Information Retrieval
INFO 4300 / CS 4300

- Instructor: Claire Cardie
  - Professor in CS and IS (and CogSci)
- Three TAs at last count
  - Liz Murnane
  - Jon Park
  - Chenhao Tan
- One dog
  - Marseille (mahr-say)

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 design!

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~~ **might be** 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).*

Manning, Schuetze, Raghavan [2009][1]

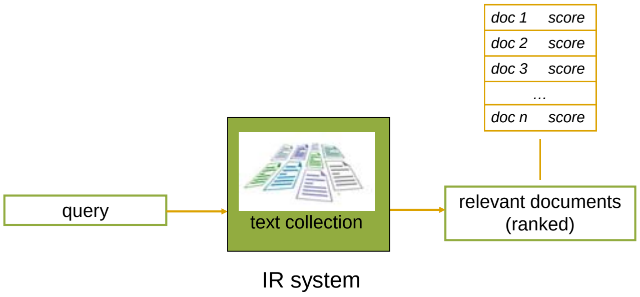
- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

[1] One of the text books we will draw from. (Freely available online.)

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

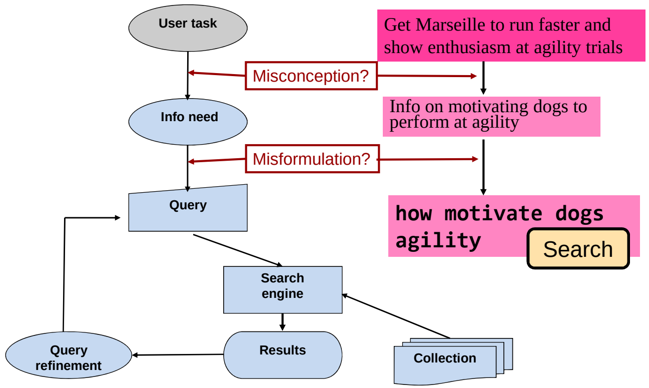


- These days we frequently think first of **web search**, but there are many other cases:
  - E-mail search, Searching your laptop, Corporate knowledge bases, Legal information retrieval
- Bush (1945) provided early, lasting inspiration for the field:

"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 Atlantic Monthly. URL: www.theatlantic.com/doc/194507/bush.

The Classic Search Model



User task

Get Marseille to run faster and show enthusiasm at agility trials

Misconception?

Info need

Info on motivating dogs to perform at agility

Misformulation?

Query

how motivate dogs agility

Search

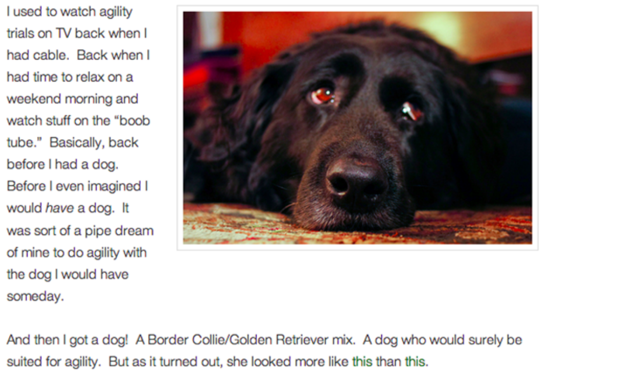
Search engine

Results

Collection

Query refinement

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.

Many Cornell Connections

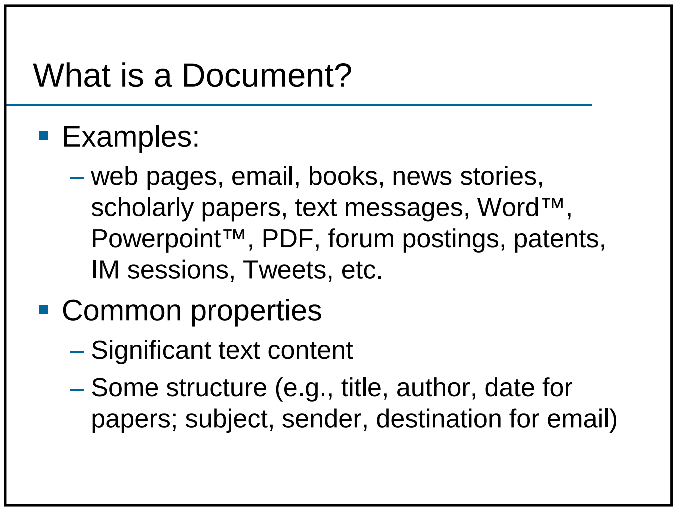
- Gerard Salton
  - Father of IR
  - Co-founded our CS department

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

Croft, Metzler & Strohman (2010)[2]

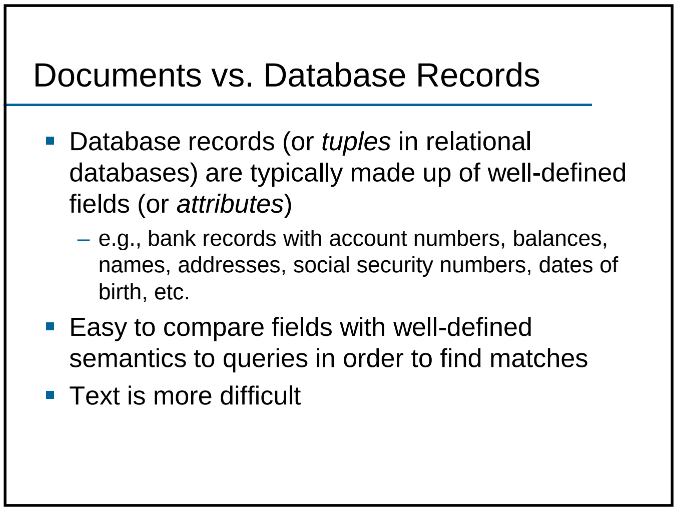
- *"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."* (Salton, 1968)
- General definition that can be applied to many types of information and search applications
- Primary focus of IR since the 50s has been on *text*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

[2] Another text book we'll draw from. (Can rent from Amazon.)



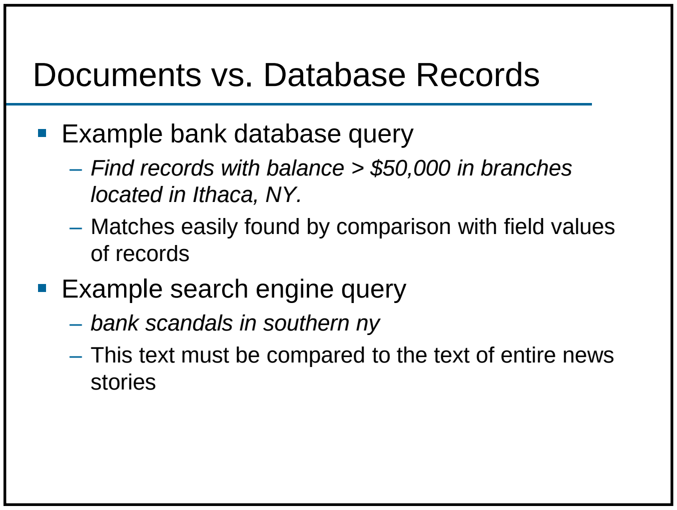
What is a Document?

- 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)



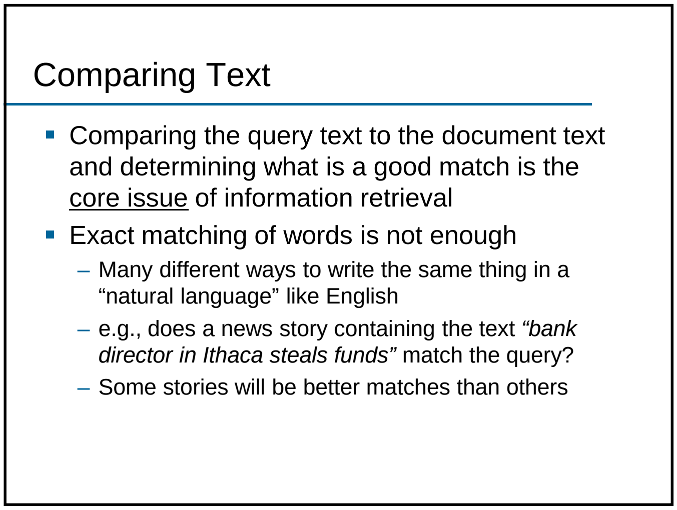
Documents vs. Database Records

- 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

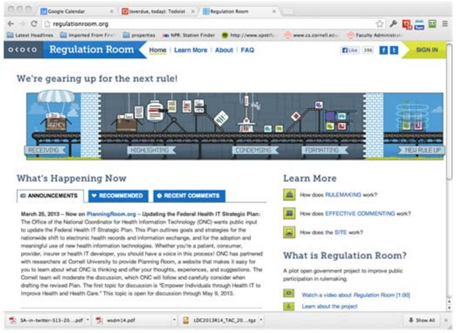
Dimensions of IR

- 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



We're gearing up for the next rule!

What's Happening Now

March 25, 2013 – Now on PlanningRoom.org – Updating the Federal Health IT Strategic Plan: The Office of the National Coordinator for Health Information Technology (ONC) wants public input to update the Federal Health IT Strategic Plan. This Plan outlines goals and strategies for the nationwide shift to electronic health records and information exchange, and for the adoption and meaningful use of new health information technologies. Whether you're a patient, consumer, provider, insurer or health IT developer, you should have a voice in this process! ONC has partnered with researchers at Cornell University to provide Planning Room, a website that makes it easy for you to learn about what ONC is thinking and offer your thoughts, experiences, and suggestions. The Cornell team will moderate the discussion, which ONC will follow and carefully consider when drafting the revised Plan. The first topic for discussion is "Empower Individuals through Health IT to Improve Health and Health Care." This topic is open for discussion through May 9, 2013.

Learn More

How does RULEMAKING work?

How does EFFECTIVE COMMENTING work?

How does the SITE work?

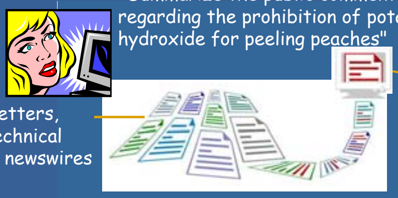
What is Regulation Room?

A pilot open government project to improve public participation in rulemaking.

Watch a video about *Regulation Room* [1:00]

Learn about the project

An E-Rulemaking Scenario

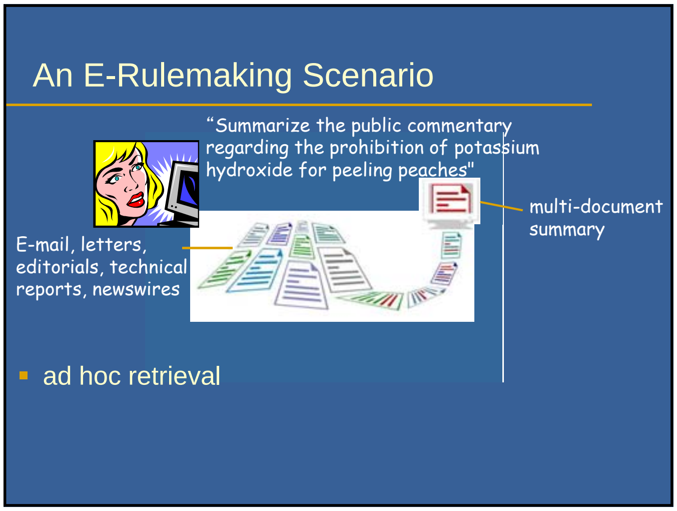


"Summarize the public commentary regarding the prohibition of potassium hydroxide for peeling peaches"

multi-document summary

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

- speech understanding



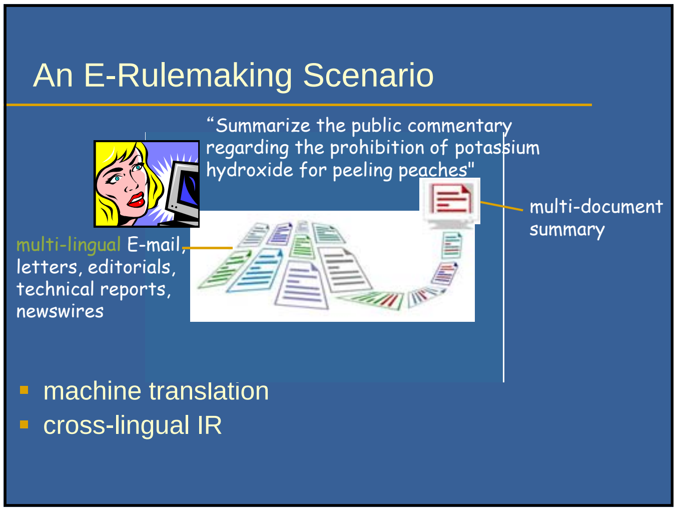
An E-Rulemaking Scenario

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

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

multi-document summary

- ad hoc retrieval



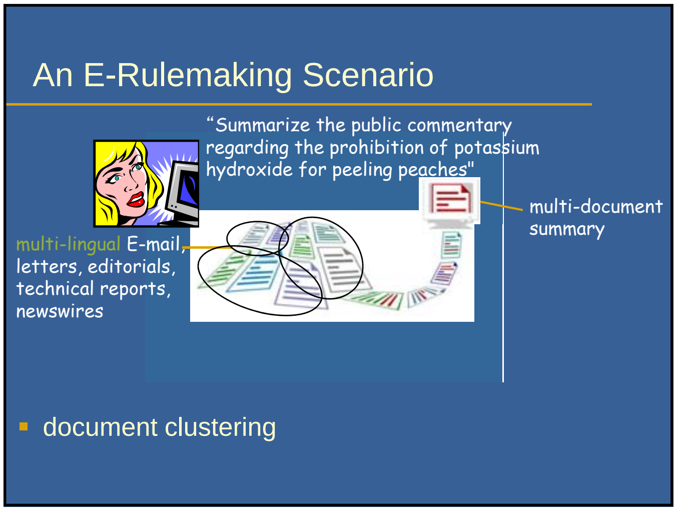
An E-Rulemaking Scenario

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

multi-lingual E-mail, letters, editorials, technical reports, newswires

multi-document summary

- machine translation
- cross-lingual IR



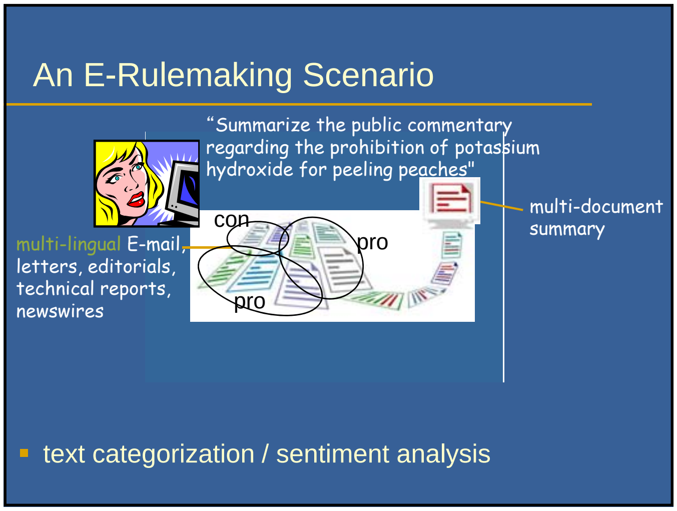
An E-Rulemaking Scenario

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

multi-lingual E-mail, letters, editorials, technical reports, newswires

multi-document summary

- document clustering



An E-Rulemaking Scenario

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

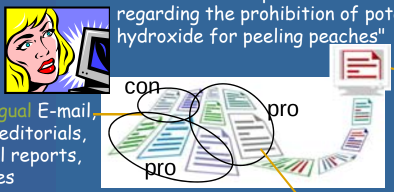
multi-lingual E-mail, letters, editorials, technical reports, newswires

multi-document summary

- text categorization / sentiment analysis

An E-Rulemaking Scenario

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



multi-document summary

multi-lingual E-mail, letters, editorials, technical reports, newswires

commenter: J. Dougherty

organization: Stonyfield Farms

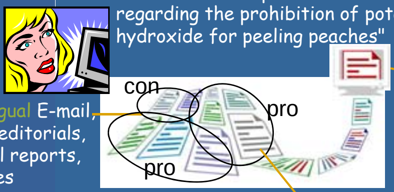
opinion: pro

reason: “potentially dangerous chemical”

- information extraction

An E-Rulemaking Scenario

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



multi-document summary

multi-lingual E-mail, letters, editorials, technical reports, newswires

commenter: J. Dougherty

organization: Stonyfield Farms

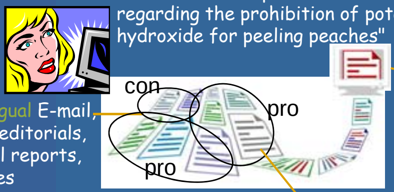
opinion: pro

reason: “potentially dangerous chemical”

- multi-document summarization

An E-Rulemaking Scenario

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



multi-document summary

multi-lingual E-mail, letters, editorials, technical reports, newswires

commenter: J. Dougherty

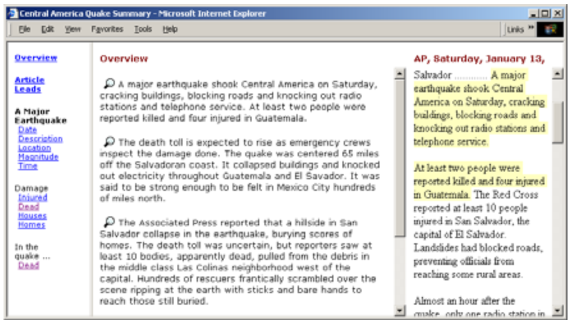
organization: Stonyfield Farms

opinion: pro

reason: “potentially dangerous chemical”

- question answering

Multi-Document Summarization


Central America Quake Summary - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links

Overview

Article Leads

A Major Earthquake
Date
Description
Location
Magnitude
Time

Damage
Injured
Dead
Houses
Homes

In the quake ...
Dead

Overview

A major earthquake shook Central America on Saturday, cracking buildings, blocking roads and knocking out radio stations and telephone service. At least two people were reported killed and four injured in Guatemala.

The death toll is expected to rise as emergency crews inspect the damage done. The quake was centered 65 miles off the Salvadoran coast. It collapsed buildings and knocked out electricity throughout Guatemala and El Savador. It was said to be strong enough to be felt in Mexico City hundreds of miles north.

The Associated Press reported that a hillside in San Salvador collapse in the earthquake, burying scores of homes. The death toll was uncertain, but reporters saw at least 10 bodies, apparently dead, pulled from the debris in the middle class Las Colinas neighborhood west of the capital. Hundreds of rescuers frantically scrambled over the scene ripping at the earth with sticks and bare hands to reach those still buried.

AP, Saturday, January 13,

Salvador .... A major earthquake shook Central America on Saturday, cracking buildings, blocking roads and knocking out radio stations and telephone service.

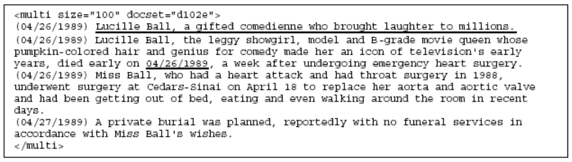
At least two people were reported killed and four injured in Guatemala. The Red Cross reported at least 10 people injured in San Salvador, the capital of El Salvador. Landslides had blocked roads, preventing officials from reaching some rural areas.

Almost an hour after the quake, only one radio station in

[White et al., 2002]

Multi-Document Summarization

- Biographical summary



```
<multi size="100" docset="d102e">
(04/26/1989) Lucille Ball, a gifted comedienne who brought laughter to millions.
(04/26/1989) Lucille Ball, the leggy showgirl, model and B-grade movie queen whose
pumpkin-colored hair and genius for comedy made her an icon of television's early
years, died early on 04/26/1989, a week after undergoing emergency heart surgery.
(04/26/1989) Miss Ball, who had a heart attack and had throat surgery in 1988,
underwent surgery at Cedars-Sinai on April 18 to replace her aorta and aortic valve
and had been getting out of bed, eating and even walking around the room in recent
days.
(04/27/1989) A private burial was planned, reportedly with no funeral services in
accordance with Miss Ball's wishes.
</multi>
```

Figure 3. Example biography summary for topic 102, "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

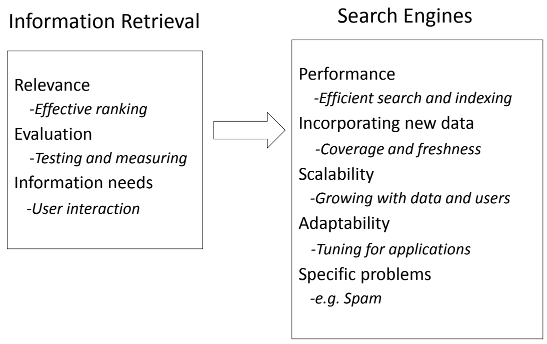
- Users and Information Needs
  - Search evaluation is user-centered
  - Keyword queries are often poor descriptions of actual information needs
  - Interaction and context are important for understanding user intent
  - Query refinement techniques such as *query expansion*, *query suggestion*, *relevance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

IR and Search Engines

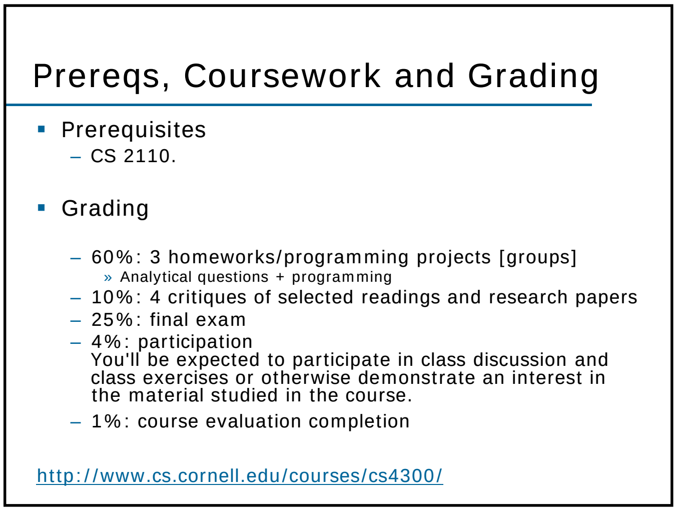


Course Goals

- 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
  - Croft, Metzler and Strohman, Search Engines: Information Retrieval in Practice, Pearson, 2010.
  - Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, An introduction to information retrieval. Cambridge University Press, 2008.



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
    You'll be expected to participate in class discussion and class exercises or otherwise demonstrate an interest in the material studied in the course.
  - 1%: course evaluation completion

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