Search Engines and Information Search

Historic Goals

"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. It is an enlarged intimate supplement to his memory." Vannevar Bush, As we may think, *Atlantic Monthly*, July 1945. (assigned week 2)

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

Vannevar Bush's 1945 vision

- Director of the Office of Scientific Research and Development (1941-1947)
- End of WW2 - what next big challenge for scientists?

Vannevar Bush,1890-1974

"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"

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."

How have we achieved search capability?

- Vannevar Bush envisioned personal index
- General open collections
  - keyword/subject-based search
  - ★ full-text search
  - ★ hypertext enhanced search

Full-text search: beginnings

- Gerald Salton, founding father of information retrieval
  - SMART retrieval system

Gerald Salton (1927-1995)

- Although search has changed, classic techniques still provide foundations – our starting point

Think first about text documents

- Early digital searches – digital card catalog:
  - subject classifications, keywords
- "Full text" : words + English structure
  - No "meta-structure"
- Classic study
  - Gerald Salton SMART project 1960's

8

Information Retrieval

- User wants information from a collection of "objects": information need
- User formulates need as a "query"
  - Language of information retrieval system
- System finds objects that "satisfy" query
- System presents objects to user in "useful form"
- User determines which objects from among those presented are relevant

9

Information Retrieval cont.

- Define each of the words in quotes
  - Information object
  - Query
  - Satisfying objects
  - Useful presentation
- Notion of *relevance* critical
  - What really want?
  - Insufficient structure for exact retrieval
- Develop algorithms for the search and retrieval tasks

10

Modeling: "*satisfying*"

- What determines if document satisfies query?
- That depends ….
  - Document model
  - Query model
  - definition of "satisfying" can still vary
- *START SIMPLE*
  - *better understanding*
  - *Use components of simple model later*

11

AND Model

- Document: *set* of terms
- Query: *set* of terms
- Satisfying:
  - document satisfies query if all terms of query appear in document

Currently used by Web search engines

12

OR Model

- Document: *set* of terms
- Query: *set* of terms
- Satisfying:
  - document satisfies query if one or more terms of query appear in document

Original IR model

why?

13

Scaling

- What are attributes changing from 1960's to online searches of today?
- How do they change problem?

14

Introducing Ranking

- Order documents that satisfy a query by how well match the query
- How capture relevance to user by algorithmic method of ordering?

15

Full-text search: beginnings

Gerald Salton (1927-1995)

- Gerald Salton, founding father of information retrieval
  - SMART retrieval system
- major idea: score documents by frequency of words of query occur
  - take into account
    - document length
    - frequency of words in collection
- one of many major contributions

Frequency Model example

Doc 1: "Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies… Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Frequencies:

science 1; knowledge 2; principles 0; engineering 0

Doc 2: "An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science …" (cos 126 description)

Frequencies:

science 2; knowledge 0; principles 1; engineering 1

Scoring documents (vector model)

	frequency-based		adjusted for word value	
	Doc 1	Doc 2	Doc 1	Doc 2
"science"	1	2	.51	1.02
"engineering"		1		1.6
"principles"		1		1.6
"knowledge"	2		3.2	
Combined SCORE	3	4	3.71	4.22

Using word occurrence and features

•word frequency in documents
• positions of words in documents

plain text

•appearance in special parts of documents
−title
−abstract
−section header
−…

marked-up text

•special features of word
−bold font
−larger font
−…

Review

- Current collections
  - Millions to trillions of documents & petabytes of information
- Queries of a few words
- Create index of collection organized by word
  - Words sorted alphabetically
  - Binary search look-up
- Design ranking method
  - Classic: word frequency
  - "Mark-up (e.g. HTML): attributes of words in doc.
    - word position, appearance, …

Revisit the Index with ranking in mind

- Retrieval systems record all info will use about document in the index
- index organized by word
  - all words in all documents = lexicon
- for each word, index records list of:
  - documents in which it appears SORTED!
    - positions at which it occurs in each doc. SORTED!
      - attributes for each occurrence
- record summary information for documents
- record summary information for words

❖Means index about as big as combination of documents!

Processing a query

- Find index entry for each word in query
- Each index entry gives list of documents containing word, usually sorted by doc. ID
- Scan through lists in parallel looking for documents containing all query words
  - Sorting makes this linear time!
- Process positions of query words and other attributes of query words for documents containing all query words
  - Allows look at how near different query words are to each other in a document

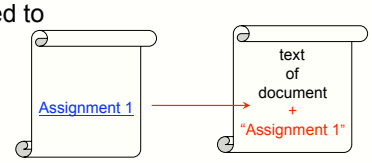
Along came the Web

- Major new element: links
  - hypertext
- Early Web search engines did not use links
  - Excite, WebCrawler, Lycos, Infoseek, AltaVista, Inktomi, Ask Jeeves (now Ask) and more
- How use links?
  - anchor text
  - link analysis

Anchor text

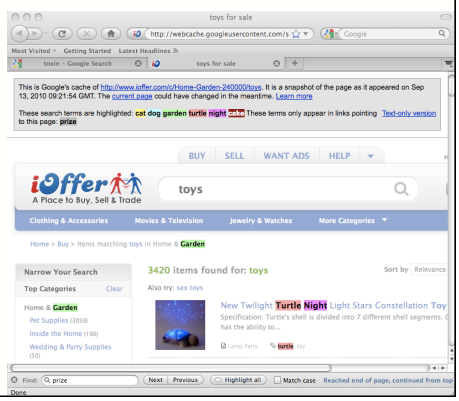
- the words used when making a link to another document:

  "Assignment 1 is now available."
- Add words of anchor text to document pointed to



Example

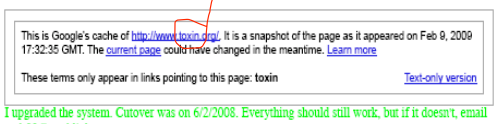
2nd result:



Example

14th result for search on Google: toxin

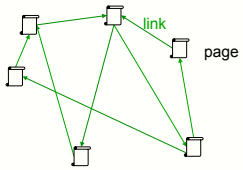
in URL



I upgraded the system. Cutover was on 6/2/2008. Everything should still work, but if it doesn't, email andr00@earthlink.net

Link analysis

- Intuition:
  when Web page points to another Web page, confers status/authority/popularity to that page
- Find a scoring of pages that captures intuition



Graph model

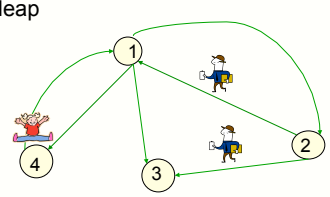
- Capture relationships between things
- Nodes represent things (Web pages)
- Edges between nodes represent that things are related (links between pages)
  - Directed or undirected (directed)
- Many many applications
- Studied mathematically
- Many algorithms for graph problems

More graph examples

- Social network:
  - node = person
  - edge if friends – directed or undirected?
- Rail system:
  - node = station
  - edge if non-stop train service between stations
- Tournament
  - node = contestant / team
  - edge from team A to team B if A beat B - directed

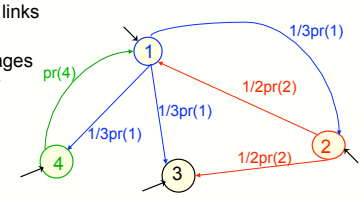
PageRank

- Algorithm that gave Google its "killer" performance
  Larry Page, Sergey Brin, R. Motwani, T. Winograd (1998)
- Random walk model
  + random leap



Define score from concept

- Each page i gets a PageRank score: pgrk(i)
- No query involved
- Equation $pgrk_{updated}(k)$ =
  α*1/n +
  (1-α)*(sum up (pgrk(i) / t_i) for all pages i pointing to page k)
  - t_i is number of links out of page i
  - n is number pages
  - α is parameter

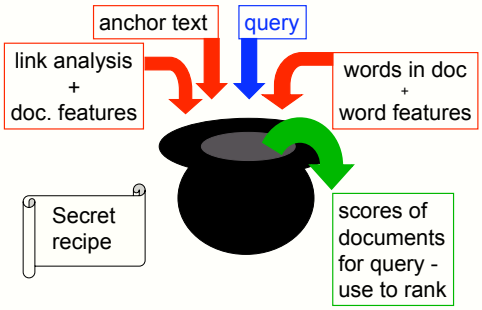


PageRank Features

- PageRank gives value to each Web page independent of queries
  - Done before any queries processed
- PageRank uses human evaluation of pages indirectly
  - Link gives recommendation
- PageRank used as strong indicator that a page should be ranked highly **if it satisfies query**

- Idea of absolute evaluation of pages was novel and highly successful!

Ranking documents w.r.t. query



Example: "peach"

Non-text objects

- images
- music
- video
- ...

Non-text: current methods

- use words around embedded objects
- use names of embedded objects
  - Search Google Images for "only": "Smokey-Bear-Only-You-Posters.jpg"
- use tagging: folksonomy flickr
- use features of object representation
  - CASS Project at Princeton Computer Science: Kai Li, Moses Charikar, Perry Cook, Olga Troyanskaya, Jennifer Rexford
  - One of several university or commercial projects

Improving Web Search?

- More
  - predicting topics without word clues
  - "deep web"
- Better
  - question answering
  - natural language queries
  - more useful presentation

Deep Web

- *Exploring a 'Deep Web' That Google Can't Grasp*, NY Times, Feb 22, 2009
- Much of info on Web behind databases
- Must query database to get info
- How search engine generate right queries on right database
- clues
  - ✳ text on front page & language analysis
  - user behavior
  - link analysis

Improving search results with user behavior

- Aggregate behavior
  - what was most popular selection?
    - Ads
- Personal behavior
  - own history of preferences
    - type of sites?
    - specific sites?
  - disambiguation
- Behavior of users like you
  - compare your behavior to others
  - see what others did in new situation

Summary

- Great search engine needs
  - Extensive collection
  - Good ranking method
    - Good word analysis
    - Good link analysis
  - Fast retrieval time
    - Use many many machines
    - Break up processing of a query
      - How?