

## 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. CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai II, 2010

## Now it's your turn

- Name, program/year, where from
- o 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?

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai Li, 2010

## Advanced Topic Course

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

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai Li, 2010

## **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.

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai LI, 2010

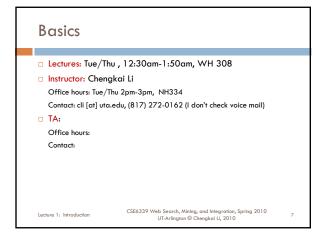
## Course Page

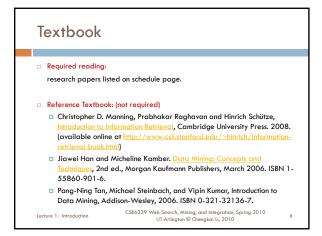
## □ <a href="http://crystal.uta.edu/~cli/cse6339">http://crystal.uta.edu/~cli/cse6339</a>

- Syllabus, Schedule (lecture notes), Resources, Accommodation based on disability.
- Course announcements will be made at WebCT.

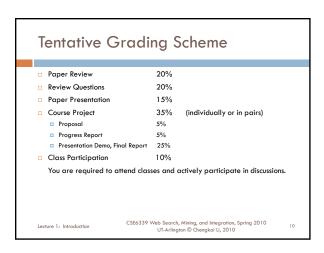
Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai Li, 2010





# 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. CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington ⊕ Chengkal II, 2010



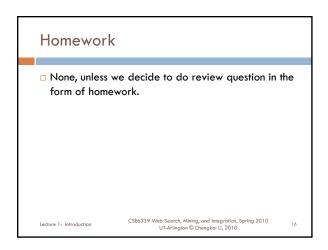
# Paper Review 20% Required for every student (except the one who is presenting) Deadline: 11:55pm, the night before the lecture. Will be skipped when we occasionally study the materials in a book chapter, instead of a paper. Each student can skip a certain number of reviews. The exact number is to be 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) Lecture 1: Introduction CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai Li, 2010

Review Question	on 20%	
<ul> <li>One question for each week</li> <li>Usually high-level, open-</li> <li>Sometime more detailed of Brief answer, not essay.</li> </ul>	ended questions	
Still need to figure out the ex In-class quiz? Homework? Online quiz (made availate)	act procedure ole and due at specific deadline)?	
<ul> <li>Another possibility is to replace it by in-class debate session.</li> </ul>		
Lecture 1: Introduction CSE63:	39 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai Li, 2010	12

# 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. CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai II, 2010



Midterm and Final Exam

None

CSE6339 Web Search, Mining, and Integration, Spring 2010
UT-Arlington © Chengkai II, 2010

WebCT		
•	ment submission (we don't accept on or hard-copy) slides	
Lecture 1: Introduction	CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai LI, 2010	18

## **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 2010 UT-Arlington © Chengkai Li, 2010

## How to Submit through WebCT

- Click button "Upload file" to upload your file.
- 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.
- 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.)
- Check your email. You must keep the notification email from WebCT.
- If you don't find your submission or don't receive notification within 10 minutes, try step 1-
- If step 6 still fails after you give it another try, email your file to the TA and yourself

CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai Li, 2010

Lecture 1: Introduction

## Regrading

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

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai Li, 2010

## **Topics**

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

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai Li, 2010

### **Topics**

- Wed Data Mining
  - $\hfill \square$  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

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, Spring 2010 UT-Arlington © Chengkai Li, 2010

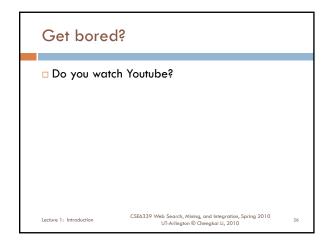
## Schedule

□ http://crystal.uta.edu/~cli/cse6339

Lecture 1: Introduction

CSE6339 Web Search, Mining, and Integration, UT-Arlington © Chengkai Li, 2010





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

CSE6339 Web Search, Mining, and Integration, Spring 2010
UT-Arlington © Chengkai U, 2010