Web Search

Introduction

Outline

- History of www and some big players
- What are the challenges?
- Crawlers
- Ranking on the web (not just content)

The World Wide Web

- Developed by Tim Berners-Lee in 1990 at CERN to organize research documents available on the Internet.
- Combined idea of documents available by FTP (File Transfer Protocol, to tranfer files between computers) with the idea of *hypertext* to link documents.
- Developed initial HTTP network protocol, URLs, HTML, and first "web server."

Web Pre-History

- Ted Nelson developed idea of hypertext in 1965.
- Doug Engelbart invented the mouse and built the first implementation of hypertext in the late 1960's at SRI.
- ARPANET was developed in the early 1970's.
- The basic technology was in place in the 1970's; but it took the PC revolution and widespread networking to inspire the web and make it practical.

Web Browser History

- Early browsers were developed in 1992 (Erwise, ViolaWWW).
- In 1993, Marc Andreessen and Eric Bina at UIUC NCSA (University of Illinois) developed the Mosaic browser and distributed it widely.
- Andreessen joined with James Clark (Stanford Prof. and Silicon Graphics founder) to form Mosaic Communications Inc. in 1994 (which became Netscape to avoid conflict with UIUC).
- Microsoft licensed the original Mosaic from UIUC and used it to build Internet Explorer in 1995.

Search Engine Early History

- By late 1980's many files were available by anonymous FTP.
- In 1990, Alan Emtage of McGill Univ. developed Archie (short for "archives")
  - Assembled lists of files available on many FTP servers.
  - Allowed regex (regular expression) search of these file names.
- In 1993, *Veronica* and *Jughead* were developed to search names of text files available through Gopher (network protocol) servers.

Web Search History

- In 1993, early web robots (spiders) were built to collect URL’s:
  - Wanderer
  - ALIWEB (Archie-Like Index of the WEB)
  - WWW Worm (indexed URL’s and titles for regex search)
- In 1994, Stanford grad students David Filo and Jerry Yang started manually collecting popular web sites into a topical hierarchy called Yahoo!.

Web Search History

- In early 1994, Brian Pinkerton developed WebCrawler as a class project at U Washington. (eventually became part of Excite and AOL).
- A few months later, Fuzzy Maudlin, a grad student at CMU developed Lycos. First to use a standard IR system as developed for the DARPA Tipster project. First to index a large set of pages.
- In late 1995, DEC developed Altavista. Used a large farm of Alpha machines to quickly process large numbers of queries. Supported boolean operators, phrases, and "reverse pointer" queries.

Web Search Recent History

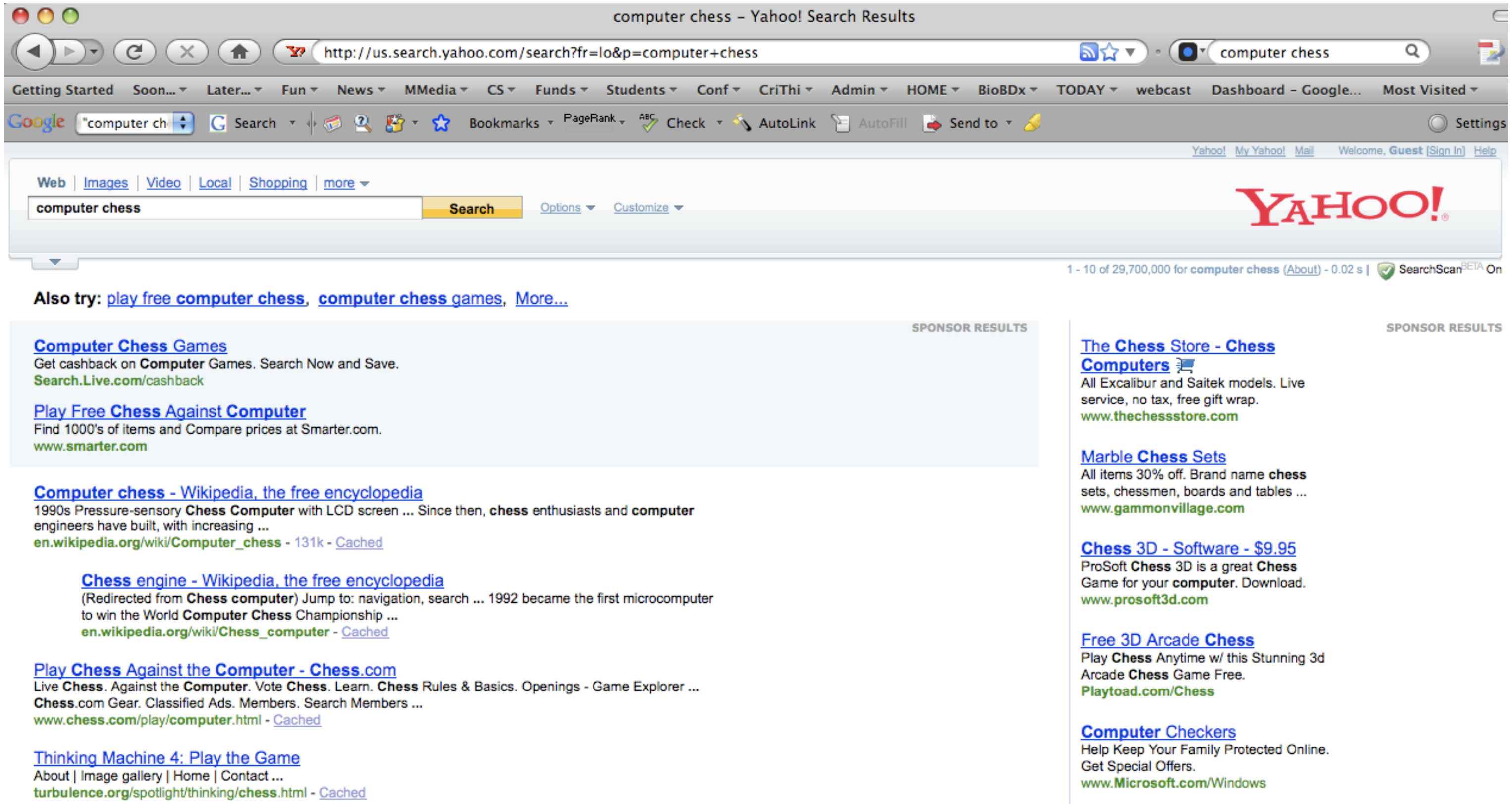
- In 1998, Larry Page and Sergey Brin, Ph.D. students at Stanford, started Google. Main advance is use of *link analysis* to rank results partially based on authority of a web page (roughly for now: the number of incoming hyperlinks).

Search Engine Wars

- The battle for domination of the web search space is heating up!
- Crucial: advertising is combined with search results!
- What if one of the search engines will manage to dominate the space?

Yahoo!

- Started off as a web directory service in 1994, acquired leading search engine technology in 2003.
- Has very strong advertising and e-commerce partners



1 - 10 of 29,700,000 for **computer chess** (About) - 0.02 s |

Also try: play free **computer chess**, **computer chess** games, More...

SPONSOR RESULTS

Computer Chess Games
Get cashback on **Computer** Games. Search Now and Save.
Search.Live.com/cashback

Play Free **Chess** Against **Computer**
Find 1000's of items and Compare prices at Smarter.com.
www.smarter.com

Computer chess - Wikipedia, the free encyclopedia
1990s Pressure-sensory **Chess Computer** with LCD screen ... Since then, **chess** enthusiasts and **computer** engineers have built, with increasing ...
en.wikipedia.org/wiki/**Computer_chess** - 131k - Cached

Chess engine - Wikipedia, the free encyclopedia
(Redirected from **Chess computer**) Jump to: navigation, search ... 1992 became the first microcomputer to win the World **Computer Chess** Championship ...
en.wikipedia.org/wiki/**Chess_computer** - Cached

Play **Chess** Against the **Computer** - **Chess**.com
Live **Chess**. Against the **Computer**. Vote **Chess**. Learn. **Chess** Rules & Basics. Openings - Game Explorer ... **Chess**.com Gear. Classified Ads. Members. Search Members ...
www.**chess**.com/play/**computer**.html - Cached

Thinking Machine 4: Play the Game
About | Image gallery | Home | Contact ...
turbulence.org/spotlight/thinking/**chess**.html - Cached

SPONSOR RESULTS

The **Chess** Store - **Chess Computers**
All Excalibur and Saitek models. Live service, no tax, free gift wrap.
www.the**chess**store.com

Marble **Chess** Sets
All items 30% off. Brand name **chess** sets, chessmen, boards and tables ...
www.gammonvillage.com

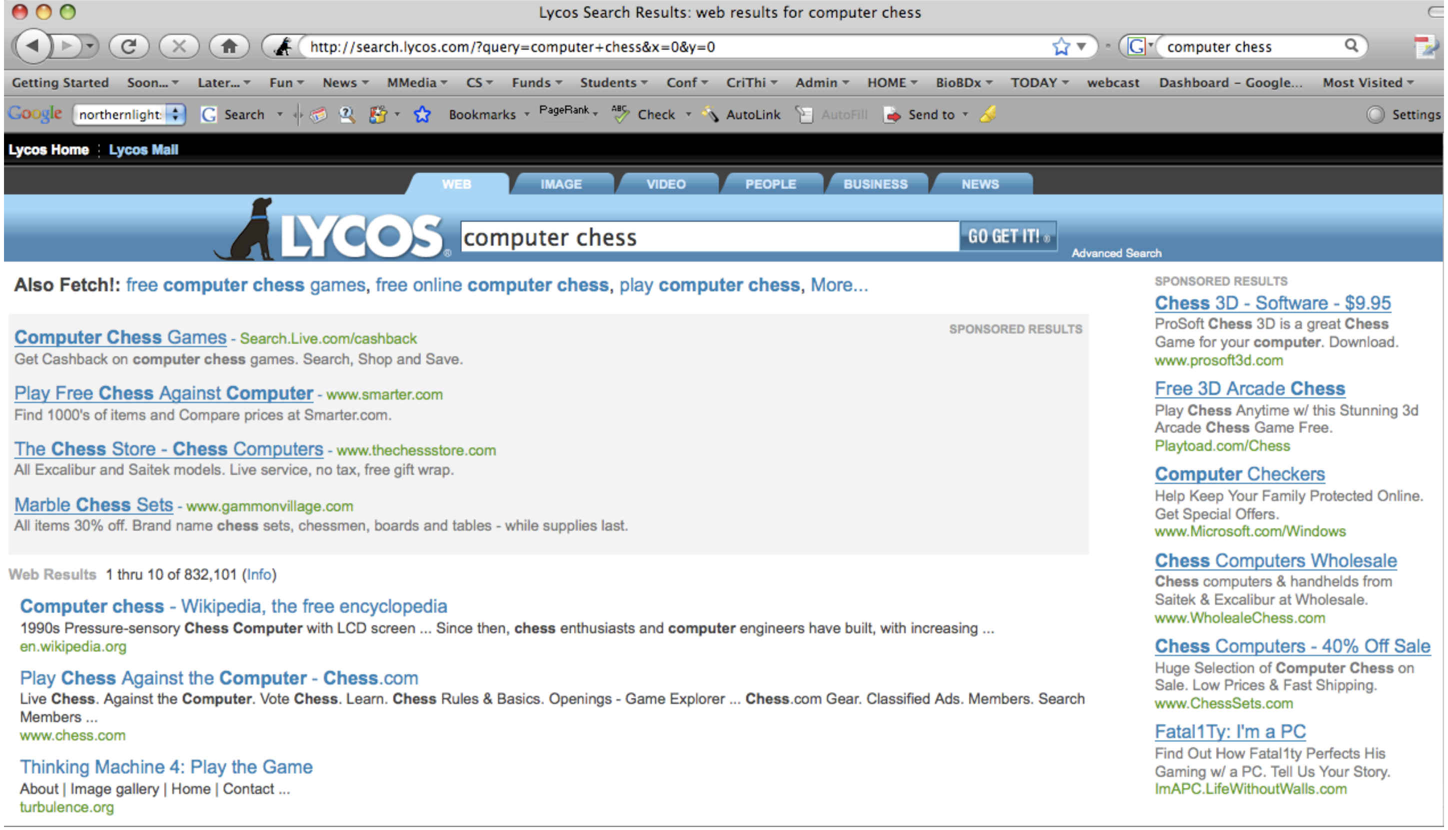
Chess 3D - Software - $9.95
ProSoft **Chess** 3D is a great **Chess** Game for your **computer**. Download.
www.prosoft3d.com

Free 3D Arcade **Chess**
Play **Chess** Anytime w/ this Stunning 3d Arcade **Chess** Game Free.
Playtoad.com/**Chess**

Computer Checkers
Help Keep Your Family Protected Online. Get Special Offers.
www.Microsoft.com/Windows

Lycos!

- One of the pioneers of the field
- Introduced innovations that inspired the creation of Google
- Currently main business are media services (phone, video etc)



Also Fetch!: free computer chess games, free online computer chess, play computer chess, More...

SPONSORED RESULTS

Computer Chess Games - Search.Live.com/cashback
Get Cashback on computer chess games. Search, Shop and Save.

Play Free Chess Against Computer - www.smarter.com
Find 1000's of items and Compare prices at Smarter.com.

The Chess Store - Chess Computers - www.thechessstore.com
All Excalibur and Saitek models. Live service, no tax, free gift wrap.

Marble Chess Sets - www.gammonvillage.com
All items 30% off. Brand name chess sets, chessmen, boards and tables - while supplies last.

Web Results 1 thru 10 of 832,101 (Info)

Computer chess - Wikipedia, the free encyclopedia
1990s Pressure-sensory Chess Computer with LCD screen ... Since then, chess enthusiasts and computer engineers have built, with increasing ...
en.wikipedia.org

Play Chess Against the Computer - Chess.com
Live Chess. Against the Computer. Vote Chess. Learn. Chess Rules & Basics. Openings - Game Explorer ... Chess.com Gear. Classified Ads. Members. Search Members ...
www.chess.com

Thinking Machine 4: Play the Game
About | Image gallery | Home | Contact ...
turbulence.org

SPONSORED RESULTS

Chess 3D - Software - $9.95
ProSoft Chess 3D is a great Chess Game for your computer. Download.
www.prosoft3d.com

Free 3D Arcade Chess
Play Chess Anytime w/ this Stunning 3d Arcade Chess Game Free.
Playtoad.com/Chess

Computer Checkers
Help Keep Your Family Protected Online. Get Special Offers.
www.Microsoft.com/Windows

Chess Computers Wholesale
Chess computers & handhelds from Saitek & Excalibur at Wholesale.
www.WholealeChess.com

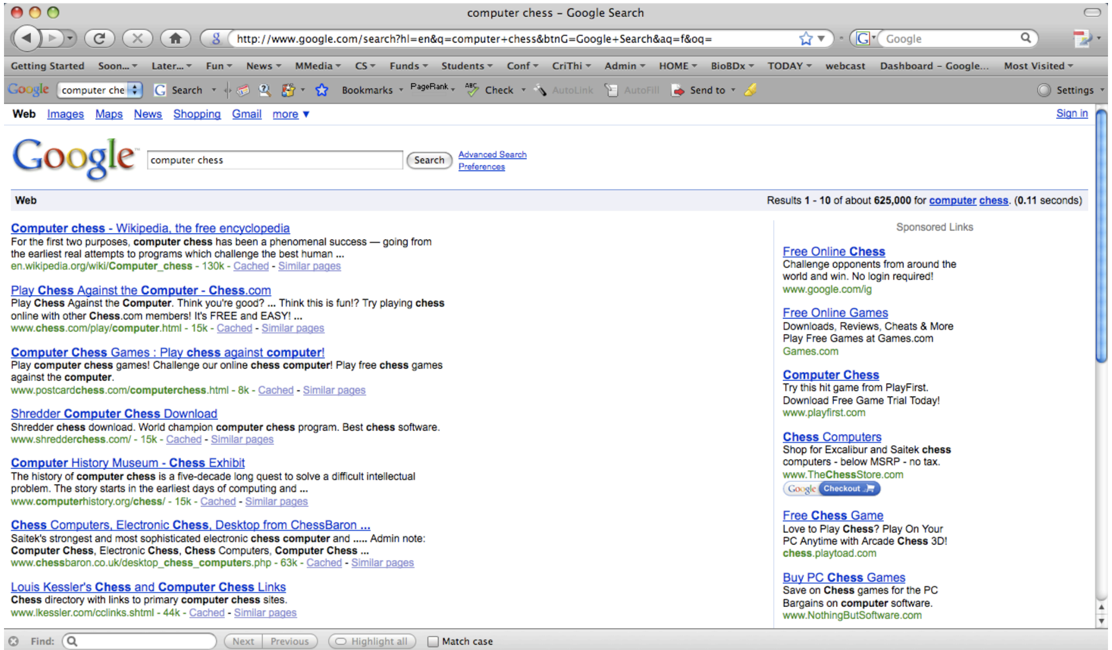
Chess Computers - 40% Off Sale
Huge Selection of Computer Chess on Sale. Low Prices & Fast Shipping.
www.ChessSets.com

Fatal1Ty: I'm a PC
Find Out How Fatal1ty Perfects His Gaming w/ a PC. Tell Us Your Story.
ImAPC.LifeWithoutWalls.com

- Verb “google” has become synonymous with searching for information on the web.

- Has raised the bar on search quality
- Has been the most popular search engine in the last few years.
- Is the most innovative and dynamic.

Google



computer chess

Web

Results **1** - **10** of about **625,000** for computer chess. (**0.11** seconds)

Computer chess - Wikipedia, the free encyclopedia
For the first two purposes, **computer chess** has been a phenomenal success — going from the earliest real attempts to programs which challenge the best human ...
en.wikipedia.org/wiki/**Computer_chess** - 130k - Cached - Similar pages

Play **Chess** Against the **Computer** - **Chess**.com
Play **Chess** Against the **Computer**. Think you're good? ... Think this is fun!? Try playing **chess** online with other **Chess**.com members! It's FREE and EASY! ...
www.**chess**.com/play/**computer**.html - 15k - Cached - Similar pages

Computer Chess Games : Play **chess** against **computer!**
Play **computer chess** games! Challenge our online **chess computer**! Play free **chess** games against the **computer**.
www.postcard**chess**.com/**computerchess**.html - 8k - Cached - Similar pages

Shredder **Computer Chess** Download
Shredder **chess** download. World champion **computer chess** program. Best **chess** software.
www.shredder**chess**.com/ - 15k - Cached - Similar pages

Computer History Museum - **Chess** Exhibit
The history of **computer chess** is a five-decade long quest to solve a difficult intellectual problem. The story starts in the earliest days of computing and ...
www.**computer**history.org/**chess**/ - 15k - Cached - Similar pages

Chess Computers, Electronic **Chess**, Desktop from ChessBaron ...
Saitek's strongest and most sophisticated electronic **chess computer** and ..... Admin note: **Computer Chess**, Electronic **Chess**, **Chess** Computers, **Computer Chess** ...
www.**chess**baron.co.uk/desktop_**chess**_**computer**s.php - 63k - Cached - Similar pages

Louis Kessler's **Chess** and **Computer Chess** Links
Chess directory with links to primary **computer chess** sites.
www.lkessler.com/cclinks.shtml - 44k - Cached - Similar pages

Sponsored Links

Free Online **Chess**
Challenge opponents from around the world and win. No login required!
www.google.com/ig

Free Online Games
Downloads, Reviews, Cheats & More
Play Free Games at Games.com
Games.com

Computer Chess
Try this hit game from PlayFirst.
Download Free Game Trial Today!
www.playfirst.com

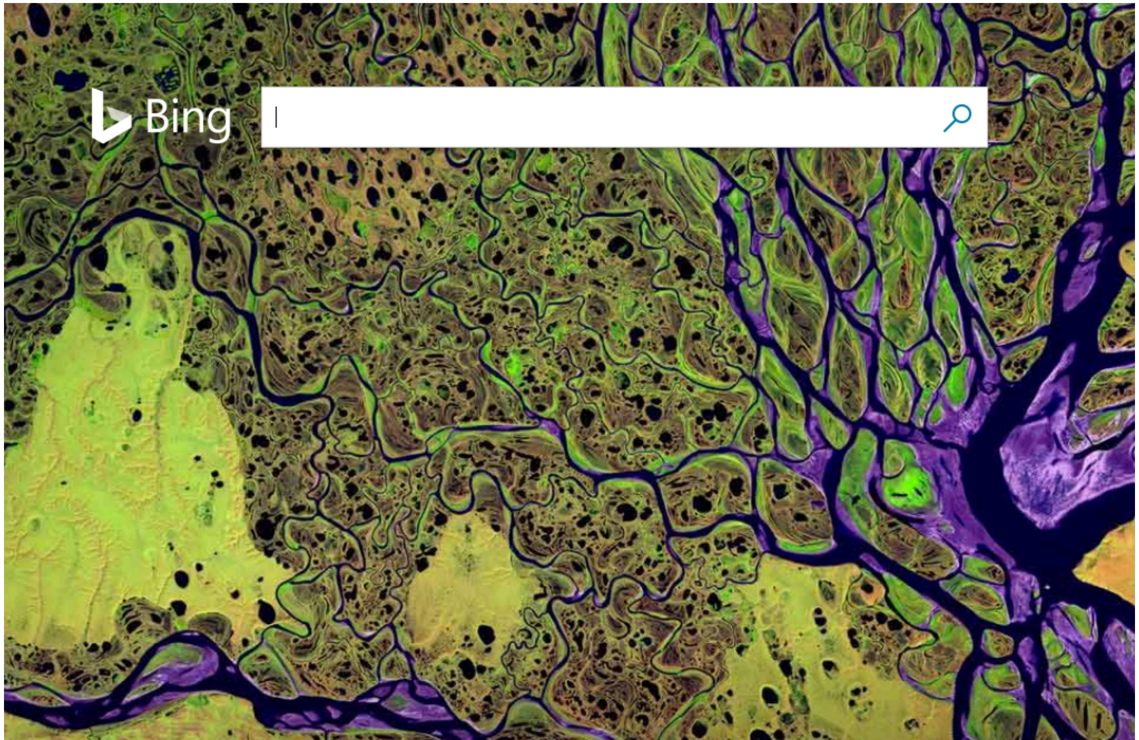
Chess Computers
Shop for Excalibur and Saitek **chess** computers - below MSRP - no tax.
www.The**Chess**Store.com

Free **Chess** Game
Love to Play **Chess**? Play On Your PC Anytime with Arcade **Chess** 3D!
chess.playtoad.com

Buy PC **Chess** Games
Save on **Chess** games for the PC
Bargains on **computer** software.
www.NothingButSoftware.com

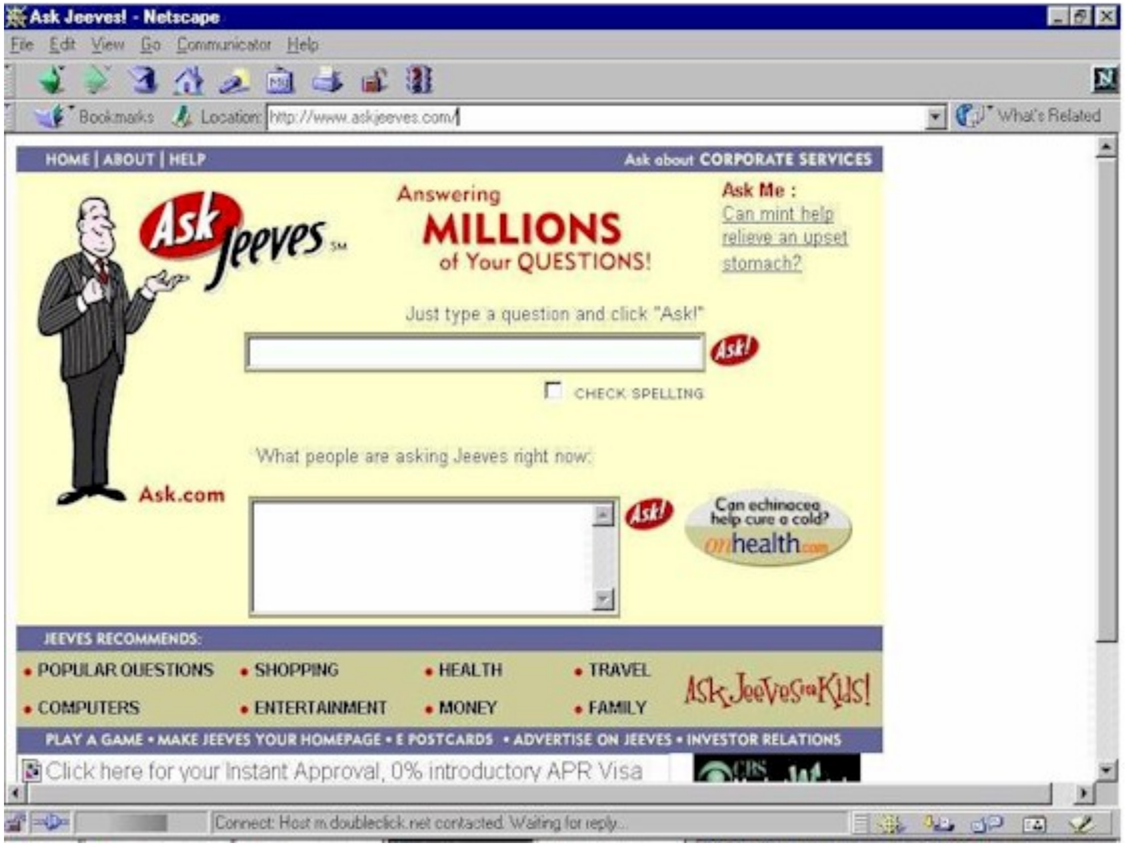
BING (was: MSN Search, Live Search)

- Bing is the second largest search engine (about 20% in US)
- Owned by Microsoft
- Main features media and imaging



Ask (Jeeves)

- Specialises in natural language question answering.



Answering MILLIONS of Your QUESTIONS!

Ask Me : Can mint help relieve an upset stomach?

Just type a question and click "Ask!"

□ CHECK SPELLING

What people are asking Jeeves right now:

Can echinacea help cure a cold? onhealth.com

Click here for your Instant Approval, 0% introductory APR Visa

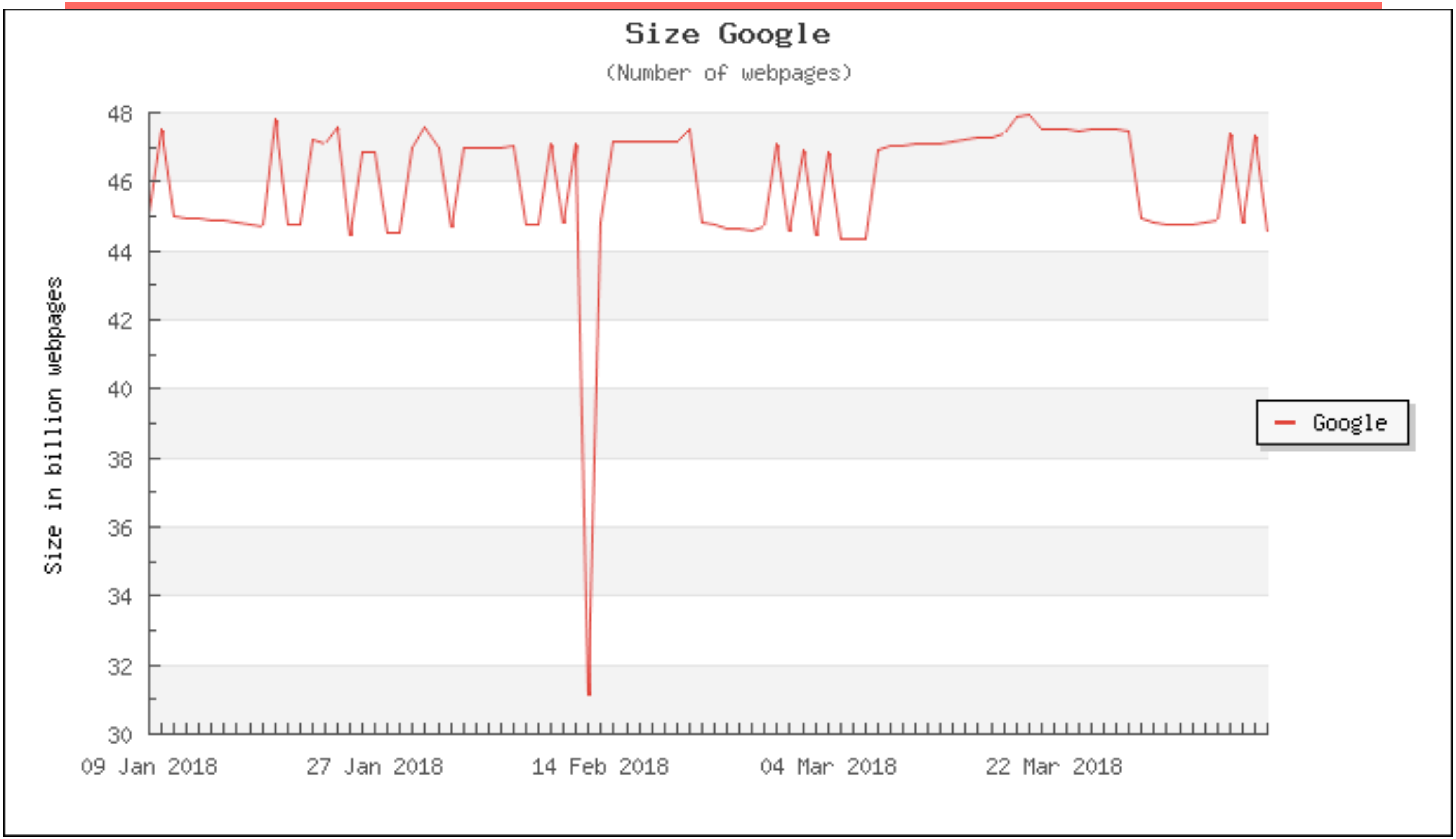
Connect: Host m.doubleclick.net contacted. Waiting for reply...

Web Challenges for IR

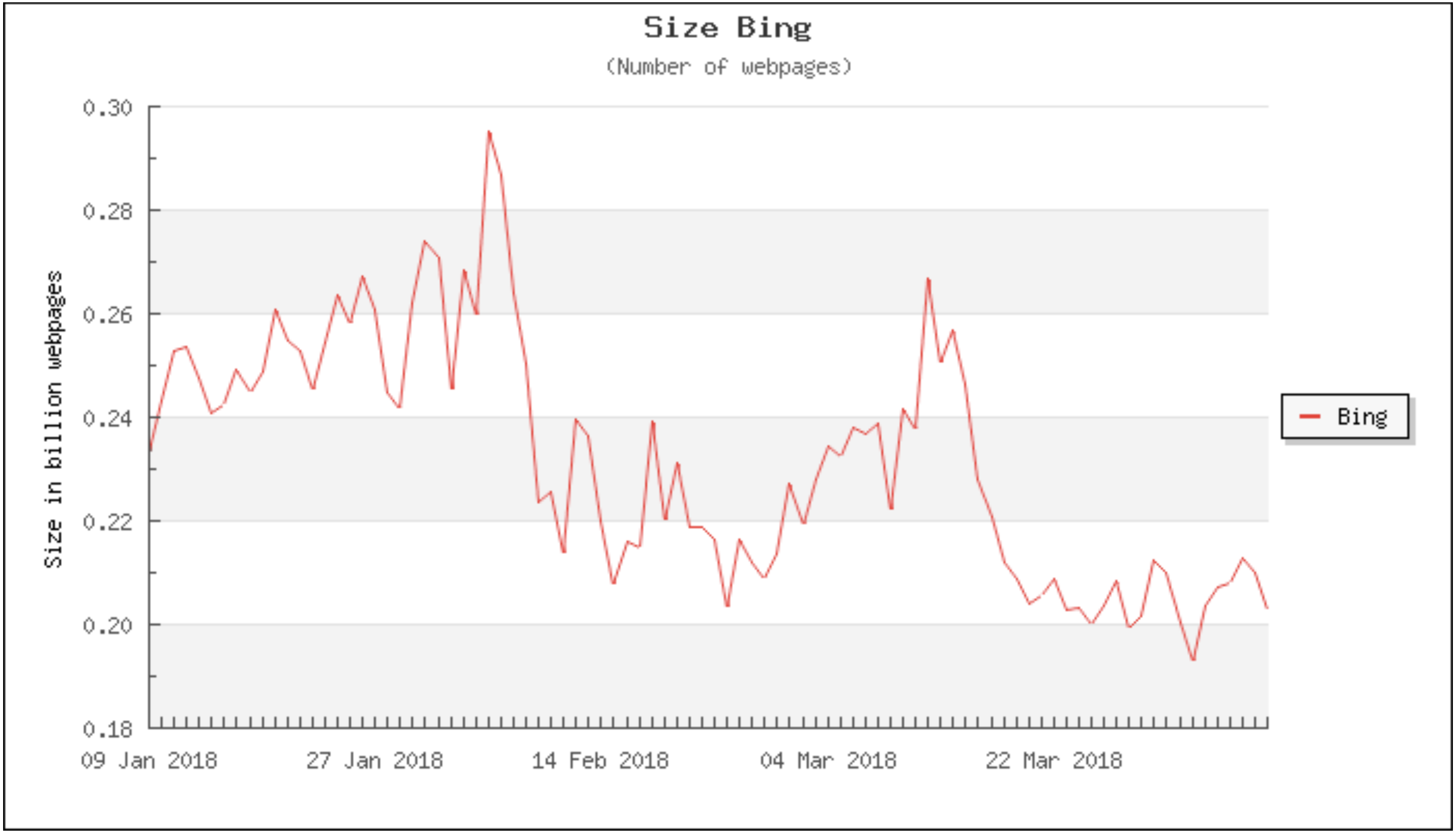
- Distributed Data: Documents spread over millions of different web servers.
- Volatile Data: Many documents change or disappear rapidly (e.g. dead links).
- Large Volume: Billions of separate documents.
- Unstructured and Redundant Data: No uniform structure, HTML errors, up to 30% (near) duplicate documents.
- Quality of Data: No editorial control, false information, poor quality writing, typos, etc.
- Heterogeneous Data: Multiple media types (images, video, VRML), languages, character sets, etc.

BIG, HOW MUCH BIG?

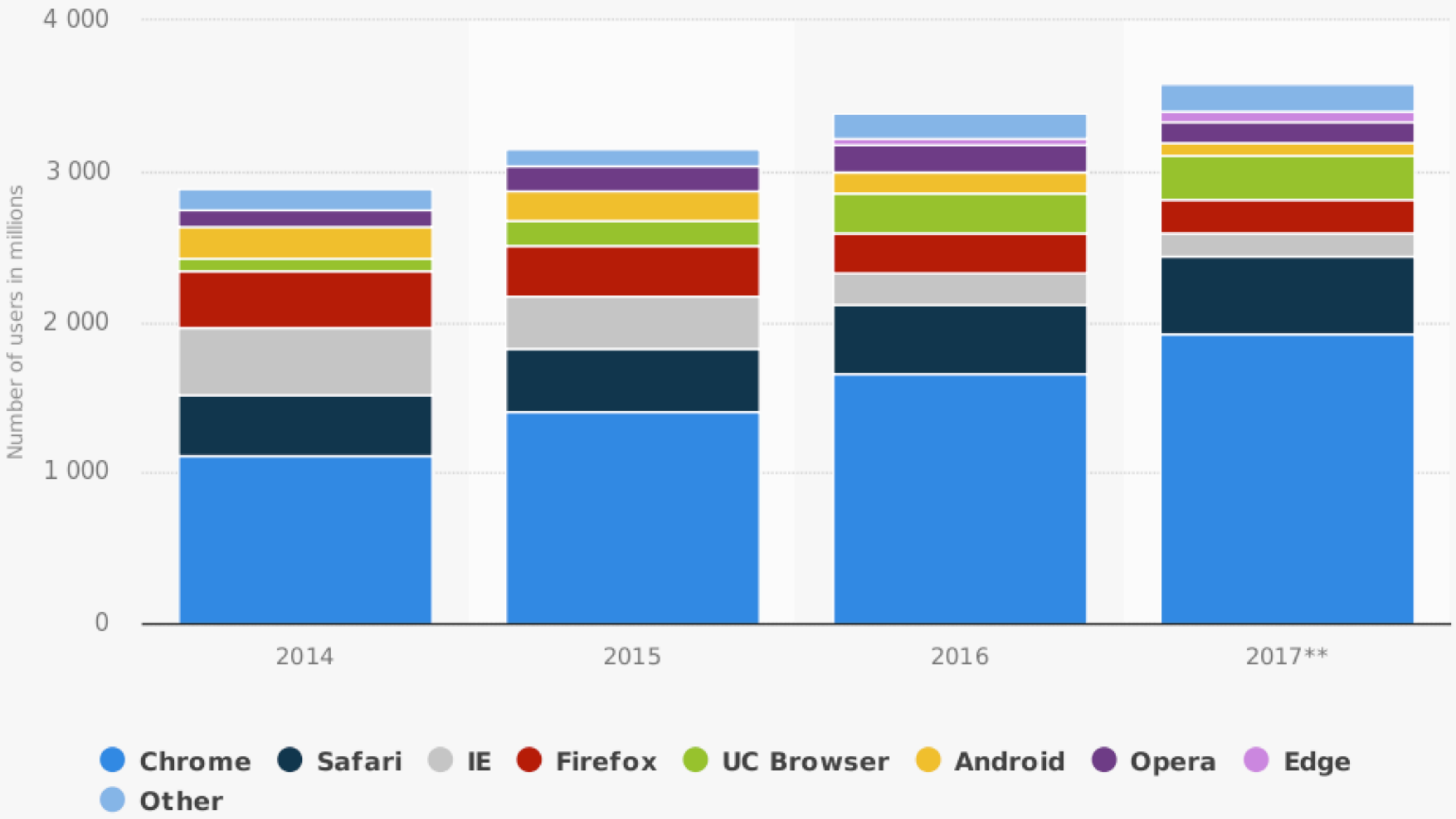
Indexed pages (Google, 2018)



Bing, last 3 months



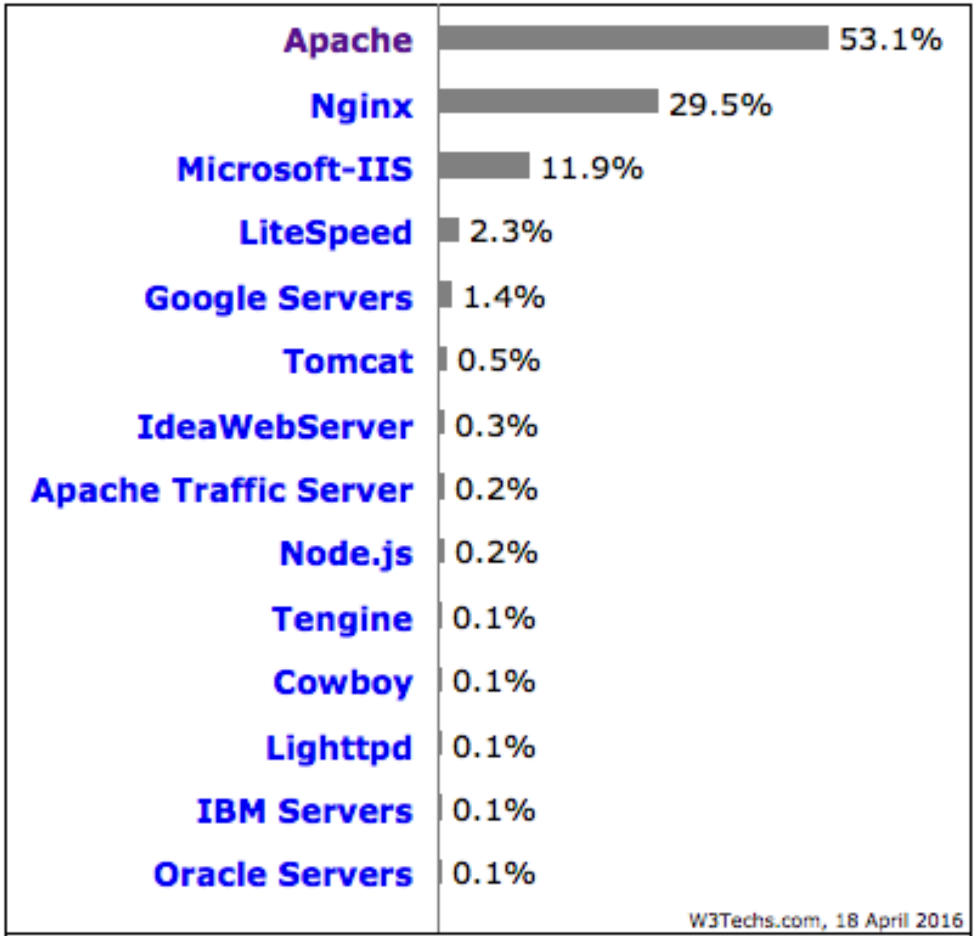
User population of selected internet browsers worldwide from 2012 to 2017 (in millions)*



Sources
StatCounter; Statista; ITU

Additional Information:
Worldwide; StatCounter; Statista; ITU; 2012 to 2017

Market shares per web servers (to store, process and deliver web pages)



W3Techs.com, 18 April 2016

Percentages of websites using various web servers
Note: a website may use more than one web server

Zipf’ s Law dominates on the Web

- Number of in-links/out-links to/from a page has a Zipfian distribution (frequency of pages with outdegree k as a % of the full population N).
- Length of web pages has a Zipfian distribution (frequency of pages with length k as a % of the full population N).
- Number of hits to a web page has a Zipfian distribution (pages accessed k times as a % of the full population N).

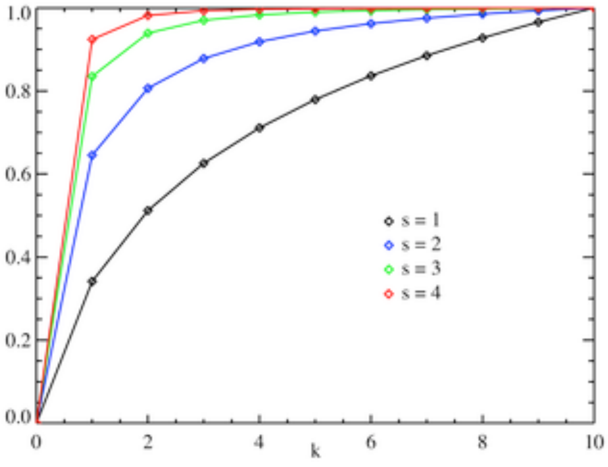
Zipf's law predicts that out of a population of N elements, the frequency of elements of “rank” k, f(k;s,N), is:

$$f(k;s,N)=\frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}.$$

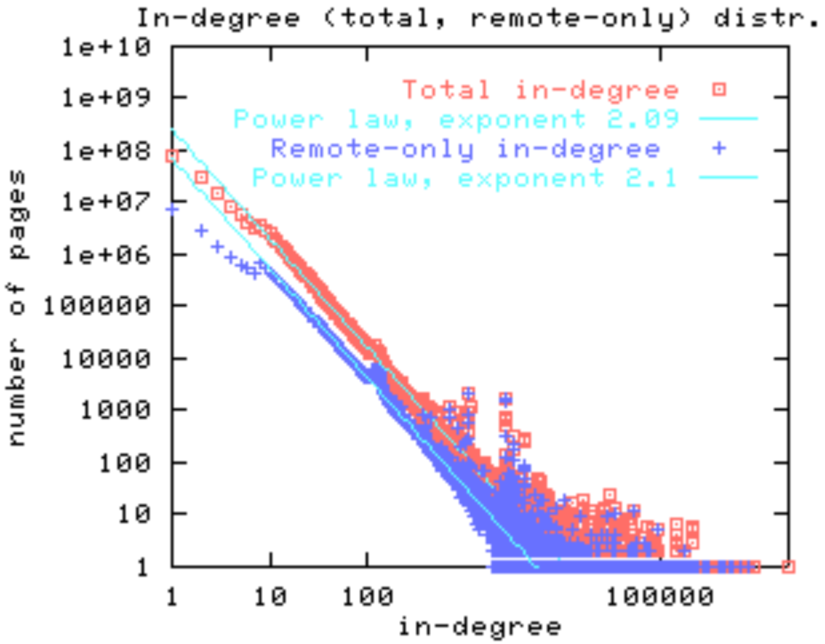
–N be the number of elements;

–k be their rank;

–s be the value of the exponent characterizing the distribution

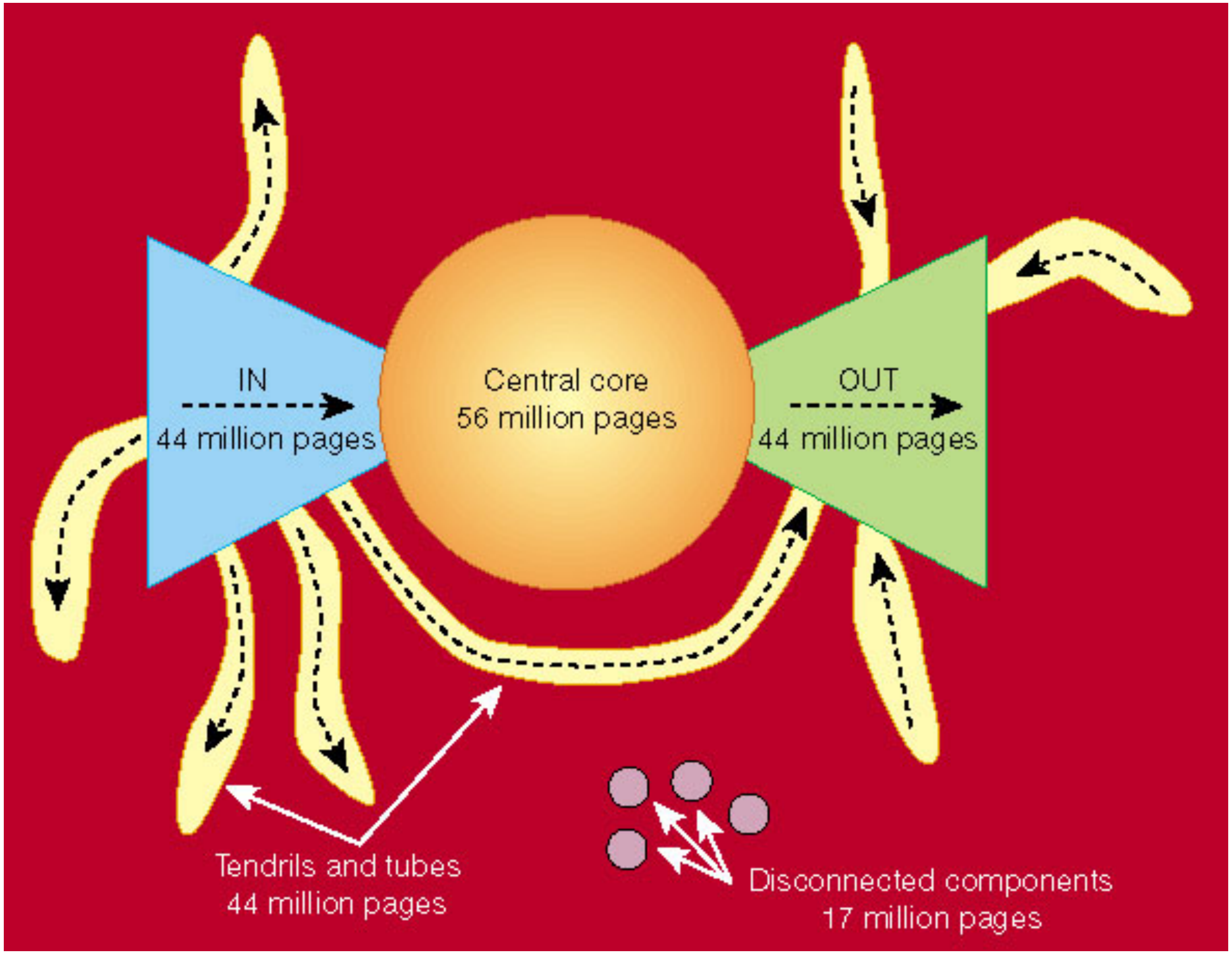


Example: Zipfian distribution of web pages in-degree



Very many pages with very low in-degree, very few pages with very high indegree

Graph Structure in the Web



http://www9.org/w9cdrom/160/160.html

THE WEB IS MOSTLY FREE ACCESS – HOW DO THE BIG PLAYERS MAKE MONEY?

Business Models for Web Search

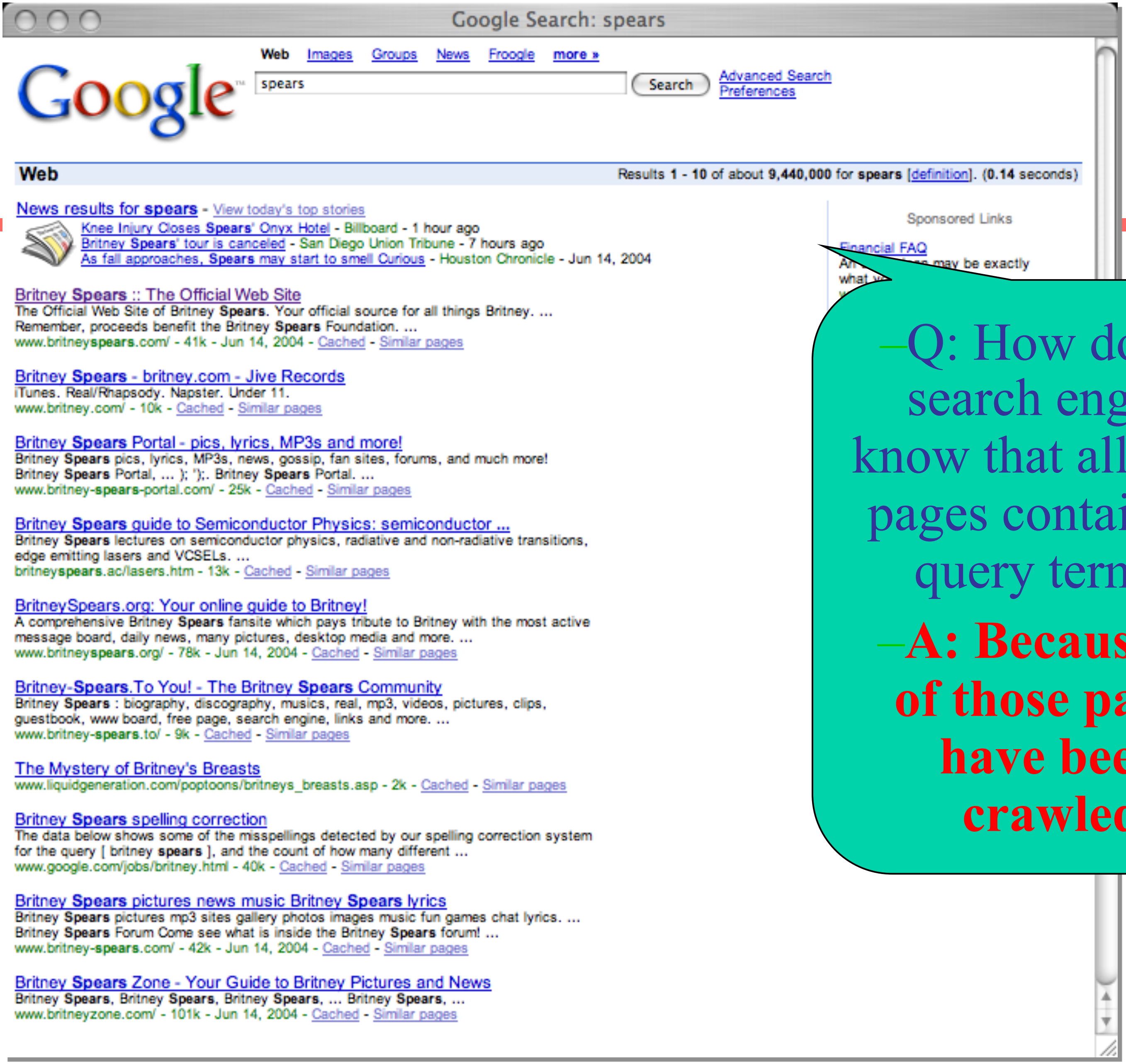
- Advertisers pay for banner ads on the site that **do not depend on a user's query**.
  - CPM: Cost Per Mille. Pay for each ad display.
    - (CPM=(cost of ad) / (#of readers/1000))
  - CPC: Cost Per Click. Pay only when user clicks on ad.
  - CTR: Click Through Rate. CTR=Number of clicks/ number of impressions*
  - CPA: Cost Per Action (Acquisition). Pay only when user actually makes a purchase on target site.
- Advertisers bid for "keywords". Ads for highest bidders displayed **when user query contains a purchased keyword**.
  - PPC: Pay Per Click. CPC for bid word ads (e.g. Google AdWords).

* **Impressions**: The number of times pages from "your site" appeared in search results,

Affiliates Programs

- If you have a website, you can generate income by becoming an *affiliate* by agreeing to post ads relevant to the topic of your site.
- If users click on your impression of an ad, you get some percentage of the CPC or PPC income that is generated.
- Google introduces AdSense affiliates program in 2003.

CRAWLING, RANKING ON THE WWW



Google Search: spears

Web

Results **1** - **10** of about **9,440,000** for **spears** [definition]. (**0.14** seconds)

News results for **spears** - View today's top stories

Knee Injury Closes **Spears**' Onyx Hotel - Billboard - 1 hour ago
Britney **Spears**' tour is canceled - San Diego Union Tribune - 7 hours ago
As fall approaches, **Spears** may start to smell Curious - Houston Chronicle - Jun 14, 2004

Britney **Spears** :: The Official Web Site
The Official Web Site of Britney **Spears**. Your official source for all things Britney. ... Remember, proceeds benefit the Britney **Spears** Foundation. ...
www.britney**spears**.com/ - 41k - Jun 14, 2004 - Cached - Similar pages

Britney **Spears** - britney.com - Jive Records
iTunes. Real/Rhapsody. Napster. Under 11.
www.britney.com/ - 10k - Cached - Similar pages

Britney **Spears** Portal - pics, lyrics, MP3s and more!
Britney **Spears** pics, lyrics, MP3s, news, gossip, fan sites, forums, and much more! Britney **Spears** Portal, ...); ');. Britney **Spears** Portal. ...
www.britney-**spears**-portal.com/ - 25k - Cached - Similar pages

Britney **Spears** guide to Semiconductor Physics: semiconductor ...
Britney **Spears** lectures on semiconductor physics, radiative and non-radiative transitions, edge emitting lasers and VCSELs. ...
britney**spears**.ac/lasers.htm - 13k - Cached - Similar pages

BritneySpears.org: Your online guide to Britney!
A comprehensive Britney **Spears** fansite which pays tribute to Britney with the most active message board, daily news, many pictures, desktop media and more. ...
www.britney**spears**.org/ - 78k - Jun 14, 2004 - Cached - Similar pages

Britney-**Spears**.To You! - The Britney **Spears** Community
Britney **Spears** : biography, discography, musics, real, mp3, videos, pictures, clips, guestbook, www board, free page, search engine, links and more. ...
www.britney-**spears**.to/ - 9k - Cached - Similar pages

The Mystery of Britney's Breasts
www.liquidgeneration.com/poptoons/britneys_breasts.asp - 2k - Cached - Similar pages

Britney **Spears** spelling correction
The data below shows some of the misspellings detected by our spelling correction system for the query [britney **spears**], and the count of how many different ...
www.google.com/jobs/britney.html - 40k - Cached - Similar pages

Britney **Spears** pictures news music Britney **Spears** lyrics
Britney **Spears** pictures mp3 sites gallery photos images music fun games chat lyrics. ... Britney **Spears** Forum Come see what is inside the Britney **Spears** forum! ...
www.britney-**spears**.com/ - 42k - Jun 14, 2004 - Cached - Similar pages

Britney **Spears** Zone - Your Guide to Britney Pictures and News
Britney **Spears**, Britney **Spears**, Britney **Spears**, ... Britney **Spears**, ...
www.britneyzone.com/ - 101k - Jun 14, 2004 - Cached - Similar pages

Sponsored Links

Financial FAQ
An may be exactly
what

- Q: How does a search engine know that all these pages contain the query terms?
- **A: Because all of those pages have been crawled**

Crawlers

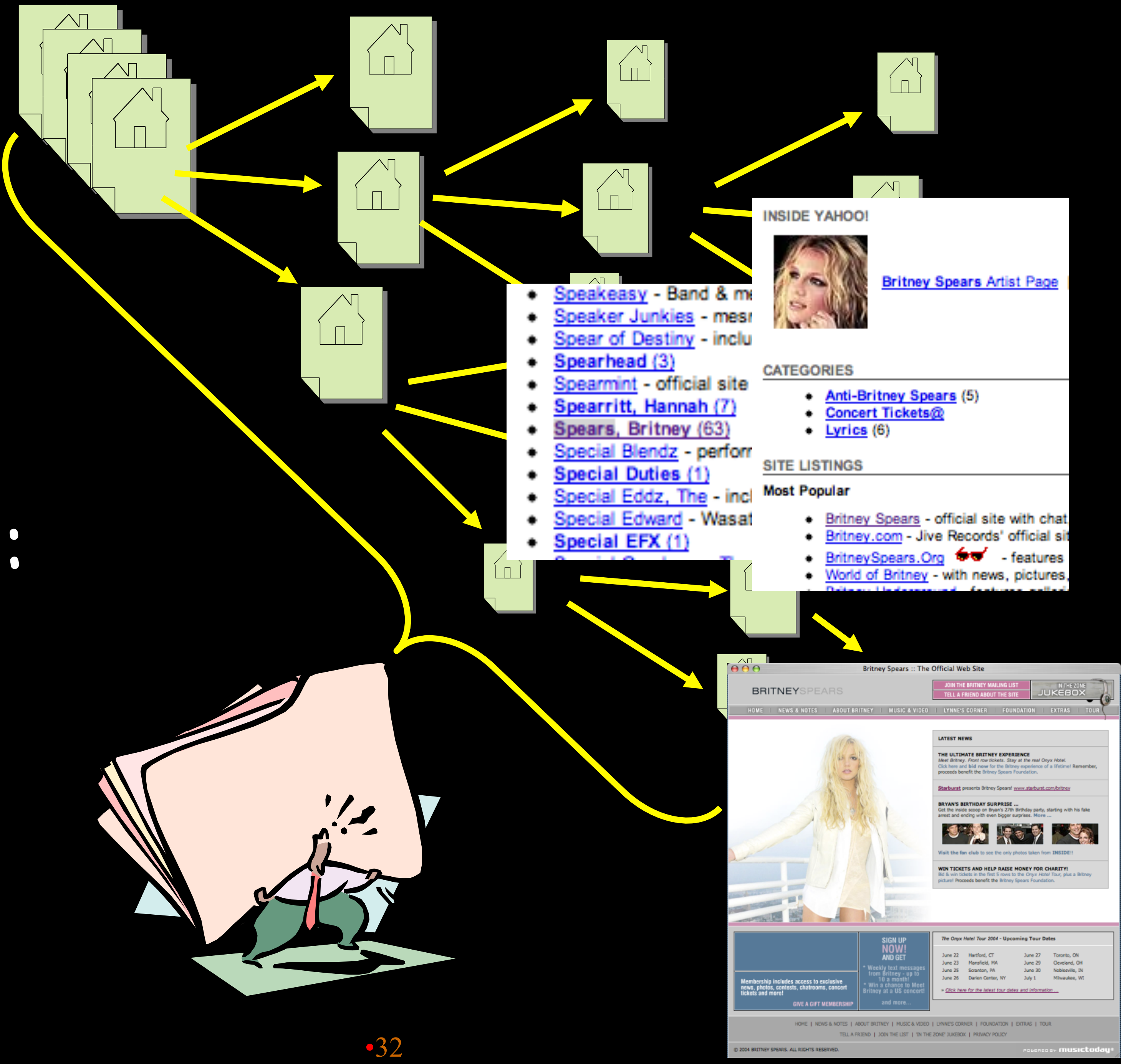
A **Web crawler** is a software application which systematically browses the Web, for the purpose of Web indexing.

Web Crawler: steps

- Starts with a set of *seeds*, which are a set of URLs given to it as parameters
- Seeds are added to a URL request queue
- Crawler starts fetching pages from the request queue
- Downloaded pages are parsed to find link tags that might contain other useful URLs to fetch
- New URLs added to the crawler's request queue, or *frontier*
- Continue until no more new URLs or disk full

starting
pages
(seeds)

Crawler: basic idea



- Speakeasy - Band & me
- Speaker Junkies - mesr
- Spear of Destiny - inclu
- Spearhead (3)
- Spearmint - official site
- Spearritt, Hannah (7)
- Spears, Britney (63)
- Special Blendz - perforr
- Special Duties (1)
- Special Eddz, The - incl
- Special Edward - Wasat
- Special EFX (1)

INSIDE YAHOO!

Britney Spears Artist Page

CATEGORIES

- Anti-Britney Spears (5)
- Concert Tickets@
- Lyrics (6)

SITE LISTINGS

Most Popular

- Britney Spears - official site with chat
- Britney.com - Jive Records' official sit
- BritneySpears.Org - features
- World of Britney - with news, pictures,

Britney Spears :: The Official Web Site

BRITNEYSPEARS

JOIN THE BRITNEY MAILING LIST

TELL A FRIEND ABOUT THE SITE

IN THE ZONE JUKEBOX

HOME | NEWS & NOTES | ABOUT BRITNEY | MUSIC & VIDEO | LYNNE'S CORNER | FOUNDATION | EXTRAS | TOUR

LATEST NEWS

THE ULTIMATE BRITNEY EXPERIENCE

Meet Britney. Front row tickets. Stay at the real Onyx Hotel.
Click here and **bid now** for the Britney experience of a lifetime! Remember, proceeds benefit the Britney Spears Foundation.

Starburst presents Britney Spears! www.starburst.com/britney

BRYAN'S BIRTHDAY SURPRISE ...

Get the inside scoop on Bryan's 27th Birthday party, starting with his fake arrest and ending with even bigger surprises. **More ...**

Visit the fan club to see the only photos taken from **INSIDE**!!

WIN TICKETS AND HELP RAISE MONEY FOR CHARITY!

Bid & win tickets in the first 5 rows to the *Onyx Hotel Tour*, plus a Britney picture! Proceeds benefit the Britney Spears Foundation.

Membership includes access to exclusive news, photos, contests, chatrooms, concert tickets and more!

GIVE A GIFT MEMBERSHIP

SIGN UP NOW! AND GET

- Weekly text messages from Britney - up to 10 a month!
- Win a chance to Meet Britney at a US concert!

and more...

The Onyx Hotel Tour 2004 - Upcoming Tour Dates

June 22	Hartford, CT	June 27	Toronto, ON
June 23	Mansfield, MA	June 29	Cleveland, OH
June 25	Scranton, PA	June 30	Noblesville, IN
June 26	Darien Center, NY	July 1	Milwaukee, WI

» *Click here for the latest tour dates and information ...*

HOME | NEWS & NOTES | ABOUT BRITNEY | MUSIC & VIDEO | LYNNE'S CORNER | FOUNDATION | EXTRAS | TOUR

TELL A FRIEND | JOIN THE LIST | 'IN THE ZONE' JUKEBOX | PRIVACY POLICY

© 2004 BRITNEY SPEARS. ALL RIGHTS RESERVED.

POWERED BY musictoday

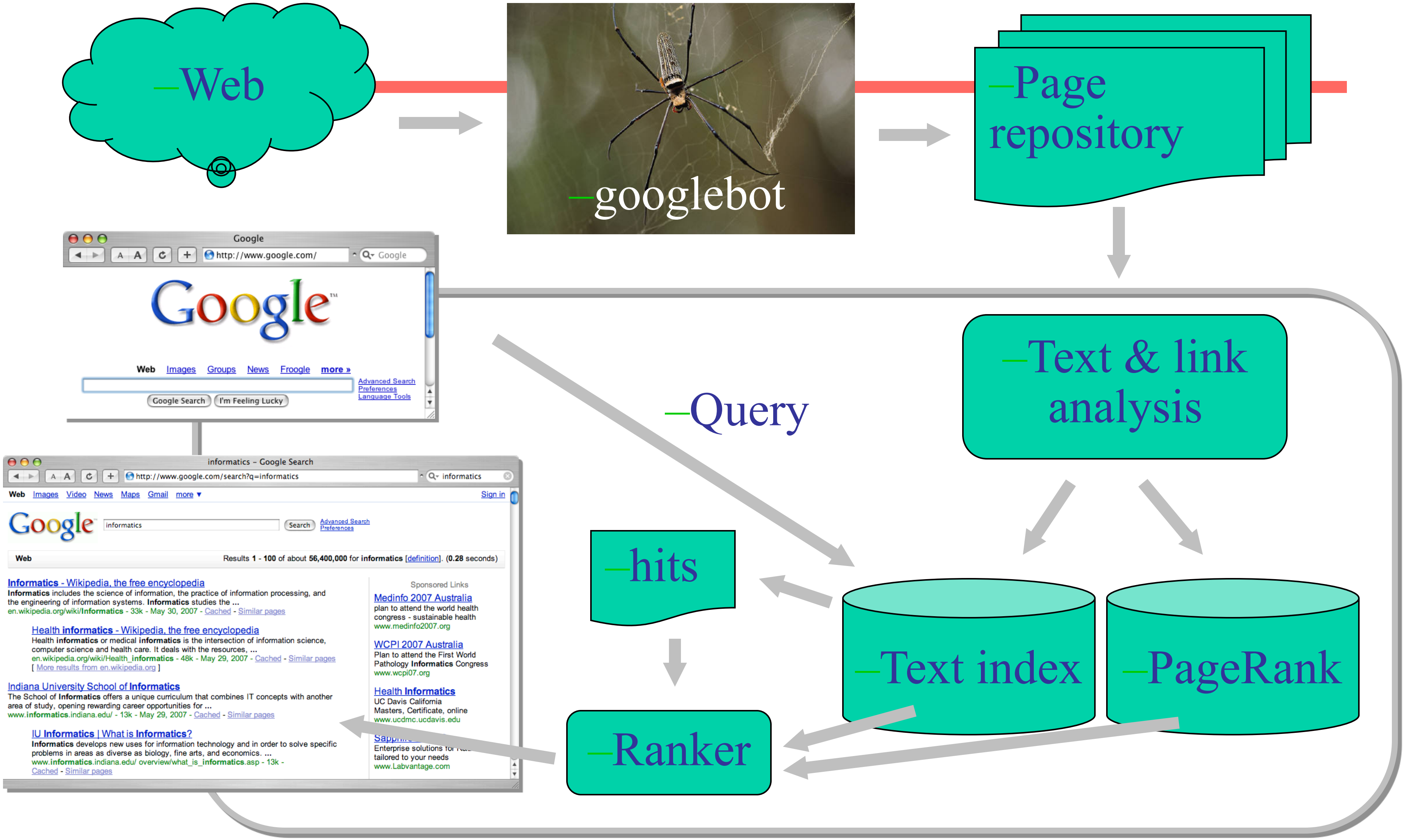
Many names

- Crawler
- Spider
- Robot (or bot)
- Web agent
- Wanderer, worm, …
- And famous “instances”: googlebot, scooter, slurp, msnbot, …

Crawlers vs Browsers vs Scrapers

- Crawlers automatically harvest all files on the web
- Browsers are manual crawlers (user search through keywords or URL names)
- Scrapers takes pages that have been downloaded, and automatically extract data from it for manipulation

A crawler within a search engine (e.g. Google)



Web Results **1** - **100** of about **56,400,000** for **informatics** [definition]. (**0.28** seconds)

Informatics - Wikipedia, the free encyclopedia
Informatics includes the science of information, the practice of information processing, and the engineering of information systems. **Informatics** studies the ...
en.wikipedia.org/wiki/**Informatics** - 33k - May 30, 2007 - Cached - Similar pages

Health **informatics** - Wikipedia, the free encyclopedia
Health **informatics** or medical **informatics** is the intersection of information science, computer science and health care. It deals with the resources, ...
en.wikipedia.org/wiki/Health_**informatics** - 48k - May 29, 2007 - Cached - Similar pages
[More results from en.wikipedia.org]

Indiana University School of **Informatics**
The School of **Informatics** offers a unique curriculum that combines IT concepts with another area of study, opening rewarding career opportunities for ...
www.**informatics**.indiana.edu/ - 13k - May 29, 2007 - Cached - Similar pages

IU **Informatics** | What is **Informatics**?
Informatics develops new uses for information technology and in order to solve specific problems in areas as diverse as biology, fine arts, and economics. ...
www.**informatics**.indiana.edu/ overview/what_is_**informatics**.asp - 13k -
Cached - Similar pages

Sponsored Links

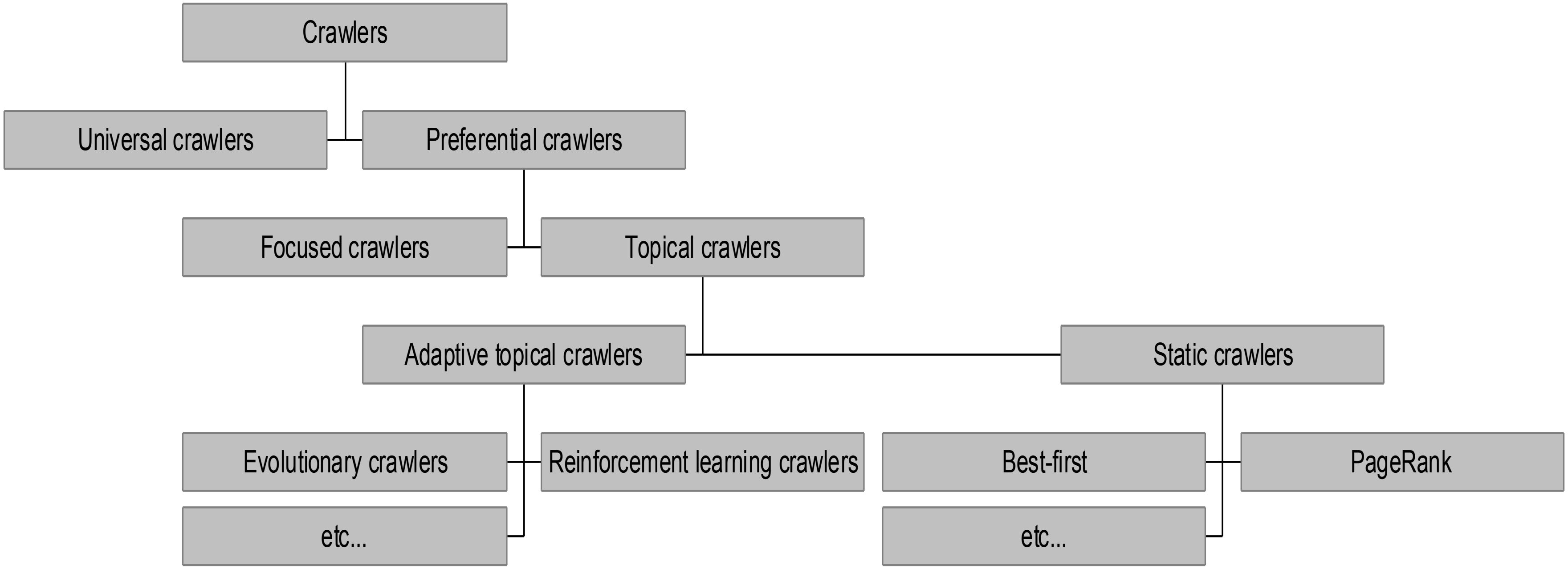
Medinfo 2007 Australia
plan to attend the world health congress - sustainable health
www.medinfo2007.org

WCPI 2007 Australia
Plan to attend the First World Pathology **Informatics** Congress
www.wcpi07.org

Health **Informatics**
UC Davis California
Masters, Certificate, online
www.ucdmc.ucdavis.edu

Sapphire
Enterprise solutions for
tailored to your needs
www.Labvantage.com

Types of crawlers

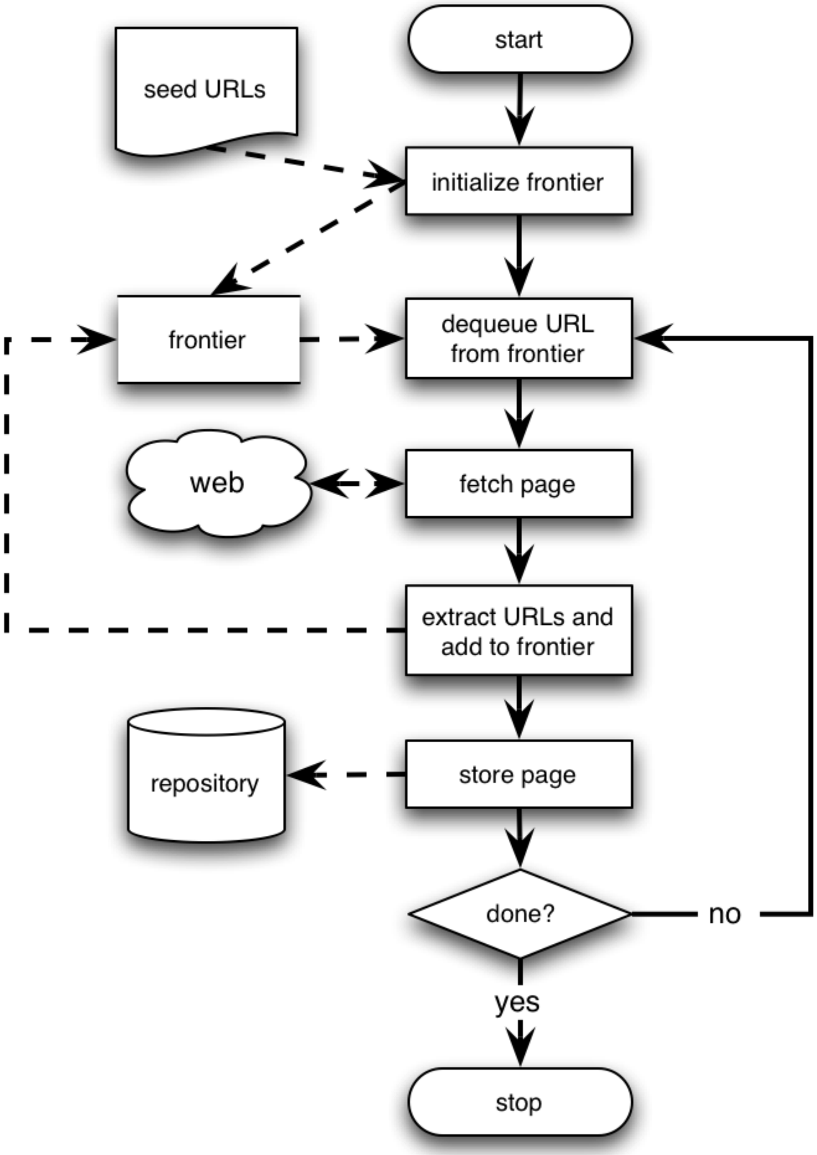


- Universal: support universal web search engines
- Preferential: Selective bias toward some pages, eg. most "relevant"/topical, closest to seeds, most popular/largest, PageRank, highest rate/amount of change, etc..

Focused Crawling

- Attempts to download only those pages that are about a particular topic
  - used by *vertical search* applications
  - E.g. Tripadvisor, PubMed, SkyScraper..
- Rely on the fact that pages about a topic tend to have links to other pages on the same topic
  - popular pages for a topic are typically used as seeds
- Crawler uses *text classifier* to decide whether a page is "on topic" before indexing it

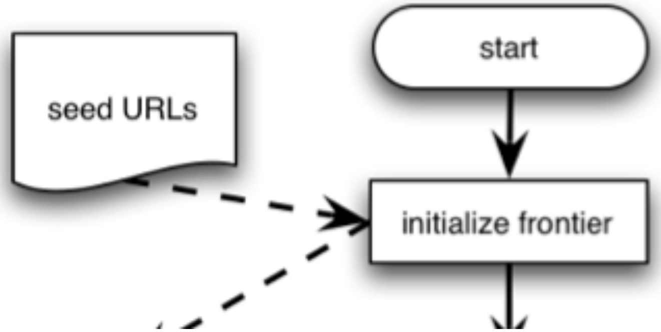
Basic crawlers



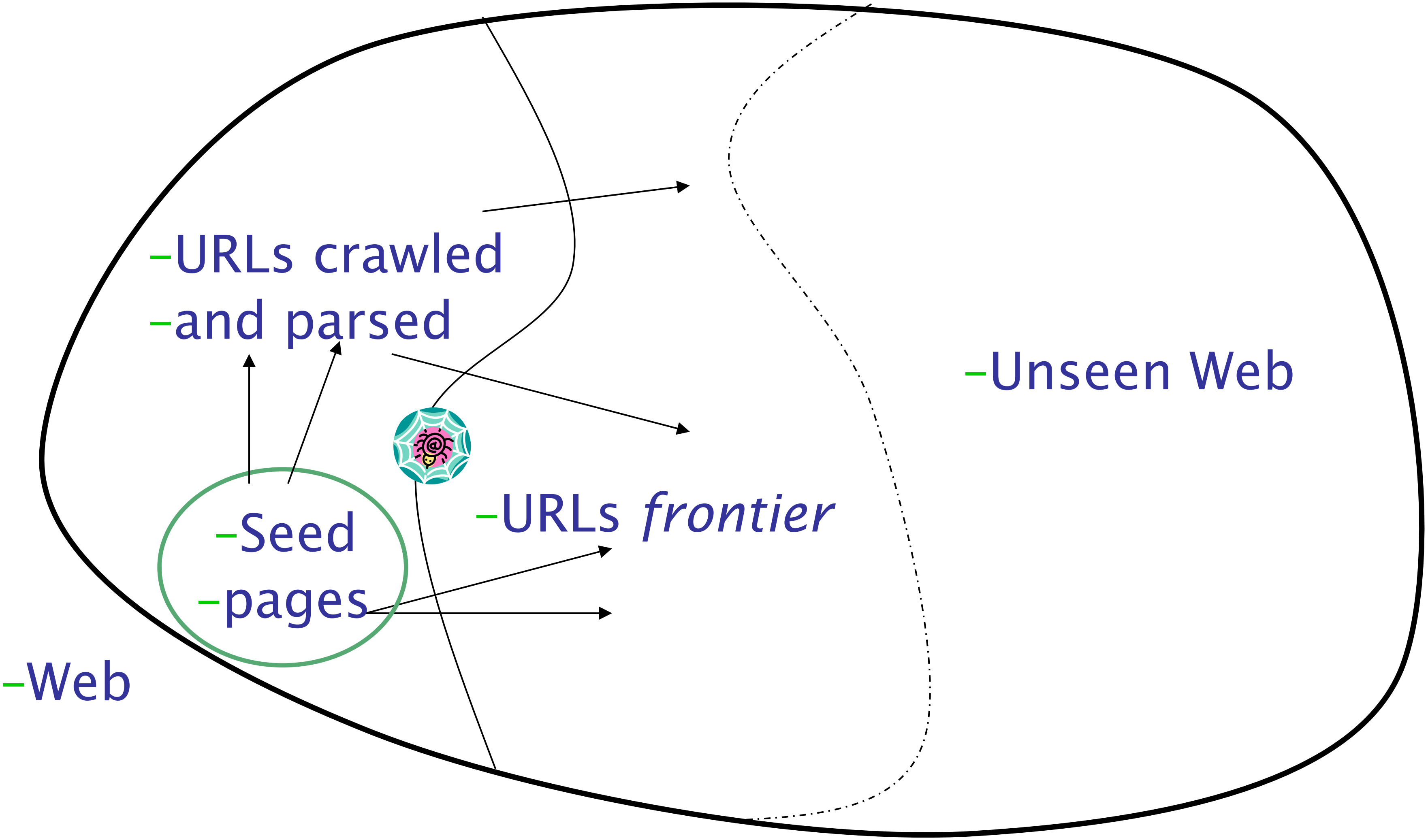
- This is a sequential universal crawler
- Seeds can be any list of starting URLs
- Order of page visits is determined by frontier data structure
- Stop criterion can be anything

URL frontier

- Frontier: The next nodes to crawl
- Crawler start from a set of seed pages (initial fronteer) and then gradually expand



Basic crawler



–URLs crawled
–and parsed

–Unseen Web

–URLs *frontier*

–Seed
–pages

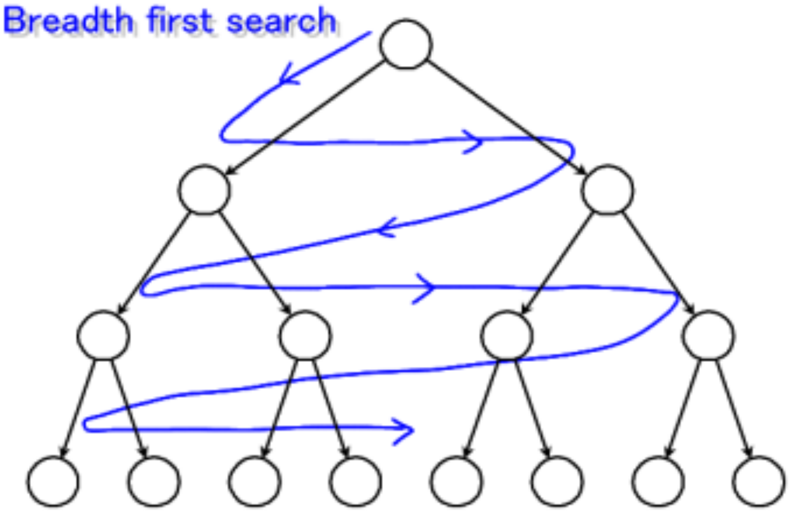
–Web

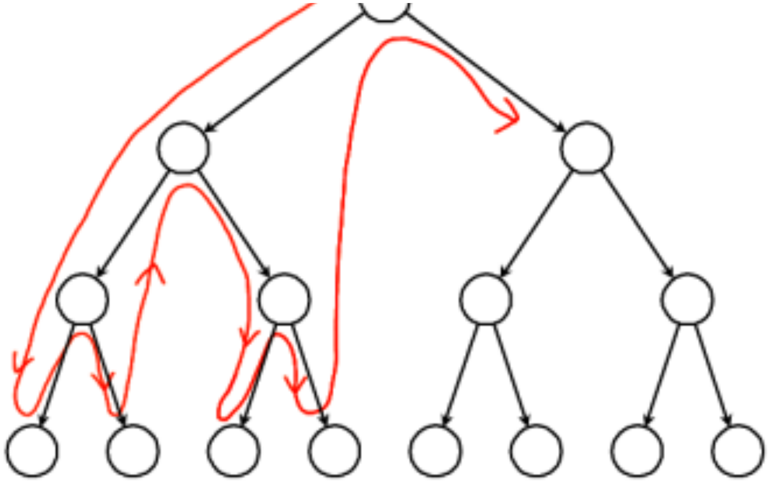
ISSUES with crawling

1. Web Graph visiting policies
2. Efficiency (multithreads and distributed crawling)
3. Ethics (accessing and scraping web pages)
4. Freshness of information
5. Coverage of the web
6. Other issues

1. Web Graph visiting policies

- Breadth First Search
  - Implemented with QUEUE (FIFO)
  - Finds pages along shortest paths
  - Important to start with "good" pages, this keeps us close: maybe we get other good stuff…
- Depth First Search
  - Implemented with STACK (LIFO)
  - Wander away ("lost in cyberspace")





2. Efficiency (1)

- Web crawlers spend a lot of time waiting for responses to requests
- To reduce this inefficiency, web crawlers use **threads** (executions that are independent and can run in parallel) and fetch hundreds of pages at once

Efficiency (2)

```
procedure CrawlerThread(frontier)
    while not frontier.done() do
        website ← frontier.nextSite()
        url ← website.nextURL()
        if website.permitsCrawl(url) then
            text ← retrieveURL(url)
            storeDocument(url, text)
            for each url in parse(text) do
                frontier.addURL(url)
            end for
        end if
        frontier.releaseSite(website)
    end while
end procedure
```

Simple Crawler Thread

Efficiency (3)

- Multithreading improves efficiency, **distributed crawling** (multiple computers) is the other method
- Three reasons to use multiple computers for crawling
  - Helps to put the crawler closer to the sites it crawls
  - Reduces the number of sites the crawler has to remember
  - Reduces computing resources required
- Distributed crawler uses a hash function to assign URLs to crawling computers

3. Crawler ethics (1)

- Crawlers can cause trouble, even unwillingly, if not properly designed to be "polite" and "ethical"
- For example, sending too many requests in rapid succession to a single server can amount to a Denial of Service (DoS) attack!
  - Server administrator and users will be upset
  - Crawler developer/admin IP address may be blacklisted

Crawler ethics (2)

- Even crawling a site slowly will anger some web server administrators, who object to any copying of their data
- **Robots**.txt file can be used to control crawlers

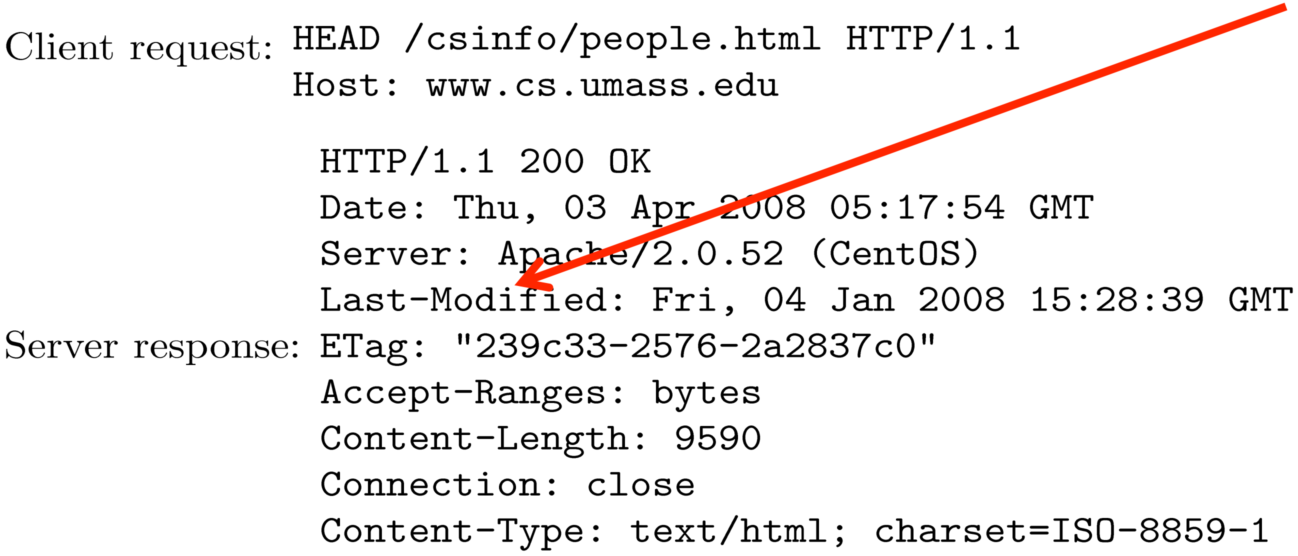
```
User-agent: *
Disallow: /private/
Disallow: /confidential/
Disallow: /other/
Allow: /other/public/

User-agent: FavoredCrawler
Disallow:

Sitemap: http://mysite.com/sitemap.xml.gz
```

4. Freshness/Age

- Web pages are constantly being added, deleted, and modified
- Web crawler must continually **revisit pages** it has already crawled to see if they have changed in order to maintain the *freshness* of the document collection
  - *stale* copies no longer reflect the real contents of the web pages

Freshness/Age (2)

- HTTP protocol has a special request type called HEAD that makes it easy to check for page changes
  - HEAD method returns (meta)information about page, not page itself



Client request:

```
HEAD /csinfo/people.html HTTP/1.1
Host: www.cs.umass.edu
```

Server response:

```
HTTP/1.1 200 OK
Date: Thu, 03 Apr 2008 05:17:54 GMT
Server: Apache/2.0.52 (CentOS)
Last-Modified: Fri, 04 Jan 2008 15:28:39 GMT
ETag: "239c33-2576-2a2837c0"
Accept-Ranges: bytes
Content-Length: 9590
Connection: close
Content-Type: text/html; charset=ISO-8859-1
```

Freshness/Age (3)

- Not possible to constantly check all pages
  - must check important pages and pages that change frequently
- Freshness is the proportion of pages that are fresh
- Optimizing for this metric can lead to bad decisions, such as not crawling popular sites who do not change frequently
- *Age* is a better metric

Freshness/Age (4)

- Expected **age** of a page *t* days after it was last crawled:

$$\mathrm{Age}(\lambda, t) = \int_0^t P(\text{page changed at time } x)(t-x)dx$$

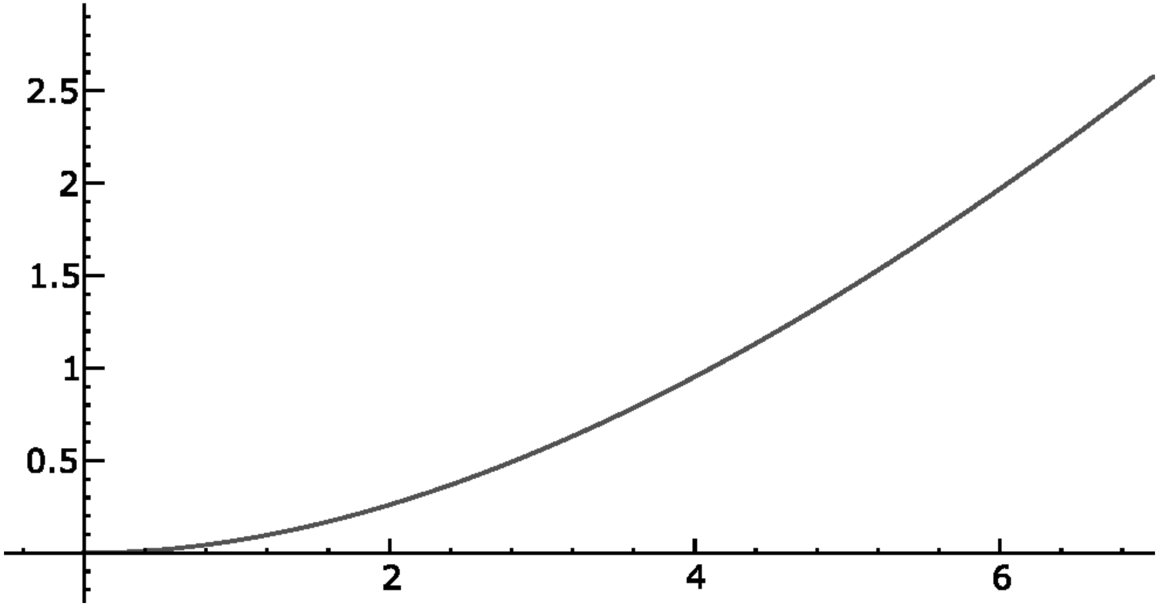
- Web page updates follow the Poisson distribution on average
  - time until the next update is governed by an exponential (Poisson) distribution :

$$\mathrm{Age}(\lambda, t) = \int_0^t \lambda e^{-\lambda x}(t-x)dx$$

λ is change rate (site-dependent)

Freshness/Age (5)

- The older a page gets, the more it costs not to crawl it
  - e.g., expected age with mean change frequency $\lambda = 1/7$ (one change per week)



Freshness/Age (6)

- Sitemaps contain lists of URLs and data about those URLs, such as **modification time** and **modification frequency (to estimate age)**
- Generated by web server administrators
- Gives crawler a hint about when to check a page for changes

Freshness/Age (7)

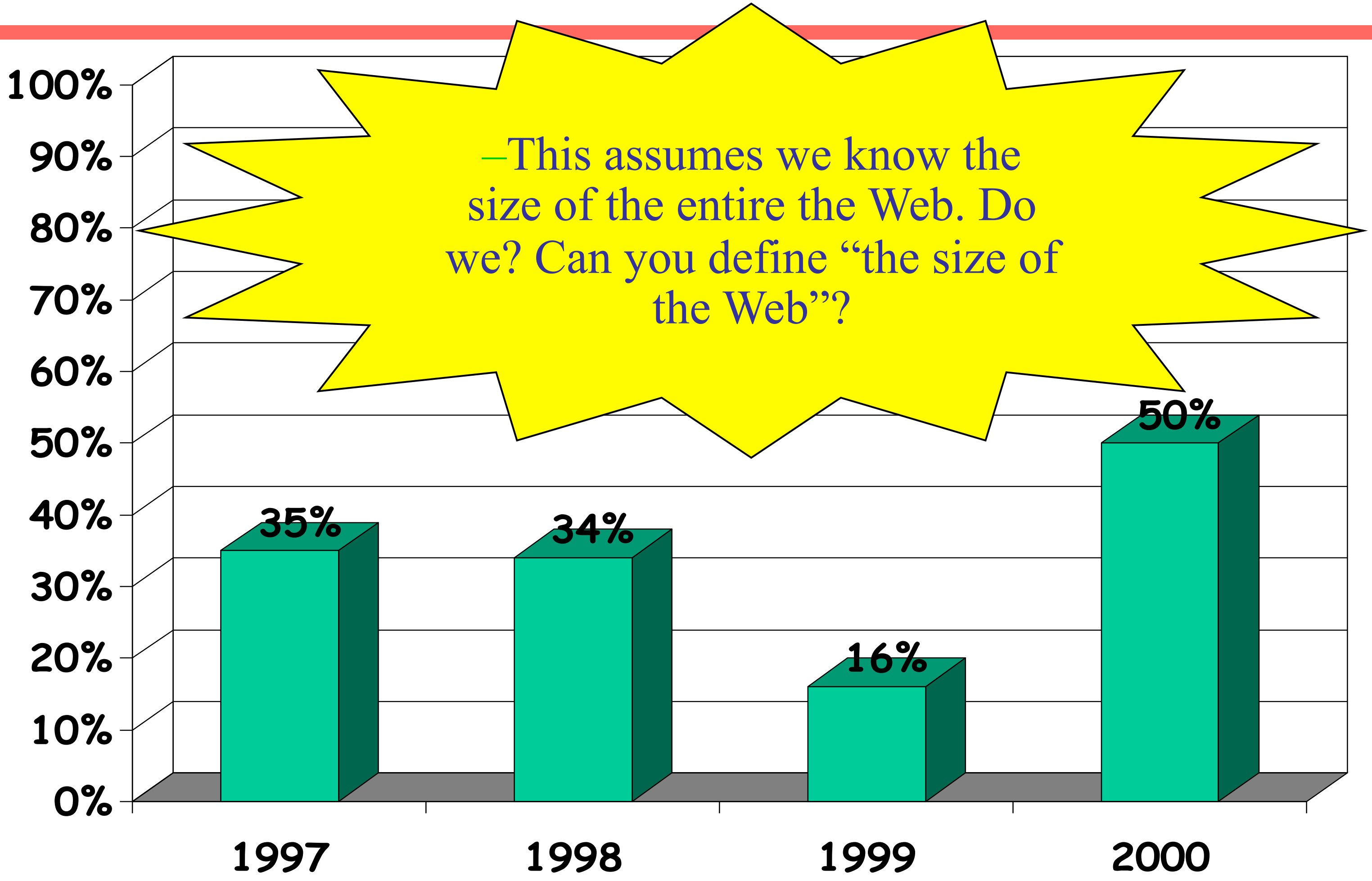
Sitemap Example

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.company.com/</loc>
    <lastmod>2008-01-15</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.7</priority>
  </url>
  <url>
    <loc>http://www.company.com/items?item=truck</loc>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.company.com/items?item=bicycle</loc>
    <changefreq>daily</changefreq>
  </url>
</urlset>
```

5. Coverage (1)

- Do we need to crawl the entire Web?
- If we cover too much, it will get stale
- There is an abundance of pages in the Web, but some are useless
- What is the goal?
  - General search engines: pages with high prestige
  - News portals: pages that change often
  - Vertical portals: pages on some topic

Coverage (2)

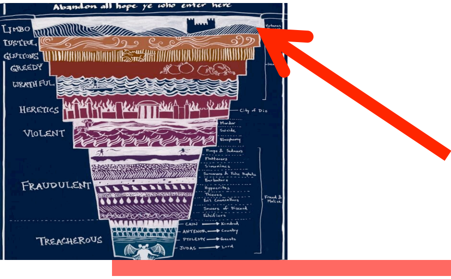


–This assumes we know the size of the entire the Web. Do we? Can you define “the size of the Web”?

Web coverage by search engine crawlers

Deep Web

Abandon all hope ye who enter here
LIMBO
Virtuous Pagans
LUSTFUL
GLUTTONS
GREEDY
Incontinence
WRATHFUL
HERETICS
City of Dis
VIOLENT
Murder
Suicide
Blasphemy
FRAUDULENT
Pimps & Seducers
Flatterers
Simoniacs
Sorcerers & False Prophets
Barbators
Hypocrites
Thieves
Evil Councellors
Sowers of Discord
Falsifiers
Fraud & Malice
TREACHEROUS
CAIN → Kindred
ANTENOR → Country
PTOLEMY → Guests
JUDAS → Lord

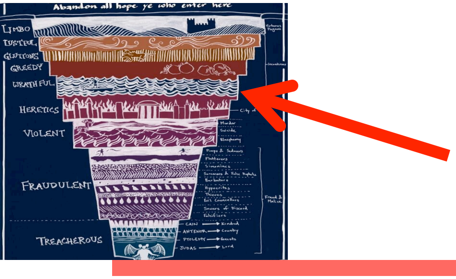


Level 1- The Surface Web

- The web that the vast majority of internet users are accustomed to.
- Accessible in any nation that does not block internet access, even places like China and Egypt.
- Social media sites like Facebook, informational websites like Wikipedia, general websites, etc.

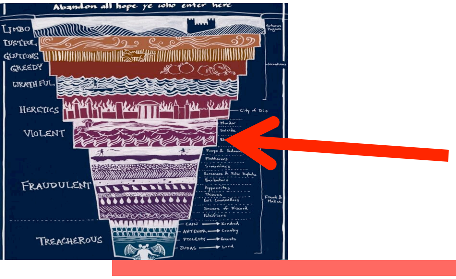
Level 2-The Bergie Web

- The layer of the Surface Web that is **blocked** in some nations. Some other information is only accessible through illegal means.
- Google locked results
- Recently web crawled old content
- Pirated Media
- Pornography



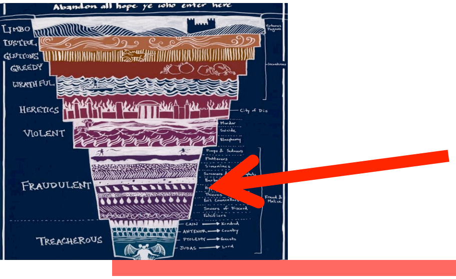
Level 3-The Deep Web

- Requires a proxy or two (namely Tor – free sw for anonymous browsing) to access.
- Contains most of the archived web pages of the 1990s Web that **did not renew** their domain names and such.
- Government/Business/Collegiate Research.
- Hackers/Script Kiddies/Virus Information.
- Illegal and Obscene Content (CP, Gore, Suicides, etc.)



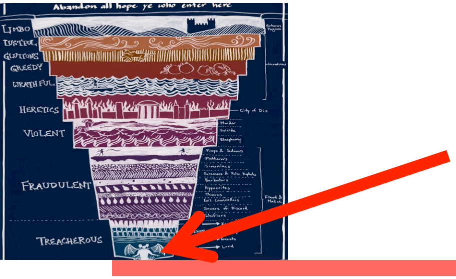
Level 4- The Charter Web

- Like the Regular Deep Web, but harder to get into and **more illegal content**.
- Advanced covert government research.
- Most of the internet black market (run on bitcoins)
- Human/Arms/Drug/Rare Animal Trafficking.
- Assassination networks , bounty hunters, illegal game hunting, and other bad stuff

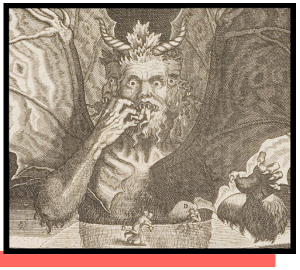


Level 5-Marianas Web

- Lowest known level of the Deep Web.
- Named after the Spanish Technician who created it.
- Extremely difficult to access, users say it is the safest part of the internet due to how private it is.
- Julian Assange and other top-level Wikileaks members are believed to have access.



Rumored Levels 6-8

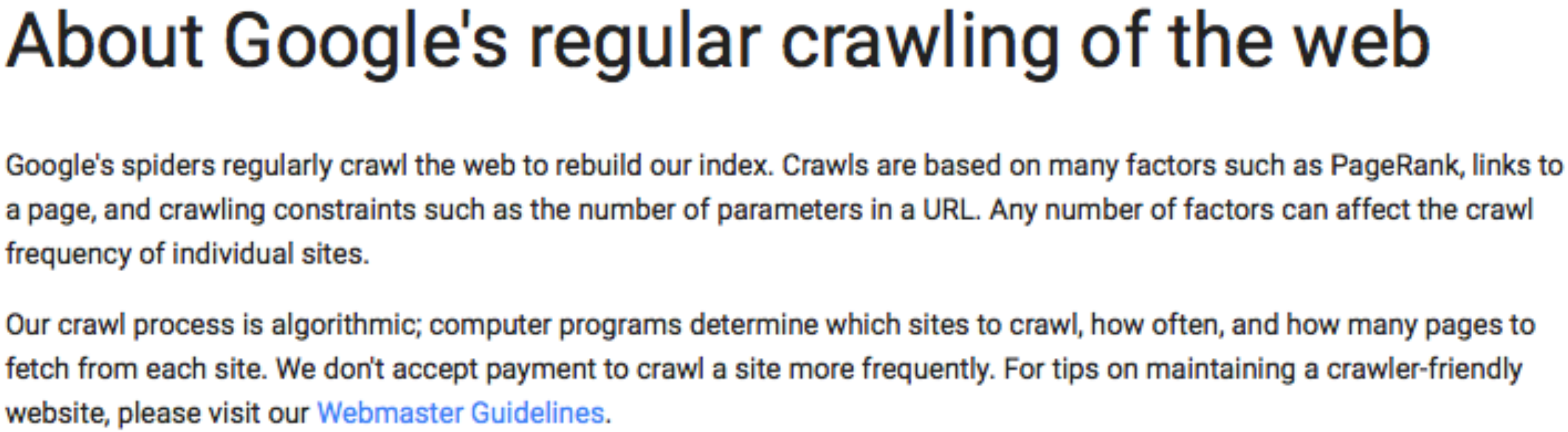


- Mostly the stuff of conspiracy theorists.
- Level 6 is a giant firewall meant to prevent people from going any further.
- Level 7 “The Fog” is where various worldwide power-players jockey for control of PrimArch. Said to be very dangerous, full of viruses and such.
- Level 8 is called PrimArch and is claimed to be controlled by an extremely powerful AI (possibly running on a quantum computer).

Other crawler implementation issues

- Duplication: Don’t want to fetch same page twice!
  - Keep lookup table (hash) of visited pages
  - What if not visited but in frontier already?
- Prioritized search: The frontier grows very fast!
  - For large crawls, need to define an exploration policy **with priorities**, rather than depth first or breadth first
- Availability: Fetcher must be robust!
  - Don’t crash if download fails
  - Timeout mechanism
- Skip policy: Determine file type to skip unwanted files
  - Can try using extensions, but not reliable
  - Can issue ‘HEAD’ HTTP commands to get Content-Type (MIME) headers, but overhead of extra Internet requests

Google



About Google's regular crawling of the web

Google's spiders regularly crawl the web to rebuild our index. Crawls are based on many factors such as PageRank, links to a page, and crawling constraints such as the number of parameters in a URL. Any number of factors can affect the crawl frequency of individual sites.

Our crawl process is algorithmic; computer programs determine which sites to crawl, how often, and how many pages to fetch from each site. We don't accept payment to crawl a site more frequently. For tips on maintaining a crawler-friendly website, please visit our Webmaster Guidelines.

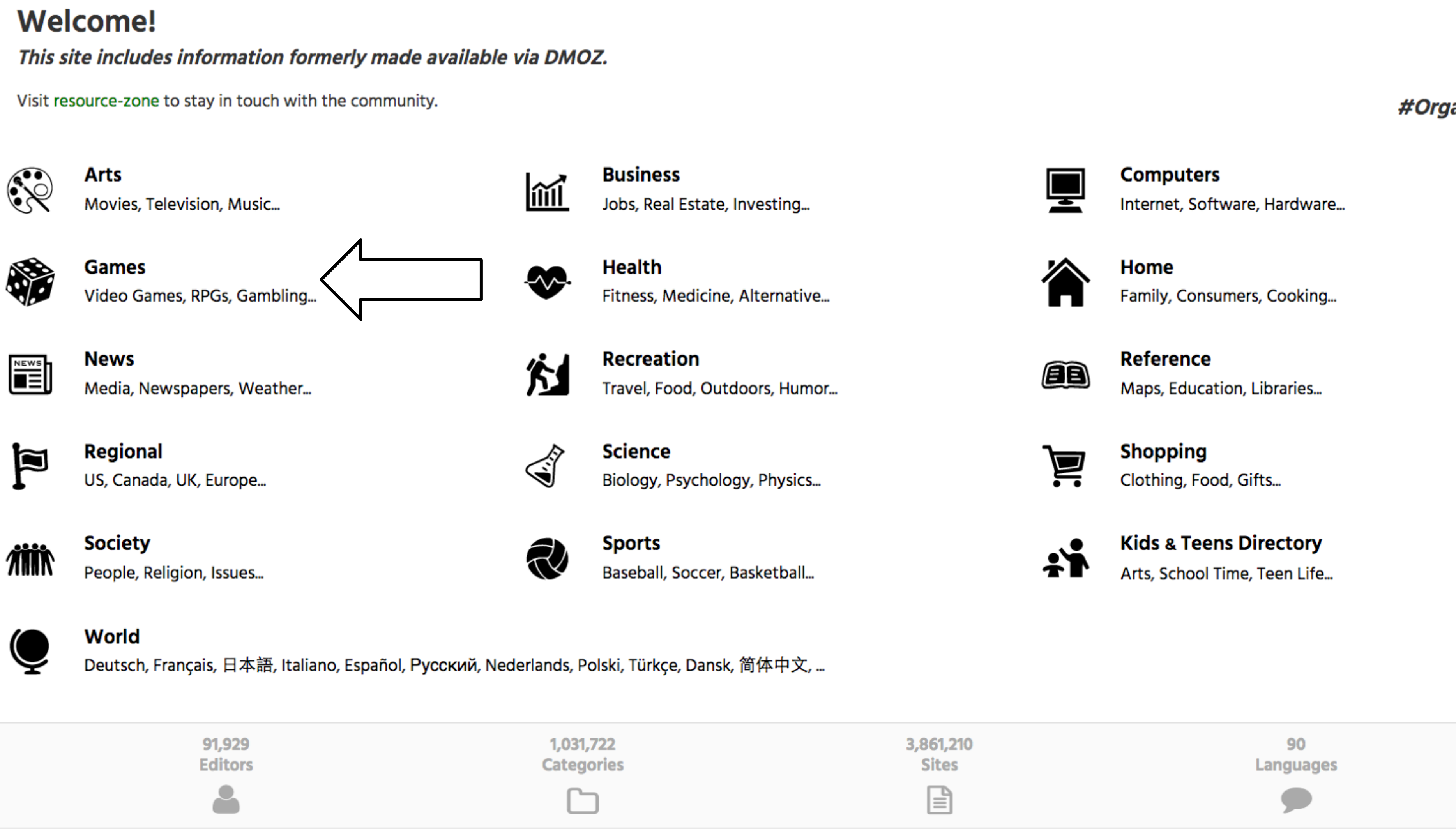
Web page ranking:
Manual/automatic classification
Link analysis

Why we need to classify pages?

- Vector Space ranking is not enough
- Queries on the web return millions hits based only on content similarity (Vector Space or other ranking methods)
- Need **additional criteria** for selecting good pages:
  - Classification of web pages into pre-defined categories
  - Assigning relevance to pages depending upon their “position” in the web graph (link analysis)

Manual Hierarchical Web Taxonomies

- Yahoo (old) approach of using human editors to assemble a large hierarchically structured directory of web pages.
  - http://www.yahoo.com/
- Open Directory Project is a similar approach based on the distributed labor of volunteer editors ("net-citizens provide the collective brain"). Used by most other search engines. Started by Netscape.
  - http://www.dmoz.org/
  - Now replaced by https://dmoztools.net/



Welcome!

This site includes information formerly made available via DMOZ.

Visit resource-zone to stay in touch with the community.

#Orga

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

Kids & Teens Directory
Arts, School Time, Teen Life...

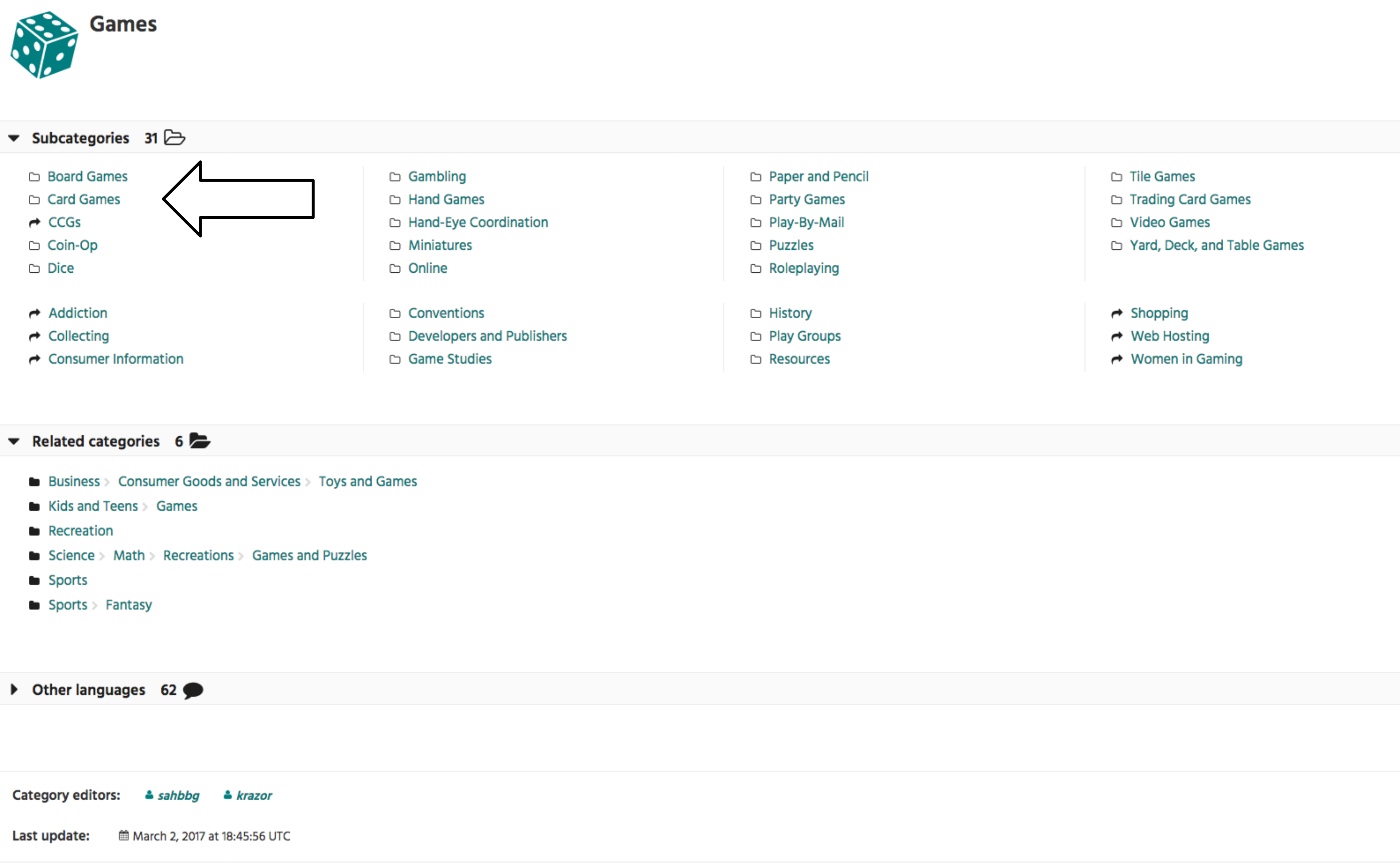
World
Deutsch, Français, 日本語, Italiano, Español, Русский, Nederlands, Polski, Türkçe, Dansk, 简体中文, ...

91,929
Editors

1,031,722
Categories

3,861,210
Sites

90
Languages



Games

Subcategories 31

- Board Games
- Card Games
- CCGs
- Coin-Op
- Dice
- Gambling
- Hand Games
- Hand-Eye Coordination
- Miniatures
- Online
- Paper and Pencil
- Party Games
- Play-By-Mail
- Puzzles
- Roleplaying
- Tile Games
- Trading Card Games
- Video Games
- Yard, Deck, and Table Games

- Addiction
- Collecting
- Consumer Information
- Conventions
- Developers and Publishers
- Game Studies
- History
- Play Groups
- Resources
- Shopping
- Web Hosting
- Women in Gaming

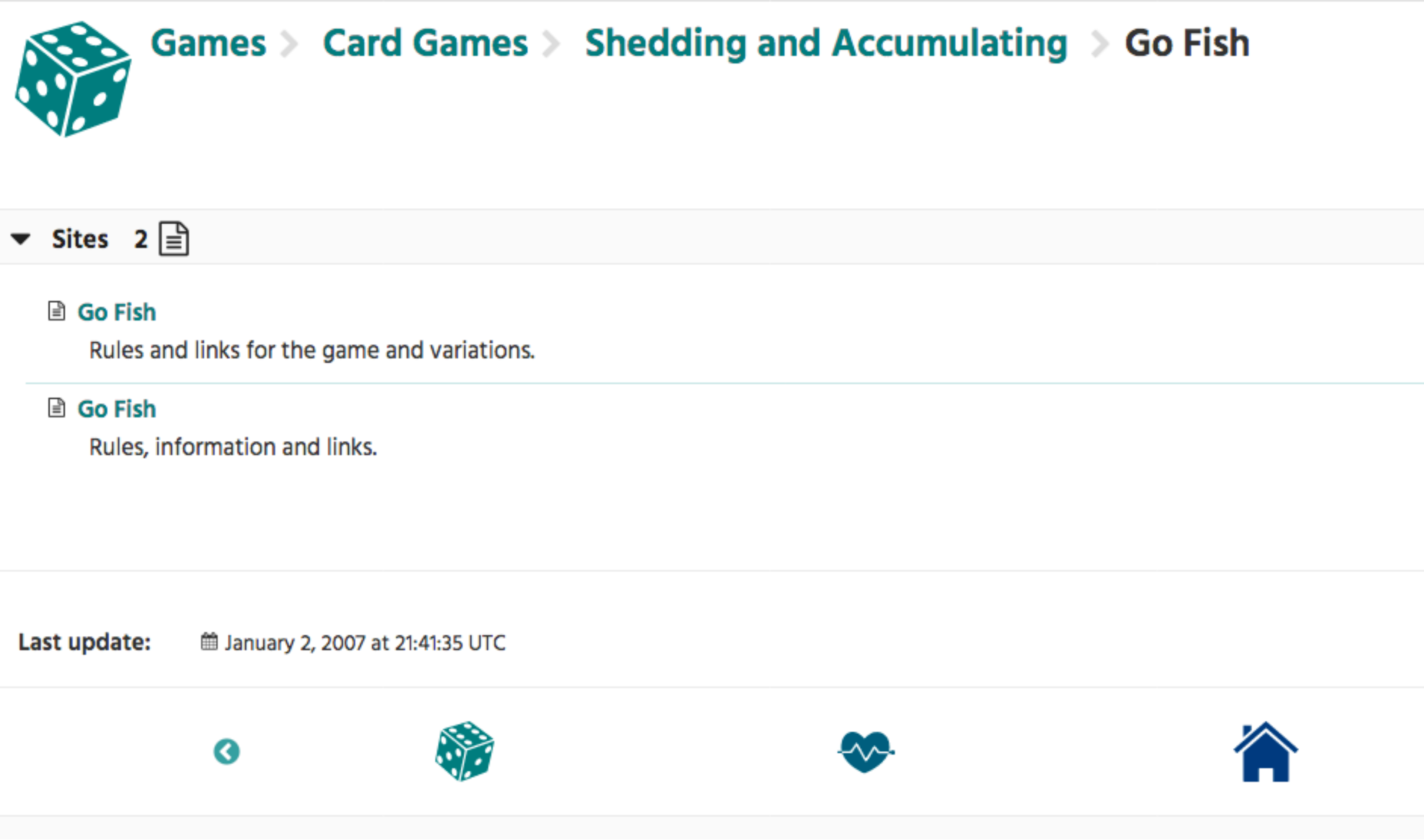
Related categories 6

- Business › Consumer Goods and Services › Toys and Games
- Kids and Teens › Games
- Recreation
- Science › Math › Recreations › Games and Puzzles
- Sports
- Sports › Fantasy

Other languages 62

Category editors: *sahbbg* *krazor*

Last update: March 2, 2017 at 18:45:56 UTC



Games > Card Games > Shedding and Accumulating > Go Fish

Sites 2

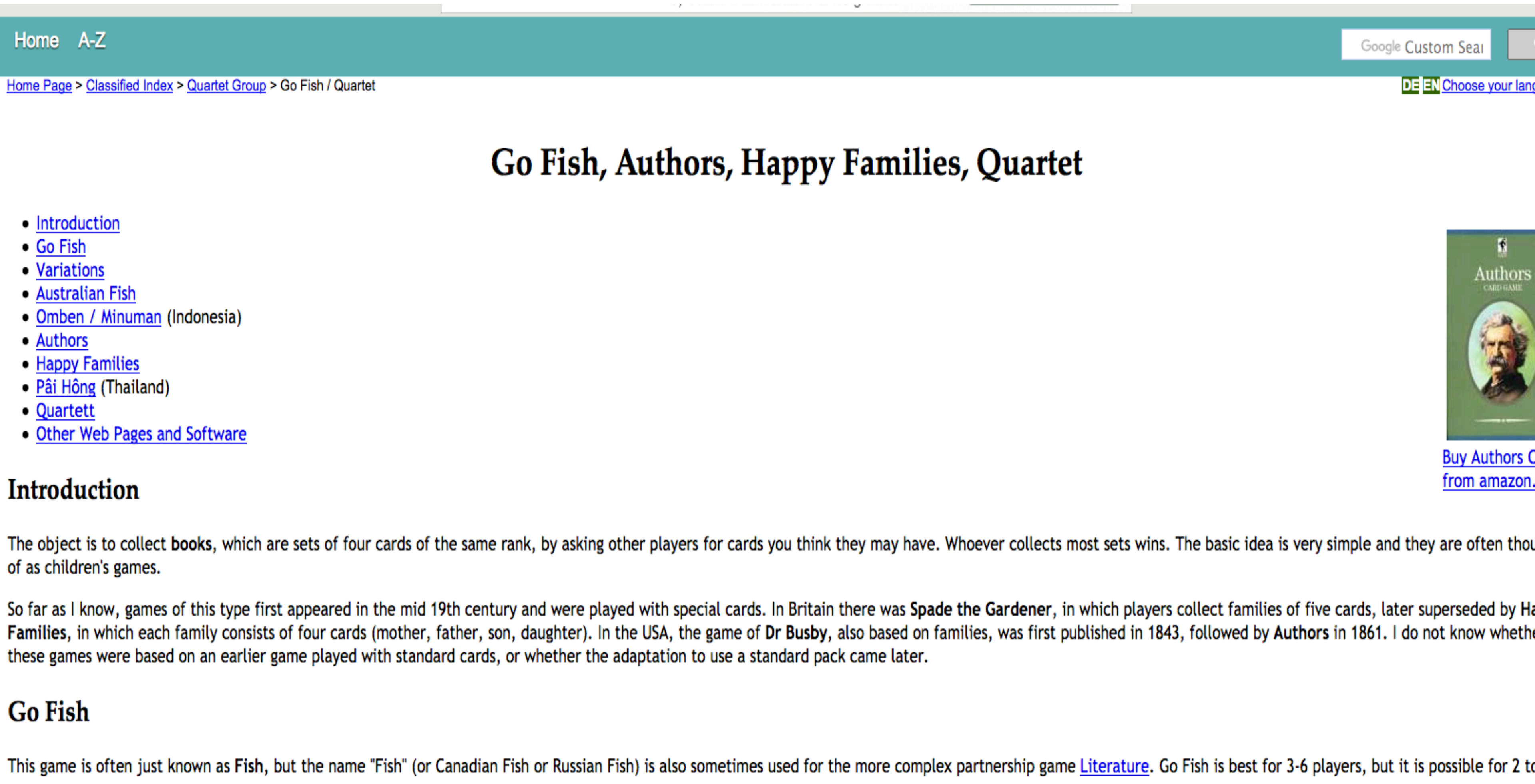
Go Fish

Rules and links for the game and variations.

Go Fish

Rules, information and links.

Last update: January 2, 2007 at 21:41:35 UTC



Home Page > Classified Index > Quartet Group > Go Fish / Quartet

DE EN Choose your lang

Go Fish, Authors, Happy Families, Quartet

- Introduction
- Go Fish
- Variations
- Australian Fish
- Omben / Minuman (Indonesia)
- Authors
- Happy Families
- Pâi Hông (Thailand)
- Quartett
- Other Web Pages and Software

Buy Authors C
from amazon.

Introduction

The object is to collect **books**, which are sets of four cards of the same rank, by asking other players for cards you think they may have. Whoever collects most sets wins. The basic idea is very simple and they are often thou of as children's games.

So far as I know, games of this type first appeared in the mid 19th century and were played with special cards. In Britain there was **Spade the Gardener**, in which players collect families of five cards, later superseded by **Ha Families**, in which each family consists of four cards (mother, father, son, daughter). In the USA, the game of **Dr Busby**, also based on families, was first published in 1843, followed by **Authors** in 1861. I do not know whethe these games were based on an earlier game played with standard cards, or whether the adaptation to use a standard pack came later.

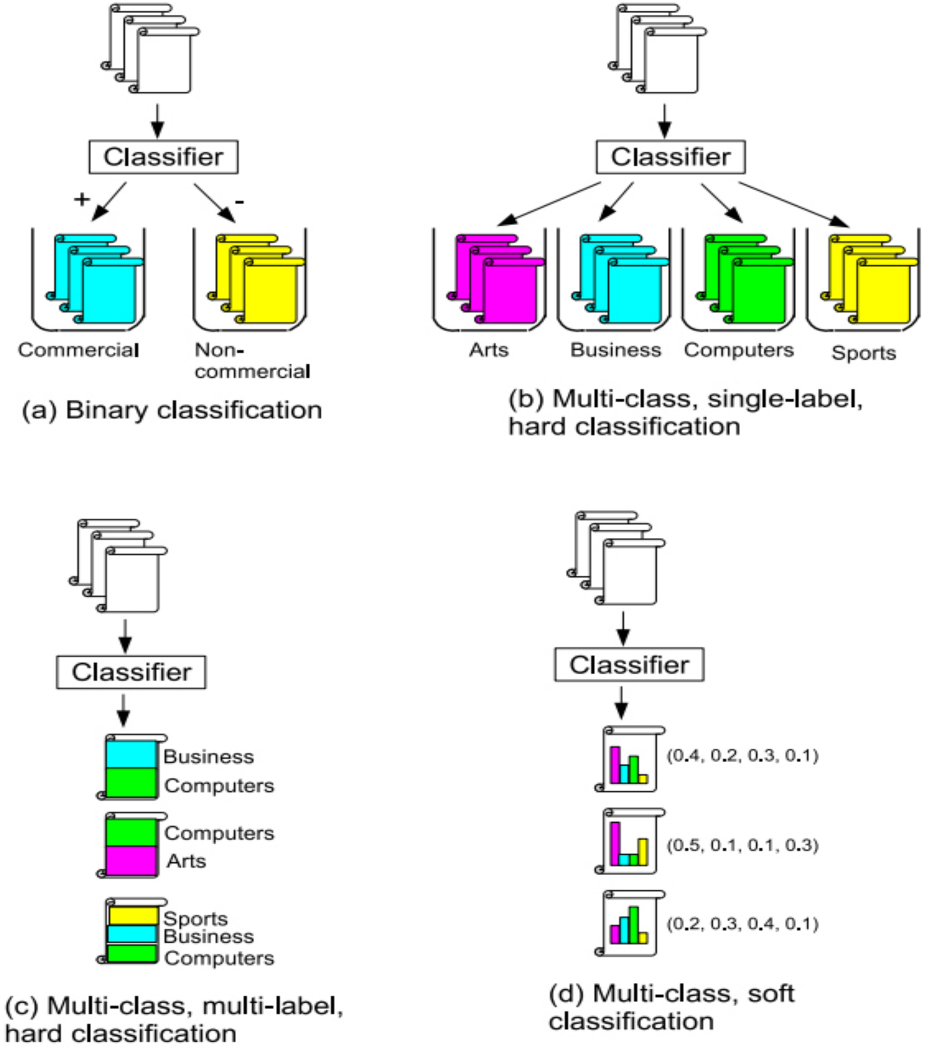
Go Fish

This game is often just known as **Fish**, but the name "Fish" (or Canadian Fish or Russian Fish) is also sometimes used for the more complex partnership game Literature. Go Fish is best for 3-6 players, but it is possible for 2 to

Web page classification

- Except for DMOZ, page categorization is "openly" used only by *focused search engines (eg Ebay, Amazon..)*
- The general problem of webpage classification can be divided into
  - **Subject classification**; subject or **topic** of webpage e.g. "Adult", "Sport", "Business".
  - **Function classification**; the **role** that the webpage play e.g. "Personal homepage", "Course page", "Commercial page".

Types of classification



(a) Binary classification

(b) Multi-class, single-label, hard classification

(c) Multi-class, multi-label, hard classification

(d) Multi-class, soft classification

– Hard vrs. Soft (multi-class) classification

Web Page Classification

- **Constructing and expanding web directories (web hierarchies)**
  - How are they doing?

Keyworder

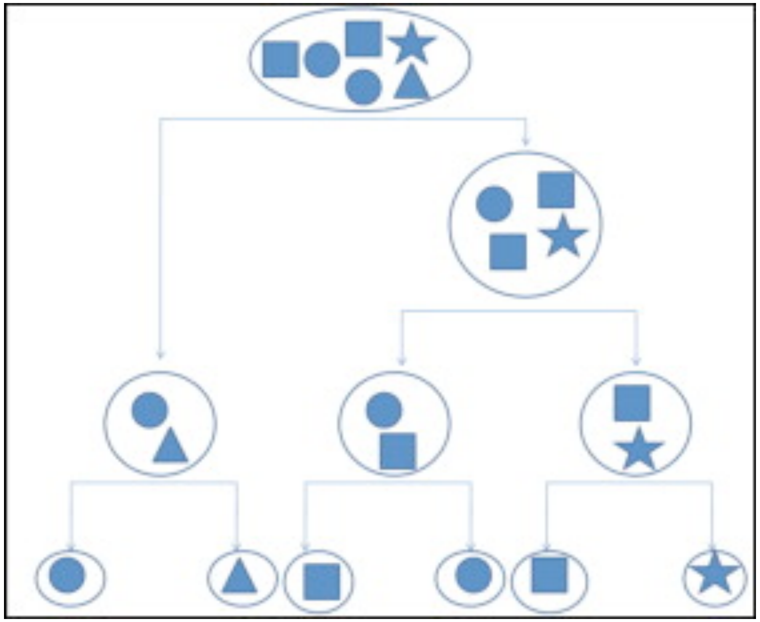
– By human effort

- July 2006, it was reported there are 73,354 editor in the dmoz ODP.

Automatic Document Classification

- Manual classification into a given hierarchy is labor intensive, subjective, and error-prone.
- Text categorization methods provide a way to automatically classify documents.
- Best methods based on training a *machine learning* (*pattern recognition*) system on a labeled set of examples (*supervised learning*).

Hierarchical Agglomerative Clustering



Strategies for hierarchical clustering generally fall into two types:

– **Agglomerative**: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

– **Divisive**: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

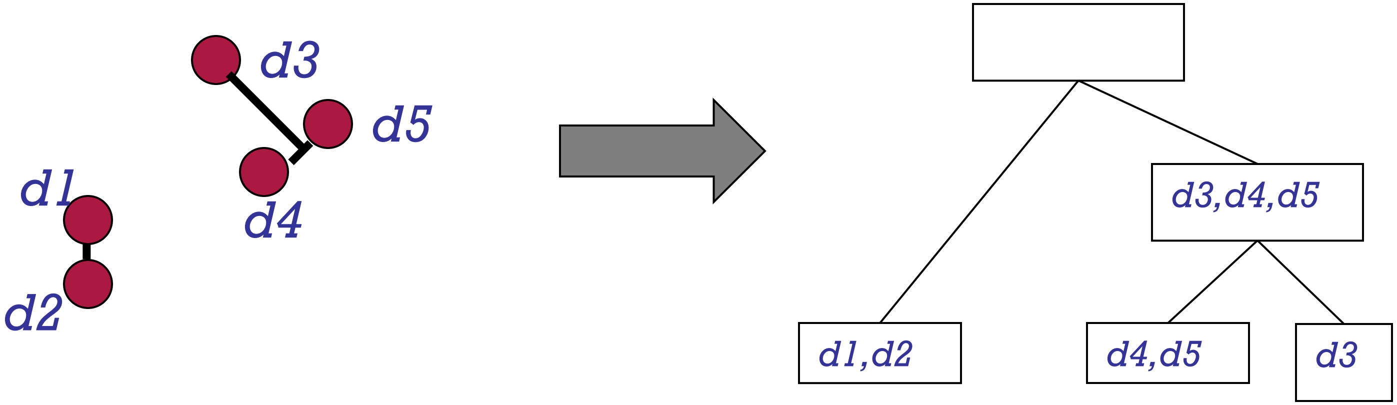
Hierarchical Agglomerative Clustering (HAC) Algorithm

- Start with all instances (web pages) in their own (singleton) cluster.
- Until there is only one cluster:
  - Among the current clusters, determine the two clusters, c_i and c_j, that are most similar.
  - Replace c_i and c_j with a single cluster $c_i \cup c_j$

"instance" is a web page, represented by a vector (see later)

Dendrogram: Document Example

- As clusters *agglomerate*, web pages are likely to fall into a hierarchy of "topics" or concepts.



Feature selection in HAC

- Problem: **how do we describe a page**?
- Bag-of-words vector not appropriate in this case (page may include off-topic information)
- Lower number of more descriptive features, based on two criteria:
  - On-page (**selected** features in the page)
  - Neibourgh features (selected features in the pages “pointing” at that page)

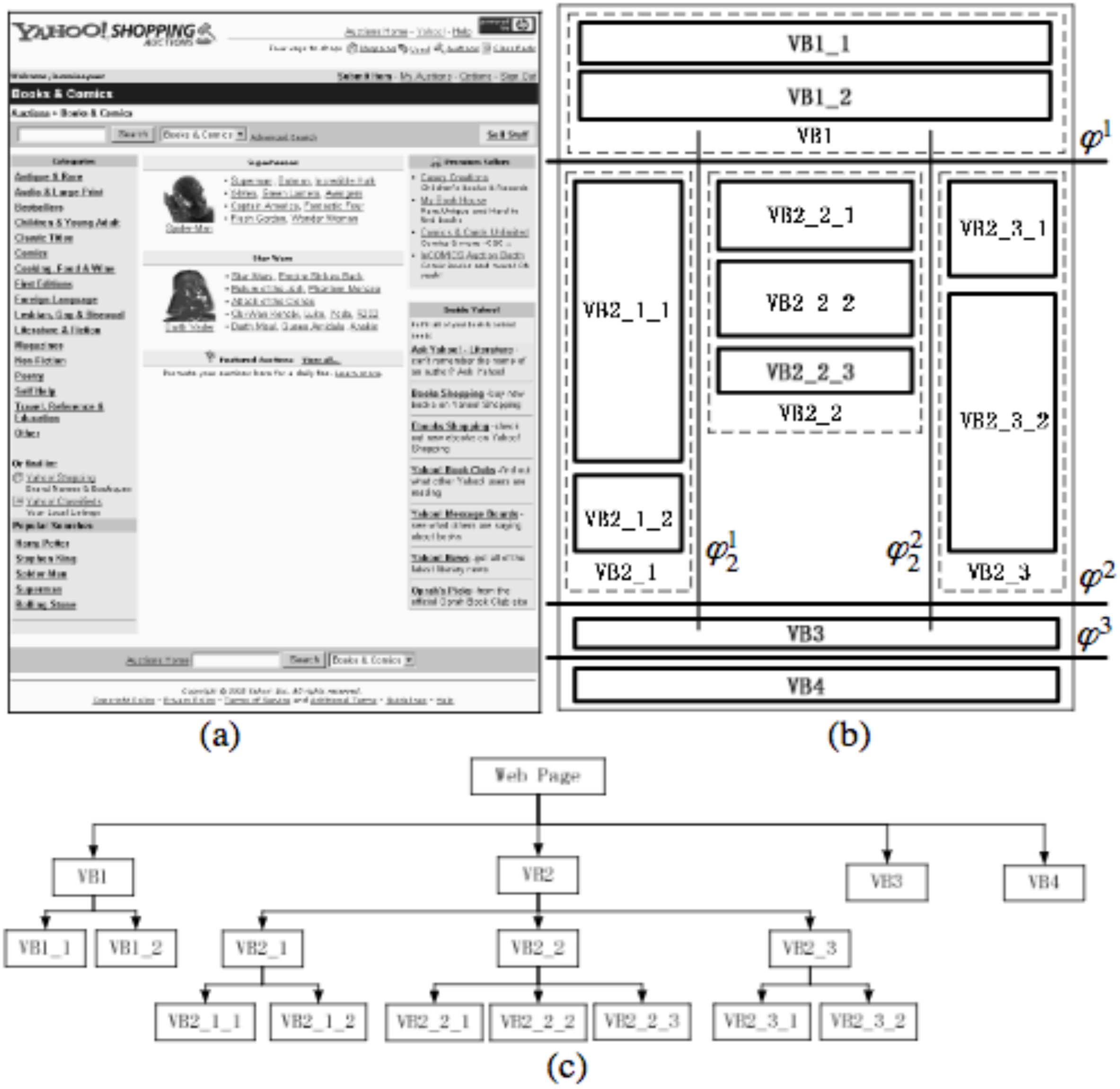
Features: On-page

- **Textual content and tags**
  - N-gram feature (n-gram= sequence of n consecutive words)
    - Also called n-words, e.g. “New York” is a biword.
    - In Yahoo!, they used **5-grams** features.
  - HTML tags or DOM (document object model)
    - Title, Headings, Metadata and Main text
      - Assigned each of them an arbitrary weight.
      - Now a day most of websites bare using Nested list (<ul><li>) which really help in web page classification (**Metatag, anchor tag**).

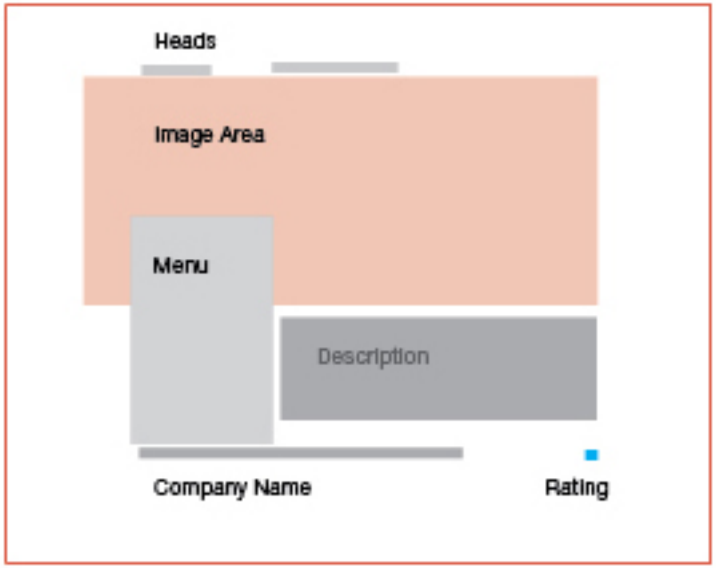
Features: On-page

- **Visual analysis**
  - Each webpage has two representations
    1. Text represented in HTML
    2. The visual representation rendered by a web browser
  - visual information is useful as well
    - Each webpage is represented as a hierarchical “Visual adjacency multi graph.”
    - In the graph each node represents an HTML object and each edge represents the spatial relation in the visual representation.
    - Challenge: web pages have templates and only a fragment of the content is relevant to the topic of the web page

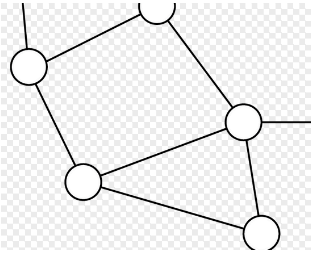
Visual graph representation

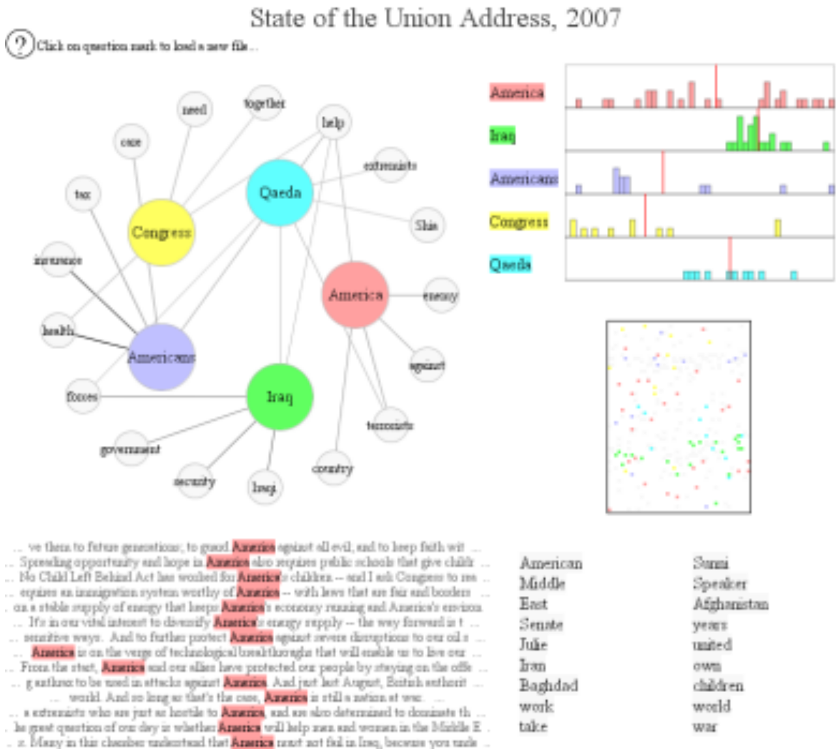


Visual analysis



Layout graph




State of the Union Address, 2007

? Click on question mark to load a new file...

... ve them to future generations; to guard America against all evil, and to keep faith wit ...
... Spreading opportunity and hope in America also requires public schools that give childr ...
... No Child Left Behind Act has worked for America's children -- and I ask Congress to rea ...
... equires an immigration system worthy of America -- with laws that are fair and borders ...
. on a stable supply of energy that keeps America's economy running and America's environ
... It's in our vital interest to diversify America's energy supply -- the way forward is t ...
... sensitive ways. And to further protect America against severe disruptions to our oil s ...
... America is on the verge of technological breakthroughs that will enable us to live our ...
... From the start, America and our allies have protected our people by staying on the offe ...
... g anthrax to be used in attacks against America. And just last August, British authorit ...
... world. And so long as that's the case, America is still a nation at war. ...
... a extremists who are just as hostile to America, and are also determined to dominate th ...
. he great question of our day is whether America will help men and women in the Middle E .
.. r. Many in this chamber understand that America must not fail in Iraq, because you unde ..

American
Middle
East
Senate
Julie
Iran
Baghdad
work
take

Sunni
Speaker
Afghanistan
years
united
own
children
world
war

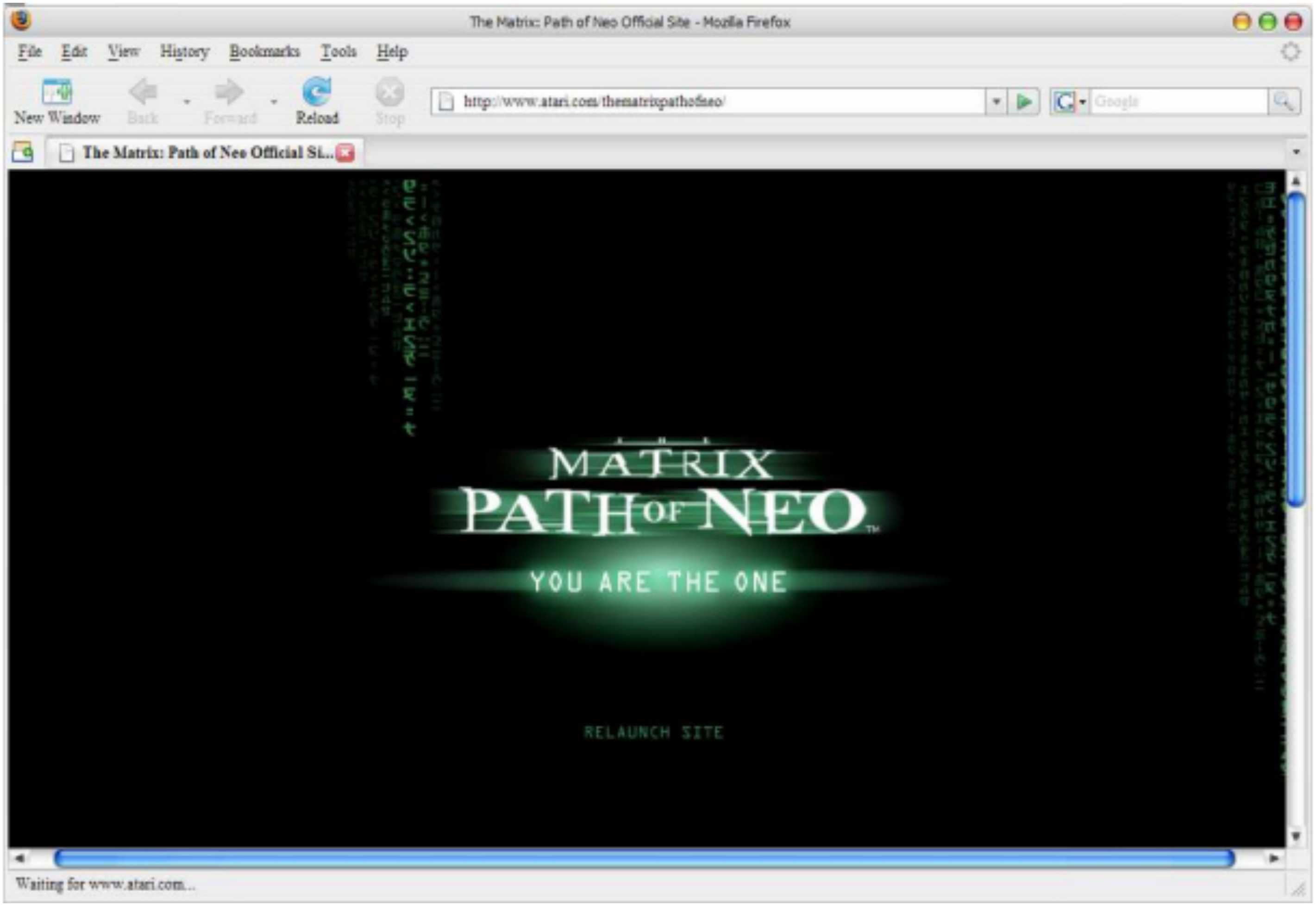
Content graph

Features: Neighbors Features

Features: Neighbors Features

- **Motivation**
  - Often in-page features are missing or unrecognizable

Example webpage which has few useful on-page features

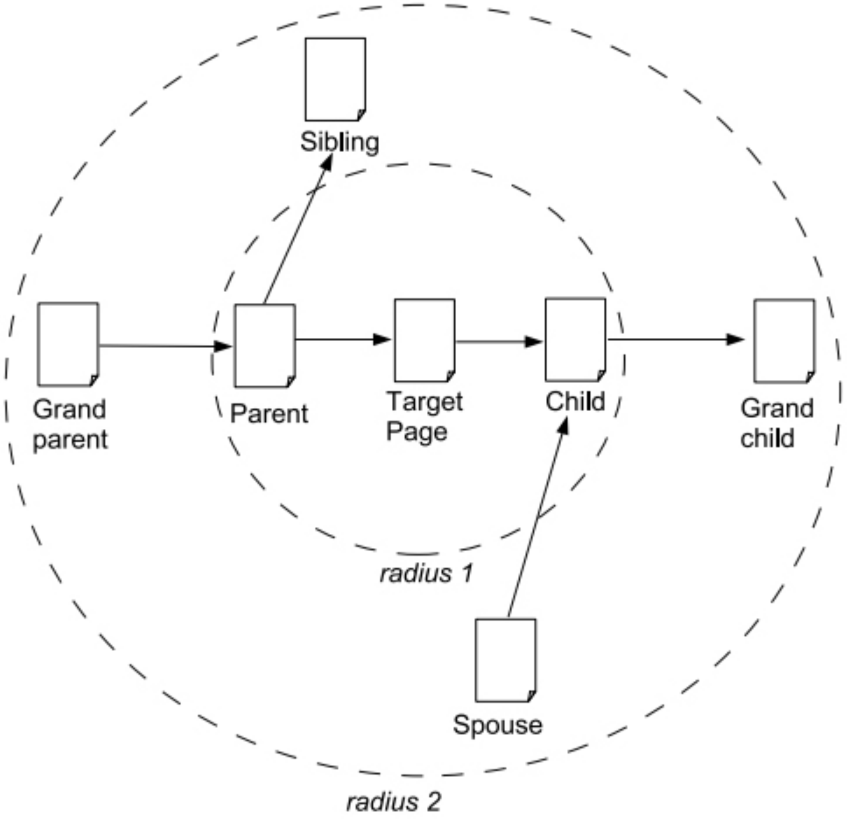

The Matrix: Path of Neo Official Site - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
New Window
Back
Forward
Reload
Stop
http://www.atari.com/thematrixpathofneo/
Google
The Matrix: Path of Neo Official Si...
MATRIX
PATH OF NEO
YOU ARE THE ONE
RELAUNCH SITE

Waiting for www.atari.com...

Features: Neighbors features

- **Underlying Assumptions**
  - When exploring the features of neighbors, some assumptions are implicitly made (like e.g. homophyly: pages point at similar ones).
  - The presence of many "sports" pages in the neighborhood of *Page-a* increases the probability of *Page-a* being in "Sport".
  - linked pages are more likely to have terms in common .
- **Neighbor selection**
  - Existing research mainly focuses on page **within two steps** of the page to be classified. (At the distance no greater than two).
  - There are six types of neighboring pages: **parent**, **child**, **sibling**, **spouse**, **grandparent** and **grandchild**.

Neighbors with in radius of two



Features: Neighbors features

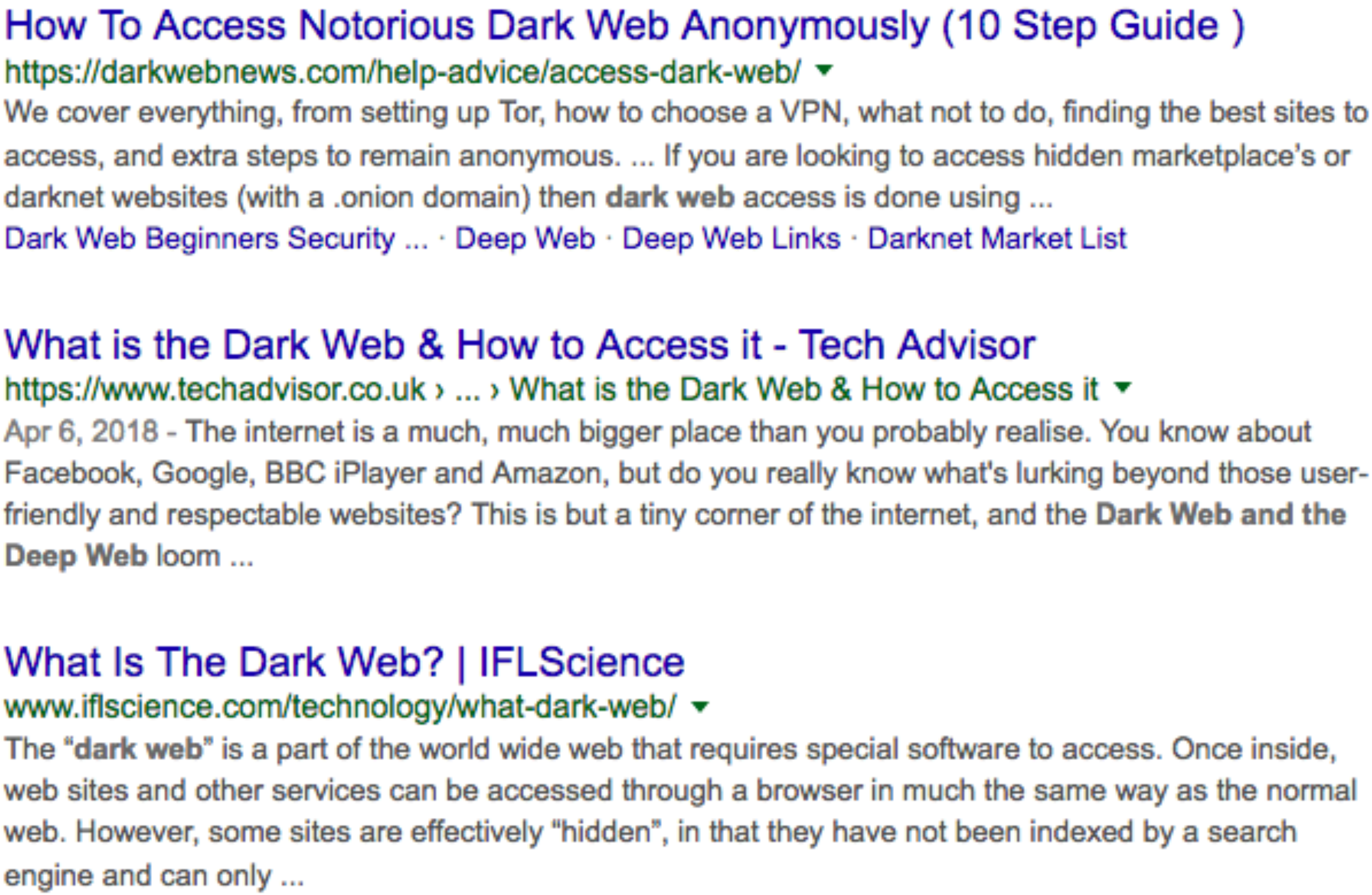
- **Neighbor selection cont.**
  - The text on the parent pages **surrounding the link** is used to train a classifier instead of text on the target page.
  - Using page title and **anchor text from parent pages** can improve classification compared a pure text classifier.

Features: Neighbors features

- **Neighbor selection cont.**
  - Summary
    - Using parent, child, sibling and spouse pages are all useful in classification, **siblings are found to be the best source.**
    - Using information from neighboring pages may introduce extra noise, should be used carefully.

Features: Neighbors features

- **Utilizing artificial links (implicit links)**
  - The hyperlinks are not the only one choice to find neighbors.
- What is implicit link?
  - **Connections between pages that appear in the results of the same query and are both clicked by users**.
- **Implicit link** can help webpage classification as well as hyperlinks.



How To Access Notorious Dark Web Anonymously (10 Step Guide)
https://darkwebnews.com/help-advice/access-dark-web/ ▾
We cover everything, from setting up Tor, how to choose a VPN, what not to do, finding the best sites to access, and extra steps to remain anonymous. ... If you are looking to access hidden marketplace's or darknet websites (with a .onion domain) then **dark web** access is done using ...
Dark Web Beginners Security ... · Deep Web · Deep Web Links · Darknet Market List

What is the Dark Web & How to Access it - Tech Advisor
https://www.techadvisor.co.uk › ... › What is the Dark Web & How to Access it ▾
Apr 6, 2018 - The internet is a much, much bigger place than you probably realise. You know about Facebook, Google, BBC iPlayer and Amazon, but do you really know what's lurking beyond those user-friendly and respectable websites? This is but a tiny corner of the internet, and the **Dark Web and the Deep Web** loom ...

What Is The Dark Web? | IFLScience
www.iflscience.com/technology/what-dark-web/ ▾
The “**dark web**” is a part of the world wide web that requires special software to access. Once inside, web sites and other services can be accessed through a browser in much the same way as the normal web. However, some sites are effectively “hidden”, in that they have not been indexed by a search engine and can only ...

Next..

- Ranking with Link Analysis (page Rank, HITS)