Needles in the Haystack: Google and other brokers in the bits bazaar
Search for "travel"
How many pages?

• Searching for "travel" on Google yields
  "About 846,000,000 results (0.49 seconds)

• In *The New Internet Navigator* by P. Gilster (1995)
  - "I found an example where searching for 'travel'
    yielded a *whopping* 2,633 documents."

• How big is the web?
  - 15 billions indexed pages
    WorldWideWebSize.com
Important Questions

- How can a search engine respond so fast?
- Does it find every relevant link?
- How does it decide which get listed first?
- Do other search engines show the same results?
  - Which is better, or more authoritative?
- Are sponsored links better or worse than "organic" links?
  - Is advertising necessary?
- What should be the role of government for indexing?
"There is no practical obstacle whatever now to the creation of an efficient index to all human knowledge, ideas and achievements, to the creation, that is, of a complete planetary memory for all mankind... The whole human memory can be, and probably in a short time will be, made accessible to every individual... This is no remote dream, no fantasy."

H.G. Wells (1937)
In the beginning...

- Early catalogues of the Internet (even before WWW) were in print
  - *The Whole Internet Catalog* (1994)
- Yahoo! was an index compiled by human editors.
- Search engines were invented in early 1990s
  - Growing popularity of the web
  - More and more content
  - More and more users
Not Just Cows

A guide to resources on the Internet and BITNET in agriculture and related subjects. Compiled by Wilfred (Bill) Drew.

Access via:
ftp ftp.sura.net; login anonymous; cd pub/nic; get agricultural.list

Information:
Email: drewwe@snymorva.bitnet

AGRICULTURE

Newsgroups:
alt.agriculture.fruit, misc., misc.rural

See also:
Forestry; Gardening; and Horticulture.

Advanced Technology Information Network

Any farmer knows that farming isn’t a “mom and pop” business any more; it’s high-tech, and it’s important to keep up with the latest developments. This resource, and the others in this group, will help you stay up-to-date. A fully complete agricultural information service offers market, news, events, weather, job listing, and safety information. Offered by the California Agricultural Technology Institute, so there is a “West Coast” bias to the information. Also contains information on trade, exports, and biotechnology.

Access via:
telnet catcat.uc.csfresno.edu; login super

Commodity Market Reports

Commodity reports compiled by the U.S. Department of Agriculture Market News Service. Twelve hundred reports covering the U.S., updated daily.

Access via:
WAIS agricultural-marketnews.src

Information:
Email: waiss@oes.orst.edu

Not Just Cows

A guide to resources on the Internet and BITNET in agriculture and related subjects. Compiled by Wilfred (Bill) Drew.

Access via:
ftp ftp.sura.net; login anonymous; cd pub/nic; get agricultural.list

Information:
Email: drewwe@snymorva.bitnet

PEN Pages

A complete information server concerning all aspects of rural life. Sections on commodity prices, family farm life, seniors on the farm, news, and nutrition. Also, provides various announcements by the USDA including its CITE-newsletter. Service provided by the Pennsylvania State University, so some information may be specific to that region.

Access via:
telnet ppsen.psu.edu, login your two-letter state abbreviation

U.C. Davis Extension 4-H Project Catalog

Intended to help members of the 4-H youth project get started in areas ranging from beekeeping to “poultry science,” these files are available in PostScript and WordPerfect 5.1 formats.

Access via:
gopher.ucr.sunset.edu /The Campus/ U.C. Cooperative Extension/4h-youth
ftp ftp.ucr.sunset.edu; login anonymous; cd pub/extension/4h-youth

USDA Extension Service Gopher

A “master gopher” for the U.S. Department of Agriculture’s activities and extension service. This includes information about the extension service, policies of the USDA and extension...
Library versus the Information Bazaar

- "Yellow Pages", directories, catalogs
- The web is not hierarchical
  - no structure like a library
- Catalogues are **out**, search engines are **in**
  - Why?
How many hosts are out there?

About how many hosts in the world are there hosting content on the internet?

A. 100,000
B. 1,000,000
C. 10,000,000
D. 100,000,000
E. 1,000,000,000
How many hosts are out there?

Internet Domain Survey Host Count

Source: Internet Systems Consortium (www.isc.org)
Library versus the Information Bazaar

- "Yellow Pages", directories, catalogs
- The web is not hierarchical
  - no structure like a library
- Catalogues are **out**, search engines are **in**
- **BUT**
  - Search engines control what you see
  - Track what you look for
"The search tools that help us find needles in the digital haystack have become the lenses through which we view the digital landscape. Businesses and governments use them to distort our picture of reality."

Blown to Bits (pg. 110)
"For the user, search is the power to find things, and for whoever controls the engine, search is the power to shape what you see."

Blown to Bits (pg. 112)
Web Search: How does it work?

1. Gather information
2. Keep copies
3. Build an index
4. Understand the query
5. Determine the relevance of results to query
6. Determine ranking of relevant results
7. Present the results
Is it really free??? Who pays?

- Users could pay a subscription fee
  - Early AOL, CompuServe
- Web sites could pay to be indexed
- The government could pay (a web tax?)
- Advertisers could pay

- It matters because it affects incentives
Page rank algorithms

• Ranking algorithms are regarded as "crown jewels" for search companies

• Factors include:
  – keywords in headings and title
  – keywords in body text
  – site is "trustworthy" (Is ucsc.edu trustworthy?)
  – links on page are to relevant locations
  – links to this page are relevant
  – age of the page
  – quality of text
Web Search: How does it work?

1. Gather information
2. Keep copies
3. Build an index
4. Understand the query
5. Determine the relevance of results to query
6. Determine ranking of relevant results
7. Present the results
1. Gathering Information: Crawling the Web

- No one controls what is published on the Web
  - totally decentralized
- To find things, search engines *crawl* the Web
- *Crawler* visits pages building an *index* of content
- *Query processor* checks user requests against the index, reports results (this is the part you use)
- Only a fraction of the Web is crawled
  - The dark web...
Crawling

- *Spiders* or *crawlers* wander the Web building indexes
  - Follow links
- Estimated that 0.02% to 3% of information on Web is indexed
- How often does a page get visited?
  - some frequently, others rarely
  - Crawler tracks which pages have frequent updates
- What about loops? Sensitive locations? Passwords?
  - Spiders and crawlers designed to "play nice" with servers
How to Crawl

● How to crawl the web
  − Start with small set of sites entered manually
  − Select a page that has not yet been visited
    ● Retrieve page's HTML
    ● Index keywords in the text
      − Find all keywords in title, body, etc.
      − For each keyword, record this page's URL
    ● Save all link URLs from the page, add to list to be crawled
2. Keep Copies

- Spider downloads page as part of "visit" in order to create the index
- Search engine caches the copy
- Is this legal? What about copyright?
  - Actually, browsing the web requires copying...
After crawling a page like:

http://nathansjslessons.appspot.com/

Crawler associates many terms with the URL
- closure, javascript, variable, interactive, ...

Terms in URL and anchor text are more important
- "Nathan has a fun lesson on closures..."

Copies are cached (why not, might need later)
3. Build an Index

- List of terms, for each term:
  - list of URLs where it appears
- More than just terms
  - terms in bold font might be more important
  - terms in title might be more important
- Must be fast to lookup
- Millions of entries
  - words, names, numbers
- Must fit in computer memory
4. Understand the Query

- Steps 1-3 happen continually "in the background" at search companies.
- User types something: what did they intend?
  - "winguardium leviosum" - Search for pages containing these terms
  - "molly ringwald" – Find movies? Bio? Pictures?
  - "are bedbugs poisonous" – Answer to question
  - "seahawks beat 49ers" – Sport scores? Articles about football?
  - "49ers beat seahawks" – Same thing?
Boolean Queries

• Structured query
  – Use logical operators AND, OR, NOT
  – monet AND water AND lillies
  – vermeer AND girl AND (NOT pearl)
5. Determine Relevance

- Lots of documents that match to some degree
- How do we know which are relevant?
- Simple relevance:
  - count number of times each term appears
  - add them up
- Some problems with simple relevance
  - longer documents get higher scores
  - "Do you like to travel to Paris? I like to travel to Paris and when I travel to Paris I visit many places. Who else likes to travel to Paris?..."
• *Recall* is percentage of relevant documents that are returned by the search

• *Precision* is the percentage of retrieved results that are relevant

• Results with 100% recall (perfect recall) include all relevant documents
  – but might also include documents that aren't relevant

• Results with 100% precision (perfect precision) include only relevant documents
  – but might have missed other relevant documents
You need a caffeine fix so you search for "coffee" in your local area to find a coffee shop. Your town has 10 coffee shops nearby (5 Starbucks, 2 Peets, Coffee Co., Express Coffee, and Java Junction).

Which result has 50% precision and 20% recall?

A. Starbucks, Peets, Coffee Co., Panera Bread, Red Onion
B. 2 Starbucks, 2 Peets, Coffee Co.
C. 2 Starbucks, Panera Bread, Waffle House
D. 5 Starbucks
6. Determine Ranking

- Which of the relevant documents should be displayed first?
- Simple answer: highest relevance first
- What about ties?
- Is relevance the only thing? What about "authority"?
  - NY Times versus random blogger
    (maybe blogger is more of an authority...)

Page rank algorithms

- Ranking algorithms are regarded as "crown jewels" for search companies

- Factors include:
  - keywords in headings and title
  - keywords in body text
  - site is "trustworthy" (Is ucsc.edu trustworthy?)
  - links on page are to relevant locations
  - links to this page are relevant
  - age of the page
  - quality of text
Mona Lisa - Wikipedia, the free encyclopedia
The Mona Lisa (Monna Lisa or La Gioconda in Italian; La Joconde in French) is a half-length portrait of a woman by the Italian artist Leonardo da Vinci, which...

News for mona lisa
DNA tests on bones found in Florence church may help ID 'Mona Lisa' model
Researchers trying to identify the model for Leonardo da Vinci's "Mona Lisa" have started DNA tests on a skeleton found in a Florence church in ...

DNA testing could reveal identity of Leonardo da Vinci's Mona Lisa
This portrait was doubtless painted in Florence between 1503 and 1506. It is thought to be of Lisa Gherardini ...

Mona Lisa - Portrait of Lisa Gherardini, wife of Francesco del ...
This portrait was doubtless painted in Florence between 1503 and 1506. It is thought to be of Lisa Gherardini ...

Mona Lisa - Smarthistory
Smarthistory.khanacademy.org/leonardo-mona-lisa.html
Smarthistory conversation about one of art history's most famous paintings, Leonardo da Vinci's Mona Lisa.

25 Secrets of Mona Lisa Revealed | LiveScience
Oct 18, 2007 - New images uncover 25 secrets about the Mona Lisa, including proof that Leonardo da Vinci gave her eyebrows, solving a long-held mystery.

WebMuseum: Leonardo da Vinci: La Joconde - ibiblio
Portrait of Mona Lisa (1479-1526), also known as La Gioconda, the wife of Francesco del Giocondo, 1503. Oil on poplar, 77 x 53 cm (30 x 20 7/8 in.)
PageRank

- Wikipedia listed at top, then Fox News, then Khan Academy
- Why not the other 29 millions results? Why these first?
- Actual algorithm closely guarded secret at Google
  - URLs get ranked higher for words that occur in URL and in anchor text
  - URL get ranked higher if more pages point to them
    - A links to B is a "vote" by A for B's content
  - URLs get ranked higher if the pages that point to them are ranked higher
The Arms Race

- Google ranking algorithm continually evolves
  - Fighting *search engine optimization*
  - Everyone wants to be ranked first...
    and they are willing to work at making it happen

- "Each year, Google changes its search algorithm around 500-600 times." *MOZ.com*

---

**Penguin 2.1 (#5) — October 4, 2013**
After a 4-1/2 month gap, Google launched another Penguin update. Given the 2.1 designation, this was probably a data update (primarily) and not a major change to the Penguin algorithm. The overall impact seemed to be moderate, although some webmasters reported being hit hard.

- [Google Penguin 2.1 Was A Big Hit (SER)](https://searchengineland.com/google-penguin-21-was-big-hit-168541)
1. Parse the query.
2. Convert words into wordIDs.
3. Seek to the start of the doclist in the short barrel for every word.
4. Scan through the doclists until there is a document that matches all the search terms.
5. Compute the rank of that document for the query.
6. If we are in the short barrels and at the end of any doclist, seek to the start of the doclist in the full barrel for every word and go to step 4.
7. If we are not at the end of any doclist go to step 4.
8. Sort the documents that have matched by rank and return the top k.

Figure 4. Google Query Evaluation
Advanced Google Queries

`link:<URL>` finds pages that link to the URL

![Google Search Results](image)

1. **Script Tags.com: functions - Unobtrusive JavaScript and Dom Scripting**
   www.scripttags.com/tag/functions/
   What's a Closure? A clever hands-on interactive mini-course by Nathan Whiteshead that guides you through bite-sized chunks from variables, to functions, ...

2. **Ned Batchelder: Online Python exercises**
   nedbatchelder.com/blog/201109/online_python_exercises.html
   Sep 9, 2011 - I've been thinking about online Python learners. There have been some cool examples of online code exercises, like Nathan's Javascript ...

3. **Variables and Values - What's a Closure?**
   nathansjslessons.appspot.com/lesson?id=1010
   Variables and Values. Congratulations! You passed your first exercise. You are awesome. To save typing, you can declare variables and give them initial values ...
7. Presenting Results

- Basic form is list of results
- Sponsored links, ads
- Snippets of content
- Related searches, local results, shopping results, ...

http://moz.com/blog/mega-serp-a-visual-guide-to-google
A search engine has two parts
- Crawler/spider to index the data
- Query processor, to answer queries using the index

Query processor must rank results; page rank does this by
- "using data differentially" – not all terms are equivalent; anchors, titles, and URLs count more
- "noting relationships of pages" – a page is important if important pages link to it
Is it really free??? Who pays?

• Users could pay a subscription fee
  – Early AOL, CompuServe
• Web sites could pay to be indexed
• The government could pay (a web tax?)
• Advertisers could pay

• It matters because it affects incentives
• Pay for position in the results
  – FTC ruled not allowed without flagging, deceptive trade practice

• Show ad when search includes certain terms
  – Google AdWords works this way (based on Overture model)
  – Advertiser pays per view or per viewer click
  – Companies bid on popular terms
    • British Airways bids $2.12 per click for "cheap travel to paris"
    • Click through rate (CTR) is 0.4%
    • Monthly search volume is 480
Summary

• Search engines enable unprecedented access to information.
• Search engines concentrate power of what we see into the hands of fewer companies.
• Search engines continue to evolve
  – Google Plus integration
  – Bing Instant Answers
  – Local results
• Next new thing?
  – blippex: measure where you spend your time browsing, rank pages by time spent on them
End