Web Search

Introduction

The World Wide Web

- Developed by Tim Berners-Lee in 1990 at CERN to organize research documents available on the Internet.
- Combined idea of documents available by FTP with the idea of *hypertext* to link documents.
- Developed initial HTTP network protocol, URLs, HTML, and first “web server.”

Pre-Web History

- Ted Nelson developed idea of hypertext in 1965.
- Doug Engelbart invented the mouse and built the first implementation of hypertext in the late 1960’s at SRI.
- ARPANET was developed in the early 1970’s.
- The basic technology was in place in the 1970’s; but it took the PC revolution and widespread networking to inspire the web and make it practical.

Web Browser History

- Early browsers were developed in 1992 (Erwise, ViolaWWW).
- In 1993, Marc Andreessen and Eric Bina at UIUC NCSA developed the Mosaic browser and distributed it widely.
- Andreessen joined with James Clark (Stanford Prof. and Silicon Graphics founder) to form Mosaic Communications Inc. in 1994 (which became Netscape to avoid conflict with UIUC).
- Microsoft licensed the original Mosaic from UIUC and used it to build Internet Explorer in 1995.

Search Engine Early History

- By late 1980’s many files were available by anonymous FTP.
- In 1990, Alan Emtage of McGill Univ. developed Archie (short for “archives”)
  - Assembled lists of files available on many FTP servers.
  - Allowed regex search of these file names.
- In 1993, Veronica and Jughead were developed to search names of text files available through Gopher servers.

Web Search History

- In 1993, early web robots (spiders) were built to collect URL’s:
  - Wanderer
  - ALIWEB (Archie-Like Index of the WEB)
  - WWW Worm (indexed URL’s and titles for regex search)
- In 1994, Stanford grad students David Filo and Jerry Yang started manually collecting popular web sites into a topical hierarchy called Yahoo.

Web Search History (cont)

- In early 1994, Brian Pinkerton developed WebCrawler as a class project at U Wash. (eventually became part of Excite and AOL).
- A few months later, Fuzzy Maudlin, a grad student at CMU developed Lycos. First to use a standard IR system as developed for the DARPA Tipster project. First to index a large set of pages.
- In late 1995, DEC developed Altavista. Used a large farm of Alpha machines to quickly process large numbers of queries. Supported boolean operators, phrases, and “reverse pointer” queries.

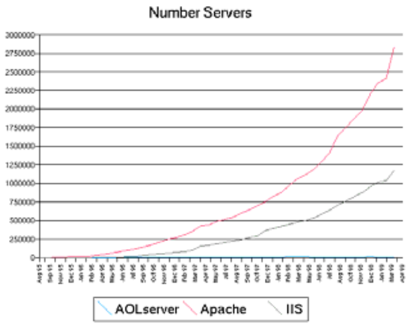
Web Search Recent History

- In 1998, Larry Page and Sergey Brin, Ph.D. students at Stanford, started Google. Main advance is use of *link analysis* to rank results partially based on authority.

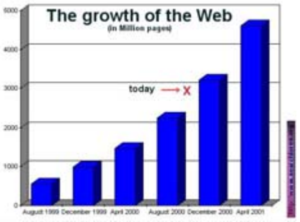
Web Challenges for IR

- Distributed Data: Documents spread over millions of different web servers.
- Volatile Data: Many documents change or disappear rapidly (e.g. dead links).
- Large Volume: Billions of separate documents.
- Unstructured and Redundant Data: No uniform structure, HTML errors, up to 30% (near) duplicate documents.
- Quality of Data: No editorial control, false information, poor quality writing, typos, etc.
- Heterogeneous Data: Multiple media types (images, video, VRML), languages, character sets, etc.

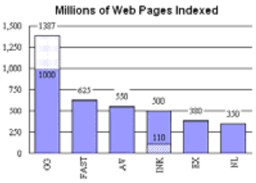
Number of Web Servers



Number of Web Pages



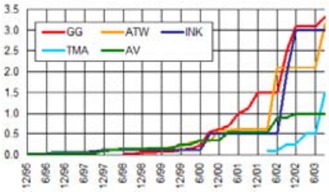
Number of Web Pages Indexed



SearchEngineWatch, Aug. 15, 2001

Assuming about 20KB per page,
1 billion pages is about 20 terabytes of data.

Growth of Web Pages Indexed



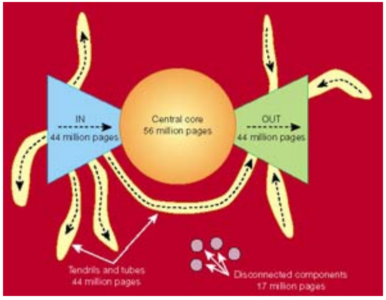
SearchEngineWatch, Jan. 28, 2005

Google lists current number of pages searched.

Some Recent Web Statistics

- As of January 2006, there are an estimated 440 million hosts on the Internet http://www.isc.org/ds/
- As of August 2006, there are an estimated 96 million Web servers on the Internet http://news.netcraft.com/archives/web_server_survey.html
- As of September 2005, yahoo.com 20 billion items (http://www.ysearchblog.com/archives/000172.html), google.com 8.1 billion web pages, search.msn.com 5 billion web pages, alltheweb.com over 3 billion web pages (August 2003)

Graph Structure in the Web



http://www9.org/w9cdrom/160/160.html

Zipf’s Law on the Web

- Number of in-links/out-links to/from a page has a Zipfian distribution.
- Length of web pages has a Zipfian distribution.
- Number of hits to a web page has a Zipfian distribution.

Zipf’s Law

- An empirical rule that describes the relation between the frequencies of appearances.
- Example -- text words: the *i*-th most frequent word appears as many times as the most frequent one divided by i^{θ}, for some $\theta \geq 1$.
- The same can be applied to in-link/out-link of a web page, length of a web page, and number of hits to a web page, among others.

Manual Hierarchical Web Taxonomies

- Yahoo approach of using human editors to assemble a large hierarchically structured directory of web pages.
  - http://www.yahoo.com/
- Open Directory Project is a similar approach based on the distributed labor of volunteer editors (“net-citizens provide the collective brain”). Used by most other search engines. Started by Netscape.
  - http://www.dmoz.org/

Automatic Document Classification

- Manual classification into a given hierarchy is labor intensive, subjective, and error-prone.
- Text categorization methods provide a way to automatically classify documents.
- Best methods based on training a *machine learning* (*pattern recognition*) system on a labeled set of examples (*supervised learning*).
- Text categorization is a topic we will discuss later in the course.

Automatic Document Hierarchies

- Manual hierarchy development is labor intensive, subjective, and error-prone.
- It would be nice to automatically construct a meaningful hierarchical taxonomy from a corpus of documents.
- This is possible with hierarchical text clustering (unsupervised learning).
  - Hierarchical Agglomerative Clustering (HAC)
- Text clustering is a another topic we will discuss later in the course.