CSCE 689 Syllabus

Internet-Scale Data Management
Texas A&M University, Fall 2010
Tu/Th 12:45pm - 2:00pm, HRBB 113

Instructor:
James Caverlee, HRBB 403, 845-0537, caverlee@cse.tamu.edu
Office Hours: 10-11am Tues/Thurs, or by appointment

Teaching Assistant:
Brian Eoff, brian.eoff@gmail.com
Office Hours: 10-11am Mon/Wed

Course Summary:
This course is an introduction to advanced research topics in Internet-scale data management, addressing the relevant theoretical foundations, methods, and tools from a wide spectrum, including (i) large-scale distributed information management; (ii) data and text mining techniques and algorithms; and (iii) data privacy and security issues in large-scale Internet systems. By the end of the semester you will be able to:

• Define and explain the key concepts, methods, and tools relevant to Internet-scale data management,
• Design, implement, and evaluate core algorithms and approaches, including distributed text mining, large graph manipulation and analysis, etc.;
• Identify the salient features and apply recent research results in Internet-scale data management.

Communication:
All course announcements will be posted to the course mailing list. If you have a specific question for either me or the TA please send us an email with 689 in the subject line. We will make our best effort to respond promptly, but we only guarantee a response within one week.

Prerequisites:
Graduate classification or approval of instructor.

Required Text:
Readings will be drawn from top computer science conferences and journals.

Class Format:
During most weeks, I will present an overview of the week's topics during the first half of class on Tuesday. The remaining time each week will be devoted to intense paper discussion.

Grading:
The course grading policy is as follows:
20% Proposal
50% Final Project Deliverable (Paper + Poster)
30% Assignments
The grading scale is A: 90-100, B: 80-89, C: 70-79, D: 60-69, F: 0-59.

The most significant portion of your grade is based on the project, for which you may work in teams of **up to three persons**. Please keep in mind that the general goal of research is to contribute new knowledge. Thus, it is important to ask yourself what research question(s) you aim at answering and what challenges you aim at solving.

In general, a project should include both: (i) an empirical evaluation of an algorithm or model on an interesting dataset, in order to better understand these methods, and possibly further improve them; and (ii) a proposal for a new algorithm and model, including a comparison with a baseline.

When picking a topic, try to ask yourself the following questions:

- Is the topic addressing an important problem? Would any one care about it if you solve the problem?
- To what extent has previous research work addressed this problem? And what remains unknown?
- Do you have any idea at all about how to solve the problem? If not, can you reformulate the problem to make it easier?
- Would you be able to evaluate your solution? That is, how can you demonstrate your solution is good and solves the problem well?

**Proposal:** The emphasis of the course is on recent and current scholarly material. We will read a number of papers from the past few years and these papers should stimulate you to think about unexplored avenues of research. For the project proposal, you should pick two or three related papers (including at least one that has been mentioned in class) as the basis for your proposal. These related papers will serve as the research grounding for your proposal. So first in your proposal, you must address these questions (in about 2 or 3 pages, single column, normal fonts and margins):

- What is the key technical content and interesting ideas behind the papers?
- How do the papers relate to the topics presented in the course?
- What are the strengths of the paper? What are its weaknesses? And how might you go about improving on the weaknesses?

I expect your writing to display a deep understanding of the papers you select; this is not merely a re-hash or a copy-and-paste job from the papers (e.g., “first they did X, then they did Y, finally they did Z”). I expect you to synthesize the main ideas and provide your own understanding and perspective on them. Provide a judgment.

In the remainder of the proposal (about 2 pages), I expect you to dig deeply into your proposed work (which should naturally build on the background materials you’ve already covered).

- What is your research question? Clearly define the research problem/ question.
- Why is this an interesting question to ask and why would we care about the answer to this question or a solution to the problem?
• Has any existing research work tried to answer the same or a similar question, and if so, what is still unknown?
• How do you plan to work out the answer to the question. (At the proposal stage, you are only expected to have a sketch of your methods.)
• How would you evaluate your solution. That is, how do you plan to demonstrate that your solution/answer is good or is reasonable.
• A rough timeline to show when you expect to finish what. List a couple of milestones.

You should plan to write a proposal of around 5 pages. The project proposal is due on September 24 (Friday) by 11:59pm. Email me and the TA your project proposal in PDF with a pithy filename. One per team.

Final Project Deliverable: At the end of the semester, you will deliver a final paper and participate in a poster session.

• The poster session will be held on December 7, 2010 and will be open to the public for around 2 or 3 hours (depending on the number of projects we have). 30% of your project grade is based on the poster session. Your team’s grade will be based on the clarity of the poster itself, audience feedback, and your answers to my questions. I will provide the poster boards and the setup materials.
• The final paper is due on December 10, 2010 by 11:59pm. You should email me and the TA your final paper in PDF with a pithy filename. The paper should be in the standard ACM conference template http://www.acm.org/sigs/publications/proceedings-templates (double-column) and up to 8 pages maximum. The paper counts for 70% of your project grade. You should write your report as if you were writing a conference paper. You should address the same questions as those you have addressed in the proposal, only with more details, especially regarding some of the challenges that you need to solve and your experimental results. You should also include your conclusions from the study and point out how your work can be further extended (i.e., future work).

Assignments: We will have around 5 homework assignments over the course of the semester. All homework assignments must be submitted by 11:59pm Central time on the due date. For the homework assignments, you may talk to any other class member or work in groups to discuss the problems in a general way. However, your actual detailed solution must be yours alone. If you do talk to other students, you must write on your assignment who it is that you discussed the problems with. Your submitted work must be written solely by you and not contain work directly copied from others.

Late Days: For the homework assignments only, you have a total of 5 late days that you can use during the semester. However, a single assignment can be submitted up to 3 days late only, so we can post solutions in a timely fashion. Once you exhaust your 5 late days, any late homework will receive a 0.

Americans with Disabilities Act (ADA) Policy Statement
The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation of their disabilities. If you believe you have a disability
requiring an accommodation, please contact the Department of Student Life, Services for Students with Disabilities, in Cain Hall or call 845-1637.

Academic Integrity Statements

AGGIE HONOR CODE: "An Aggie does not lie, cheat, or steal or tolerate those who do."

Upon accepting admission to Texas A&M University, a student immediately assumes a commitment to uphold the Honor Code, to accept responsibility for learning, and to follow the philosophy and rules of the Honor System. Students will be required to state their commitment on examinations, research papers, and other academic work. Ignorance of the rules does not exclude any member of the TAMU community from the requirements or the processes of the Honor System.

For additional information please visit: http://www.tamu.edu/aggiehonor/