COS 435: Information Retrieval, Discovery, & Delivery

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

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.

Vannevar Bush's 1945 vision

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

Vannevar Bush,1890-1974

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

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."

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

" This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system." Tim Berners-Lee, *Information Management: A Proposal,* first circulated at CERN March 1989

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*

Information from Data?

- Structured data : database system
  - tagged, typed
  - well-defined semantic interpretation
- Semi-structured data: tagged
  - XML, HTML?
  - some help with semantic interpretation
- Unstructured:
  - Text
  - Graphics: 2D, 3D
  - Music
  - Video
  - any help with semantic interpretation?

What information do you want?

- Database queries
  - structured data
  - precise queries
    - Database query languages like SQL
  - precise response
    - data matches query or not
- Information retrieval queries
  - semi-structured or unstructured data
  - experiment to get right query
  - "Know it when see it" correctness
- "Surprise me"
  - data mining

Information retrieval queries

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

How do you retrieve relevant information?

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

Information discovery

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

Content delivery

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

Information Delivery

Broadly construed can mean:
- user interfaces
- protocols
- sources
- other ?

Big question: what is mode of interaction?
compare handheld wireless, desktop

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?

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 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
- Evaluating retrieval systems
- Web crawling

Topics 2

- system design of search engines: distributed storage and computing
- adding structure to information: databases, XML, the semantic Web
- Document similarity
- Clustering
- Non-text media search
- Real-time search

Course logistics

- TA: Siyu Yang
- Web site: www.cs.princeton.edu/cos435
  - ***Schedule and Assignments*** has all reading, deadlines, and links to problem sets
- Text: ***Introduction to Information Retrieval***
  - available online
- Test – two, take-home
- Homework, approx. 6
- Project – your choosing with approval