

Agenda

EITF25 Internet - Web Information (search, browse, ...)

Anders Ardö

EIT – Electrical and Information Technology, Lund University

December 10, 2014

- 1 Web search
- 2 Web search engines
- 3 Web robots, crawler, focused crawling
- 4 Web search vs Browsing
- 5 Privacy, Filter bubble

A. Ardö, EIT

EITF25 Internet - Web Information (search, browse, ...)

December 10, 2014

1 / 42

A. Ardö, EIT

EITF25 Internet - Web Information (search, browse, ...)

December 10, 2014

2 / 42

Outline

- 1 Web search
- 2 Web search engines
- 3 Web robots, crawler, focused crawling
- 4 Web search vs Browsing
- 5 Privacy, Filter bubble

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

A. Ardö, EIT

EITF25 Internet - Web Information (search, browse, ...)

December 10, 2014

3 / 42

A. Ardö, EIT

EITF25 Internet - Web Information (search, browse, ...)

December 10, 2014

4 / 42

- no one knows
- estimates (text pages)
 - 2005 'more than 11.5 billion'
 - 2007 'more than 20 billion'
 - 2010 ' 20 - 55 billion '
 - 435 billion web pages saved over time (Wayback Machine - <https://archive.org/web/>)
- Google claims (2008) to know 10^{12} unique URLs (text, images, ...)
- Size of the Internet 31st Dec 2013 (<http://www.factshunt.com/>)
 - 14.3 Trillion - Webpages, live on the Internet.
 - 672 Exabytes - 672,000,000,000 GB of accessible data.
 - Over 1 Yotta-byte - Total data stored on the Internet.

1 Yotta-byte = 1,000,000,000,000,000,000,000,000 Bytes!

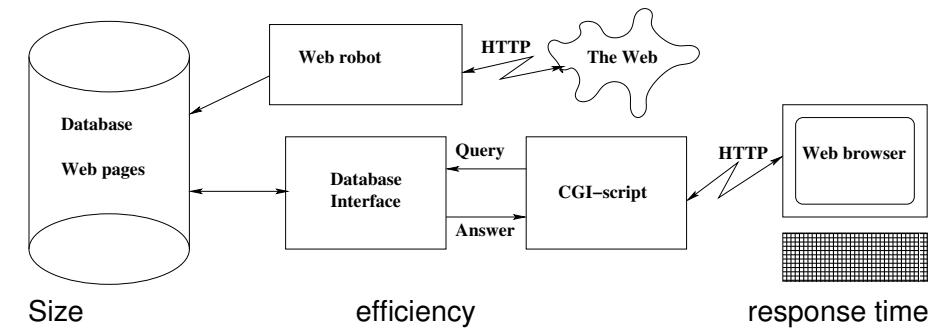
Digital Libraries

- 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

- 1 Web search
- 2 Web search engines
- 3 Web robots, crawler, focused crawling
- 4 Web search vs Browsing
- 5 Privacy, Filter bubble

Search Engine - Basic structure



Size

efficiency

response time

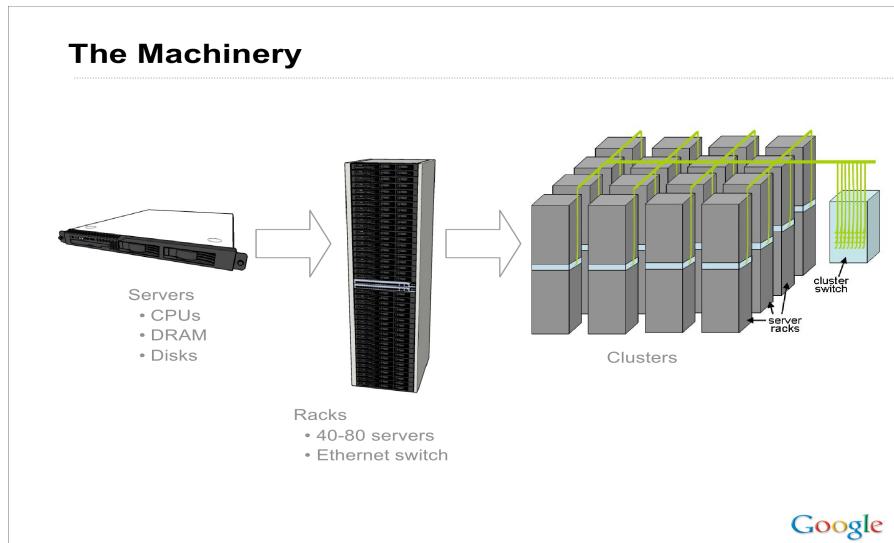
- software crawling the web (much like a human clicking on links)
- collect all found web-pages into a database (IR system)
- offer a web-interface to that database

IR = Information Retrieval: search and rank (sort)

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

- started late 1990:s
- estimated 450,000 low-cost commodity servers (2006)
- estimated 900,000 low-cost commodity servers (2010)
- Over 9,000,000 Servers (2013 - <http://www.factshunt.com/>)
- 1 trillion links to web pages (July 2008)
- “over 8 billion web pages”
- estimate 40 - 50 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

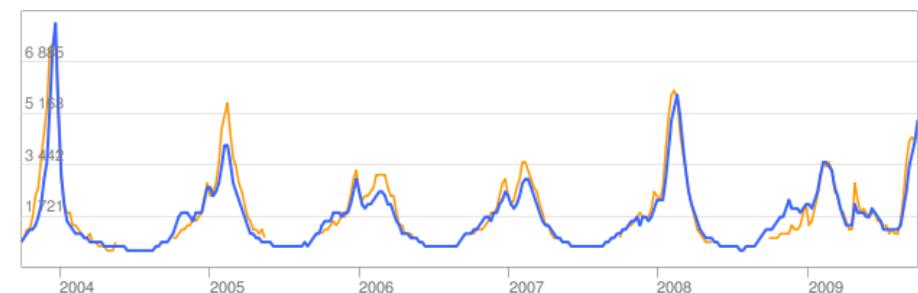


The Joys of Real Hardware

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.**



Twitter

Have an account? [Sign in](#)

new to Twitter?
Easy, free, and instant updates. Get access to the information that interests you most.

[Sign Up >](#)

Top Tweets [View all >](#)

- safety** Worth remembering: the accounts of scantily-clad women claiming to link to their "home videos" are generally linking to malware.
about 1 hour ago
- MCRofficial** Danger Days: The True Lives Of The Fabulous Killjoys is out now! Do it now and do it loud! [#dangerdays](http://wbr.fm/dangerdays)
57 minutes ago
- OldHossRadbourn** Hoss was forced to watch "Dora the Explorer" today. I began to suspect it was fictional when I saw it involved a woman who could read a map.
4 hours ago

See who's here

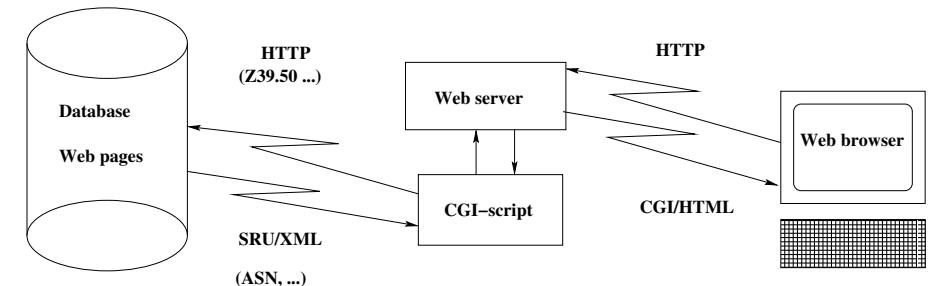
Friends and industry peers you know. Celebrities you watch. Businesses you frequent. Find them all on Twitter.

Top Tweets [View all >](#)

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)
- 500 M tweets per day (2014)
- Twitter moods
- (J. Bollen, H. Mao, X. Zeng: "Twitter mood predicts the stock market" <http://arxiv.org/abs/1010.3003>)

Google, Bing, Ask



Overlap between search engines

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

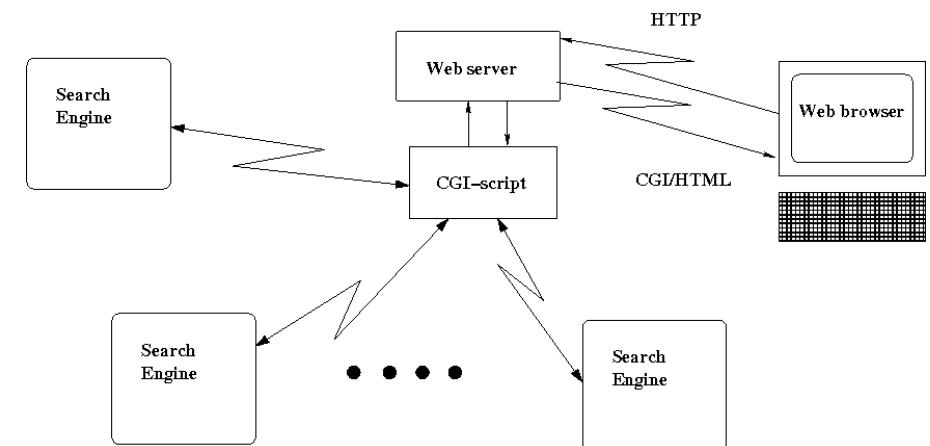
Search results		
Only in 1	Shared by 2	In all 3
85 %	12 %	3 %

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



- 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

- 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/>

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>

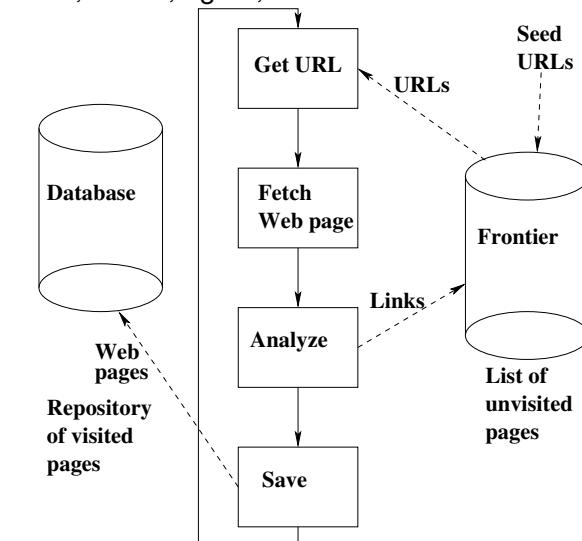
The screenshot shows the Wolfram Alpha search interface. The search bar at the top contains the query "proton". Below the search bar, a note says "Assuming 'proton' is a particle | Use as an ion or a word instead". The main content area is titled "Input interpretation: p (proton)". It provides several pieces of information: "Mass: 938.27203 MeV/c² ≈ 1.672622 × 10⁻²⁷ kg (kilograms)", "Electric charge: +1 e (elementary charge) ≈ +1.602177 × 10⁻¹⁹ C (coulombs)", "Particle type: unflavored baryon", "Quark content: duu (constituent quarks)", and "Quantum numbers: spin-parity (J^P) 1/2⁺". There are also "More" buttons for each section.

Outline

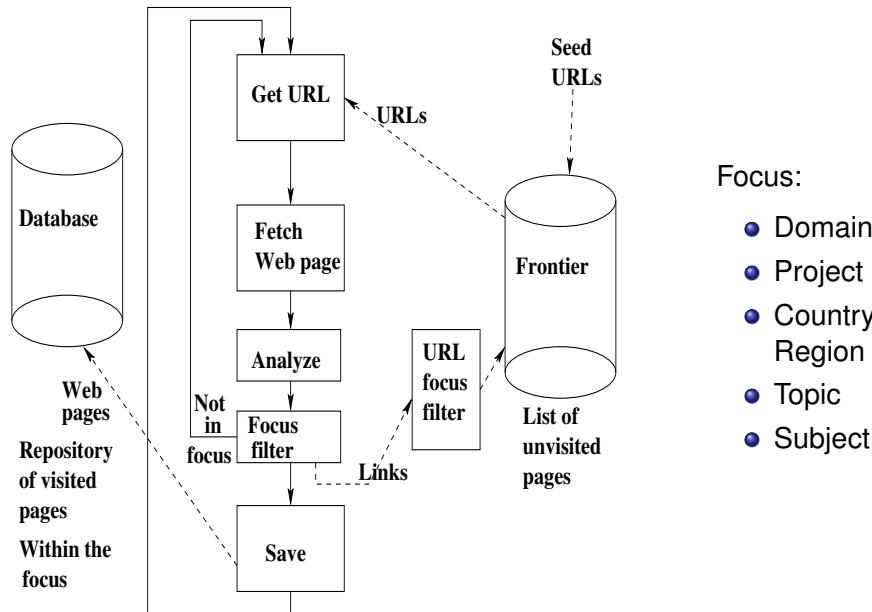
- 1 Web search
- 2 Web search engines
- 3 Web robots, crawler, focused crawling
- 4 Web search vs Browsing
- 5 Privacy, Filter bubble

Web Robot - Basic architecture

Spider, Crawler, Robot, agent, ...



Focused Crawling



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

- 1 Web search
- 2 Web search engines
- 3 Web robots, crawler, focused crawling
- 4 Web search vs Browsing
- 5 Privacy, Filter bubble

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

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

Dmoz (ODP), Yahoo! Directory

Outline

- 1 Web search
- 2 Web search engines
- 3 Web robots, crawler, focused crawling
- 4 Web search vs Browsing
- 5 Privacy, Filter bubble

Filter bubble

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

Problem?

Filter bubble example I

The image shows two side-by-side Google search results for the query "Egypt".
The left search, labeled "Scott: Egyptian Protests", shows results related to the Arab Spring and protests in Egypt. A large oval highlights a link to "Egypt - Wikipedia, the free encyclopedia" and another to "Egypt News - The protests of 2011 - The New York Times".
The right search, labeled "Daniel: Travel Information", shows results related to travel and tourism in Egypt. A large oval highlights a link to "Egypt - Wikipedia, the free encyclopedia" and another to "Egypt Daily News, Egypt News".
Both searches include standard Google filters like "Everything", "Images", "Videos", "News", "Shopping", "Books", and "More".

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

A. Ardö, EIT

EITF25 Internet - Web Information (search, browse, ...)

December 10, 2014

37 / 42

Filter bubble example II

The image shows two side-by-side Bing search results for the query "climate change".
The left search, labeled "US: Informational Sites", shows results from informational websites. A large oval highlights a link to "Climate Change - Wikipedia, the free encyclopedia" and another to "Climate Change - The Tel for Our Civilization".
The right search, labeled "EU: Climate Action Sites", shows results from climate action websites. A large oval highlights a link to "Stop Climate Change - Wikipedia, the free encyclopedia" and another to "Stop Climate Change - Das Zentraleinsystem für den Klimaschutz".
Both searches include standard Bing filters like "Web", "Images", "Videos", "Shopping", "News", "Maps", "Mehr", and "Hotmail".

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

A. Ardö, EIT

EITF25 Internet - Web Information (search, browse, ...)

December 10, 2014

38 / 42

ToS-DR

Terms-of-Service – Didn't Read; <http://tosdr.org/>

- 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).

- 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!

För att inte tala om NSA ...

A. Ardö, EIT

EITF25 Internet - Web Information (search, browse, ...)

December 10, 2014

39 / 42

A. Ardö, EIT

EITF25 Internet - Web Information (search, browse, ...)

December 10, 2014

40 / 42

????

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>

“Well, officer, the coffee pot at home tried to tell my PDA to buy some Colombian beans on the way home, but the car overheard the message and took it as a command to turn for the grocery store right away...”

