Spring 2014
[Th, 2:00-4:45]
Manning Hall, RM 304

INLS 509

Information Retrieval

Instructor: Diane Kelly, Ph.D.
E-Mail: dianek@email.unc.edu
Phone: 919.962.8065
Office: Manning Hall, RM 10
Office Hours: by appointment

Description

Study of information retrieval and question answering techniques, including document classification, retrieval and evaluation techniques, handling of large data collections, and the use of feedback.

Goals

The field of information retrieval (IR) is concerned with the analysis, organization, storage, and retrieval of unstructured and semi-structured data. In this course, we will focus on mostly text. While IR systems are often associated with Web search engines (e.g., Google), IR applications also include digital library search, patent search, search for local businesses, and expert search, to name a few. Likewise, IR techniques (the underlying technology behind IR systems) are used to solve a wide range of problems, such as organizing documents into an ontology, recommending news stories to users, detecting spam, and predicting reading difficulty. This course will provide an overview of the theory, implementation, and evaluation of IR systems and IR techniques. In particular, we will explore how search engines work, how they "interpret" human language, what different users expect from them, how they are evaluated, why they sometimes fail, and how they might be improved in the future.

Pre-requisites and General Requirements

- There are no pre-requisites for this course if you are a graduate student. If you are an undergraduate, it is useful to have had either INLS 101 or 200.
- This is a small class so your participation is critical.
- You are expected to read the assigned readings by the dates listed on the schedule. Be prepared to ask questions and discuss the readings in class.
- You should have access to the readings during class.
- Please let me know in advance if you will not be able to attend class.

Evaluation

Your major assignments for this course are: Participation (10%), Homework (2@15% each), Literature Review (40%) and Final Exam (20%).

Materials

Readings will be placed in the Sakai site or you will need to obtain them from UNC Library's online collections or from the open web. Unfortunately, when I convert to a PDF, the character encoding for the URLs often gets messed-up, so you might not be able to click on the link or even copy and paste it into the URL box on your web browser. In case of failure, enter the title of the resource in a search box or manually enter the URL.

Milestones

February 6, 2014
Homework 1 Due

March 20, 2014
Homework 2 Due

April 3, 2014
Literature Review Proposal

April 24, 2014
Literature Review and Presentation

May 5, 2014 (Monday, Noon)
Final Exam

Assignments

Participation (10%)

You are an important part of this course. There are few things worse than teaching to a half-empty classroom. One thing worse than teaching to a half-empty classroom is teaching to a room full of people engaged in conversations with others or with their eyes glued to their computers! Because your presence is key to a successful class, your participation will be 'graded.' Your grade is based on several things: attendance, engagement and general participation.

- *Attendance.* You are expected to attend all classes and to arrive to class before it starts. You will lose participation points for excessive and unexcused absences and for arriving late to class. You should be seated and ready to start at 2:00 PM.
- *Behavior.* Be courteous to your classmates and course instructor by not conversing with others during class lectures. Turn off cell phones, pagers, and other devices that might disrupt class. Use laptops and other devices to support current course activities only. Pay attention. Daydream infrequently.
- *General Participation.* Class participation consists of doing the following: being prepared for class, making observations about the readings, asking questions, taking notes and actively listening.

Homework (2@15% each)

You will have two homework assignments that ask you questions about the readings and give you an opportunity to practice some of the problems we work in class. Homework is due at the start on the class period on which it due. Late submissions will be penalized. The penalty will be assessed based on the reason(s) that it was not turned in on time.

Literature Review and Presentation (40%)

Note: Assignment description borrowed, in part, from Prof. Jaime Arguello's INLS 509 syllabus.

This assignment consists of three parts: proposal (10%); presentation (10%) and review (20%). The objective of this assignment is for you to gain an in-depth understanding of a particular area of Information Retrieval (IR) that is of interest to you and for you to share what you found with your classmates.

Your **proposal** is due on April 3. In this proposal, you should identify the general topic of your literature review and provide a *preliminary* list of research papers you will review (no less than 10, no more than 15). For at least five of the papers, you should provide a one-paragraph summary, including 2-3 sentences about why you chose the paper. If citation information is available for the paper, note this as well. You should try to pick the most important and well-known pieces of research to include and this can be a good indicator of impact.

In your proposal, you should also include a narrative description of your search process and the evolution of your thinking about the topic (you might even structure this part as a series of dated diary entries). Keep track of the dates and times you search, the databases and journals your search, the queries you issue and the search strategies you employ (e.g., citation chaining, journal run, speaking with other people). What were your goals at different points? What problems did you experience? What did you do to overcome these problems? How did your thinking about your topic change over time? How did you feel at different points?

This proposal is an important opportunity for you to get feedback from me about your topic, scope, resources and search strategies. **Your proposal should be emailed to me.** You are encouraged to come by my office to discuss your preliminary ideas and findings BEFORE this assignment is due.

Your **presentation** is due on April 24. You will present your literature review to the class. This is mainly so that your classmates can learn about the topic, but it also provides you with an opportunity to practice synthesizing a lot of literature, presenting it to an audience that does not know much about the topic and leading discussion. Your presentation should be around 15 minutes. Following this, there will be 5-7 minutes for questions and discussion. You should prepare at least two discussion questions. I assume that most

people will use slides (it doesn't matter what type of software is used). If you'd like to do something different, please let me know. **You should upload your slides to our course Sakai site.**

Your **review** is also due on April 24. Your review should be about 10-15 single spaced pages (or 20-30 double spaced) and discuss between 10-15 research papers. Your paper should include the following information:

1. A description of the topic and the specific problem/issue you explored. What is the problem? Why is it important? Why is it difficult?
2. A survey of how others have attempted to solve the problem. Organize and present the reviewed literature in a way that shows you understand the different approaches to the problem. How are different solutions similar? How are they different? Do different solutions make different assumptions? What are those assumptions? How have solutions evolved with time? How have different approaches built on each other?
3. A survey of evaluation. How are solutions to the problem typically evaluated? Are there agreed upon metrics that are suitable to the task? What assumptions do these metrics make? Are there any specific conferences or workshops that specialize on the task?
4. What do you think? How do you see this area of IR progressing? What are the key problems that remain to be solved? How do you think this research will change people's lives? Your own view of the problem is very important. Make sure to allocate about 20% of your paper and presentation to this.

Around mid-February, I will distribute a list of suggested topics for this assignment, a starter-paper for each topic and a list of resources. We will also have the opportunity to discuss this assignment in class if you have questions.

Final Exam (20%)

The Final Exam is cumulative and integrative, and will be closed-book and closed-note. The format of the exam questions will be varied. I like open-ended questions; in particular, I like to present stimuli (for example, a passage from a piece of research or a figure from the readings) and ask you questions about the stimuli. You might also have a few multiple-choice questions. You will also have problems similar to those from the homework. During the final exam, you will be allowed to bring one-page of notes: these should be hand-written notes on one side of a blank piece of 8X10 paper.

Grading

Grading is based on UNC Registrar Policy for undergraduate and graduate-level courses (http://registrar.unc.edu/academic-services/grades/explanation-of-grading-system/).

Graduate (H: 95-100%) • Superior work: complete command of subject, unusual depth, great creativity or originality Undergraduate (A: 95-100%; A-: 90-94%)) • Mastery of course content at the highest level of attainment that can reasonably be expected of students at a given stage of development.
Graduate (P: 80-94%) • Satisfactory performance that meets course requirements (expected to be the median grade of all students in the course) Undergraduate (B: 80-89%) • Strong performance demonstrating a high level of attainment for a student at a given stage of development.
Graduate (L: 70-79%) • Unacceptable graduate performance: substandard in significant ways Undergraduate (C: 70-79%) • A totally acceptable performance demonstrating an adequate level of attainment for a student at a given stage of development.
Undergraduate (D: 60-69%) A marginal performance in the required exercises demonstrating a minimal passing level of attainment.
Graduate (F: 69% or less) • Performance that is seriously deficient and unworthy of credit Undergraduate (F: 59% or less) • For whatever reason, an unacceptable performance. The F grade indicates that the student's performance in the required exercises has revealed almost no understanding of the course content.

January 9: Introductions and Course Overview; Introduction to Information Retrieval; Architecture of a Search Engine

- READ: Croft, W. B., Metzler, D., & Strohman, T. (2009). *Search engines: Information Retrieval in Practice*. Cambridge University Press. (Chapters 1 & 2)

January 16: Getting Text: Crawls and Feeds; Processing Text

- READ: Croft, et al., Chapter 3; Chapter 4.0-4.3
- READ: Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval*. New York, NY: Cambridge University Press. (Chapter 2). Online at: http://nlp.stanford.edu/IR-book/information-retrieval-book.html. (pgs. 19-34)

January 23: Ad-hoc Retrieval; Indexing and Query Processing

- READ: Manning, et al., Chapter 1
- READ: Croft, et al., Chapter 7.0-7.1.1
- READ: Croft, et al., 5.0-5.3.6; 5.7-5.7.2

January 30: Retrieval Models: Vector Space, Term Weighting and Relevance Feedback

- READ: Rasmussen, E. (2011). Access models. In I. Ruthven & D. Kelly (Eds.) *Interactive Information Seeking, Behaviour and Retrieval*. Facet Publishing.
- READ: Croft, et al., Chapter 7.0-7.1.2

February 6: Retrieval Models: Language Modeling and Query-likelihood

- READ: Croft, et al., 7.3
- **DUE: HWK 1**

February 13: Class Cancelled

February 20: Web Search, Link Analysis & Spam

- READ: Baeza-Yates, R., Ribeiro-Neto, B. & Maarek, Y. (2012). Web Retrieval (Chapter 11). *Modern Information Retrieval (2nd Edition)*. Addison-Wesley. Online: **http://www.mir2ed.org**. (**PAGES 449-480)**
- READ: Gyongyi, Z., & Garcia-Molina, H. (2005). Spam: It's not just for inboxes anymore. *IEEE Computer, 38*(10), 28-34. Online: **http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.387.7549&rep=rep1&type=pdf**
- REVIEW: **https://www.google.com/insidesearch/howsearchworks/fighting-spam.html**

February 27: User Interaction & Interactive IR

- READ: Baeza-Yates, R., Ribeiro-Neto, B. & Maarek, Y. (2012). Web Retrieval (Chapter 11). *Modern Information Retrieval (2nd Edition)*. Addison-Wesley. Online: **http://www.mir2ed.org**. (**PAGES 482-507)**
- READ: Cool, C. & Belkin, N.J. (2011). Interactive information retrieval: history and background. In I. Ruthven & D. Kelly (Eds.) *Interactive Information Seeking, Behaviour and Retrieval*. Facet Publishing.

March 6: Evaluation I: History, Measures; Significance Tests; Relevance & Assessors

- READ: Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval, 4*(4), 247-375. (PAGES: 248-350) (UNC Library)

March 13: SPRING BREAK

March 20: Evaluation II: User Studies

- READ: Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*(1-2), Chapters 9-10. (UNC Library)
- READ: Qvarfordt, P., Golovchinsky, G., Dunnigan, T., & Agapie, E. (2013). Looking ahead: Query preview in exploratory search. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*, 243-252.
- **DUE: HWK 2**

March 27: Log Analysis and User Behavior

- READ: Dumais, S., Jeffries, R., Russell, D. M., Tang, D. & Teevan, J. (forthcoming). Understanding user behavior through log data and analysis. J.S. Olson and W. Kellogg (Eds.), *Human Computer Interaction Ways of Knowing*. New York: Springer, 2014.
- READ: Bateman, S., Teevan, J., & White, R. W. (2012). The search dashboard: How reflection and comparison impact search behavior. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*, Austin, TX, 1785-1794.

April 3: Log Analysis and Click Bias (Example Study)

- READ: Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, 154-161.
- **DUE: Literature Review Proposal**

April 10: Log Analysis: Search Trails and Snippet Caption Features (Example Studies)

- READ: White, R. W. & Huang, J. (2010). Assessing the scenic route: Measuring the value of search trails in Web logs. *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR '11)*, Geneva, Switzerland, 587-594.
- READ: Clarke, C. L. A., Agichtein, E., Dumais, S., & White, R.W. (2007). The influence of caption features on clickthrough patterns in web search. *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*, 135-142.

April 17: Catch-up; The Future; Wrap-up and Review

- READ: Allan, J., Croft, B., Moffat, A., & Sanderson, M. (2012). Frontiers, challenges and opportunities for information retrieval: Report from SWIRL 2012. SIGIR Forum, 46(1), 2-32. Online: http://sigir.org/forum/2012J-TOC.html
- READ: Baeza-Yates, R., Ribeiro-Neto, B. & Maarek, Y. (2012). Web Retrieval (Chapter 11). *Modern Information Retrieval (2nd Edition)*. Addison-Wesley. Online: http://www.mir2ed.org. (Section 11.11)

April 24:

- Your Presentations
- Due: Literature Review & Presentation

May 5: Final Exam: 12:00-3:00 PM