Google Deconstructed

Marie desJardins
Google CS4HS Teacher Development Workshop
July 19, 2011
Everyone Knows What Google Is…

♦ Right?
♦ So how does it work?
What’s So Great About Google?

♦ Abstraction (of course!)
♦ Crawl/index/search model
♦ PageRank algorithm and complex ranking model
♦ Scalability and robustness
Let’s Try It…

♦ Google demo/activity:

• Find the market capitalization of google
• Find out where Marie went to college & worked before coming to UMBC
• Find a zero-hit “word”
  - Let’s add it to the course website
  - Now does Google notice it?
  - How long do you think it will it take for Google to “learn” it?
• Find a satellite picture of this building
• Show the PageRank of several different pages
PreHistory of the Web: Computation

- First, there were **numbers** (early counting systems)
- Then there were “**computers**” (abaci, Babbage, Difference Engine, Jacquard loom)
- Then there was the **theory of computation** (Turing)
- Then there were **computers** (Eniac, PCs, Macs; explosion of computing power in the latter half of the 20th century)
History of the Web: Medieval Times

♦ Internet (1960s – government sites and a few universities)
♦ Talk/email and Usenet newsgroups used by a limited population in the ‘80s
♦ 1988: First computer worm (created by Robert Morris)
♦ 1989: World Wide Web invented by Tim Berners-Lee
♦ 1992: Netscape browser released
♦ 1993: World Wide Web Worm (not a virus: an early search engine); 300K documents
♦ 1995: eBay goes live
♦ 1996: Larry Page invents PageRank
♦ 1997: Google.com domain registered; 2M-100M documents; 20M queries in AltaVista
History of the Web: The Renaissance

♦ Notable quotes:
  • “It is foreseeable that by the year 2000, a comprehensive index of the Web will contain over a billion documents”;  
  • “It is likely that search engines will handle hundreds of millions of queries per day by the year 2000.” [Brin & Page, “The Anatomy of a Search Engine”]

♦ 1998: Google incorporated with initial investment of $1M
♦ 1999: RSS (automated web feeds) are invented
♦ 1999: blogger.com goes live; blogs explode in popularity
♦ 2000: GoogleAds introduced
♦ 2001: PageRank patented
♦ 2003: MySpace.com created in ten days (following the success of Friendster)
♦ 2004: Facebook launched
♦ 2004: Google IPO, with a market capitalization of $23B
♦ 2004: Mass-market VOIP (Voice Over Internet Protocol)
♦ 2006: Google buys YouTube
History of the Web: The Modern Era

♦ 2007: Google has a 53.6% market share (Yahoo has 20%)
♦ 2008: Facebook overtakes myspace as the #1 social networking site
♦ Today:
  • Google has a market capitalization of $145B, over 20K employees, and a 72% market share
  • “As of October 31, 2007, [Google] is the largest American company… that is not part of the Dow Jones Industrial Average.” [Wikipedia, “Google”]
  • The planet has 6.7B people, of whom 1.6B have Internet access
♦ Tomorrow:
  ? The Semantic Web will enable more meaningful information retrieval
  ? Service-based computing and intelligent agent technology will let you perform all sorts of tasks automatically (travel planning, etc.)
  ? Mobile phones will become the primary Internet platform
  ? Google’s Chrome browser will move most of our computing into the “cloud”
  ? Continuing increased global access to the Web will lead to increased democracy, freedom of information, and protection of human rights

Sources: [investor.google.com, google-system.blogspot.com, internetworldstats.com]
Google: Crawling the Web

♦ Googlebot: Start with a few “well connected” pages, follow the links. Lather, rinse, repeat.
♦ Deepbot: Crawl the entire web once a month
♦ Freshbot: Visit frequently updated sites more often
♦ In 2005, Google claimed to index more than 8B documents.
  • Yahoo claimed to index 19.2B documents, but Google disputed this (and claimed it indexes three times as many documents as its closest competitor, but wouldn’t say how big its index actually was)
♦ As of July 2008, Google had identified 1T (that’s one trillion) unique URLs (but not all of them are interesting enough to index)
♦ Google handles “several hundred million” queries a day (avg in July 2009: 330M) [searchengine.com]
Google Queries and Tools

♦ Demo:

♦ Important concepts: stemming, stop words

♦ Parsing

♦ Term reordering for efficiency

♦ More apps:
  • Google products on “Even More” link
  • Google Maps
  • Google Earth
  • Google Mail, Calendar, Documents, … you name it
  • ...and coming soon, Google Plus!
Ranking the Web

♦ Googlefight: When queries duke it out

♦ Simple keyword-based similarity:
  • tfidf (term frequency – inverse document frequency): # times a word (term) appears in a document / # documents in the collection in which the word appears
  • vector space model: represent each document as a vector (tfidf of each possible word that could be indexed); document (or query-document) similarity is computed by cosine similarity: \( \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} \)

♦ Google uses keyword similarity but includes other criteria in “scoring” documents, most notably PageRank
  • PageRank assigns a “score” to a web page based on how many other pages point to it
  • A page’s PageRank depends on the PageRanks of the “referring” pages, so it is a recursive definition!
  • Other criteria in Google’s ranking scheme include “extra credit” for keywords that appear earlier in the body of the document, in the title tag, in H1 HTML headers, and in anchor text (on links to the page being ranked)

♦ “Search engine optimization” (legitimate) vs. “Google bombing” (not allowed)
Indexing the Web

♦ Googlewhack: Query that returns exactly one page (can you find one?)

♦ Efficient data structures are needed!
  - How do you find a phone number in the yellow pages, or a word in the dictionary?
  - Examples of successively more efficient data structures: Unordered list, ordered list, ordered binary tree, hash table

♦ Indexes: sorted by keyword, with “pointer” to pages containing that keyword

♦ Document server: store the cached versions of the actual pages
How much storage is needed for (say) 10B documents?

Server farms

- As of 2006, Google had an estimated 450,000 (that’s half a million) servers in locations around the globe
- Issues: Energy consumption, heat generation/dissipation, environmental impact (server farms are typically located near, and dissipate their heat into, water sources)
- Google’s June 2007 goal: carbon neutrality by 2008 (via carbon offsets, not actual renewable energy sources… although they are actively interested in solar energy and energy-efficient machines and architectures)

Design philosophy: Use many cheap, redundant machines, plus failure detection and handling → robust, scalable, affordable performance

Cloud computing (21st-century computation paradigm: use a very loosely connected network of many servers to manage enormous quantities of data and huge numbers of queries)
Privacy and Security

◆ Is it good that Google provides ready access to “all” information?
  - Do you want all of your publicly available data to be readily available to any random inquirer?
  - Do you want all of your data to be permanently available via caching, even if you’ve taken it off your facebook page?
  - Do you want “leaked” (legally or not, but against your wishes) private information to become readily, permanently available public data?
  - If you create a story/book/painting/drawing/idea, do you want other people to be able to get it and use it as much as they want, for any purpose, for free?
    - Do you have any music you didn’t pay for on your ipod?
    - Do you have movies you didn’t buy?

◆ Google Earth: security and privacy issues
◆ China/censorship – filtering search results
◆ Digitizing books – copyright issues
◆ Cookies/user data
◆ Tabulating and selling search queries
◆ Copyright and privacy law are way behind the technology curve!
Why is it free?

- Advertising: $10.6B ad revenue in 2006.
- How Google Ads work:
  - AdWords: “pay-per-click” for “sponsored links” using Google’s technology to identify relevant ads
  - AdSense to put sponsored advertising on your own
- Big issue: Click fraud
What Doesn’t Google Do Well?

♦ What doesn’t Google do well?
  • Non-digitized documents (handwritten/scanned data); non-text documents (audio, images, video, multimedia (e.g., PowerPoint presentations)
  • Databases and dynamically generated content
  • Semantic knowledge (beyond keywords and PageRank)