Web Mining Seminar
Spring 2008
MWF 11:10-12:00pm Maginnes 113

Instructor: Dr. Brian D. Davison
Dept. of Computer Science & Engineering
Lehigh University

davison@cse.lehigh.edu
http://www.cse.lehigh.edu/~brian/course/webmining/
Course Objectives

- To gain a background in web mining techniques
- To become proficient at reading technical papers
- To gain knowledge of important current web mining research
- To gain experience presenting technical material
- To learn to write critical reviews of research papers
- To explore a research project in some depth and write and present a technical paper summarizing that work
Teaching materials

- **Required Text:**

- **Optional Text:**
  - *Data Mining: Practical Machine Learning Tools and Techniques, 2nd Ed.* By Witten and Frank, Morgan Kaufmann

- **Papers:**
  - Most (perhaps all) available online
    - Author's homepages
    - Citeseer/ResearchIndex
    - Google Scholar
    - ACM Digital Library
    - IEEEExplore
Seminars are less formal

- We have a small class
- Introduce yourselves!
Introduction to

Web Mining
What is data mining?

- Data mining is also called *knowledge discovery and data mining* (KDD)
- Data mining is
  - extraction of useful patterns from data sources, e.g., databases, texts, web, images, etc.
- Patterns must be:
  - valid, novel, potentially useful, understandable
Classic data mining tasks

- **Classification:**
  mining patterns that can classify future (new) data into known classes.

- **Association rule mining**
  mining any rule of the form $X \rightarrow Y$, where $X$ and $Y$ are sets of data items. E.g.,
  Cheese, Milk $\rightarrow$ Bread [sup =5%, confid=80%]

- **Clustering**
  identifying a set of similarity groups in the data

- **Sequential pattern mining:**
  A sequential rule: $A \rightarrow B$, says that event $A$ will be immediately followed by event $B$ with a certain confidence
What is web mining?

- The process of discovering knowledge from web page content, hyperlink structure, and usage data.
- Builds on existing data and text mining techniques, but adds many new tasks and algorithms.
- Three types, based on sources of data (often combined in practice):
  - Web structure mining
  - Web content mining
  - Web usage mining
Importance of web data mining

- The web is unique!
  - Amount of information is huge and still growing, on almost any topic, and changes continuously
  - No single editorial control: significant variations in quality, much duplication, and data formats vary widely
  - Significant information is linked (within and between web sites)
  - Web reflects a virtual society --- interactions among people, organizations, and automated systems, no longer limited by geography

- The Web presents challenges and opportunities for mining
Importance of web data mining

- Online organizations generate a huge amount of data
  - How to make best use of data?
- Knowledge discovered from web data can be used for competitive advantage.
  - Online retailers (e.g., amazon.com) are largely driven by data mining.
- Web search engines are information retrieval (text mining) and data mining companies
- Web surfers/searchers need tools to find, recommend, organize, and extract useful information from the Web
Why not?

- The data is abundant.
- Computing power is not an issue.
- Data mining tools are available
- The competitive pressure is very strong.
  - Almost every online company is (or should be) doing it.
Related fields

- Web mining is a multi-disciplinary field, with contributions from:
  - Data mining
  - Machine learning
  - Statistics
  - Databases
  - Information retrieval
  - Visualization
  - Natural language processing
  - Graph theory
  - etc.
Organization of course

- The course has three components:
  - Lectures - introduction to many of the main topics
  - Papers
    - Foundational and recent
    - Presented by students
    - Written critiques
    - In-class discussion
  - Semester-long web mining research project
- See online syllabus
Semester Research Project

- Individual, or groups of two (will grade each other)
  - Plus formal and informal feedback from instructor
- Should be the beginning of what could be a publishable project.
  - On some aspect of web mining
- Topic will be proposed by student
  - and approved by instructor
- Students present
  - Ideas early in the semester for feedback
  - Completed project at the end of the semester
- Write a scientific paper at the end.
  - Publish as a technical report if not more (some have been published at WWW + CIKM)
Grading

- Midterm exam: 20%
  - Covering background material
- Paper critiques: 20%
  - Weekly critiques of one paper to be presented
- Presentations: 10%
  - Short (no more than 20 minutes)
- Participation: 20%
  - Attendance, discussion, involvement
- Project: 30%
Expected topics to cover

- Classification (supervised learning)
- Clustering (unsupervised learning)
- Web information retrieval
- Web content mining
- Web structure mining
- Web usage mining
Initial Course Approach

- Except for background material, most days will have:
  - a student presentation of a paper
    - usually 20 minutes max, at most 8 slides
  - a student critic (devil's advocate)
    - 5 minutes to say why the paper should never have been published, or at least why it is not useful now
  - class discussion of topics in paper

- Most weeks
  - you will need to write one review/critique of a paper.

- The paper presentation slides and best review will be posted online
Feedback and suggestions

- Your feedback and suggestions are most welcome!
  - I need it to adapt the course to your needs.
  - Let me know if you find any errors in the textbook.

- Share your questions and concerns with the class – very likely others may have the same.

- No pain no gain
  - The more you put in, the more you get
  - Your grades are proportional to your efforts.
Paper Sources

- World Wide Web conferences
  - WWW, WSDM
- Information retrieval and database confs.
  - SIGIR, ECIR, CIKM, VLDB, SIGMOD, ICDE
- Data mining conferences
  - KDD, ICDM, SDM, PKDD, WSDM
- Other related conferences
  - ICML, ECML, UM, CHI, AAAI, IJCAI, etc.
- Journals
  - TWEB, TOIS, TOIT, JACM, CACM, IEEE...
- Other workshops, symposia (WebKDD, AIRWeb, etc.)
Why read scientific papers?
Why read scientific papers?

- Avoid reinventing the wheel
- See examples of successful research
- Stay or become current in a technical field
- Get ideas to improve or refute papers
- To explain or teach the concepts to others
How should you read a scientific paper?
How should you read a scientific paper?

- **Skim to decide whether worthwhile**
  - Determine credibility
  - Find out if it relates to your work

- **Read in detail**
  - Be skeptical
  - Challenge assumptions, arguments, methods, statistics, data

- **Take notes**
  - Write summary
  - What did authors not think to do?
  - Consider how to use approach in your work
Why review papers?
Why review papers?

- Comes naturally from a critical reading
- To contribute to the peer-review process
  - Expected of members of the community
- To learn about new work before published
  - Still confidential until published
- To get a better feel for the threshold for acceptance
- To have a voice in determining what gets published
How to write good papers?
How to write good papers?

- Read good and bad papers, and note the differences
- Provide strong motivation for work
  - Explain why exciting or important
- Demonstrate expertise
  - Connect work to foundational and recent papers
- Clearly present argument or experimental work
  - Provide sufficient detail to reproduce equivalent results
- Show significance of results
- Defuse potential criticisms
- Describe clearly the contributions of the paper

*Tell 'em what you'll tell 'em; tell 'em; tell 'em what you told 'em*
Homework

- Read Chapter 1 (online) for today
- Read meta papers for Wednesday and Chapter 3 when you get the book
- Propose a paper for us to read
  - Browse through a few conference proceedings from the past few years
  - Send me URL and bibliographic reference for your preferred paper
  - Due in one week by email (anytime Monday Jan 21)