COS 435: Information Retrieval, Discovery, & Delivery

Questions about how we ***find***, ***organize***, ***evaluate*** and ***deliver*** information

## Historic Goals

" to organize the world's information and make it universally accessible and useful"

" 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."

## Historic Goals

"Google's mission is to organize the world's information and make it universally accessible and useful" Google's mission statement, ~ 1998.

"A memex is a device in which an 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." Vannervar Bush, As we may think, *Atlantic Monthly*, July 1945.

## *As We May Think*

- How does Vannevar Bush's 1945 vision match our vision in 2009?
- What of his vision have we achieved?
- What do you think we will eventually achieve?
- Are there any parts of his vision that you think are impossible?

- "This is a much larger matter than merely the extraction of data for the purposes of scientific research; it involves the entire process by which man profits by his inheritance of acquired knowledge"

- "Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified."
- "associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing."

## Concepts

- Data ?
- Information ?
- Content?
- Knowledge ?

One definition

- *Data*: 0's and 1's stored, with or without structure
- *Information*: *Data* with semantic interpretation
- *Content*: all *information* in a document or collection
- *Knowledge*: a functional understanding of *information*

## Data help us?

- Structured data : data base
  Tagged, typed
- Semi-structured data: tagged – XML
  HTML?
- Unstructured:
  – Text
  – Graphics: 2D, 3D
  – Music
  – Video

## What do you want?

- Know it there – Data Bases - data retrieval
- Know it when see it – Information Retrieval
- Surprise me – Data Mining (COS 424)

## Information Retrieval vs Search

discovery of content
+
retrieval of content relevant to query

= search

SEARCH ENGINES

## Delivery of content

- in *digital libraries*, search tool and content repository over one umbrella organization: e.g. Library of Congress
- on *Web*, actual Web pages not provided by search engines (although can get cached copy sometimes)
  – Where Web pages stored affects delivery

## What do you want, Part 2

- information need v.s. query form
  - *User* has information need
  - *Retrieval system* has query form
- Does query capture information need?
- Relevance
  - A *judgment* by user
  - Compare: *no* sense of relevance in data retrieval

## How do you do it?

- Model
  - Contents
  - Query
  - Matching of contents to query - results
- Algorithms
  - Effectiveness
  - Efficiency

## What are performance issues?

- Effectiveness: does search return relevant results ?
- Large amounts data – disks I/O! or not?
- Networking
  - Where is data?
  - Should data be somewhere else?
- Web
  - How find information?
  - How use Web structure?

## Information Delivery

Broadly construed can mean:

- User Interfaces
- Protocols
- Storage Management
- Bandwidth management

Big question: what is model of interaction?

compare handheld wireless, CS Dept machine

## Information Delivery cont.

Focus on latter two:

- Storage management
  - Distributed storage
  - Permanence
- Bandwidth management
  - Caching
  - Prefetching
  - Content distribution networks

## Topics 1

- query models for searching (keyword-based)
- models of documents
- Indexing and inverted files
- Ranking documents
- Using linking structure for Web content analysis
- Semantic and feedback techniques
- User behavior-based relevance criteria; privacy issues
- Manipulating search engine results (SEOs)
- Evaluating retrieval systems

## Topics 2

- Web crawling
- Document similarity
- Clustering
- Non-text media search: e.g. music, images
- adding structure to information: databases, XML, the semantic Web

## Topics 3

- system design of search engines: distributed storage and computing
- Information caching
- Content distribution networks
- Reliability and permanence of information

## Course logistics

- Web site www.cs.princeton.edu/cos435
  - *Schedule and Assignments* has all reading, deadlines, and links to problem sets
- Texts
  - ***Introduction to Information Retrieval***
  - available online
- Test – two, take-home
- Homework, approx. every couple of week
- Presentation – more later
- Project – your choosing with approval