

CSE6339 WEB SEARCH, MINING, AND INTEGRATION

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
Department of Computer Science and Engineering, University of Texas at Arlington
©Chengkai Li, 2009

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.

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

2

Now it's your turn

- Name, program/year, where from
- Research/focus area
- Course 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?

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

3

Disclaimers

- It is not an easy course
 - My courses are always heavy. Likely you will spend more time on this course than any other ones.
 - There will be Bs and Cs.
 - It is a research course:
 - Not every question has a textbook answer.
 - Be prepared to explore.
 - Exam questions would require through understand of materials and creative thinking.

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

4

Background Checking

- Prerequisite:
 - CSE 4331/5331 Database Systems II or
 - CSE 5334 Data Mining or
 - consent of instructor
 - (If you have not taken these two courses, you need to get my permission. Talk to me after class)
- Background:
 - We assume you already have some background knowledge of data mining, information retrieval, Web search. This is an advanced research course.
 - We will spend several initial lectures to review the background.
 - It is your responsibility to make up.

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

5

Course Page

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

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

6

Basics

- **Lectures:** Tue/Thu , 9:30am-10:50am, NH 229
- **Instructor:** Chengkai Li
Office hours: Tue/Thu 11am-12pm, NH229
Contact: cli [at] uta.edu, (817) 272-0162 (I don't check voice mail)
- **TA:**
Office hours:
Contact:

Textbook

- **Required Textbook:**
Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, [Introduction to Information Retrieval](#), Cambridge University Press. 2008.

(available online at <http://www.cs.stanford.edu/~hinrich/information-retrieval-book.html>)
- **Required reading:**
research papers listed on schedule page.

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.
- You won't do well in the course if you just read the slides
 - You need to read the book and papers.

Tentative Grading Scheme

- | | | |
|-----------------------------------|-----|----------------------------|
| □ Midterm | 20% | |
| □ Final | 30% | |
| □ Paper Presentation | 15% | |
| □ Course Project | 25% | (individually or in pairs) |
| □ Proposal | 2% | |
| □ Progress Report | 5% | |
| □ Presentation Demo, Final Report | 18% | |
| □ Bonus Points | 10% | |
- You are required to attend classes and actively participate in discussions (both in-class and WebCT).

Paper Presentation 15%

- Starting from Week 5
 - 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.

Course Project 25%

- **Be prepared to get hands dirty.**
- Individually or in pairs (the group members should contribute to the project evenly).
- Several stages:
 - P0: Team information
 - 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 (**Time and Location TBD**)
- I will provide sample project topics.
 - Will be research-type and exploratory.

Midterm (20%) and Final Exam(30%)

- Midterm Exam
(Thursday, March 12th, 9:30-10:50am, NH 229)
- Final Exam
(Thursday, May 14th, 8-10:30am, NH 229)

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

13

Bonus points: class participation (10%)

- In-class discussion
- WebCT discussion group
 - discussion topics will be posted from time to time.
 - You are highly encouraged to initiate your own thread.

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

14

WebCT

- Announcement
- Student assignment submission (we don't accept email submission or hard-copy)
- Discussion Group
- Grades

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

15

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).

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

16

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.

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

17

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.

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

18

Topics

- Web Search
 - search engine architecture
 - crawling
 - indexing
 - link analysis (HITS, PageRank)
 - large-scale systems (MapReduce)
- Web Data Mining
 - classification and clustering
 - support vector machines
 - latent semantic indexing

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

19

Topics in Textbook

- Web Data Integration
 - information extraction
 - answering queries using views
 - schema matching
 - Deep Web
- Other Topics
 - Web 2.0
 - social networks
 - Semantic Web

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

20

Schedule

- <http://crystal.uta.edu/~cli/cse6339/schedule.htm>

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

21

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](#)

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

22

Get bored?

- Do you watch Youtube?

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

23

<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

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2009
UT-Arlington © Chengkai Li, 2009

24