading There will be:

1. **Two** in class presentation (75pts each) given individually. Students will be graded for the presentation skills, the content, organization, clarity, and amount of work put into preparation.

3. Final Paper (50pts)

Final grade computation During the semester you can earn 200pts or more (in the case of extra points). The grade will be determine in the following way: $\#$ of earned points divided by 2 = % grade.

The grade will be determine in the following way: of earned points = % grade. The % grade which is translated into letter grade in a standard way i.e. 100 - 90 % is A range, 89 - 80 % is B range, 79 - 70 % is C range, 69 - 60 % is D range and F is below 60%.

Presentations This is an opportunity for students to learn how to deliver

1. a well structured and prepared lecture based on the textbook material,

2. understand and present a newest data mining application.

Students will have to prepare power point based lecture slides and explain in detail, with examples the material.

The hard copy (black and white in slide spread format) of the slides and the CD containing the presentation is to be delivered to the Professor before the presentation starts.

Presentation 1 (75 points). It is a detailed, lecture type presentation. It has to last at least one hour.

It will be based on, or extending the content of the book, or other books or lecture notes posted on the web. I will provide you my own slides on the subject and some students presentations from previous courses. You must IMPROVE on them and add your own material.

This is an **educational part**. The main goal of this presentation is to **teach others** the material. Students have to put time and effort into **understanding the material**, present it slowly and be prepared to answer students questions.

Remember that "I don't understand" is also an answer, but don't over-use it! The better answer is: "the book is not very clear, I think that it is ..., or I understood it as ...".

Presentation 1 subjects are to be discussed with Professor.

Presentation 2 (75pts) It is a shorter presentation (20-30 minutes). It is a presentation of a Data Mining major application. The goal of this part is to bring an update to what is being done on the subject today.

Students have a freedom of choice of their subjects. Presentation 2 CAN BE COMBINED with the Presentation 1 as a second part of it, or be done separately.

This presentation must consists of two parts.

Part 1 Short overview of the techniques, methods used as taught during the course or found in the literature.

Part 2 Detailed presentation of the application. If you present an application that is a part of a paper you must include on your slides authors names, title and place (journal, conference) where it was published and the date of the publication, or any other source of the paper you use. MAKE a copy of the paper and distribute it in class.

If you present a commercial application you must find relevant data about the application. You must distribute in class all materials you find relevant to the presentation.

General Principles of Presentations

First slide must contain your names, student ID, professor name, course number and the title.

Second slide must contain ALL sources you used for the LECTURE part of your presentation. The book is included. In the case of the book the reference you have to put are title of the chapter, sections and pages numbers.

Third slide is an OVERVIEW of your presentation.

Fourth slide include the title and bibliography of your sources.

Please, e-mail a text file containing information included on these 4 slides to me.

Remember to include give a source of any picture, of slides copied from a source or any DIRECT citation on the bottom of each of your slides where it appears.

Final Paper Here is the procedure:

Step 1 FIND (Web or other sources) two or three articles or research papers on DATA MINING subject of your choice.

Step 2 Write motivation why you have chosen this particular articles.

Step 3 Write at least one page summary of each article. In your own words.

Step 4 Write you own evaluation of each article. Address the following:

1. How well the article is written: motivation, the clarity of the subject and the goals and relevance to the field.
2. Any other remarks and your own reflections.

PAPER GENERAL PRINCIPLE Any direct citations (even of ONE SENTENCE!) must have a standard form of a citation: give the page of the paper and show clearly when it start and when it finishes.

Course Contents and Schedule

The course will follow the Lecture Notes and the book very closely and in particular we will cover the following chapters and subjects. The order does not need to be sequential.

Chapter 1 Introduction. General overview: what is Data Mining, which data, what kinds of patterns can be mined.

Chapter 3 Data preprocessing: data cleaning, data integration and transformation, data reduction, discretization and concept hierarchy generation.

Chapter 5 Concept Descriptions: Characteristic and Discriminant rules. Data Generalization.

Chapter 6 Mining Association Rules in Large Databases. Transactional databases and Apriori Algorithm .

Chapter 7 Classification and prediction. Decision Tree Induction ID3, C4.5).

Genetic algorithms and other methods.

Chapter 8 Cluster Analysis. A Categorization of major Clustering methods.

TRENDS and Developments - newest research and applications presentation.