CS 4501: 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 undergraduate-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, 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) and Linux operating system. 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. You are supposed to be familiar 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 are several good textbooks for the topic of information retrieval. The first book listed below is our official textbook, and the others are recommended references.

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 (∼1 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 (∼2 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 (∼3 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 (∼3 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 (∼2 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 (∼2 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. Search applications (∼2 week): We will introduce modern applications in search systems, including recommendation, personalization, and online advertising, if time allows.

5 Communications

5.1 Meeting Times

We will have our lecture on every Monday and Wednesday afternoon from 5:00pm to 6:15pm, at Olsson Hall 120.

5.2 Office Hours

The instructor's office hour will be held on Monday and Wednesday afternoon from 4pm to 5pm, Rich Hall 408. The TA's office hour will be announced later.

5.3 Course Web Site

The course web site is under construction and will be announced 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 feeling anything you don't really want to share with your classmates.

The Piazza site for this class is under construction and will be announced soon.

6 Gradings

The course is lecture-based. Grading is based on a set of homework assignments (55%), a midterm exam (25%), and a final exam (25%). Extra credits are given to the last homework assignment (it will be a competition among all the students in building the most effective search engine).

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. Any late submission will incur a 15% penalty for that assignment.

6.2 Exams

Both midterm and final exams will be closed-book written exams. The coverage of each exam will be discussed before the exam. A review session will be given one week before the exam. The format of exam questions include True/False question, short answer questions, and short essay questions. The length of the exams will be 75 minutes in class.

6.3 Grade Cutoffs

We will use the standard grade cutoff points:

Table 1: 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.