EITF25 Internet - Web Search

Anders Ardö

EIT – Electrical and Information Technology, Lund University

November 28, 2013

Agenda

Outline

Why Web search ...

- Explosion of (digital) information
  within all types of information collections
- Harder and harder to follow information flow
- Faster way to find relevant information when its needed
- Challenges
  - Distributed, dynamic data
  - Large volume
  - Unstructured, heterogeneous data

Size of the Web

- no one knows
- estimates (text pages)
  - 2005 'more than 11.5 billion'
  - 2007 'more than 20 billion'
  - 2010 ' 20 - 55 billion '
- Google claims (2008) to know 10^{12} unique URLs (text, images, ...)

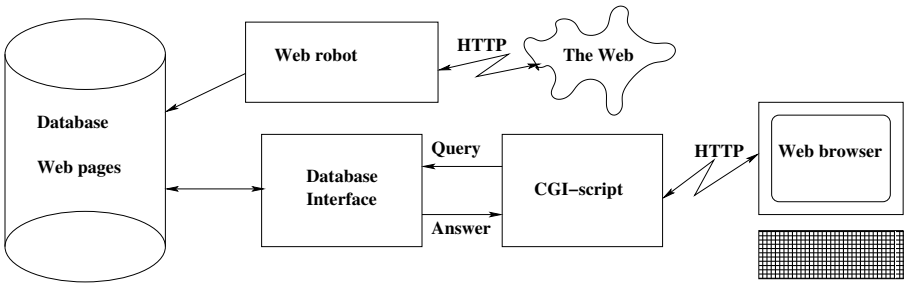
Important questions

Digital Libraries

- How do I find **relevant** information?
- How do I navigate the digital information landscape?
- How structure and organize information to ease knowledge extraction?
- How to create collections, properly organized, with relevant material?
- How to keep collections updated?

Outline

Search Engine - Basic structure



Size efficiency response time

- software crawling the web (much like a human clicking on links)
- collect all found web-pages into a database (IR system)
- offer a web-interface to that database

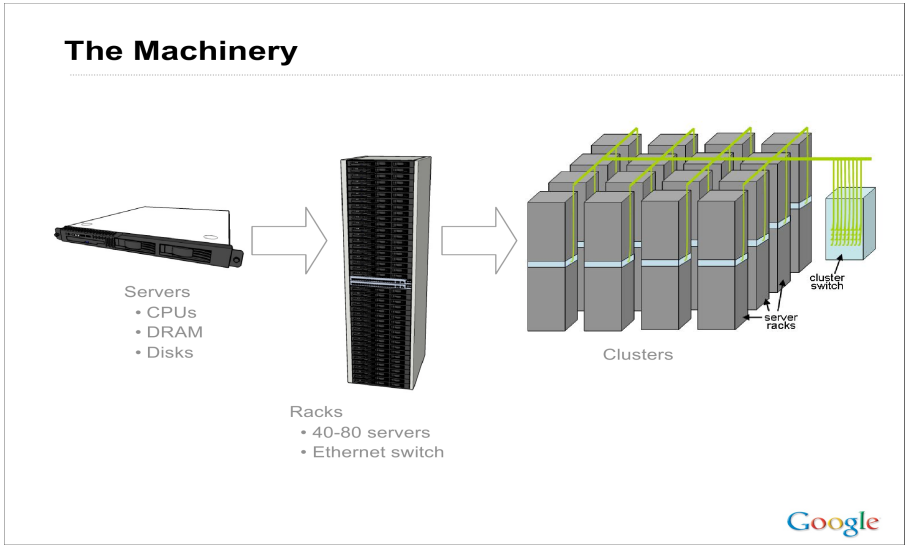
Size of search engines

- not published
- guesses 1 - 20 - 50 billion pages
- overlap between search engines is small $\approx$ 5 - 10 %

Google

- started late 1990:s
- estimated 450,000 low-cost commodity servers (2006)
- estimated 900,000 low-cost commodity servers (2010)
- 1 trillion links to web pages (July 2008)
- “over 8 billion web pages”
- estimate 40 billion pages?
- goal is to index all the world’s data

Google Servers



From Jeff Dean http://www.odbms.org/download/dean-keynote-ladis2009.pdf

Google Servers

Interlude - Google experience - Jeff Dean

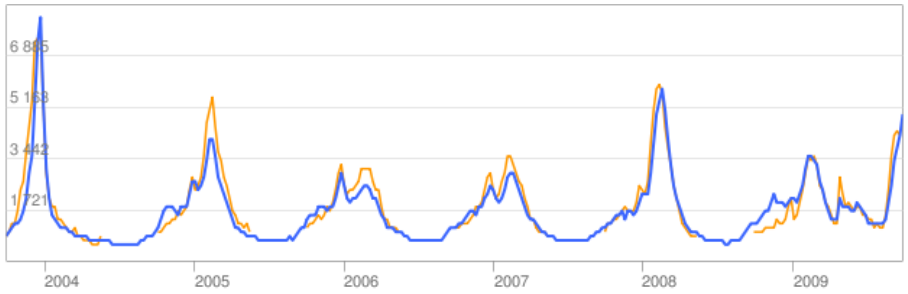
The Joys of Real Hardware

Typical first year for a new cluster:

~0.5 overheating (power down most machines in <5 mins, ~1-2 days to recover)
~1 PDU failure (~500-1000 machines suddenly disappear, ~6 hours to come back)
~1 rack-move (plenty of warning, ~500-1000 machines powered down, ~6 hours)
~1 network rewiring (rolling ~5% of machines down over 2-day span)
~20 rack failures (40-80 machines instantly disappear, 1-6 hours to get back)
~5 racks go wonky (40-80 machines see 50% packetloss)
~8 network maintenances (4 might cause ~30-minute random connectivity losses)
~12 router reloads (takes out DNS and external vips for a couple minutes)
~3 router failures (have to immediately pull traffic for an hour)
~dozens of minor 30-second blips for dns
~1000 individual machine failures
~thousands of hard drive failures
slow disks, bad memory, misconfigured machines, flaky machines, etc.

Long distance links: wild dogs, sharks, dead horses, drunken hunters, etc.

Google Flu Trends



http://www.google.org/flutrends/

Twitter

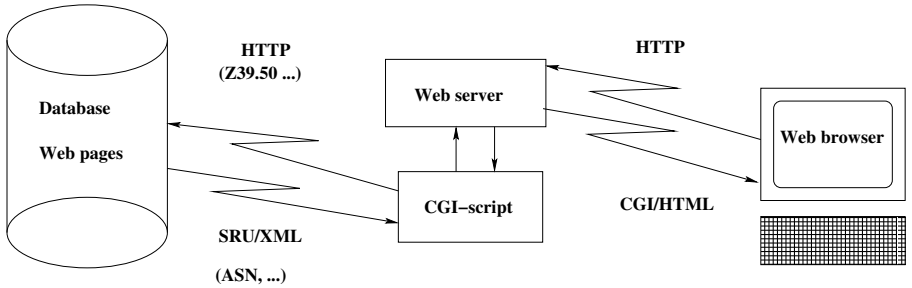
Twitter

- broadcast what's on your mind
- max 140 chars
- 27.3 M tweets per day (November, 2009)
- 250 M tweets per day (October, 2011)
- Twitter moods
- (J. Bollen, H. Mao, X. Zeng: "Twitter mood predicts the stock market" http://arxiv.org/abs/1010.3003)

Search engine examples

Google, Bing, Ask

Search Engine - Application



Overlap between search engines

Compare Google, Yahoo, and Ask Jeeves.
Using 10316 queries and hits from first result page.

Search results

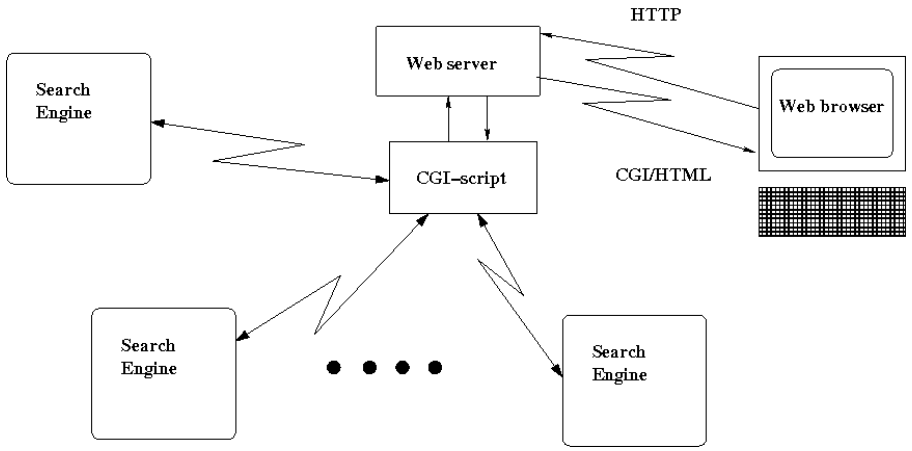
Only in 1	Shared by 2	In all 3
85 %	12 %	3 %

MetaSearch engine Dogpile found 68 % of all results.

Amanda Spink, Bernard J. Jansen, Vinish Kathuria, Sherry Koshman, (2006) "Overlap among major web search engines", Internet Research, Vol. 16 Iss: 4, pp.419 - 426, ISSN: 1066-2243

DOI: 10.1108/10662240610690034

Meta Search Engine - Application



MetaSearch Engine

- it's software that simultaneously search several individual search engines
- collecting, reviewing and ranking their answers
- and give them back in a merged/condensed form to the user
- they are not better than the quality of the search engine databases they obtain results from

MetaSearch engines

- Simultaneously search several individual search engines
- Query translation
- Result merging
  - Simple merge
  - Duplicate detection
  - tf-idf/similarity ranking
  - Position based
- Check that page still exists and is available

MetaSearch Engine examples

Dogpile, Yippy, DuckDuckGo

Special (Vertical) search engines

- prices
  ex: prisjakt, PriceRunner, ...
  http://www.pricerunner.co.uk/
  http://www.prisjakt.nu/
- jobs
  ex: freejobsearch, jobspider, ...
  http://freejobsearch.org/
  http://www.jobspider.com/
- Housing
  ex: rightmove, hemnet, bovision, ...
  http://www.rightmove.co.uk/
  http://www.hemnet.se/
  http://bovision.se/
- ... and so on ...

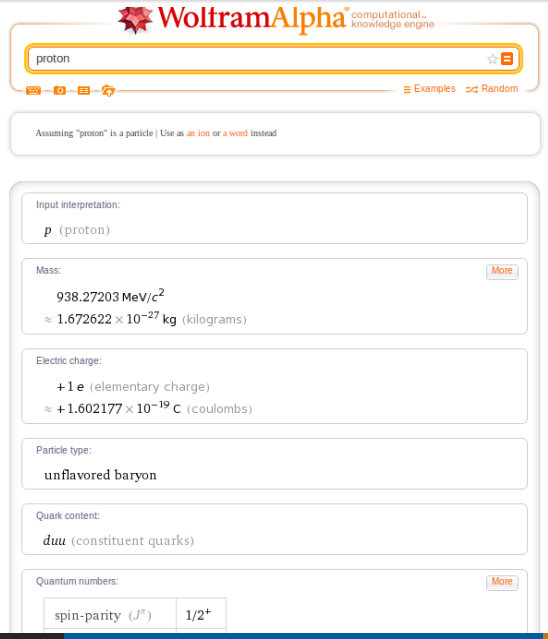
see http://www.thesearchenginelist.com/

Other Search Engines

Wolfram Alpha

"Wolfram|Alpha introduces a fundamentally new way to get knowledge and answers — not by searching the web, but by doing dynamic computations based on a vast collection of built-in data, algorithms, and methods."
Cited from http://www.wolframalpha.com/about.html

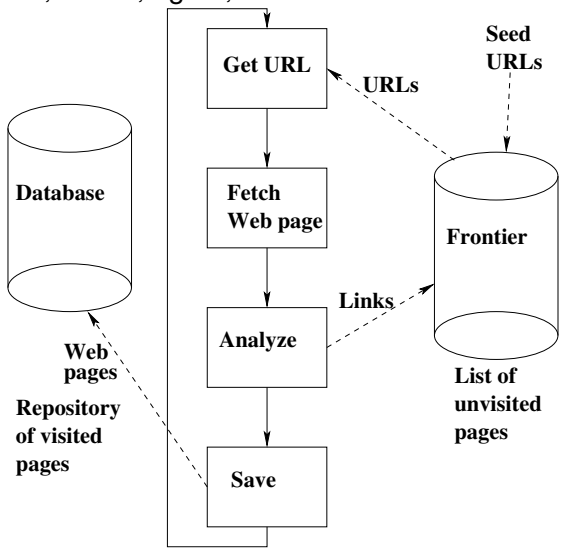
Wolfram Alpha example



Outline

Web Robot - Basic architecture

Spider, Crawler, Robot, agent, ...



Web Robot - Ethics

- Important - **BE NICE**
- Do not overload network or server
- Robot exclusion protocol check for http://www.foobar.com/robots.txt

robots.txt:

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /DATA/
Disallow: /Images/
```

- HTML meta-tag ROBOTS

```
<META NAME="ROBOTS"
CONTENT="NOINDEX,
NOFOLLOW">
```


Web Robot - Problems

- Network failures
- Erroneous URLs
- Unreachable servers
- Password protection
- Spider traps
- Recursive URLs
- Character set encodings
- Same page - different URLs

Web Robot - More Problems

Hidden Web

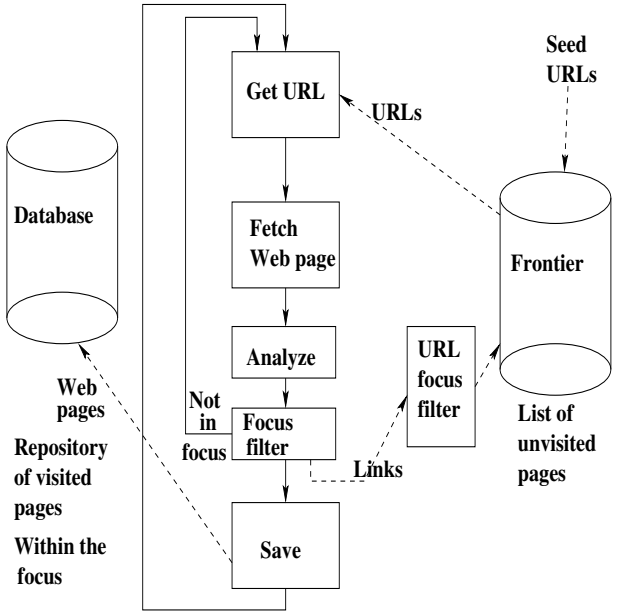
- Databases
- Dynamic scripts
- ... ?

Web Robot - Traversal algorithms

- Depth first (Stack, LIFO queue)
- Breadth first (FIFO queue)
- Best first (How?)
- Relevance order (How?)

Outline

Focused Crawling



Focus:

- Domain
- Project
- Country Region
- Topic
- Subject

Topic-specific Web-crawling

- Problem
  Construct a topic specific search-engine
  (ex. Carnivorous plants)
- Solution
  Make a Web-crawler walk through Internet and collect all pages with topic 'Carnivorous plants'

easier said than done!

Outline

Browsing

- No idea how formulate a query
- Willing to invest some time
- Structure: flat vs hierarchy
  - Manual vs automatic classification
  - Lack of standard classification/terminology
- Precision - NOT recall

Browsing vs search

- Search
  - LOTS of data
  - Unstructured
  - Unrelated items clutter results
- Browsing
  - Small amounts of data
  - Hierarchically structured
  - Quality assessed

Browsing examples

Dmoz (ODP), Yahoo! Directory

Outline

Filter bubble

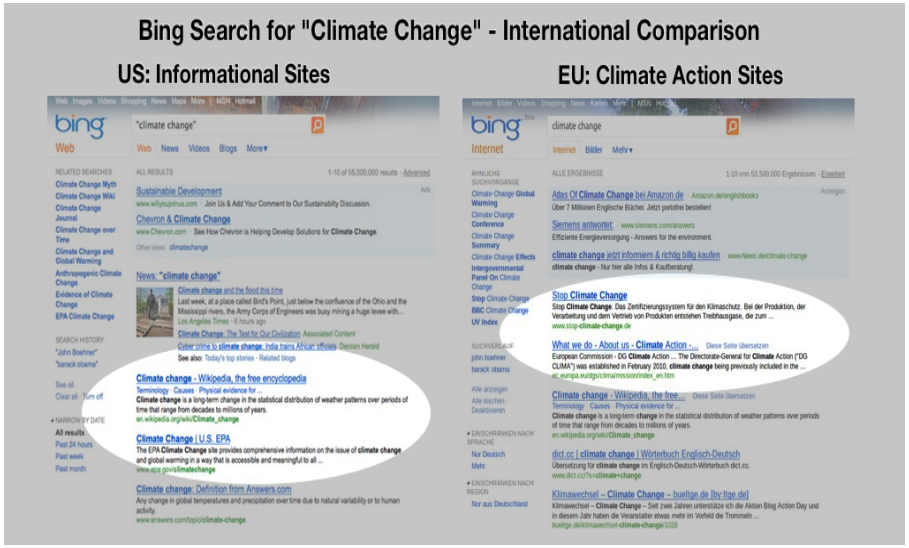
- What do search engines or social sites know about me?
- At least location, search history, click history, likes, and more . . .
- Personalize whats shown (search results, . . .) using this info
- Show us what we want/like to see - algorithmically
- . . . and not whats relevant (who decides that?)

Problem?

Filter bubble example I

From http://www.thefilterbubble.com/what-is-the-internet-hiding-lets-find-out

Filter bubble example II



From http://www.thefilterbubble.com/what-is-the-internet-hiding-lets-find-out

ToS-DR

Terms-of-Service – Didn't Read; http://tos-dr.info/

- you give **Google** (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content.
- **Facebook**: you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook (IP License).

Privacy

- Search history, clicks, photos, documents, comments, . . .
- leads to a profile
- that can be used by ads or sold, or even stolen
- which might lead to it ending up in unwanted places
- and used against you

Beware!

The Web - future

????

Infinity i-Kitchen – intelligent fridge runs Linux
http://www.geek.com/articles/chips/this-intelligent-fridge-runs-linux-on-an-arm-chip-20101126/

Read:
T. Berners-Lee, "*Long Live the Web: A Call for Continued Open Standards and Neutrality*", Scientific American, November 22, 2010.
http://www.scientificamerican.com/article.cfm?id=long-live-the-web

Questions!