COS 435: Information Retrieval, Discovery, & Delivery

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

Concept of Information in Digital Age

- What is information?
- Where do we find it?
- How do we extract it?

Concept of Information in Digital Age

- What is information?
- How is it different from data?
- How is it different from knowledge ?

Historic Vision

"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." Vannevar Bush, As we may think, *Atlantic Monthly*, July 1945.

Vannevar Bush

- Director of the Office of Scientific Research and Development (1941-1947)

Vannevar Bush,1890-1974

- End of WW2 - what next big challenge for scientists?

Vision

" 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" Vannevar Bush, As we may think, *Atlantic Monthly*, July 1945

Prophetic: Hypertext

✳ "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."

Prophetic: Wikipedia et al

- "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."

Historic Goals

"Google's mission is to organize the world's information and make it universally accessible and useful" Larry Page, Sergey Brin, 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." Vannevar Bush, As we may think, *Atlantic Monthly*, July 1945.

Retrieval

- Collection of "information objects"
  - "information object" is unit of information
    - think "document" or "image"
  - want precise model: representation
- language for asking for info want
  - query
- retrieval system to find relevant info
  - return "info objects" best satisfy query
  - experiment to get right query
  - "Know it when see it" correctness

Unstructured information objects

- Information retrieval usually refers to unstructured objects:
  - Text
  - Graphics: 2D, 3D
  - Music
  - Video
  - any help with semantic interpretation?

Compare

- Structured information: database system
  - tagged, typed
  - well-defined semantic interpretation
  - precise queries
    - database query languages like SQL
  - precise response
    - data matches query or not
- Semi-structured objects: tagged
  - XML, HTML?
  - some help with semantic interpretation

Discovery

- Content discovery
  - constructed collections: *digital libraries*
    - all in one (conceptually) place
    - curated?
  - harvested collections
    - Web crawling
  - databases behind Web pages
    - "deep Web"
  - temporal issues

Discovery

- Information discovery
  - combinations
  - content analysis
    - clustering
  - relationship analysis
    - network analysis
      - metadata

Delivery

- Content delivery
  - search tool and content repository over one umbrella organization
    - e.g. Facebook, Library of Congress
  - *Web* search engines: actual Web pages not provided by search engines
    - freshness issue
    - can get cached copy sometimes
  - where content stored affects delivery
    - Storage Management
    - Bandwidth management

Delivery

- Information delivery - broadly construed:
  - mode of interaction?
    - compare handheld, desktop
  - user interfaces
  - visualization
    - analysis
  - protocols
  - sources
  - other ?

What are efficiency issues?

- Large amounts data
  - build indexes
  - disks I/O! or not?
  - distributed data
- Large volume of queries
  - distributed computing
- Expensive analysis
  - algorithm design

Search Engine

A system that implements information retrieval methods for a collection

- May create the collection
  - discovery of content
- Has a query language and retrieval model
- Has methods for presenting query results

system architecture + algorithms + implementation

Topics

- Information retrieval models for text documents
- **Indexing and inverted files**
- Ranking documents
- Using linking structure for Web content analysis
- User behavior-based relevance criteria
- Evaluating retrieval systems
- **Social networks as sources of meta-info**
- **Social networks as sources of information**
- **Recommender systems**

Topics cont.

- **Privacy issues**
- Web crawling
- system design of search engines: distributed storage and computing
- Document similarity
- Clustering
- Non-text media search
- **Searching dynamic information sources**

Course logistics

- TA: Logan Stafman
- Web site:

COS home page -> academics -> courses -> COS 435

  - General Information
  - Schedule and Assignments
  - Project description
- Communication: using Piazza
  - announcements
  - Q&A
- Text: ***Introduction to Information Retrieval***
  - available online
  - 2 other online texts – see general info

Course Work

- Tests – two, take-home
- Homework, 6
  - first one due next Wed.
- Project – your choosing with approval
  - Pairs or singles