Policy

CS512 "Data Mining: Principles and Algorithms"

About the Course

This is a graduate-level (or advanced) course on data mining. It introduces the principles, algorithms and applications of data mining, including algorithms, methods, implementations and applications. Topics to be covered include introduction to networks, mining information networks, mining sequence and graph data, advanced clustering methods, outlier analysis methods, mining stream data, mining spatiotemporal data, and mining text data, Web data, when time available. The course will serve mainly CS graduate students interested in data mining. Also, the course may attract students from other disciplines who need to understand, implement, and/or use data mining methods to analyze massive amount of data.

Prerequisites

Background: CS 412 or consent of instructor (good statistics and machine learning knowledge will help understand the course materials). We strongly encourage students to take the undergraduate level data mining course CS412, which is offered every semester. If you have missed it, please download CS412 course slides and class-notes to work on the materials along with the progress of the course.

Textbook

1. Jiawei Han, Micheline Kamber, and Jian Pei. "Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2011.
   - Note: Two chapters of the course will use this 3rd edition of the book (Chapters 11 & 12 of 3rd ed. of the book). The remaining chapters will use network mining materials plus the 2nd edition of the book (2006) starting at Chapter 8 (mainly replying on slides or related papers).
   - See the book's home page for errata, course slides, and other reference materials.
2. Yizhou Sun and Jiawei Han, Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool, 2012 (which will be used mainly in the first second half of the course).

References

The following texts are recommended for reference, and are also on reserve at Grainger Engineering Library. There are numerous other books or online resources on data mining available. The books marked in red are highly recommended textbooks.


Major conference proceedings that will be used in class, including ACM SIGKDD (KDD), ACM SIGMOD, VLDB, ICDM, SDM (SIAM Data Mining conference), ICDE, ICML, WWW, and other related conferences. For students who take courses, you can obtain passwords in class for some sets of conference proceedings resources.

A set of readings can be accessed at the list of readings

Course Format, Activities, and Evaluation

This course will draw materials mainly from the textbook and some recent research literature. Students will study the materials and complete all the course requirements.

Assignments and Course Project

1. There will be about two homework assignments (no programming assignment)
2. One theme-based survey report (plus a set of companion slides), and
3. One research project (which will be first presented in class and then hand in written report)

- The theme-based survey report will be on the specified sections and will be determined based on your assignments: It will be evaluated with a similar standard as a survey paper publishable in a journal or magazine (Note: We encourage each survey paper be done in a small group, and each will need to identify the major portion you have worked on: attaching author names to sections.)
- The course project will be evaluated in a similar standard as a research paper publishable in a conference.

Examination

There will be two midterm exams (Each midterm exam will be 75 minutes in length, possibly 90 minutes long based on the number of questions given, and one page of cheat sheet is allowed/encouraged in exams).

1. The first is around the 7th week of the course, and
2. The second is around the 12th week of the course (right before the project presentation)

Evaluation

We plan to determine final grades of the course in the following way:

1. Two Midterm exams: 40% in total (20% each)
2. Assignments: 10% (2 assignments in total)
3. Class attendance (3%): Max misses w/o penalty: 3, then 0.3% for each miss
4. Class presentation and/or research survey (12% total)
   1. Class presentation: May use 10 min. class survey presentation to replace the survey report (consent of instructor)—contents must closely aligned with the class content and in very high technical quality
   2. Hand-in together with companion presentation slides [due at the end of 12th week]
   3. Encourage to align up with your research project topic domain
   4. Survey report [expect to be comprehensive and in high quality, 15-20 pages]
5. Research project proposal (one-page): 0% (due at the end of 4th week)
6. Final course project: 35% (due at the end of semester)
   1. Evaluated by class (50%) and TA + instructor (50%) collectively!
   2. The final project will be evaluated based on (1) technical innovation, (2) thoroughness of the work, and (3) clarity of presentation
   3. The final project will need to hand in: (1) project report (length will be similar to a typical 8-12 page double-column conference paper), and (2) project presentation slides (which is required for both online and on-campus students)
   4. Each course project for every on-campus student will be evaluated collectively by instructor (plus TA) and other on-campus students in the same class
   5. The course project for online students will be evaluated by instructors and TAs only