Instructor: Chengkai Li

- Office hours: Friday 10am-12pm
- Office: Nedderman Hall 334
- Phone: (817) 272-0162
- E-mail: cli [AT] uta [DOT] edu

Course Description

We will study the current trends in the research of data and information management. It involves popular topics in the areas of databases, data mining, information retrieval, and the intersections of these areas. The goals of the course are: to expose graduate students to the cutting-edge of research in these areas; to equip them with the necessary skill sets for finding hot jobs; to help them identify research topics and come up with preliminary works through course projects; and to prepare those new students for doing research with faculty in databases and data mining. Detailed topics include:

- **Data and Information Management for Web 2.0**
  - social network
  - recommendation systems
  - online advertising
  - spam
  - extremely large-scale database systems (Google BigTable, Hadoop, ...)
  - structured Web queries
- **Data Retrieval and Exploration**
  - ranking and top-k queries
  - keyword search
  - facets over databases
  - interactive query formulation
  - query results organization and navigation
- **Novel Data Management Systems**
  - provenance and lineage
  - uncertainty and probability
  - deep-Web
  - column-oriented database
  - privacy and security
  - biological data management
- scientific data management
- architectural support

**Prerequisites**

- CSE 3330/5330 Database Systems I  or  CSE 4331/5331 Database Systems II  or  similar courses  or  consent of instructor

**Textbook**

- No textbook required. We will use the state-of-art literature.

**Grades**

- **Paper Review** 25%  must be done independently. Each student review 2 papers each week, and finish totally about 16 reviews.
- **Presentation** 15%  must be done independently. Each student present one designated paper in one lecture. And must submit the presentation file before the class.
- **Course Project** 50%  must be done in groups of size 2 or 3 (or individually with the consent of instructor).
  - Proposal 5%
  - Progress Report 1 5%
  - Progress Report 2 5%
  - Presentation and Demo 15%
  - Final Report 20%
- **Class participation** 10%  Students are expected to attend classes and actively participate in discussions.

**Announcements**

Stay tuned and make sure to check the [announcement page](#) frequently. Important announcements will be posted there.

**Assignments and Deadlines**

- All the assignments (*Paper Reviews, Presentation Slides, Proposal, Progress Reports, and Final Report*) must be submitted through WebCT. We will NOT take email submission, unless the school verifies that WebCT was malfunctioning or unavailable.
- Everything is due by 11:59pm on the due date (listed in the [schedule page](#)). The deadline is automatically managed by WebCT. You can still turn in assignment after the deadline. However, you automatically lose 5 points per hour after the due time, till you get 0. (Each individual assignment is 100 points.) We cannot waive the penalty, unless there was a case of illness or other substantial impediment beyond your control, with proof in documents from the school.
- Before the deadline, the WebCT will allow you to submit as many times as you want. However, you will only be allowed to submit once after the deadline.
- Every student must equally contribute to the group project. Only one student in a group needs to upload project-related assignments into WebCT.

**Last Day to Drop or Withdraw**

Apr 04, 2008
WebCT

The WebCT page is [http://www.uta.edu/webct](http://www.uta.edu/webct). Use your NetID and password to login. The item "Start Here!" links to WebCT tutorials. Please consult them for help if necessary. Try to post a message in "Junk". Let me know if there is any problem. We use WebCT for the following activities:

- **Posting assignment instructions and files.**

- **Student submit assignments electronically.**
  - When submit on WebCT, click button "Upload file" to upload your file first. **Then you must click button "submit assignment"**. Otherwise, your file will not be submitted.
  - If you want to receive **automatic confirmation** after you submit your assignments, please follow the steps below:
    1. Go to WebCT.
    2. Go to Communication Tools > Mail.
    3. Click button "Message Settings", which is above the Folders table.
    4. Check the box next to "Forward my mail to:", which is in the section "Mail forwarding" under "Message Settings".
    5. Enter your UTA email address in the box. Automatic confirmation of your homework submission will be delivered to this email address. Don't use emails outside UTA.
    6. Click "Update" at the bottom of the page.
  - You can also fill in your email address for each individual assignment, in the "Notification" box.

- **Discussion group (bulletin board).**
  - Various topics have been created for your discussion. Most of the names of the topics are self-describing.
    - The Announcements on the course homepage will also be replicated in the discussion group.
    - If your post doesn't fit into any topic, use "Miscellaneous".
    - The "Junk" is for testing purpose only.

- **Posting students scores and grades.**

---

**Ethics Policies and Academic Integrity**

The College cannot and will not tolerate any form of academic dishonesty by its students. This includes, but is not limited to cheating on examinations, plagiarism, or collusion (explained in the document below).

Students are required to read the following document carefully, sign it, return the signed copy to the instructor, and keep a copy for their own records. Hardcopies of this document will be provided to the students in the first class, and also can be picked up in the instructor's office. If you print by yourself, please make it double-sided.

[Statement on Ethics, Professionalism, and Conduct for Engineering Students](http://www.uta.edu/)

**Miscellaneous**

If you require accommodation based on disability, I would like to meet with you in the privacy of my office during the first week of the semester to ensure that you are appropriately accommodated. Please read the page of [the office for students with disabilities](http://www.uta.edu/).
**Syllabus**  | **Announcements**  | **Schedule**  | **Resources**  | **Project Requirements**
---|---|---|---|---

## Schedule

The tentative schedule is as follows. We may change the schedule as necessary.

**Paper Review:**

Due at 11:59pm, two days before the scheduled paper is discussed. For example, R1 (the review of paper P1) is due at 11:59pm, Sunday, Feb. 3rd.

**Paper Presentation Slides:** (Only the presenter of the scheduled paper is required to submit the slides.)

Due at 11:59pm, two days before the scheduled paper is discussed. For example, S1 (the slides of paper P1) is due at 11:59pm, Sunday, Feb. 10th.

### Date  | Paper#  | Lecture/Activities  | Presenter  | Due  | Le
---|---|---|---|---|---
01/15  |  | Course Overview and Introduction  | Chengkai Li  |  | I
01/17  |  | Review of Database Management Systems  | Chengkai Li  |  | I
01/22  |  | Review of Data Mining  | Chengkai Li  |  | I
01/24  |  | Review of Web Information Retrieval  | Chengkai Li  |  | I
01/29  |  | Course Project Topics  | Chengkai Li  | slid  | I
01/31  |  | Paper Review, Presentation, Research Resources  | Chengkai Li  |  | I

**Structured Querying of the Web**

02/12  | P3  | Tao Cheng, Xifeng Yan, Kevin Chen-Chuan Chang: *EntityRank: Searching Entities Directly and Holistically*, VLDB 2007: 387-398  | Arjun Dasgupta  |  | I

**Retrieval, Exploration, and Navigation of Databases**

02/21  | P6  | Cong Yu, H. V. Jagadish: *Querying Complex Structured Databases*, VLDB 2007: 1010-1021  | Raghu Srinivasan  |  | I
03/06  | P10  | Zhiyuan Chen, Tao Li: *Addressing diverse user preferences in SQL-query-result navigation*, SIGMOD Conference 2007: 641-652  | Shivkumar Chandrashekhar  |  | I

**Social Network Analysis and Collaborative Filtering**

03/11  | P11  | Lars Backstrom, Cynthia Dwork, Jon M. Kleinberg: *Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography*, WWW 2007: 181-190  | Xin Jin  |  | I

**Happy spring break**

03/25  |  | Group 1 (Supreeth Chakravarthy, Aditya Telang) and 2 (Shivkumar Chandrashekhar, Aniruddha Deshpande)  |  |  | I
03/27  |  | Group 3 (Arjun Dasgupta, Xin Jin, Raghu Srinivasan), 4 (Rahul Dhar, Robin Michael), and 5 (Naved Kazi, Muhammad Safiullah)  |  |  | I
04/01  | P13  | Nikolay Archak, Anindya Ghose, Panagiots G. Ipeirotis: *Show me the money!: deriving the pricing power of product features by mining consumer reviews*, KDD 2007: 56-65  | Arjun Dasgupta  |  | I

**Systems and Architecture**


**Last day to drop or withdraw**

04/08  | P15  | Seung-Taek Park, David M. Pennock: *Applying collaborative filtering techniques to movie search for*  | Muhammad Safiullah  |  | I
<table>
<thead>
<tr>
<th>Date</th>
<th>Page</th>
<th>Authors and Paper Title</th>
<th>Presentation and Demo Slides</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/10</td>
<td>P16</td>
<td>Panagiotis G. Ipeirotis, Eugene Agichtein, Pranay Jain, Luis Gravano: To search or to crawl?: towards a query optimizer for text-centric tasks. SIGMOD Conference 2006: 265-276</td>
<td>Robin Michael</td>
</tr>
<tr>
<td>04/15</td>
<td>P17</td>
<td>Jayant Madhavan, Shirley Cohen, Yin Luna Dong, Alon Y. Halevy, Shawn R. Jeffery, David Ko, Cong Yu: Web-Scale Data Integration: You can afford to Pay as You Go. CIDR 2007:342-350</td>
<td>Naved Kazi</td>
</tr>
<tr>
<td>04/24</td>
<td></td>
<td>Group 1 (Supreeth Chakravarthy, Aditya Telang) and 2 (Shivkumar Chandrashekhar, Aniruddha Deshpande)</td>
<td></td>
</tr>
<tr>
<td>04/29</td>
<td></td>
<td>Group 3 (Arjun Dasgupta, Xin Jin, Raghu Srinivasan) and 4 (Rahul Dhar, Robin Michael)</td>
<td></td>
</tr>
<tr>
<td>05/01</td>
<td></td>
<td>Group 5 (Naved Kazi, Muhammad Safiullah) and Summary of course</td>
<td></td>
</tr>
<tr>
<td>05/08</td>
<td></td>
<td></td>
<td>Final Report</td>
</tr>
</tbody>
</table>

University calendar: **Spring 2008**
Requirements on project-related submissions:

**Progress Report 1**

By the cornerstone of progress report 1, your project should be in the following status:

- well-defined problem documented
- research challenges clearly identified and explained
- architecture/components has been carefully thought out
- methods and algorithms have been thoroughly designed in details
- initial implementation partially done
- has a clear evaluation plan

In your progress report 1, write regarding the above bullets. Moreover, highlight the following:

- change of initial plan
- difficulties encountered
- expected challenges in later stages

**Progress Report 2**

By the cornerstone of progress report 2, your project should be in the following status:

- well-defined problem documented
- research challenges clearly identified and explained
- architecture/components designed and documented in details, and finalized
- algorithms implementation finished, some improvements undergoing
- experiments and evaluation undergoing

In your progress report 2, write regarding the above bullets. Moreover, highlight the following:

- changes made since progress report 1
- progress achieved since progress report 1
- difficulties encountered
- things to be accomplished in final stage

Project Presentation and Demo

Final Report