CIS 833 – Information Retrieval and Text Mining

Lecture 15

Web Search & Crawling

October 22, 2015

Credits for slides: Allan, Arms, Manning, Lund, Noble, Page.

Example

- Document collection (2 documents)
  - d_1: Xerox reports a profit but revenue is down
  - d_2: Lucent narrows quarter loss but revenue decreases further
- Model: MLE unigram from documents; $\lambda = ½$
- Query: *revenue down*
  - $P(Q|d_1) = ?$
  - $P(Q|d_2) = ?$
- Ranking: ?

$$p(Q \mid d) = \prod_{t \in Q} ((1-\lambda)\, p(t \mid M_c) + \lambda p(t \mid M_d))$$

Example

- Document collection (2 documents)
  - d_1: Xerox reports a profit but revenue is down
  - d_2: Lucent narrows quarter loss but revenue decreases further
- Model: MLE unigram from documents; $\lambda = ½$
- Query: *revenue down*
  - $P(Q|d_1) \sim [(1/8 + 2/16)/2] \times [(1/8 + 1/16)/2]$
    $= 1/8 \times 3/32 = 3/256$
  - $P(Q|d_2) \sim [(1/8 + 2/16)/2] \times [(0 + 1/16)/2]$
    $= 1/8 \times 1/32 = 1/256$
- Ranking: $d_1 > d_2$

Next

- Web Search
  - Textbook Chapter 19 – Web search basics
  - Textbook Chapter 20 – Web crawling
  - Textbook Chapter 21 – Web analysis
  - Monika R. Henzinger, Hyperlink Analysis for the Web. IEEE Internet Computing, vol. 5, no. 1, pp. 45-50, Jan/ Feb., 2001.

Search Engine Early History

- By late 1980's many files were available by anonymous FTP.
- In 1990, Alan Emtage of McGill University 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 URLs:
  - 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)

Keyword-based search engines

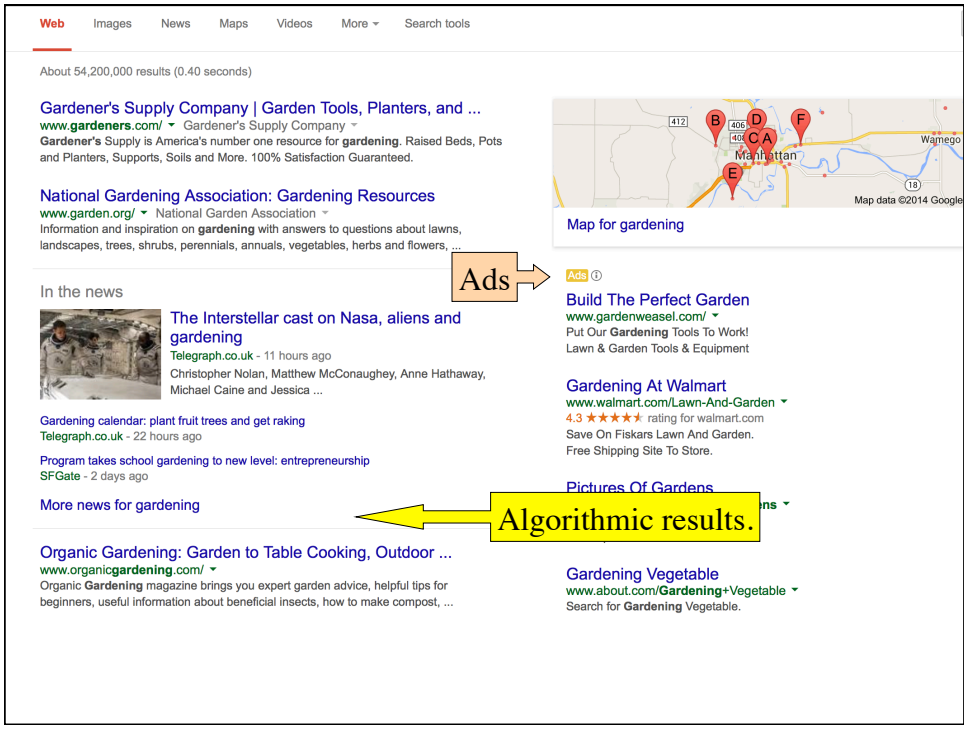
- In early 1994, Brian Pinkerton developed WebCrawler as a class project at University of Washington (eventually became part of Excite and AOL).
- A few months later, Fuzzy Maudlin, a grad student at CMU developed Lycos. First to use a standard IR system as developed for the DARPA Tipster project. First to index a large set of pages.
- In late 1995, Digital Equipment Corporation (DEC) developed Altavista. Used a large farm of Alpha machines to quickly process large numbers of queries. Handled 2 million searches a day.

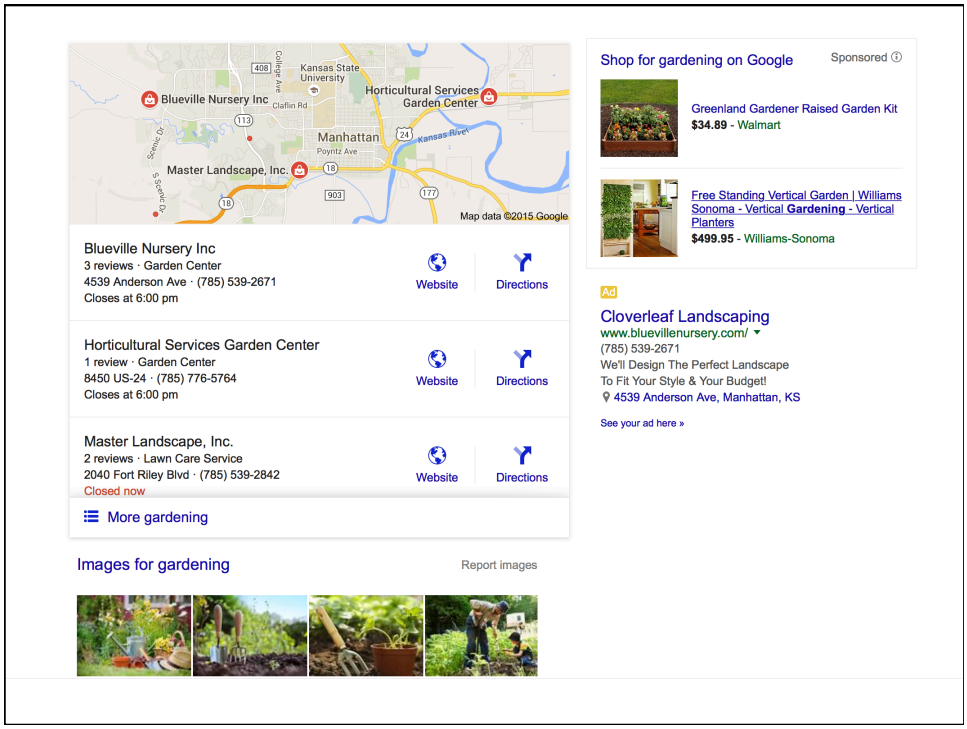
Paid placement ranking: Goto.com (morphed into Overture.com → Yahoo!)

- Your search ranking depended on how much an advertiser paid
- Auction for keywords: ***casino*** was expensive!

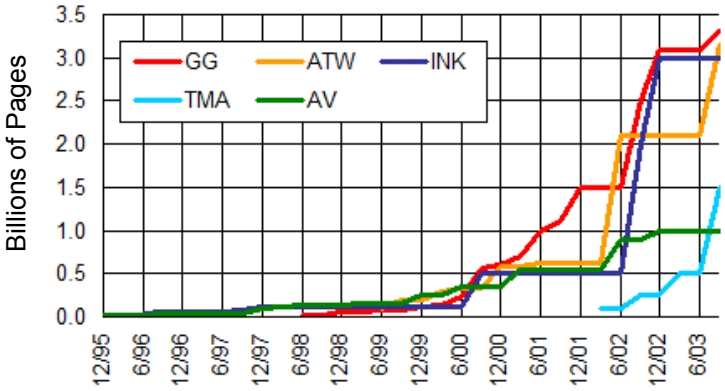
Web Search Recent History

- 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.
- Meanwhile Goto/Overture's annual revenues were nearing $1 billion.
- Result: Google added paid-placement "ads" to the side, independent of search results


Web
Images
News
Maps
Videos
More
Search tools
About 54,200,000 results (0.40 seconds)
Gardener's Supply Company | Garden Tools, Planters, and ...
www.gardeners.com/ Gardener's Supply Company
Gardener's Supply is America's number one resource for gardening. Raised Beds, Pots and Planters, Supports, Soils and More. 100% Satisfaction Guaranteed.
National Gardening Association: Gardening Resources
www.garden.org/ National Garden Association
Information and inspiration on gardening with answers to questions about lawns, landscapes, trees, shrubs, perennials, annuals, vegetables, herbs and flowers, ...
In the news
The Interstellar cast on Nasa, aliens and gardening
Telegraph.co.uk - 11 hours ago
Christopher Nolan, Matthew McConaughey, Anne Hathaway, Michael Caine and Jessica ...
Gardening calendar: plant fruit trees and get raking
Telegraph.co.uk - 22 hours ago
Program takes school gardening to new level: entrepreneurship
SFGate - 2 days ago
More news for gardening
Organic Gardening: Garden to Table Cooking, Outdoor ...
www.organicgardening.com/
Organic Gardening magazine brings you expert garden advice, helpful tips for beginners, useful information about beneficial insects, how to make compost, ...
Ads
Algorithmic results.
Map data ©2014 Google
Map for gardening
Ads
Build The Perfect Garden
www.gardenweasel.com/
Put Our Gardening Tools To Work!
Lawn & Garden Tools & Equipment
Gardening At Walmart
www.walmart.com/Lawn-And-Garden
4.3 rating for walmart.com
Save On Fiskars Lawn And Garden.
Free Shipping Site To Store.
Gardening Vegetable
www.about.com/Gardening+Vegetable
Search for Gardening Vegetable.


Kansas State University
Blueville Nursery Inc
Horticultural Services Garden Center
Manhattan
Master Landscape, Inc.
Map data ©2015 Google
Blueville Nursery Inc
3 reviews · Garden Center
4539 Anderson Ave · (785) 539-2671
Closes at 6:00 pm
Website
Directions
Horticultural Services Garden Center
1 review · Garden Center
8450 US-24 · (785) 776-5764
Closes at 6:00 pm
Website
Directions
Master Landscape, Inc.
2 reviews · Lawn Care Service
2040 Fort Riley Blvd · (785) 539-2842
Closed now
Website
Directions
More gardening
Images for gardening
Report images
Shop for gardening on Google
Sponsored
Greenland Gardener Raised Garden Kit
$34.89 - Walmart
Free Standing Vertical Garden | Williams Sonoma - Vertical Gardening - Vertical Planters
$499.95 - Williams-Sonoma
Ad
Cloverleaf Landscaping
www.bluevillenursery.com/
(785) 539-2671
We'll Design The Perfect Landscape To Fit Your Style & Your Budget!
4539 Anderson Ave, Manhattan, KS
See your ad here »

Early Growth of Web Pages Indexed

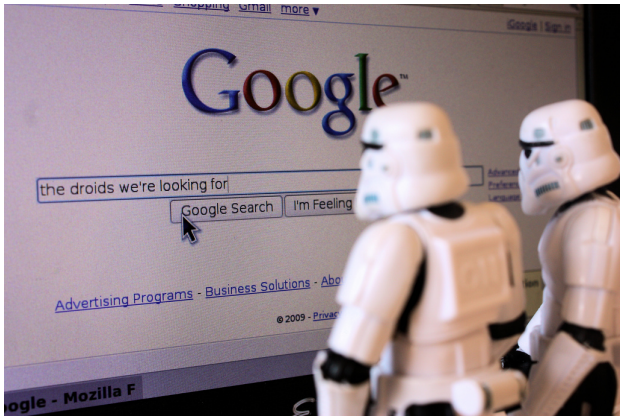


Google
Inktomi
AllTheWeb

Teoma
Altavista

SearchEngineWatch http://blog.searchenginewatch.com/blog/041111-084221

More Recent Numbers



Google's index more than 1 trillion (1,000,000,000,000) pages in 2008, and more than 30 trillion pages in 2013! It handles 100 billion searches a month!
http://news.softpedia.com/news/Google-Explains-How-Search-Works-and-Makes-Sense-of-30-Trillion-Pages-333874.shtml

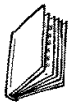
Assuming 20KB per page, 1 billion pages is about 20 terabytes of data.

The Web (Corpus) by the Numbers

1 Kilobyte = a very short story

"Jack and Jill went up the hill to fetch a pail of water. Jack fell down and broke his crown and Jill came tumbling after."

1 Megabyte = a short book



1 Gigabyte = 20 meters of shelved books

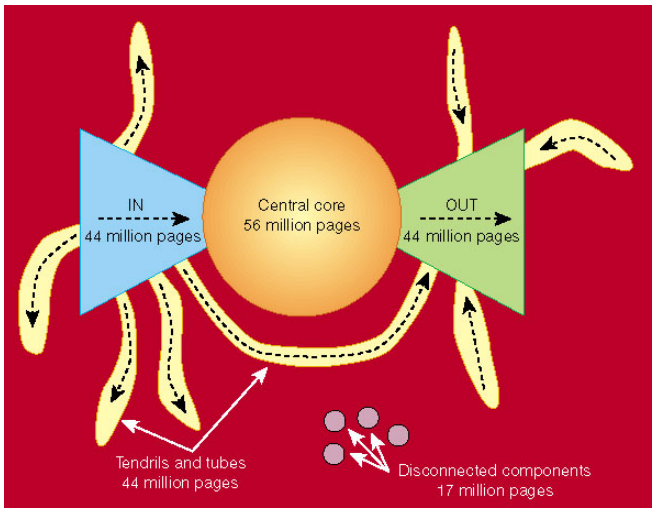
1 Terabyte = an academic research library

20 Terabytes of text on surface Web?

⇩

20 academic research libraries (with some 20,000 meters of shelved books each!)

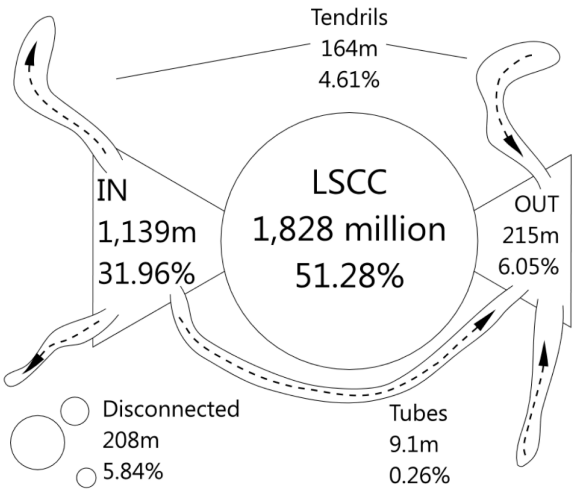
The Web Graph



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

[Broader et al., 2000]

The Web Graph (2014)



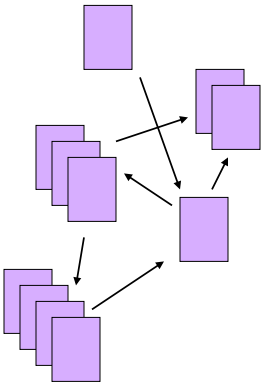
[Meusel et al., 2014]

Without search engines the web wouldn't scale

1. No incentive in creating content unless it can be easily found – other finding methods haven't kept pace (taxonomies, bookmarks, etc.)
2. The web is both a technology artifact and a social environment

   "The Web has become the "new normal" in the American way of life; those who don't go online constitute an ever-shrinking minority." – [Pew Foundation report, January 2005]

The Web Corpus



- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- Scale much larger than previous text corpora … but corporate records are catching up.
- Growth – slowed down from initial "volume doubling every few months" but still expanding
- Content can be *dynamically generated*

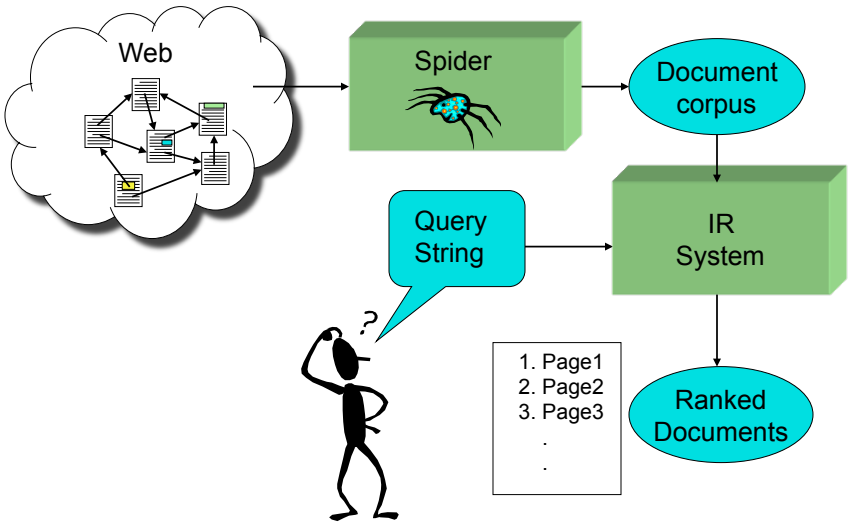
Web Search Users

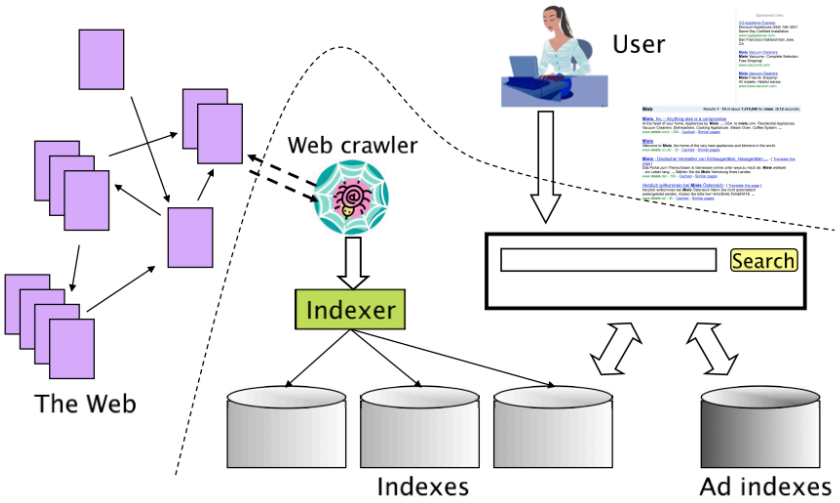
- Make ill defined queries
  - Short
    - AV 2001: 2.54 terms avg, 80% < 3 words
    - AV 1998: 2.35 terms avg, 88% < 3 words
  - Imprecise terms
  - Sub-optimal syntax (most queries without operator)
  - Low effort
- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth
- Specific behavior
  - 85% look over one result screen only (mostly above the fold)
  - 78% of queries are not modified (one query/ session)
  - Follow links – "the scent of information" ...

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 40% (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.

Web Search Using IR





What any spider *must* do

- Be Polite: Respect implicit and explicit politeness considerations for a website
  - Only crawl pages you are allowed to
  - Respect *robots.txt*
- Be Robust: Be immune to spider traps and other malicious behavior from web servers

What any spider *should* do

- Be capable of distributed operation: designed to run on multiple distributed machines
- Be scalable: designed to increase the crawl rate by adding more machines
- Performance/efficiency: permit full use of available processing and network resources

Spiders (Robots/Bots/Crawlers)

- Start with a comprehensive set of root URLs (seeds) from which to start the search.
- Follow all links on these pages recursively to find additional pages.
- Index/Process all **novel** found pages in an inverted index as they are encountered.
- May allow users to directly submit pages to be indexed (and crawled from).