Name that tune.

Song title? Performer(s)?
Topics in Information Retrieval and Web Search

“Existing Information”

2/9/10
Tell Me . . .

- . . . What’d you learn at the library on Monday?
Objectives

After this class you will be able to (it is my hope!):
- Know how information scientists study information retrieval.
- Say that you have contributed to the discussion today.
- Be able to make it to your next class, or the dorm, or wherever, without slipping on the ice.
- Big “thanks” to School of Information Asst. Prof. Matt Lease for most of this content.

- Randolph – remember to take roll.
# Dimensions of IR

<table>
<thead>
<tr>
<th>Content</th>
<th>Applications</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>Web search</td>
<td>Ad hoc search</td>
</tr>
<tr>
<td>Images</td>
<td>Vertical search</td>
<td>Filtering</td>
</tr>
<tr>
<td>Video</td>
<td>Enterprise search</td>
<td>Classification</td>
</tr>
<tr>
<td>Scanned docs</td>
<td>Desktop search</td>
<td>Question answering</td>
</tr>
<tr>
<td>Audio</td>
<td>Forum search</td>
<td></td>
</tr>
<tr>
<td>Music</td>
<td>P2P search</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literature search</td>
<td></td>
</tr>
</tbody>
</table>
Search/Retrieval Landscape

- **Verticals/content**: news, sports, classifieds, ...

- **Format**: text, images, audio, video
  - text: html/xml, text, email, chat, transcribed, blog, ...

- **Repository/archive/collection**
  - desktop/mobile, enterprise, Web

- **Query**: descriptive (textual/spoken), by example
  - Typically inexact (NOT ISBN, barcode, etc.)

- Typically both content & query are **unstructured** or only semi-structured (e.g., not database)
Some IR Tasks

- **Ad-hoc search**
  - Find relevant documents for an arbitrary text query

- **Filtering**
  - Identify relevant user profiles for a new document

- **Classification**
  - Identify relevant labels for documents

- **Question answering**
  - Give a specific answer to a question
### Beginnings

- **Vannevar Bush** (1890 – 1974). **1945**: “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. ...** Books of all sorts, pictures, current periodicals, newspapers... Business correspondence... longhand notes, photographs, memoranda, all sorts of things... There is, of course, provision for consultation of the record by the usual scheme of indexing. If the user wishes to consult a certain book, he taps its code on the keyboard, and the title page of the book promptly appears...”

- **Calvin N. Mooers** (1919–1994). **1951**: IR “embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, technique, or machines that are employed to carry out the operation”
Relevance

- What is it?
  - **Simplistic definition**: A relevant document contains the information that a person was looking for when they submitted a query to the search engine.
  - Many factors influence a person’s decision about what is relevant: e.g., task, context, novelty, style.
  - *Topical relevance* vs. *user relevance*
Modeling Relevance

- **Retrieval models** define a view of relevance.
- **Ranking algorithms** used in search engines are based on retrieval models.
- Ranking is typically statistical and more often (for text) based on its observable properties rather than latent linguistics.
  - i.e., counting simple text features such as words instead of inferring underlying linguistic syntax.
  - However, both kinds of *features / evidence* can be incorporated into a statistical model.
Evaluation

- Experimental procedures and measures for comparing system output to user expectations
  - Originated in Cranfield experiments in the 60s
- Experiments often use one or more pre-defined test collections of documents, queries, and relevance judgments
- Recall and precision are two examples of effectiveness measures
"The Cranfield indexing experiments in the 1960s are often cited as the beginning of the modern era of computer-based retrieval system evaluation (Cleverdon, Mills and Keen, 1966). In the Cranfield studies, retrieval experiments were conducted on a variety of test databases in a controlled, laboratory-like setting. . . . The aim of the research was to find ways to improve the relative retrieval effectiveness of IR systems through better indexing languages and methods (Cleverdon, 1970). The components of the Cranfield experiments were: a small test collection of documents, a set of test queries, and a set of relevance judgments, that is a set of documents judged to be relevant to each query. Human searchers, their interaction with the system, their interpretation of the query, and their process-formed relevance judgments were not factors included in these experiments. For purposes of performance comparisons, it was necessary to select quantitative measures of relevant documents output by the system under various controlled conditions. The measures used in the Cranfield II experiments are recall and precision, derivatives of the concept of relevance." (Hildreth, 2001).
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 inferring user intent
- Query refinement techniques such as *query expansion*, *query suggestion*, *relevance feedback* improve ranking
IR and Search Engines

Information Retrieval

- Relevance
  - Effective ranking
- Evaluation
  - Testing and measuring
- Information needs
  - User interaction

Search Engines

- Performance
  - Efficient search and indexing
- Incorporating new data
  - Coverage and freshness
- Scalability
  - Growing with data and users
- Adaptability
  - Tuning for applications
- Specific problems
  - e.g. Spam
Web Search Today II

Did you mean: **state of texas**

*TexasOnline: Official Portal of Texas*
Texas Secretary of State Esperanza Hope Andrade is reminding Texans that February 1 is the deadline to register to vote in the 2010 Primary Elections. ...
www.texasonline.com/ - Cached - Similar - ☐ ☐ ☒

*TExES*
Significant changes to Texas Administrative Code (TAC) §227.10 (a)(3)(C) were approved by the State Board for Educator Certification (SBEC) on October 10,...
www.texas.ets.org/ - Cached - Similar - ☐ ☐ ☒

*In the state of Texas*
In the state of Texas, "highly qualified" administrators and teachers are in great demand and in short supply. According to current census information,...
texasreview101.com/History.htm - Cached - Similar - ☐ ☐ ☒

*Local business results for state of texes near Austin, TX* - Change location

<table>
<thead>
<tr>
<th>Location</th>
<th>Name</th>
<th>Website</th>
<th>Phone</th>
<th>Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>The Bob Bullock Texas State History Museum: IMAX Theatre</td>
<td><a href="http://www.thestoryoftexas.com">www.thestoryoftexas.com</a></td>
<td>(512) 936-4639</td>
<td>50 reviews</td>
</tr>
<tr>
<td>B</td>
<td>State of Texas: Capitol Visitors Center</td>
<td><a href="http://www.tspb.state.tx.us">www.tspb.state.tx.us</a></td>
<td>(512) 463-0063</td>
<td>19 reviews</td>
</tr>
<tr>
<td>C</td>
<td>State of Texas: General Information</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Web++ Search Today III

Haiti Earthquake Relief
Donate $25 to Help Children and Families Hurt by the 7.0 Earthquake.
www.worldvision.org/haiti

Haiti Earthquake Disaster
www.foodforthepoor.org

Aid Haiti Quake Victims
Habitat for Humanity is working on shelter for victims. Donate today!
www.habitat.org/Haiti-Earthquake

Haiti - Latest News
Haiti earthquake, Monday news updates
- CNN - 10 minutes ago
12:20 p.m. Monday, January 18 -- Former Senate Majority Leader Bill Frist, a medical doctor, arrived in Port-au-Prince, Haiti, on Monday to help in the relief... full story

U.S. troops boost Haiti aid security as looters swarm
- Reuters via Yahoo! News - 9 minutes ago

Why does Haiti suffer so much?
- CNN - 14 minutes ago

More US troops, UN peacekeepers expected for Haiti
- AP via Yahoo! News - 15 minutes ago

more Haiti news...
Web Search Today IV

HAITI
News about Haiti
Images of Haiti
Haiti Economy
Haiti Food
Haiti Tourism
Haiti Map
Haiti Travel
Reference Articles on Haiti

RELATED SEARCHES
Haiti Radio
Haiti Music
Haiti News Network
Haiti News Today
Haiti Sakaplet
Haiti People
Police Haiti
Haiti News Picture

SEARCH HISTORY
Now you can go back further with search history. Learn More.

HAITI - Wikipedia, the free encyclopedia
Haiti (Haitian Creole: Ayiti), officially the Republic of Haiti (République d'Haiti; Republik Ayiti) is a Haitian Creole- and French-speaking Caribbean country. Along with ... History - Politics - Departments ... - Geography
en.wikipedia.org/wiki/Haiti - Wikipedia on Bing
Indexing Process

- Text Acquisition
- Text Transformation
- Index Creation
- Document data store
- Index

E-mail, Web pages, News articles, Memos, Letters
Query Process

- User Interaction
- Ranking
- Evaluation
- Log Data
- Document data store
- Index

R.G. Bias | rbias@ischool.utexas.edu
Who and Where?
Query formulation reflects an ongoing dialog between users and search engines

- Users formulate queries *for the* search engine, based on a mental model of what it “understands”
- Search engines optimize their “understanding” *for the* (most frequent) submitted queries
- Individual session and long term, personal and aggregate

Result: query “language” is continually evolving
Verbosity and Complexity

- Complex information requires complex description
  - Information theory [Shannon’51]
  - Human discourse implicitly respects this [Grice’67]
- Simple searches easily expressed in keywords
  - navigation: “alaska airlines”
  - information: “american revolution”
- Verbosity naturally increases with complexity
  - More specific information needs [Phan et al.’07]
  - Iterative reformulation [Lau and Horvitz’99]
Query Disambiguation

- Given (typically terse like “apple”) query, infer possible underlying intents / needs / tasks
- With longer queries, detect key concepts and/or segment (e.g. “new york times square”)
Vertical Search

- Aka/related: federated / distributed / specialty
- Searching the “Deep” web
- One-size-fits-all vs. niche search
  - Query formulation, content, usability/presentation
Cross-Lingual IR

- 2/3 of the Web is in English
- About 50% of Web users do not use English as their primary language
- Many (maybe most) search applications have to deal with multiple languages
  - monolingual search: search in one language, but with many possible languages
  - cross-language search: search in multiple languages at the same time
Cross-Lingual IR -- Ideal

- Let user express query in native language
- Search information in multiple languages
- Translate results into user’s native language
Spoken Search

Longer and more natural queries emerge given support for spoken input [Du and Crestiani’06]

See also: studies by Nick Belkin
Location-based Search

Cyclopedia: Augmented Wikipedia

Layar
Browse The World!
Mobile Augmented Reality Browser

Wikutitude
World Browser
The Driskill Hotel

The Driskill Hotel is a historic Austin hotel built in 1886, featuring 188 guest rooms and suites, dynamic meeting space perfect for events of any size, and an award-winning restaurant. - 604 Brazos Street, Austin, TX (800) 252-9367

More info
Write a review

Uchi Restaurant
Hotel San Jose
Stars in Your Eyes
Finn and Porter Restaurant
ZTejas
Fonda San Miguel Restaurant
See something you want to remember
When you notice an item to remember, tap "Remembers" in the Amazon App.

The Amazon app includes Amazon Remembers

Snap photo & send
Your iPhone camera will open. Take a photo of the item and it will be sent to Amazon.

See reminders
Your photos & any similar products that Amazon finds are stored in the app and on Amazon.com.
Content-based music search

Shazam
Spoken “Document” Retrieval
Other Visual Interfaces
Retrieving Information, not Documents

Events related to "haiti"

Timeline

- 1804: Haiti attempted to establish closer ties with the United States
- January 1, 1804: It is also known as Haiti's "independence city" because it was there that Gen. Jean-Jacques Dessalines declared Haiti's independence on January 1, 1804.
- January 1, 1804 (from W.Gonâvèes) "Gonâvèes is also known as Haiti's City of Independence because it was there that Jean-Jacques Dessalines declared Haiti, the former Saint-Domingue, independent from France on January 1, 1804 by reading the Act of Independence, drafted by Boisrond Tornerre, on the Place d'Armes of the town."
- January 1, 1804 (from List of French Governors of Saint-Domingue) "January 1, 1804, Independence of Haiti. Jean-Jacques Dessalines is Provisional Chief of the Haitian Government to September 22, 1804 and then Emperor of Haiti until October 17, 1806"
Community QA

Politics and Society

- Why is Latin America called 'Latin' America? (Edit categories)
  - Popularity: 335
- What is the difference between Saudi Arabia and Arabia? (Edit categories)
  - Popularity: 76
- What were the Maya achievements? (Edit categories)
  - Popularity: 48

Science

- What is the scientific name for dragonfly ants? (1 response)
- What are three uses of boron chemicals? (1 response)
- Does nucleus only contain protons and neutrons? (2 responses)
- What colour is violet? (3 responses)
- How has globalisation had an impact on the role of the state and the concept of sovereignty?
- Is s. equi sub specie equi an aerobic bacteria?
- When did the sugar act occur? (1 response)
- Does it rain sperm at a baby shower? (3 responses)
- How much stronger is the new vitreous carbon material invented by the Tokyo Institute of Technology compared with the material made from cellulose?
- Where can I go to find answers about alcohol related questions? (2 responses)
- If the earth suddenly stopped spinning, would we go flying through the air? (4 responses)
Social Media

Most Helpful Customer Reviews

7,905 of 8,083 people found the following review helpful:

⭐⭐⭐⭐⭐ Kindle 2: First Impressions from a Kindle 1 owner, February 25, 2009

By P. Inhofer - See all my reviews[

This review is from: Kindle Wireless Reading Device (6" Display, U.S. Wireless) (Electronics)

On the one hand I love my Kindle 1... I use it everyday, I subscribe to my favorite n...
Images are organized by topic and flow in from the left as they are submitted or dugg. Click each row to expand. Once expanded, drag the row to scroll and browse.
Washington Tackles Health Care Reform

Overview: The House and Senate have approved sweeping legislation that would provide health care insurance for most Americans, at huge cost to the government. The House plan, approved Nov. 7 in an almost strictly party-line vote, would spend $1.06 trillion to extend coverage to about 36 million Americans. The Senate bill, passed on Dec. 24, went through several iterations before attracting the support of a filibuster-proof coalition of 60 votes. It would cost $871 billion and give coverage to 31 million people who lack it now.

Read more...

All Coverage

Obama stumps in Massachusetts with health-care reform at risk
By Karl Vick, Paul Kane

The president tries to resurrect the candidacy of Democratic Senate candidate Martha Coakley in the surprisingly close race to succeed the late Edward M. Kennedy in Massachusetts. A Coakley loss to state Sen. Scott Brown (R) would deprive Democrats in the Senate of the super-majority they need to pass Obama's top domestic priority.

Related

Obama pleads for pragmatism on health-care overhaul - Feature

White House, unions reach deal on 'Cadillac' tax
By Lori Montgomery, Michael D. Shear

A bargain on taxing high cost health insurance policies puts negotiators close to an overall deal on a health-care reform plan. Under the agreement, family plans that cost more than $24,000 and individual policies that cost more than $8,300 would be subject to a 40 percent surtax. The tax would be imposed on the insurance company, but economists believe it would be passed on to workers.

Last year, the average family policy in America cost $13,375, according to a survey by the Kaiser Family Foundation.

The changes would cut revenue for the health reform package by $60 million over 10 years, a sum likely to be made up by
Memetracker
“Hyper-local” Search

EveryBlock: A news feed for your block

Track and discuss what’s new in your neighborhood.

Choose a city...

- Atlanta
- Boston
- Charlotte
- Chicago
- Dallas
- Detroit
- Houston
- Los Angeles
- Miami
- New York
- Philadelphia
- San Francisco
- San Jose
- Seattle
- Washington, DC

Find news near you
Enter an address and see nearby crime, news coverage, neighbor announcements and more.

Browse news by topic
Get a citywide overview of news by category, then click on topics to find what you’re looking for.

Explore your city
Choose a neighborhood or ZIP code and see what’s happening nearby or track trends over time.

Check out the free EveryBlock iPhone app.
Collaborative Search
Book Search

- Find books or more focused results
- Detect / generate / link table of contents
- Classification: detect genre (e.g. for browsing)
- Detect related books, revised editions
- Challenges
  - Variable scan quality, OCR accuracy
  - Copyright
  - Monetary model
The Information’s Out There

Datamob

Public data put to good use.

Datamob highlights the connection between public data sources and the interfaces people are building for them. Read all about it.

hot new datasets

- Failure Trace Archive
- Census 2000 Planning Database (for Census 2010)
- NYC DataMine
- New York City Baby Name Data

hot new interfaces

- The Full Wiki Maps
- NYC Way
- Casual Earnings of German Parliamentarians
- The Decline: The Geography of a Recession
Crowdsourcing

Page Hunt: Rules of the Game
1. Take a look at the web page shown.
2. Guess the words to 'hunt' it down.
3. If the web page is in the top 5 results, you win!
4. Get bonus points by avoiding frequent queries!

Play!

reCAPTCHA IS A FREE ANTI-BOT SERVICE THAT HELPS DIGITIZE BOOKS.

LEARN HOW reCAPTCHA WORKS

USE reCAPTCHA ON YOUR SITE

STRONG SECURITY
ACCESSIBLE TO BLIND USERS
30+ MILLION SERVED DAILY

NEW See how accurate reCAPTCHA is at digitizing content!
Mechanical Turk

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:
- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you’re satisfied with the results
Big Tent Challenges and Opportunities

Your guide
3 Multi-touch kiosks
Focus: designing software interfaces to support students and scholars in their tasks.
Cognitive IR

Yan Zhang
Assistant professor
School of Information
Guest Lecture: March 26, 2010

Research interests in main two areas:

- Psychological processes involved in people’s interaction with information systems, particularly web-based IR
- Consumer health information needs and information seeking, as well as the design of consumer health information systems.
Field Work

Lecia Barker
Research Associate Professor
lecia@ischool.utexas.edu
512-232-8364
Room: UTA 5.451
http://www.ischool.utexas.edu/~lecia/

My current research is in uses and social implications of information and communication technologies in education; socio-educational contexts of computing; and underrepresentation in information technology.

Diane Bailey
Assistant Professor
diane.bailey@ischool.utexas.edu
(512) 471-1626
Room: UTA 5.438
http://www.ischool.utexas.edu/~debailey/

Research examines the interplay among technology, work and organization. employs techniques for data collection and analysis that borrow heavily from ethnography but altered to suit the study of highly technical work and workers.
Content-based Image Retrieval

Kristen Grauman
Assistant Professor
Department of Computer Sciences

Guest Lecture: April 16, 2010
Natural Language Processing

Raymond J. Mooney
Professor of Computer Science
The University of Texas at Austin

- Syntactic and Semantic Analysis
- Named-Entity Recognition
- Also Machine Learning
Machine Learning

- E.g. Learning to Rank (automatically learning statistical ranking functions)
Who, what TYPES, . . .

- Friday – Bring your MBTI sheets.
- Also be prepared to discuss the Foss book.
- Questions?