CS 572: Information Retrieval

Lecture 1: Course Overview and Introduction
11 January 2016
Lecture Plan

• What is IR? (the big questions)

• Course overview

• Logistics

• Toy example
Information Retrieval

- *Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information*
  
  *Gerald Salton, 1968*

- Compare to Google corporate mission: *Google's mission is to organize the world's information and make it universally accessible and useful.*
Lots of Data

Ex: Large Hadron Collider: 36 Billion events per minute.

How much of this data is useful information?
Fundamental Mismatch

• Computers deal with **data**
• Humans (and systems) want **useful information**
• Examples:
  – **Structured Data** (easiest): bits $\rightarrow$ schema $\rightarrow$ query
  – **Text**: bits $\rightarrow$ represented as words
  – **Audio**: bits $\rightarrow$ speech recognition $\rightarrow$ words
  – **Images**: bits $\rightarrow$ image processing $\rightarrow$ computer vision $\rightarrow$ object recognition $\rightarrow$ `<tags>`?
  – **Video**: image processing+audio+activity recognition $\rightarrow$ ?
Search Challenges (2002)

UMass CIIR report, 2002

• **Global information access:** Satisfy human information needs through natural, efficient interaction with a system ... [over the world’s data]... in any language.

• **Contextual retrieval:** Combine search technologies and knowledge about query and user context ... to provide the most “appropriate” answer for a user’s **information** needs.
A “classic” IR Task

• Given:
  – A corpus of textual natural-language documents.
  – A user query in the form of a textual string.

• Find:
  – A ranked set of documents that are relevant to the query.

Contrast with database/SQL queries
A search model

Task

Info Need

Verbal form

Query

Search engine

Results

Corpus

Get rid of mice in a politically correct way

Info about removing mice without killing them

How do I trap mice alive?

Misconception?

Mistranslation?

Misformulation?

Find this: mouse trap any language

Search

Query refinement

Information Retrieval Challenges

• Understand user’s query (information need)

• Interpret and organize data (indexing)

• Rank “documents” by expected utility for user

• Find answers to show to user

• Evaluate, improve search, repeat
Specific Challenges (2012)

- **Beyond ranked list**: enrich querying & results
- **IR for all**: empower the user to search & learn
- **Capture context**: current task, time, etc.
- **Beyond document retrieval**: complex data & result integration
- **Domain search**: Verticals, apps, restricted data
- **Evaluation**: for new search types, tasks
History of the (IR) World: the Ancients

- Vanevar Bush: “As we may think”: “Memex”, 1945 → proto-IR → desktop search

- Gerald Salton: founder of IR as field
History: the mainframe era (50s-70s)

- High-level programming languages
- Fortran, Ada, Cobol, ...
- Data structures
- Sorting
- Searching
- Routing
- Graph theory
- Internet (ARPA-net)
- Salton develops IR
IR System (80’s, 90’s)

- **Query String** → **IR System** → **Ranked Documents**

- **Document corpus**

- **Results:**
  1. Doc1
  2. Doc2
  3. Doc3
  ...
History: the Web (90s)

• Distributed computing
• Mosaic, Netscape, IE
• Web search
  – Alta Vista, Excite...
Google architecture (circa 1998)
Web 2.0: Now
Search Engine Architecture: Indexing

- Text Acquisition
- Index Creation
- Text Transformation

Input: E-mail, Web pages, News articles, Memos, Letters

Output: Document data store, Index
Search Engine Architecture: Querying

![Diagram of search engine architecture](image)

- User Interaction
- Ranking
- Evaluation
- Log Data
- Document data store
- Index
And in this course...

"I think you should be more explicit here in step two."
Course Logistics

• Lectures:
  – Monday, Wed: 11:30am-12:45pm, W301
  – Following dates will be rescheduled or canceled:
    • Feb 22, 24 (WSDM 2016); March 16 (CHIIR 2016); April 18.

• Office hours:
  – TBD, most likely Tue and/or Thu afternoons.

• Communication: Piazza (experiment):
  https://piazza.com/emory/spring2016/sp16_cs_572_000/home

• Course Website (not yet updated):
  http://www.mathcs.emory.edu/~eugene/cs572/
Logistics: Course structure

• Two parts:
  – Part 1: (roughly through end of February):
    • Fundamentals: indexing, retrieval, ranking, evaluation
  – Part 2: Research topics in IR and Web search:
    • Web, web 2.0, social networks, ... (more later)

• Grading:
  – Two implementation homeworks: 30%
  – Midterm exam: 25%
  – Final project: 45% (proposal, implementation, presentation, report).
Texts


- **SUI**: Search User interfaces, Marti Hearst, CUP, free online: [http://searchuserinterfaces.com/](http://searchuserinterfaces.com/)

- Additional readings will be posted online as needed.
Approximate course outline

- Week 1: Introduction, Boolean retrieval
- Week 2: Foundations (Information Science/Theory): Information seeking/user models; Text representation
- Week 3: Text representation: Tokenization, phrases, entities. Indexing
- Week 4: Indexing (cont’d): data structures; Retrieval algorithms; HW 1 due
- Week 5: Evaluation (metrics, test collections, statistical tests).
- Week 6: Ranking: Boolean and vector-space retrieval models, Probabilistic retrieval models
- Week 7: Ranking: Language modeling, Learning-To-Rank; HW 2 due
- Week 8: Loose ends; Midterm Exam
- Week 9: Semantics, Knowledge bases; Final Project Proposals
- Week 10: Latent semantic indexing (LSA/pLSA, LDA).
- Week 11: Result Presentation/User Interfaces: Snippet generation, summarization
- Week 12: Interactive IR, user feedback
- Week 13: Question answering/spillover
- Week 14: Final project presentations
Sample final projects from previous courses

• Query personalization
• Similar people finder in CQA
• Search satisfaction prediction
• Question Answering
• Course recommendation
Your Turn

• What do you want to get out of this course?