CS 6501: Information Retrieval

Hongning Wang
Department of Computer Science
University of Virginia

1 Course Overview

Search engine systems, such as Google and Bing, are now an essential tool of everyone's life to deal with explosively growing online information (e.g., web pages, tweets, video, images, news articles, forum discussions, and scientific literature). In this course, you will learn the underlying technologies of modern information retrieval system, and obtain hands-on experience by using existing information retrieval toolkits to set up your own search engines and improving their search accuracy.

This is a graduate-level introductory course for information retrieval. It will cover algorithms, design, and implementation of modern information retrieval systems. Topics include: retrieval system design and implementation, text analysis techniques, retrieval models (e.g., Boolean, vector space, probabilistic, and learning-based methods), search evaluation, retrieval feedback, advanced text mining, search log mining, and applications in web information management.

2 Prerequisites

It is recommended you have taken CS 2150 (or equivalent course) and have a good working familiarity with at least one programming language (Java is recommended). Significant programming experience will be helpful as you can focus more on the algorithms being explored rather than the syntax of programming languages.

Basic mathematics background is also required. Since this is a graduate-level course, you should know basic concepts of probability (e.g., Bayes's theorem), linear algebra (e.g., vector, matrix and inner product). Good knowledge in mathematics will help you gain in-depth understanding of the methods discussed in the course and develop your own idea for new solutions.

3 Text Books

There is no official textbook for this course. However, we do recommend the following books for your reference (especially the first one).

1. ***Introduction to Information Retrieval***. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.
2. ***Search Engines: Information Retrieval in Practice***. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.
3. ***Modern Information Retrieval***. Baeza-Yates Ricardo and Berthier Ribeiro-Neto. 2nd edition, Addison-Wesley, 2011.

4. ***Information Retrieval: Implementing and Evaluating Search Engines***. Stefan Buttcher, Charlie Clarke, Gordon Cormack, MIT Press, 2010.

4 Course Content & Schedule

In this course, we will introduce a variety of basic principles, techniques and modern advances for searching, managing, and mining information. Topics to be covered include (the schedules are tentative and subject to change, please keep track of it on the course website):

1. Introduction (0.5 week): We will highlight the basic structure and major topics of this course, and go over some logistic issues and course requirements.

2. Search engine architecture (1.5 week): We will briefly discuss the basic building blocks of a modern search engine system, including web crawler, basic text analysis techniques, inverted index, query processing, search result interface.

3. Retrieval models (2 weeks): Retrieval model, a.k.a., ranking algorithm, is arguably the most important component of a retrieval system, and it directly determines search effectiveness. We will discuss classical retrieval models, including Boolean, vector space, probabilistic and language models. We will also introduce the most recent development of learning-based ranking algorithms, i.e., learning-to-rank.

4. Retrieval evaluation (1 week): Assessing the quality of deployed system is essential for retrieval system development. Many different measures for evaluating the performance of information retrieval systems have been proposed. We will discuss both the classical evaluation metrics, e.g., Mean Average Precision, and modern advance, e.g., interleaving.

5. Relevance feedback (1 week): User feedback is important for retrieval systems to evaluate the performance and improve the effectiveness of their service strategies. However, in most practical system, only implicit feedback can be collected from users, e.g., clicks, which are known to be noisy and biased. We will discuss how to properly model implicit user feedback, and enhance retrieval performance via such feedback.

6. Link analysis (1 week): We will discuss the unique characteristic of web: inter-connection, and introduce Google's winning algorithm PageRank. We will also introduce the application of link analysis techniques in a similar domain: social network analysis.

7. Text mining (3 week): Text information takes a major portion of online information. Properly modeling text documents is essential for improving search effectiveness and discovering actionable knowledge. We will introduce basic text mining techniques, including text categorization, clustering and topic models.

8. Search applications (2 week): We will introduce modern applications in search systems, including recommendation, personalization, and online advertising, if time allows.

9. Final project presentation (1 week): We will ask you to present your final project in class.

There will be one week near the end of semester for midterm review and exam. Detailed date to be announced later.

5 Communications

5.1 Meeting Times

We will have our lecture on every Tuesday and Thursday morning from 9:30am to 10:45am, at Rice Hall 340.

5.2 Office Hours

The lecture's office hour will be held on Thursday morning from 11am to 12pm, Rich Hall 408. The TA's office hour will be on Tuesday afternoon from 2pm to 4pm, Rich Hall 532.

5.3 Course Web Site

The course web site is `http://sifaka.cs.uiuc.edu/~wang296/Course/IR_Fall/`, and it will be migrated to a UVa site soon.

5.4 Piazza

The most important forum for communicating in this class is the course's Piazza. Piazza is like a newsgroup or forum – you are encouraged to use it to ask questions, initiate discussions, express opinions, share resources, and give advice.

We expect that you will be courteous and post only material that is somehow related to the topic of Information Retrieval or course content. The posts will be lightly moderated.

Note that private posts to Piazza can be used for things like conflict requests, or for letting us know that you have that sinking feelinganything you don't really want to share with your classmates.

The Piazza site for this class is `http://piazza.com/virginia/fall2014/cs6501/home`.

6 Gradings

The course is lecture-based. Grading is based on a set of homework assignments (25%), a late midterm exam (25%), a paper presentation (15%) and a final course project (40%). Since this is a graduate-level course, more credits are given towards paper presentation and course project (with 5% extra credit).

6.1 Homework

Homework assignments will be a mix of paperwork and machine problems. Written homework should be finished individually, discussions with peers or instructor is allowed, but copying or any other type of cheating is strictly prohibited. You will be given one week to finish the written homework. Some of the machine problems are designed for teamwork and due day may vary. Everyone will have one chance to ask for extension (extra three days from the deadline). After that, no extension will be granted. And please inform the instructor at least one day prior to the deadline, if you want an extension.

6.2 Midterm

Midterm will be a closed-book written exam. It will cover most of basic content discussed in class. The format of exam questions include True/False question, short answer questions, and short essay questions. The length of the exam will be 75 minutes in class.

6.3 Paper Presentation

After each lecture, there will be two to three assigned readings. Everyone is asked to select one paper from the list, and prepare a 15-minutes presentation for the class. One paper can only be presented by one student. Students are required to prepare the slides by themselves (the original authors' slides are not allowed to be used for this presentation). The purpose of this paper presentation is to help students to practice giving talks in front of public at conferences or other situations.

Both the instructor and other students will grade the presentation. The detailed grading criteria are as follows.

Table 1: Evaluation criteria for paper presentation

Aspects	Score		
Slides content was clearly visiable and self-explainable	1	5	10
Important messages of the paper were properly highlighted	1	5	10
Organization and logic of the presentation were easy to follow	1	5	10
Presenter did not just read off of the slides	0	5	10
Explained approaches/methods clearly	0	5	10
Responded to audience's questions well	0	5	10

6.4 Course Project

The course project is to give the students hands-on experience on solving some novel information retrieval and/or text mining problems. The project thus emphasizes either research-oriented problems or "deliverables." It is preferred that the outcome of your project could be publishable, or tangible, typically some kind of novel research problem or prototype system that can be demonstrated (where bonus points applied). Group work is strongly encouraged, but not required.

More details about the project will be discussed on the course website, including suggested topics and available resources, but it consists of these major parts:

1. Project proposal (20%): State your motivation, research problem, and expected outcome of your course project. Due on the end of 5th week of semester. Discussion with instructor prior to deadline is encouraged.

2. Project presentation (40%): 20 minutes presentation about what you have done for this course project. Format could be tailored according to the nature of the project, e.g., slides presentation and/or system demo.

3. Project report (40%): Detail documentation of your project. Quality requirement is the same as research papers, i.e., in formal written English and rigorous paper format. Due on the last week of course (***before*** project presentation).

6.5 Grade Cutoffs

We will use the standard grade cutoff points:

Table 2: Grade cutoff points

Letter Grade	Point Range
A	[93,105]
A-	[90, 93)
B+	[87, 90)
B	[83, 87)
B-	[80, 83)
C+	[77, 80)
C	[73, 77)
C-	[70, 73)
D+	[67, 70)
D	[63, 67)
D-	[60, 63)
F	[0, 60)

7 Acknowledgements

Thanks to Professor ChengXiang Zhai from University of Illinois at Urbana-Champaign; some teaching materials borrowed from his course site for CS410. And special thanks to Sean Massung from University of Illinois at Urbana-Champaign for his invaluable help in preparing this course.

Thanks to you for reading the entire syllabus. Hopefully it makes your experience a bit easier and less stressful.