CS 4501: Information Retrieval

Hongning Wang
Department of Computer Science
University of Virginia

1 Course Overview

> *There's nothing that cannot be found through some search engine or on the Internet somewhere.*
>
> – Eric Schmidt, executive chairman of Alphabet Inc (formerly known as Google)

Given search engines have become an indispensable 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), aren't you already eager to know how Google understands our inquires, how Google retrieves the best answer among millions of candidates in just milliseconds, and how I could build a better Google?

In this course, you will learn the underlying technologies of modern information retrieval systems and obtain hands-on experience by using existing toolkits to set up your own search engine system to solve real-world problems. We will disclose the secrets behind a typical search engine system to you step by step, such that you will be able to:

- Get familiar with core search engine techniques ranging from back-end indexing, query processing and document ranking components to front-end result display, query recommendation and personalization modules;
- Explore a specific frontier topic in information retrieval on your own and present your thoughts and ideas to your instructor and peers in panel discussions;
- Team up with others to practice collaborative problem solving;
- Get a sense of basic research activities: formulating a real-world problem in an abstract and mathematic way, and develop principled solutions for it;
- Build your customized search engine for a specific domain you care about and deploy it to help others.

2 Prerequisites

Significant programming experience will be helpful as you can focus more on the algorithms being explored rather than the syntax of programming languages. 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.

On the other hand, 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. You are supposed to be familiar basic concepts of probability (e.g., probability distributions, Bayes's theorem and expectation), linear algebra (e.g., vector, matrix and inner product).

If you are not sure if you have met such prerequisites, please feel free to contact the instructor.

3 Course Content & Schedule

To help you build up your own search engine in the end, we will introduce a variety of basic principles, techniques and modern advances for searching, managing, and mining information. You will learn through lectures, in-class discussions, homework assignments and course projects. Topics to be covered include (the schedules are tentative and subject to change, please keep track of it on our course website):

1. *What is in this course?* 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. *What are the basic building blocks of Google?* Search engine architecture (∼2 week): A good system architecture is vital for a search engine to work efficiently and scalably. We will 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. (Machine Problem 1)

3. *How does Google find the best answer to my query?* 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. (Written Assignment 1 and Machine Problem 2)

4. *Is Google really better than Bing?* Retrieval evaluation (∼3 weeks): 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. (Machine Problem 3)

5. *How does Google improve itself?* Relevance feedback (∼2 weeks): 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. (Machine Problem 4)

6. *What is Google's big secret of early success?* Link analysis (∼2 weeks): Web is not simply a collection of documents, but the documents are all interlinked. 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. *What else can Google do?* Search applications (∼2 weeks): Search techniques are now far beyond the application of search engines, but it is actually all over our daily life. We will introduce modern applications in search systems, including recommendation, personalization, and online advertising, if time allows.

4 Assessments

This course will be mostly delivered through lecture-driven classes, but together we will explore the frontier new topics in the field by in-class discussions, mock panel discussions, homework assignments, and course projects. This helps you obtain a comprehensive understanding of the course materials.

Reading Assignments (10%) To help you get a comprehensive picture of information retrieval studies, we have compiled a list of recommended readings for you. They are all most influential and well-cited research papers in the field. On our course website, you can find the listed papers for each of these major topics. Please carefully read them *before* you come to class, prepare your answers to the listed questions, and post them on our course forum. Most of the questions are open-ended, and we discuss about them in class. You are encouraged to form groups in reading the papers. And you are highly welcome to post questions on our Piazza page to initiate any IR-related discussions. Your answers will be peer graded, and the answers get most of upvotes from your peers will receive *extra* bonus.

Homework Assignments (35%) We have prepared a set of homework assignments to guide you through the core pipeline of a search engine system step by step. The assignments will be a mix of written assignments and machine problems. Written assignments cover basic math concepts that are highly related to information retrieval study, and machine problems cover implementation details of different components in a retrieval system. As computational artifacts, implementation efficiency and readability of your machine problem solutions will be prioritized in our evaluation.

In-class Panel Discussions (5%) Each student is responsible for leading a in-class 10-minutes panel discussion about the frontier topic in information retrieval. You will have the freedom to define the topic: for example, demonstrate a new feature in Snapchat and discuss how to implement it efficiently with architectures we have learnt, or what should be the next big thing in information retrieval. Post your proposed topic of panel discussion on our course forum first, so that students with similar interest could form a group of panelists and host a longer session of discussion. The evaluation will be primarily participation based, and extra bonus points will be given to the discussion leaders who attract the most attention from the rest of class.

The sign-up of discussion topics should be performed before the end of 6th week of the semester, and then the instructors will assign the time slots of each proposed panel discussion.

Exam (20%) We will have one late mid-term exam to cover the important concepts and techniques we learn in the class. The format of exam questions include True/False question, short answer questions, and short essay questions. The exam will contain both fact-based questions (what you have learned in this class) and research-like open discussions (what you have read and thought about in this class). The length of the exam will be 75 minutes in class. A review session will be given one week before the exam.

Course Project (30%) Teamwork is very important in computer science: it is hardly to imagine any large computer system is built by one person. Our course project gives you hands-on experience on solving some novel information retrieval problems that you care about. The project appreciates either research-oriented problems or “deliverables.” You need to identify the problem on your own, apply the knowledge learned (or even beyond) in class, and work in a group of 3-4 students to solve it. It is preferred that the outcome of your project could be publishable, e.g., your (unique) solution to some (interesting/important/new) problems, or tangible, e.g., some kind of prototype system that can be demonstrated. Bonus points will be given to the groups meet either one of above criteria. Discuss with the instructor and TAs about your project idea and progress is an important way to ensure your success in the end. Every group needs to present their work to the class and submit a written report to summarize their results.

5 Resources

The best resource for this course is, of course, Google! There are already various types of informative documentations, technical reports, research papers, and open implementations out there on the Internet. Form the right keyword queries, properly use the advanced search functions, and realize the limitation of current retrieval paradigm are also important steps for you to learn in this course.

We have an official text book for this course, "***Introduction to Information Retrieval***. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007." The online free version of this book is available at `http://nlp.stanford.edu/IR-book/information-retrieval-book.html` You can download the PDF version for your reference.

In addition, there are several other good textbooks for the topic of information retrieval, and reading them will give you a more comprehensive understanding of information retrieval.

- ***Search Engines: Information Retrieval in Practice***. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.
- ***Modern Information Retrieval***. Baeza-Yates Ricardo and Berthier Ribeiro-Neto. 2nd edition, Addison-Wesley, 2011.
- ***Information Retrieval: Implementing and Evaluating Search Engines***. Stefan Buttcher, Charlie Clarke, Gordon Cormack, MIT Press, 2010.

We have also list several good online resources to help you master the course material, including similar courses offered in other top computer science departments, public toolkits and libraries, and tutorial papers and reports. You are also welcome to share any material you found helpful in our course forum.

6 Policies

What should not you do in this course? Plagiarism is considered as a serious misconduct in computer science in both industry and academia: it hurts your credibility and might also cause legal matters in copyright and intellectual property. As a result, the written assignments should be finished individually, discussions with peers or instructor is allowed, but copying or any other type of cheating is strictly prohibited. For the machine problems, copying others' code or implementation is also prohibited, and using third-party public library is allowed (unless explicitly introduced not to) but it has to be clearly documented and explained in your assignment report.

When to start working on my assignments? 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 within two weeks passing the due date will incur a 15% penalty for that assignment; 50% penalty afterwards. Start early is always recommended: given the nature of computer engineering, exceptions and errors always happen in the last step.

Evaluation Rubrics The detailed evaluation rubrics will be carefully discussed in the instruction for your homework assignments, project presentation and report. And you find them on our course web accordingly.

Grade Cutoffs We will use the standard grade cutoff points and no curing will be applied to your final grades, such that you can keep track of and predict your final letter grade on the fly:

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 Communications

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

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 held on Tuesday and Thursday afternoon from 2pm to 3pm, Rich Hall 414.

Course Web Site The course web site is `http://www.cs.virginia.edu/~hw5x/Course/IR2015/_site`. All the course announcements and materials will be posted on this website. Our Collab site will only be used for homework submission and grades releasing.

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. The Piazza site for this class is `https://piazza.com/virginia/fall2015/cs4501/home`. Please enroll yourself at the beginning of this semester.

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.

8 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.