CS6093
Advanced Database Systems
(Topics on the Web)
Spring 2011
Introduction

• Fernando Diaz
  – Ph.D. in CS, 2007, University of Massachusetts Amherst
  – Research Scientist, Yahoo! Research
  – Interests: Information Retrieval, Information Extraction, Time-Sensitive Information
  – http://ciir.cs.umass.edu/~fdiaz/

• Cong Yu
  – Ph.D. in CSE, 2007, Univ. of Michigan
  – Research Scientist, Google Research
  – Interests: Information Extraction, Social Data Mining, Structured Data Management
  – http://www.eecs.umich.edu/~congy/
Go Around the Room

- Name
- Program (Undergrad, MS, PhD)
- Year
- Interests

- Also fill the sign up sheet
Outline

• Introduction

• Course Overview
  – Logistics
  – Grading
  – Deadlines
  – Projects

• Course Content
  – Lecture outline

• How to read and review a research paper
Course Overview

• Research-oriented (i.e., papers and projects)
  – Lectures involve discussion of research papers
  – Group projects involve significant coding, as well as more research-ish efforts (i.e., understanding intellectual novelty and technical significance)

• Keep in mind: the course can often be exploratory in nature and many discussions may not lead to definitive answers

• Don’t worry, we will guide you through

• We will focus on the following five main topics
  – Search and Ranking
  – Evaluation
  – Personalization
  – Information Extraction
  – Data Mining

  Information Retrieval

  Data Web
Logistics (see handout)

- **Time and Location**
  - Mondays 6:00p – 8:25p, RH602
  - 10 minute break mid-class

- **Office Hour**
  - M 4:50p – 5:50p, LC246 (starting Feb 7th)

- **Website**
  - Currently @ [http://www.eecs.umich.edu/~congy/cs6093/s11/](http://www.eecs.umich.edu/~congy/cs6093/s11/)
  - Moving to Poly CIS server soon

- **Mailing List**
  - to be determined (will be posted on the web)
Grading

• Not curve based
  – You do your job well, you will get a good grade

• Group project accounts for 75%
  – 30% mid-term report
  – 45% final report and demo
  – Your contribution to the project will be judged by your group members and submitted as part of the report!

• Class participation accounts for 25%
  – Reading material will be posted on site a week before class
  – Quizzes (4)
  – Participation in class discussion throughout the term
    • Giving brief overview of assigned reading material
    • Asking and answering questions
    • Participating in the final project demo presentation by critiquing other group’s projects
More on Quizzes

• Questions on materials covered in recent lectures

• Questions on core concepts of the assigned reading material

• Possible short technical review of the assigned paper
  – three strong points and three weak points
An Immediate Deadline

• **2/7: Project selection and group formation**
  – Two weeks from today!
  – Group size: 3 (ideal) or 4 (discouraged)
  – Smaller groups are also possible with approval from us
  – Ten project topics are listed on the website
    • Detailed description for each project can be requested by emailing the corresponding instructor
    • Request descriptions for only the projects you are seriously interested, too many choices often lead to *analysis paralysis* 😊
  – Other topics are possible, especially with students in the Ph.D. program; please request approval from us
  – Once decided, email us with a list of 3 projects (in order of preference) and your group members
    • *First come first serve*, the earlier you let us know, the more likely you get assigned to your favorite project
    • Project can only be selected *after* your group has been formed
Future Important Dates (Deliverables)

• 3/21: mid-term report due
  – Motivation: novelty and significance
  – Project design and preliminary study

• 5/9: project presentation and demo in class
  – Presentation of the design, implementation and experiment
  – Demonstration of the working prototype

• 5/13: final report due
  – Motivation
  – Project design and implementation details
  – Description of experimental analysis
  – Lesson learned and conclusions drawn
Successful Project Requires:

• **Step 1:** Find other students who can work well with you
  – Divide & Conquer or Extreme Programming
  – Early Bird or Night Owl
  – Etc.

• **Step 2:** Choose a topic of interest to your group
  – Interest drives commitment

• **Step 3:** Survey relevant literature first!
  – Estimate the novelty and significance of your proposed approach
  – Understand what’s required to carry out your project

• **Step 4:** Collect and understand datasets when required

• **Sometimes, you may need to iterate between steps 2 and 3**
  – So get started as soon as possible!
Successful Project Requires:

- **Step 5: Design phase**
  - System architecture
  - Key algorithms
  - Preliminary study to verify the approach
  - You should have this ready for your mid-term report

- **Step 6: Implementation phase**
  - A working prototype
  - Detailed experimental evaluation on reasonable datasets
  - Analysis and explanation of results

- **Step 7: Writing phase**
  - Don’t overlook it!
  - Even the greatest idea is useful only when others can understand it
Current Project Topics

1. Personalize a Search Engine (Personalization)
2. Performance Comparison of Multiple Search Engines (Evaluation)
3. Detecting Trends in Facebook/Twitter Feeds (Ranking / Data Mining)
4. Detecting Trends in Multilingual News (Ranking / Data Mining)
5. A New Search Engine for NYU-Poly CS (Ranking)

6. Structured Search over Wikipedia (Ranking / Structured Search)
7. Domain Specific Knowledge Reconstruction using Wikipedia (Information Extraction)
8. Site Specific Example Driven Information Extraction (Information Extraction)
9. Mining Interesting Differences in Social Networks (Data Mining)
10. Mining Patterns from Status Updates (Data Mining)
<table>
<thead>
<tr>
<th>Date</th>
<th>Lecturer</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/24</td>
<td>Both</td>
<td>Overview &amp; How to Read a Research Paper</td>
</tr>
<tr>
<td>1/31</td>
<td>Cong</td>
<td>IR: Structured Search</td>
</tr>
<tr>
<td>2/7,2/14</td>
<td>Fernando</td>
<td>IR: Ranking</td>
</tr>
<tr>
<td>2/21</td>
<td>no class</td>
<td>President’s Day</td>
</tr>
<tr>
<td>2/28,3/7</td>
<td>Cong</td>
<td>DW: Information Extraction</td>
</tr>
<tr>
<td>3/14</td>
<td>no class</td>
<td>Spring Break</td>
</tr>
<tr>
<td>3/21,3/28</td>
<td>Cong</td>
<td>DW: Data Mining</td>
</tr>
<tr>
<td>4/4,4/11</td>
<td>Fernando</td>
<td>IR: Evaluation</td>
</tr>
<tr>
<td>4/18,4/25</td>
<td>Fernando</td>
<td>IR: Personalization</td>
</tr>
<tr>
<td>5/9</td>
<td>Both</td>
<td>Project presentation and demo</td>
</tr>
</tbody>
</table>
IR: Ranking

- Lecture 1: Structured Search
  - Introduction to semi-structured data
  - Introduction to structured search (FlexPath)
  - Semi-structured Database Management System (Timber)

- Lecture 2: Introduction to Information Retrieval Ranking
  - Basic problem and instances
    - e.g. web search, enterprise search, question answering)
  - Text-based ranking
  - Web ranking

- Lecture 3: Advanced Ranking
  - Machine learning and ranking
  - Aggregated ranking
  - Freshness and trend detection
Information Extraction

• **Lecture 1: Overview of Information Extraction**
  – Basic extraction technologies
  – Wrapper technology (DBLife)
  – Extraction platform (UIMA)

• **Lecture 2: Web-Scale Information Extraction**
  – Extraction planning (Snowball)
  – Open information extraction (TextRunner/KnowItAll)
  – WebTable
Data Mining

• Lecture 1: Introduction to Data Mining
  – Frequent item set mining
  – Association rule mining
  – Data cube analysis

• Lecture 2: Social Data Mining
  – Community analysis and graph mining
  – Analysis of collaborative reviewing sites
  – Social data and web search
IR: Evaluation

• Lecture 1: Introduction to IR Evaluation
  – Cranfield paradigm
  – Metrics
  – Dealing with limited data

• Lecture 2: IR Evaluation in Production Systems
  – Exploiting clicks
  – Modeling clicks
  – Explore/exploit
IR: Personalization

• Lecture 1: Modeling Interests
  – Modeling a population’s interests
  – Modeling group interests
  – Modeling individual interests

• Lecture 2: Incorporating Interests
  – Personalized ranking
  – Personalized news
Any Questions?
Where to find the paper?

- **DBLP**
  - [http://www.informatik.uni-trier.de/~ley/db/](http://www.informatik.uni-trier.de/~ley/db/)

- **Google Scholar**
  - [http://scholar.google.com/](http://scholar.google.com/)

- **CiteSeer**
  - [http://citeseer.ist.psu.edu/](http://citeseer.ist.psu.edu/)

- **ACM Digital Library**
  - [http://portal.acm.org/dl.cfm](http://portal.acm.org/dl.cfm)

- **IEEE Xplore**
  - [http://ieeexplore.ieee.org/Xplore/guesthome.jsp](http://ieeexplore.ieee.org/Xplore/guesthome.jsp)
What kind of paper?

- **Survey**
  - What happened before: e.g., SIGRecord

- **Vision**
  - What will happen next: e.g. CIDR

- **Engineering style**
  - System and Algorithm: e.g., SIGMOD
    - Demonstrate the proposed system and/or algorithm is superior to previous ones
  - Experiment: e.g., ICDE
    - Compare previous approaches and draw a (anti-conventional wisdom) conclusion

- **Scientific style**
  - Analyze and explain the underlying reason behind a phenomenon
  - Hypothesis testing
What kind of paper?

• **Theory:** e.g., PODS
  - Prove a theorem
  - Analyze the complexity of an algorithm

• **Different kinds of paper require different kinds of appreciation**
  - Don’t judge a system paper by how many theorems in the paper
  - Don’t judge a scientific style paper by how immediately useful it is
The structure of a paper

- **Title and Abstract**
  - Highlight the contributions of the paper

- **Introduction**
  - Motivation
    - Why is this problem interesting?
    - Why is solving this problem necessary?
    - What kind of impact solving this problem can have?
  - Contribution
    - Is this a new problem?
    - If this is an old problem, how is the proposed solution different from earlier ones?
    - What formal properties the proposed solution has?

- **Related Work**
  - Demonstrate that you have a good understanding of the field
  - Distinguish your work from earlier ones
The structure of a paper

- **Methodology (the main body)**
  - Problem definition
  - Description of solution
  - Explanation of why the solution works
  - Applicability of the solution

- **Experimental evaluation**
  - Compare with earlier approaches
  - Compare with naïve approaches

- **Conclusion**
  - Again, highlight the contributions of the paper

- **References**
  - Give credits to earlier works
How to evaluate a paper

• **Understand it first!**
  - Can you use your own word to summarize what problem is being addressed in the paper?
  - What are the key contributions of the paper, again, in your own words?

• **Motivation**
  - Is the problem novel?
  - Is the work significant?
  - 2009 Ig Nobel Biology Prize winning research: “demonstrating that kitchen refuse can be reduced more than 90% in mass by using bacteria extracted from the feces of *giant pandas*”
How to evaluate a paper

• **Methodology**
  – Is the solution technically deep?
  – Is the solution applicable to a broad range of scenarios?
  – Is the solution repeatable?

• **Evaluation**
  – Is the solution better than earlier solutions?
  – Is the solution better than simpler/cheaper alternative solution?

• **Presentation**
  – Is the paper understandable by someone with enough technical background?
Traps in judging a paper

• All-around good papers are rare
  – Even those appeared in highly competitive conferences
  – Almost all papers have flaws in one aspect or another
    • It is up to your own judgment to know when the good parts out-weigh the bad ones

• Papers in good conferences are on average better than those in less competitive conferences, but not absolutely

• Don’t be intimidated by the reputation of the authors or which institutes they are associated with
The more you read, the better you are at it

- We will read around 40 papers throughout the term
- You will read additional 5-10 papers in connection with your group project
- You will become an expert!
Questions?