CSE6339 WEB SEARCH, MINING, AND INTEGRATION

Self Introduction

- Chengkai Li
- http://ranger.uta.edu/~cli
- Research interests:

  databases, data mining, information retrieval, Web
- Looking for students

  Master/PhD project/thesis topics available.

Now it's your turn

- Name, program/year, where from
- focus area
- Courses taken, skills/experiences related to this course
- Why do you want to take this course?
- What do you want to get from the course?
- What would make you like/hate this course?

Advanced Topic Course

- This is a research course:
  - Not every question has a textbook answer.
  - Be prepared to explore.

Background Checking

- Prerequisite:

  CSE 3330/5330 Database Systems I or
  CSE 5334 Data Mining or
  consent of instructor
- Background:
  - you already have some background knowledge of data mining, information retrieval, Web search.
  - We will spend several initial lectures to review the background.
  - It is your responsibility to make up.

Course Page

- http://crystal.uta.edu/~cli/cse6339
  - Syllabus, Schedule (lecture notes), Resources, Accommodation based on disability.
- Course announcements will be made at WebCT.

Basics

- Lectures: Tue/Thu , 12:30am-1:50am, WH 308
- Instructor: Chengkai Li
  - Office hours: Tue/Thu 2pm-3pm, NH334
  - Contact: cli [at] uta.edu, (817) 272-0162 (I don't check voice mail)
- TA:
  - Office hours:
  - Contact:

Textbook

- Required reading:
  - research papers listed on schedule page.
- Reference Textbook: (not required)
  - Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. (available online at http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html)
  - Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6.
  - Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Addison-Wesley, 2006. ISBN 0-321-32136-7.

The slides

- The slides highlight the gist of the most important concepts and techniques.
- But
  - It is not meant to be complete. Details may not be included.
  - It may be simplified for ease of explanation in limited time and space.

Tentative Grading Scheme

Paper Review	20%	
Review Questions	20%	
Paper Presentation	15%	
Course Project	35%	(individually or in pairs)
Proposal	5%	
Progress Report	5%	
Presentation Demo, Final Report	25%	
Class Participation	10%	

You are required to attend classes and actively participate in discussions.

Paper Review 20%

- Required for every student (except the one who is presenting)
  - Deadline: 11:55pm, the night before the lecture.
  - Will be skipped when we occasionally study the materials in a book chapter, instead of a paper.
  - Each student can skip a certain number of reviews. The exact number is to be determined.
- What is in the review
  - About 800 words.
  - Summarize the problem, approach, and contribution (200 words)
  - Critiques (on important things, rather than trivial points) (300 words)
  - You thoughts on improvements/contradictions/comparison (300 words)

Review Question 20%

- One question for each week or two weeks
  - Usually high-level , open-ended questions
  - Sometime more detailed and specific.
  - Brief answer, not essay.
- Still need to figure out the exact procedure
  - In-class quiz?
  - Homework?
  - Online quiz (made available and due at specific deadline)?
- Another possibility is to replace it by in-class debate session.

Paper Presentation 15%

- Starting from next Thursday (we need volunteer.)
  - Study one paper (sometimes more) in each lecture.
  - One student will present the paper.
- Presentation slides:
  - Deadline: 11:55pm, the night before the lecture.
  - Should be carefully designed.
  - Cover 80 minutes.
  - The presentation should be interactive: present the papers, raise questions, and moderate discussions.
  - The more discussions/debates, the better.
- Each student may need to present twice, depending on the number of registered students.

Course Project 35%

- Be prepared to get hands dirty.
- Individually or in pairs (the group members should contribute to the project evenly).
- Several stages:
  - P1: Proposal (problem definition and motivation.)
  - P2: Progress Report (revised problem definition and motivation, initial architecture and algorithm design.)
  - P3: Final Report (in the format of a research paper.)
  - P4: Presentation and Demo
- I will provide sample project topics.
  - Will be research-type and exploratory.

class participation (10%)

- Mostly In-class discussion
- WebCT discussion is encouraged
  - You are highly encouraged to initiate discussion thread.

Homework

- None, unless we decide to do review question in the form of homework.

Midterm and Final Exam

- None

WebCT

- Announcement
- Student assignment submission (we don't accept email submission or hard-copy)
  - Presentation slides
  - Review
  - Review question
  - Project deliverables
- Grades
- Discussion

Deadlines

- Everything will be submitted through WebCT.
- Due time: 11:55pm
- Late submission: 5-point deduction per hour, till you get 0. (The raw score of each assignment is 100. So there is no point to submit it after 20 hours).

How to Submit through WebCT

1. Click button "Upload file" to upload your file.
2. Fill in your email address (UTA email address only) in the "Notification" box.
3. Then you must click button "submit assignment". Otherwise, your file will not be submitted.
4. Verify that your file is indeed submitted into WebCT. (You should see the file name after "Student files". Click the link to download the file and verify it.)
5. Check your email. You must keep the notification email from WebCT.
6. If you don't find your submission or don't receive notification within 10 minutes, try step 1-5 again.
7. If step 6 still fails after you give it another try, email your file to the TA and yourself immediately.

Regrading

- 7 days after we post scores on WebCT. TA will handle regrade requests. Won't consider it after 7 days.
- If not satisfied with the results, 7 days to request again. Instructor will handle it, and the decision is final.
- We usually even change score of your review, since its grading is subjective by nature, unless unfair grading is obvious.

Topics

- Other Topcis
  - Structured querying of the Web
  - social networks
  - Semantic Web
- Web Search
  - search engine architecture
  - crawling
  - indexing
  - link analysis (HITS, PageRank)

Topics

- Wed Data Mining
  - classification and clustering
  - Clustering web search results
  - large-scale data processing (MapReduce)
- Web Data Integration
  - information extraction
  - answering queries using views
  - schema matching
  - Deep Web

Schedule

- http://crystal.uta.edu/~cli/cse6339

Where to find papers

- Google
- Google Scholar
- DBLP Bibliography
- CiteSeer
- Services through UTA Library

http://library.uta.edu/JDBC/DBs/dbejournal.jsp

- ACM Digital Library
- IEEE Xplore
- Other Computer Science articles

Get bored?

- Do you watch Youtube?

http://www.youtube.com/watch?v=gC2ew6qLa8U

http://www.youtube.com/watch?v=463gKcXDVzQ

Don't do it. It's not worth it.

We are very serious about this.

read & sign the statement