INFO 4300 Courses of Study

Prerequisite: CS 2110/ENGRD 2110 or equivalent.

Studies the methods used to search for and discover information in large-scale systems. The emphasis is on information retrieval applied to textual materials, but there is some discussion of other formats. The course includes techniques for searching, browsing, and filtering information and the use of classification systems and thesauruses. The techniques are illustrated with examples from web searching and digital libraries.

INFO 4300 Courses of Study

Prerequisite: CS 2110/ENGRD 2110 or equivalent.

We will focus on search engine designt.

Studies the methods used to search for and discover information in large-scale systems. The emphasis is on information retrieval applied to textual materials, but there is some discussion of other formats. The course includes techniques for searching, browsing, and filtering information and the use of classification systems and thesauruses. The techniques are illustrated with examples from web search and digital libraries.

Why the switch? Why study IR?

• THEN: As recently as the 1990s, studies showed that most people preferred getting information from other people rather than from information retrieval systems.

• 2004 Pew Internet Survey: 92% of Internet users say the Internet is a good place to go for getting everyday information.

The field of computer science that is most involved with R&D for search is information retrieval (IR).
Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Basic assumptions of IR

- **Collection**: A set of documents
  - Assume it is a static collection for the moment

- **Goal**: Retrieve documents with information that is relevant to the user’s information need and helps the user complete a task

Ad hoc retrieval

These days we frequently think first of web search, but there are many other cases:

- E-mail search, Searching your laptop, Corporate knowledge bases, Legal information retrieval

Bush (1945) provided early, lasting inspiration for the field:

"Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, ‘memex’ will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

The Classic Search Model

User task

Info need

Query

Search engine

Query refinement

Results

Collection

Misconception?

Info on motivating dogs to perform at agility

Misformulation?

how motivate dogs agility

Search

Get Marseille to run faster and show enthusiasm at agility trials

Many Cornell Connections

- Gerard Salton
  - Father of IR
  - Co-founded our CS department

- Amit Singhal
  - PhD student of Salton's
  - Head of “search” at Google
  - Totally rewrote the search code at Google in 2001

Agility: Not just for high energy dogs

I used to watch agility trials on TV back when I had cable. Back when I had time to relax on a weekend morning and watch stuff on the “boob tube.” Basically, back before I had a dog. Before I even imagined I would have a dog. It was sort of a pipe dream of mine to do agility with the dog I would have someday.

And then I got a dog! A Border Collie/Golden Retriever mix. A dog who would surely be suited for agility. But as it turned out, she looked more like this than this.

Croft, Metzler & Strohman (2010)²

- “Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.” (Salton, 1968)
- General definition that can be applied to many types of information and search applications
- Primary focus of IR since the 50s has been on text and documents

² Another text book we’ll draw from. (Can rent from Amazon.)
What is a Document?

- Examples:
  - web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum postings, patents, IM sessions, Tweets, etc.
- Common properties
  - Significant text content
  - Some structure (e.g., title, author, date for papers; subject, sender, destination for email)

Documents vs. Database Records

- Database records (or tuples in relational databases) are typically made up of well-defined fields (or attributes)
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics to queries in order to find matches
- Text is more difficult

Documents vs. Database Records

- Example bank database query
  - *Find records with balance > $50,000 in branches located in Ithaca, NY.*
  - Matches easily found by comparison with field values of records
- Example search engine query
  - *bank scandals in southern ny*
  - This text must be compared to the text of entire news stories

Comparing Text

- Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval
- Exact matching of words is not enough
  - Many different ways to write the same thing in a "natural language" like English
  - e.g., does a news story containing the text "bank director in Ithaca steals funds" match the query?
  - Some stories will be better matches than others
Dimensions of IR

- IR is more than just text, and more than just web search
  - although these are central
- People doing IR work with different media, different types of search applications, and different tasks

Other Media

- New applications increasingly involve new media
  - e.g., video, photos, music, speech
- Like text, content is difficult to describe and compare
  - text may be used to represent them (e.g. tags)
- IR approaches to search and evaluation are appropriate

Different tasks: regulationRoom.org

An E-Rulemaking Scenario

"Summarize the public commentary regarding the prohibition of potassium hydroxide for peeling peaches"

E-mail, letters, blogs, technical reports, newswires

- speech understanding
An E-Rulemaking Scenario

"Summarize the public commentary regarding the prohibition of potassium hydroxide for peeling peaches."

- ad hoc retrieval
- machine translation
- cross-lingual IR
- multi-lingual E-mail, letters, editorials, technical reports, newswires
- multi-document summary
- document clustering
- text categorization / sentiment analysis
An E-Rulemaking Scenario

"Summarize the public commentary regarding the prohibition of potassium hydroxide for peeling peaches"

multi-document summary

commenter: J. Dougherty
organization: Stonyfield Farms
opinion: pro
reason: "potentially dangerous chemical"

- information extraction
- multi-document summarization
- question answering

"Summarize the public commentary regarding the prohibition of potassium hydroxide for peeling peaches"

multi-document summary

commenter: J. Dougherty
organization: Stonyfield Farms
opinion: pro
reason: "potentially dangerous chemical"

Multi-Document Summarization

(White et al., 2002)
Multi-Document Summarization

- Biographical summary

![Image](image.png)

Figure 3. Example biography summary for topic 162, "Lucille Ball".

[Lin and Hovy, DUC 2002]

Big Issues in IR

- Relevance
  - Retrieval models define a view of relevance
  - Ranking algorithms used in search engines are based on retrieval models

We will cover these...

Big Issues in IR

- Evaluation
  - Long tradition of using empirical procedures and measures to compare system output with user expectations
  - Typically use test collection of documents, queries, and relevance judgments
    » Most commonly used are TREC collections

We will cover these...

Big Issues in IR

- Users and Information Needs
  - Search evaluation is user-centered
  - Keyword queries are often poor descriptions of actual information needs
  - Interaction and context are important for understanding user intent
  - Query refinement techniques such as query expansion, query suggestion, relevance feedback improve ranking

We will cover these...
IR and Search Engines

- **A search engine** is the practical application of information retrieval techniques to large scale text collections
- Web search engines are best-known examples, but many others
  - Open source search engines are important for research and development
    - e.g., Lucene, Lemur/Indri, **Galago**
- Big issues include main IR issues but also some others

**Course Goals**

- To help you to understand search engines, evaluate and compare them, and modify them for specific applications
- Provide broad coverage of the important issues in information retrieval and search engines
  - includes underlying (mathematical) models and current research directions

**Reference Material**

- No specific required text book
- Many lectures are derived from these sources
Prereqs, Coursework and Grading

- Prerequisites
  - CS 2110.

- Grading
  - 60%: 3 homeworks/programming projects [groups]
    - Analytical questions + programming
  - 10%: 4 critiques of selected readings and research papers
  - 25%: final exam
  - 4%: participation
  - 1%: course evaluation completion

You'll be expected to participate in class discussion and class exercises or otherwise demonstrate an interest in the material studied in the course.

http://www.cs.cornell.edu/courses/cs4300/