Needles in the Haystack: Google and Other Brokers in the Bits Bazaar

Derived from Blown To Bits Chapter 4 with the same title

“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

Number of pages?

Searching for “travel” in Google today yielded “About 2,910,000,000 results (0.28 seconds)”


At least 40-50 billion pages
(www.worldwidewebsize.com)

The Library and the Bazaar

- “Yellow pages”, directories, and catalogues
- The “Web” is not hierarchical (no structure like a library)
- Catalogues are out - search engines are in.
- But - search engines control what you see

“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
Where to look?

Google is not necessarily the first place to look!
  • Go directly to a Web site -- www.irs.gov
  • Go to your bookmarks -- dictionary.cambridge.org
  • Go to the library -- library.ucsc.edu
  • Go to the place with the information you want -- www.npr.org
  • Browser address window remembers where you’ve been.

Ask, “What site provides this information?”
Different search engines will produce different results.

Keyword Search

Search Engine words are independent
  – Words don’t have to occur together
Use Boolean queries and quotes
  – Logical Operators: AND, OR, NOT
    monet AND water AND lilies
    “van gogh” OR gauguin
    vermeer AND girl AND NOT pearl

Keyword Search

Searching strategies …
  – Limit by top level domains or format … .edu
  – Find terms most specific to topic … ibuprofen
  – Look elsewhere for candidate words, e.g. bio
  – Use exact phrase only if universal, … “Play it again”
  – If too many hits, re-query … let the computer work
  – “Search within results” using “-” … to get rid of junk

Lilly and Zyprexa

• treatment for schizophrenia
• possible serious side effects
• lawyer leaked a confidential internal Lilly memo
• Judge ruled lawyer acted improperly but “the bits had escaped and could not be recaptured,” (Blown to bits pg 116)
“Web sites are primarily fora for speech… the risk of unlimited inhibitions of free speech should be avoided when practicable.”

Judge Weinstein - in refusing to order web sites to remove copies of the leaked memo.

The Fall of Hierarchy

- Early catalogues both on the Internet (even before the WWW) and in print.
- “The Whole Internet Catalog” (1994)
- Yahoo was one compiled by human editors.
- Search engines started to be used in early 90s with the growing popularity of the web.

“The difficulty seems to be, not so much that we publish unduly … but rather that publication has been extended far beyond our present ability to make real use of the record. The summation of human experiences is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships… Our ineptitude in getting at the record is largely caused by the artificiality of systems of 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.”

It Matters How It Works

1. Gather information.
2. Keep copies.
3. Build an index.
4. Understand the query.
5. Determine the relevance of each possible result to the query.
6. Determine the ranking of the relevant results.
7. Present the results.
### 1. Gather Information

- Spiders or web crawlers wander the web building indices
- Estimates range from .02% to 3% of information is indexed
- How often does a page get visited?
  - Some frequently (daily), others rarely (determined by the crawler to not be changing)
- How does the crawler find it’s way and not go in circles?
- Login’s keep bots/crawlers out.

### 2. Keep Copies

- Spider downloads the page as part of the “visit” in order to create the index.
- Search engine may “cache” the copy.
- Is this legal? What about copyright?
- But wait, browsing requires copying as well.

*(AFP) – Sep 15, 2011

NEW YORK — Google and publishers told a US judge Thursday they are close to settling a lawsuit over the Internet giant's controversial book-scanning project…*

### 3. Build an Index

- List of terms and for each term a list of where it appeared
- More than just the terms
  - Terms in bigger font might be more important
  - Terms in the title might be more important
- Must be very fast to lookup
- Could be millions of entries (not just words, but names, special numbers, etc.) requiring Gigabytes of memory
- Must fit in the computers memory (see next slide)

### Accessing RAM vs Disk

How much longer does it take to access a random location in your computer's RAM memory compared to accessing a random location on the disk?

A. 10 times as long
B. 100 times as long
C. 10,000 times as long
D. 100,000 times as long
E. More than 1,000,000 times as long

### Binary Search

- Pick a number between 1 and 1000. How many guesses will I need?
- What about that index with 25 million entries?

- The search happens in memory, but the list of URLs associated with the “term” will likely be on disk.
4. Understand the Query
- Steps 1-3 happen in “the background”
- Not much “understanding” in today’s search engine’s but that could change soon.
- Advanced search engine features help

Cardinal’s beat Rangers vs Ranger’s beat Cardinal’s

What about a business called “THE”?  

5. Determine Relevance
- “Recall” - what percentage of relevant documents are returned by the search?
- Simple relevance calculation -
  - count the number of times each search word appears in the document, add them all up
- Long documents get higher scores.
- Uninteresting words like “the” contribute to the score.
- All word occurrences are not equal (title words should count more).

6. Determine Ranking
- Which of the relevant documents should be displayed first?
- Simple solution - put one with highest relevance score first.
- What if many have the same score?
- Are ones with the highest relevance score really the most important? What about the source of the document (e.g. NY Times vs some random blog post).

Page Rank Algorithms
- The “crown jewels” of search engines lie in their page rank algorithms.
- Factors include:
  - keywords in heading or titles
  - keyword only in the body text
  - site is “trustworthy”
  - links on this page are to relevant pages
  - links to this page are relevant
  - age of the page
  - quality of the text (e.g. absence of misspellings)

Google’s PageRank Algorithm
- If lots of pages point TO this page, this must be a “more important” page
- Tweaking the page rank algorithm can make or break a small business.

7. Presenting Results
- Mostly just a list.
- Maybe there are better forms.
- Those sponsored links…
Who Pays for What?

- Users could pay a subscription fee (early AOL and CompuServe)
- Web sites could pay for being indexed.
- The government could pay (taxes?).
- Advertisers could pay.

Placements, Clicks, and Auctions

- Buy higher position in the ranking - FTC said don’t do it without flagging it as such.
- Banner ads displayed when search included certain terms.
  - pay for view or pay for click throughs?
- Companies bid for popular terms.
- Companies exercise editorial power (censorship?) by refusing certain ads.

Search is Power

“For the user, search is the power to find things, and for whoever controls the search engine, search is the power to shape what you see.”

Blown to Bits pg 112

Did you mean adoption?

- Amazon search for “abortion” gave results that included “Did you mean adoption?”
- This was algorithmic, not editorial.
  Adoption is spelled similarly to abortion.
- They tweaked their “algorithm” to not do this when a pro-choice group complained.

Did you mean adoption?

They tweaked their “algorithm” to not do this when a pro-choice group complained.

A. I agree with this “tweak”, searching for abortion should not ask if I meant adoption.
B. I’m undecided
C. I disagree, this type of “tweak” is inappropriate.

Manipulating Search Results

- White text on white background with words that will raise your rank.
- Google Bombing - “miserable failure” search in 2000 yielded white house biography of George Bush
- Companies that will help you move up in the ranking with changes to your web site.
<table>
<thead>
<tr>
<th>Search Engines Don’t See Everything</th>
<th>Google in China</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Spiders don’t index the contents of many document formats found on the web.</td>
<td>• Google agreed in 2006 to censor its Chinese search results.</td>
</tr>
<tr>
<td>• They don’t reach into databases. (If you type something on a page then the content is probably in a database - e.g. my.ucsc.edu, ecommons.)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A Day in The Life</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.cs.washington.edu/WhyCSE">www.cs.washington.edu/WhyCSE</a></td>
<td>• Search engines offer unprecedented access to information.</td>
</tr>
<tr>
<td></td>
<td>• Search engines place the power to shape what we see into the hands of a few companies.</td>
</tr>
<tr>
<td></td>
<td>• You can count on search engines to continue to evolve.</td>
</tr>
<tr>
<td></td>
<td>• Why not be one of the people to drive those changes?</td>
</tr>
</tbody>
</table>