Information Retrieval Systems

Engineering and Optimization of Search Engines

MSCI 541/720

Winter 2018

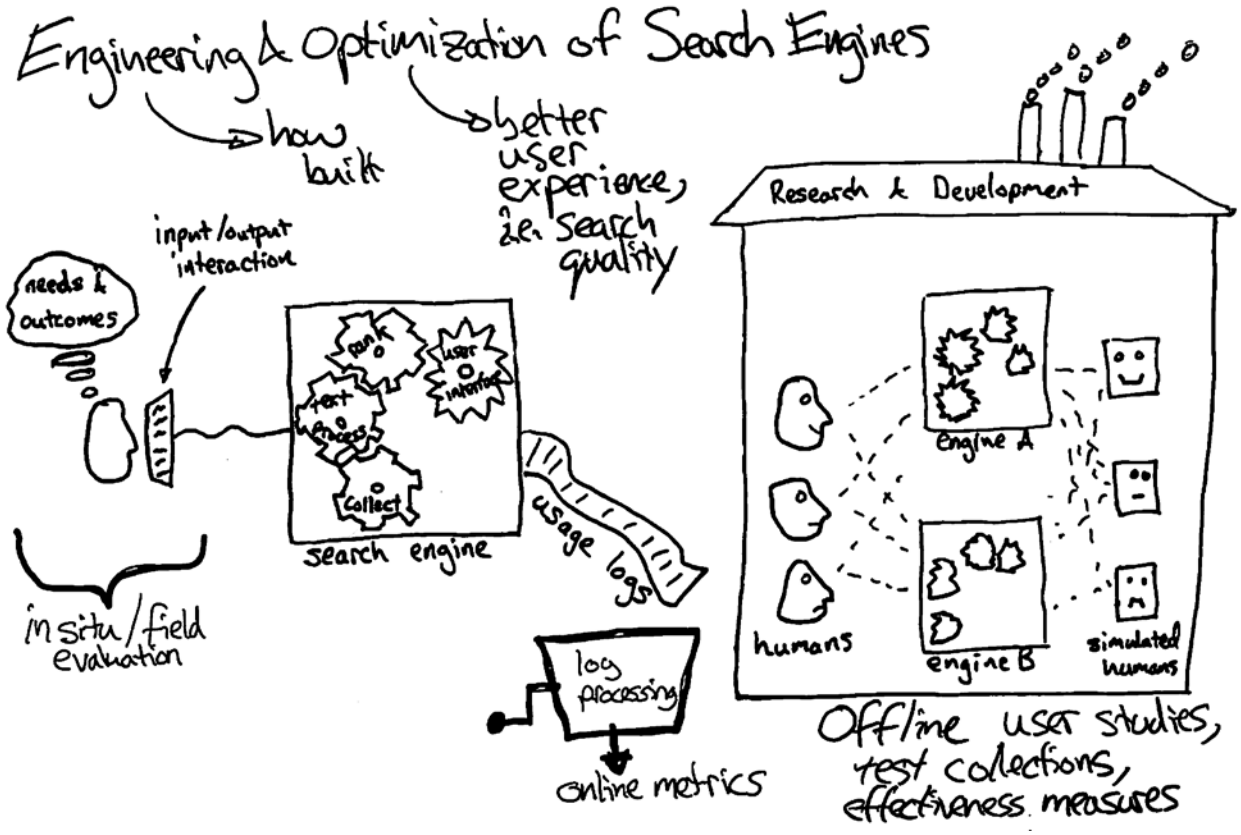
Instructor: Mark D. Smucker

Introduction

Search engines are prevalent in society. We use web search engines such as Google and Bing for many reasons ranging from finding simple facts to researching literature about leading treatments for diseases. Many web sites have their own search engines, e.g. Twitter, Amazon, and the University of Waterloo. Companies also often use search engines internally to help employees find information across the company. Certain products also contain search engines for a variety purposes including the search of personal email archives, help pages, and product documentation.

I have designed this course to provide students the opportunity to learn the engineering behind search engines and how to optimize search engines to provide higher quality user experiences.

Course Concepts



Intended Learning Outcomes

By the end of the course, a student should be able to:

1. Identify, explain, and implement the key components of a search engine.
2. Explain the advantages and disadvantages of in-situ, online, and offline evaluation methods.
3. Implement and compute offline effectiveness measures using a custom or existing test collection.
4. Make and justify decisions based on the outcome of experiments.
5. Diagnose search quality problems and suggest areas of engine improvement for future experiments.
6. Design and conduct an experiment to test a proposed system change.

Learning Activities

This course places significant weight on the homework assignments. The majority of the homework assignments require software design and implementation based on high level requirements and algorithms. A portion of the homework will also include problems similar to those that will be found on the final exam. The programming assignments are cumulative and require working software from the previous assignment to make the subsequent assignment also work.

Tutorials will be used to tutor and help students successfully complete the assignments.

Foundation for Successful Participation

Course prerequisites are MSCI 240 (Algorithms and Data Structures) and MSCI 252 (Probability and Statistics for Engineers) but as a 500 level course, students need to be comfortable with less structure and higher expectations for self-regulated learning. Past experience has shown that students who are weak in their information systems skills, in particular computer programming, have significant difficulty with course expectations.

Assessment of Intended Learning Outcomes

Outcome	HW	Exam
Identify, explain, and implement the key components of a search engine.	X	X
Explain the advantages and disadvantages of in-situ, online, and offline evaluation methods.	X	X
Implement and compute offline effectiveness measures using a custom or existing test collection.	X	X
Make and justify decisions based on the outcome of experiments.	X	X
Diagnose search quality problems and suggest areas of engine improvement for future experiments.	X	X
Design and conduct an experiment to test a proposed system change.	X	

There will be approximately 5 homework assignments. Homework not handed in on time will be late and receive a zero. Please see the late policy in the course details section.

MSCI 720 (Graduate) students are required to meet all the requirements of MSCI 541 plus they must complete a 3000-5000 word term paper. The paper is a chance to explore an area of information retrieval in more depth than we will in the course and its homework. In addition to the final term paper, MSCI 720 students are required to complete the following related work: submission of paper ideas, submission of an annotated bibliography. The paper can either be a serious survey of a research area of information retrieval (a literature review), or it may be an actual research paper that investigates some aspect of information retrieval and measures its performance.

541 Grading
Homework: 50% - Each assignment is 10%.
Final Exam: 50%

720 Grading
Homework: 40% - Each assignment is 8%.
Term paper: 10% (Breakdown: ideas: 1%, annotated bibliography: 4%, paper: 5%)
Final Exam: 50%

The instructor reserves the right to modify the weighting of all course components in the final computation of the course grade. All grades are final one week after work has been returned. If you have an issue with graded homework, please see the TA first. If you cannot resolve your problem with the TA, please see the instructor. All graded material will be returned in class or tutorial. The final exam is not returned.

The final exam is open book and notes. Any paper book and paper notes are allowed. All electronic devices other than a simple scientific calculator are banned for the final exam. Thus, electronic copies of the text book, etc. are not allowed for the final exam.

Course Administration Details

Instructor: Mark D. Smucker, mark.smucker@uwaterloo.ca, http://www.mansci.uwaterloo.ca/~msmucker/
TA: TBD

Lecture: TTh 11:30-12:50, CPH 3681
Tutorial: F 12:30-1:20, CPH 3681
Office Hours

By appointment, CPH 3624. See http://www.mansci.uwaterloo.ca/~msmucker/appointments.html

Canceled and Rescheduled Lectures

The following lectures are cancelled:

- Tuesday, Feb 13
- Thursday, Feb 15
- Tuesday, March 13
- Thursday, March 15

We need 6 hours to make up these lectures. Scheduled makeup lectures will be announced in class.

UW-Learn

We will use UW-Learn for email messaging to the class, distribution of materials, and other uses. You are responsible for email messages sent to you via UW-Learn, i.e. email messages sent to your uwaterloo.ca address.

Texts

Recommended: Search Engines: Information Retrieval in Practice by Croft, Metzler, and Strohman. This text is available online from the authors (free), but you may not use the electronic version during the final exam.

Lectures and Tutorials

You should come to all lectures. Tutorials will be a chance for you to obtain help from the TA. Lectures will be your primary source for this course with the text (and the multitude of other texts and the web) as your supplementary sources. Lecture notes will not be provided. You must obtain lecture notes from fellow students if you miss lecture. The instructor's notes are for his use and are not suitable for students.

Late policy

Excused late work: Extensions will be granted for serious illness or other serious life events (e.g. death in family). As soon as possible, contact the instructor to obtain a new due date.

No other extension will be granted for late work except as follows: Each student is granted 3 late days that can be used without excuse for homework only. Students are allowed to spend their late days as they wish. Weekends and holidays are not counted as late days. Students are advised to use their late days wisely (or not use them at all). All unexcused late work will be given a zero.

Etiquette / Rules

We have some simple rules that are all based on respect for the educational process, respect for the privilege and honor to attend university, and respect for others and yourself. Failure to follow these classroom rules will result in a 3 stage escalation process:

1. Warning. The instructor will remind you to follow the rule in question.
2. Leave the classroom. The instructor will ask you to leave the classroom for the remainder of the day's lecture or tutorial.
3. Letter to Management Sciences Associate Chair of Undergraduate Studies or Department Chair. The instructor will ask you to leave the classroom and will also send a letter to the Management Sciences Associate Chair of Undergraduate Studies or Department Chair detailing the discipline problem.

The classroom rules are:
Rule 1: Be on time.
Rule 2: Raise your hand to speak.
Rule 3: No distractors. Details: No distracting materials. This includes all non-class materials. For example, working on homework for another class. Reading a newspaper. Laptops, iPods, phones, etc. are all banned. You cannot do two things at once. You must reserve 50 minutes each lecture for MSCI 541/720 or choose to not attend lecture.

Outside the classroom etiquette:

1. Do not contact the instructor by phone except in the case of emergency.
2. Do not "drop in for a quick question". All office visits to the instructor should be via appointment or during office hours.

Academic honesty / working with others / use of other sources

All work must be your own. You must acknowledge assistance from the web, books, and others. No material may be copied from any source. Turning in the work of others as your own; failure to acknowledge sources and assistance and discussions; and providing answers to others are violations of academic integrity. All violations of academic integrity will be reported to the Associate Dean.

You may discuss homework with others, but you may not exchange work or answers. For work involving computer code (e.g. Java, C#, etc.), this means that your work may not contain the code, or portions of code, of another person or book or website etc. unless otherwise specified by the homework assignment. Discussions of computer programs should be at the level of pseudo-code and not at the level of real computer code.

How do you work with others etc. and still be academically honest (ethical)?

1. Attempt to do all work by yourself first. If you solve a problem using only your brain and hard work, then you have no one else to acknowledge.
2. Work with others, but never copy. Working in groups can be a great way to learn material better and faster. You can also turn to original sources and other books, but you must do ALL of the following (a, b, c, and d):
    a. You must put away all group notes, books, papers, internet webpages, solution guides, etc. and write-up your own solution. If you cannot write up a solution from start to finish without looking at notes, then the work is not your own, but someone else's. Go back, study the problem some more, then try to write up your solution. Repeat until you can do it without using the notes. All students must be able to pass an oral examination that demonstrates that they have submitted their own work.
    b. You must acknowledge your sources whether a written source or a person. It is often obvious to your instructor and TAs when you have copied an answer, no matter how obscure you might think your source is. You must acknowledge anyone you discussed a problem or assignment with.
    c. You must also acknowledge the nature of your source. If you uncovered a solution to a problem, you must acknowledge that!
    d. Write your acknowledgments either at the beginning or end of your problem write-up. Examples:
        i. I discussed this problem with Bob and Beth.
        ii. Working with Sue, we solved this problem together.
        iii. I found an answer to this problem in XXXX, by X Y, page X.

Other Important Notices

Responsibilities: All students must be familiar with their responsibilities as outlined in this syllabus and as outlined by the Faculty of Engineering and the university. Please see the following url for more details: https://uwaterloo.ca/engineering/current-undergraduate-students/academic-support/course-responsibilities

Academic Integrity: In order to maintain a culture of academic integrity, member of the University of Waterloo community are expected to promote honesty, trust, fairness, respect and responsibility. Refer to Academic Integrity website (https://uwaterloo.ca/academic-integrity/) for details.

Grievance: A student who believes that a decision affecting some aspect of his/her university life has been unfair or unreasonable may have grounds for initiating a grievance. Read Policy 70 (https://uwaterloo.ca/secretariat/policies-procedures-guidelines/policy-70) Student Petitions and Grievances, Section 4. When in doubt, please contact the department's administrative assistant who will provide further assistance.

Discipline: A student is expected to know what constitutes academic integrity to avoid committing an academic offence, and to take responsibility for his/her actions. A student who is unsure whether an action constitutes an offence, or who needs help in learning how to avoid offences (e.g. plagiarism, cheating) or about "rules" for group work/collaboration should seek guidance from the course instructor, academic advisor, or the undergraduate Associate Dean. For information on categories of offences and types of penalties, students should refer to Policy 71

(https://uwaterloo.ca/secretariat/policies-procedures-guidelines/policy-71) Student Discipline. For typical penalties check Guidelines for the Assessment of Penalties (https://uwaterloo.ca/secretariat/policies-procedures-guidelines/guidelines/guidelines-assessment-penalties).

Appeals: A decision made or penalty imposed under Policy 70 (Student Petitions and Grievances) (other than a petition) or Policy 71 (Student Discipline) may be appealed if there is a ground. A student who believes he/she has a ground for an appeal should refer to Policy 72 (Student Appeals) http://www.adm.uwaterloo.ca/infosec/Policies/policy72.htm.

Note for Students with Disabilities: AccessAbility Services (http://uwaterloo.ca/disability-services/), located in Needles Hall, Room 1132, collaborates with all academic departments to arrange appropriate accommodations for students with disabilities without compromising the academic integrity of the curriculum. If you require academic accommodations to lessen the impact of your disability, please register with the office at the beginning of each academic term.

Plagiarism software and alternatives: Plagiarism detection software will be used to screen assignments in this course. This is being done to verify that use of all material and sources in assignments is documented. Students will be given an option if they do not want to have their assignment screened by this software. Any student wanting to opt out of this plagiarism detection methodology, must ask the professor for an alternative.

Intellectual Property: Students should be aware that this course contains the intellectual property of their instructor, TA, and/or the University of Waterloo. Intellectual property includes items such as:

- Lecture content, spoken and written (and any audio/video recording thereof);
- Lecture handouts, presentations, and other materials prepared for the course (e.g., PowerPoint slides);
- Questions or solution sets from various types of assessments (e.g., assignments, quizzes, tests, final exams); and
- Work protected by copyright (e.g., any work authored by the instructor or TA or used by the instructor or TA with permission of the copyright owner).

Course materials and the intellectual property contained therein, are used to enhance a student's educational experience. However, sharing this intellectual property without the intellectual property owner's permission is a violation of intellectual property rights. For this reason, it is necessary to ask the instructor, TA and/or the University of Waterloo for permission before uploading and sharing the intellectual property of others online (e.g., to an online repository). Permission from an instructor, TA or the University is also necessary before sharing the intellectual property of others from completed courses with students taking the same/similar courses in subsequent terms/years. In many cases, instructors might be happy to allow distribution of certain materials. However, doing so without expressed permission is considered a violation of intellectual property rights.

Please alert the instructor if you become aware of intellectual property belonging to others (past or present) circulating, either through the student body or online. The intellectual property rights owner deserves to know (and may have already given their consent).