

# Web Search

## Introduction

# Outline

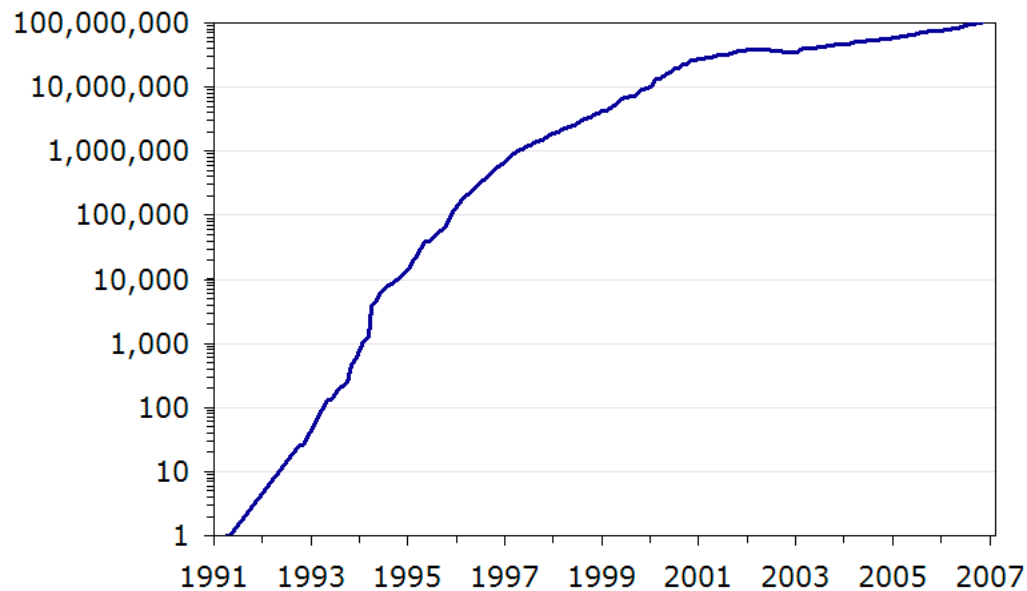
- History of www and some big players
- What are the challenges?
- Crawlers
- Ranking on the web (not just content)

# The World Wide Web

- The first prototype of the Internet was available since 1960, with the creation of ARPANET, or the Advanced Research Projects Agency Network. ARPANET used packet switching to allow multiple computers to communicate on a single network. With TCP/IP protocol, ARPANET in 1983 began to assume the shape of a «network of networks»
- The WWW was developed by **Tim Berners-Lee** in 1990 at CERN to organize research documents available on the Internet.
- Berners-Lee combined idea of documents available by FTP (File Transfer Protocol, to transfer files between computers) with the idea of *hypertext* to link documents. He created the first web site.
- Developed initial HTTP network protocol, URLs, HTML, and first “web server.”

# First web sites

- In 1990 there was one (Berners Lee)
- The year after, 9 more (they are millions now)



# What was the problem?

- Berners-Lee allowed to easily put «accessible» content on a computer (create a web site), and to simplify navigation from one computer to the other (through hyperlinks)
- However, as the number of websites were growing, it was very hard for people to find information by simply following hyperlinks.
- The subsequent «invention» was the **web search engine**.

# Web Search Engines, Web Browsers, Web Servers

- Browsers are software applications (programs) that are installed on computers and **allow users to access and view any web page** (previously indexed in the search engine), acting as intermediaries between the user and the Internet. To access, you need the URL!
- A Search engine is a **website** that allows to look for specific websites or information based on keywords, dates and other criteria.
- **Without a Web Search engine, we would have to know the exact URL of each page to be able to access them.** Web browsers allow us to access information on the Internet, but it is the search engine that tells it **how to access this information**.
- Finally, Web **servers** are computers that deliver (serves up) Web pages. Every Web server has an IP address and possibly a domain name.

# 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, and phrase 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).
- What was extraordinary is that their solution was almost **independent** from the number of available web pages!

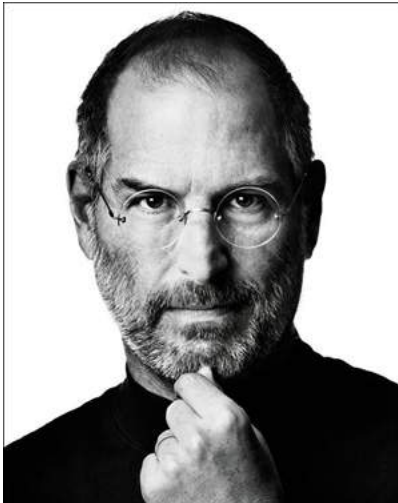




# Has it been only a technological revolution? Let's recap the steps



—Bill Gates: A PC on the table of everyone



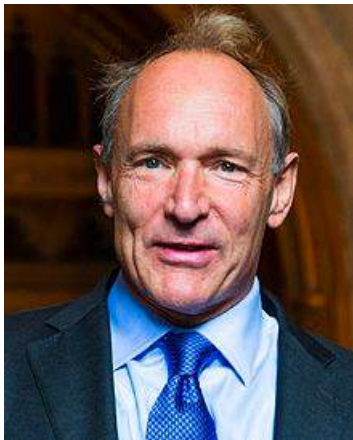
—Steve Jobs: A smartphone in the hands  
of everyone

# Has it been only a technological revolution?

## Let's recap the steps



- Robert Kahn invented INTERNET:  
How to connect devices worldwide (TCP/IP)



- Tim Berners-Lee invented the WWW:  
How to easily make information available  
on a device (web sites), and share it with  
others (hyperlinks)

# Has it been only a technological revolution? Let's recap the steps



—Brin and Page invented Google:  
How to easily FIND the information  
in the WWW (regardless of how  
many sites and links are there)



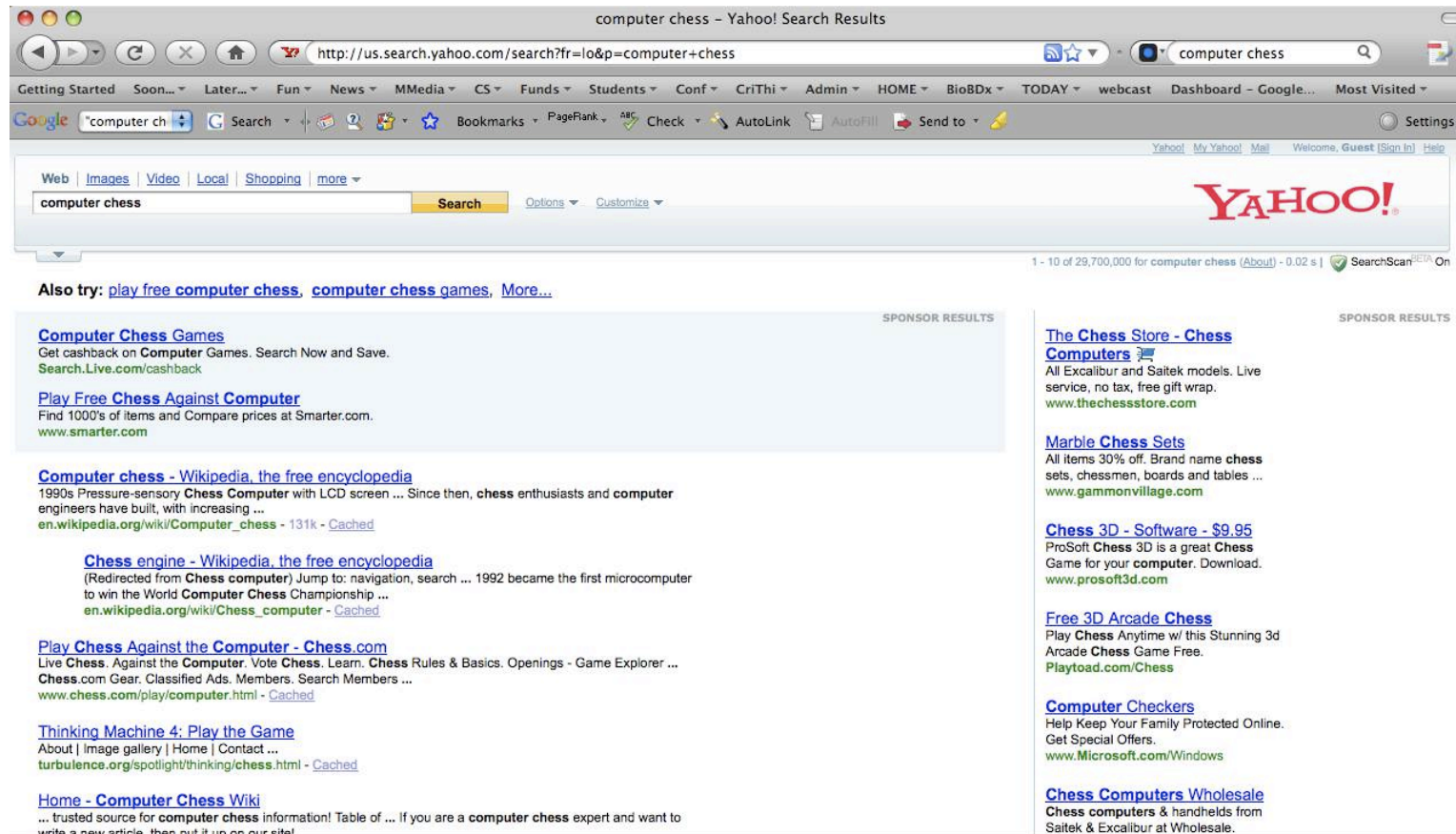
—Mark Zuckerberg invented Facebook:  
How to share comments and opinions  
with any of your friends worldwide

# It has been infact a CULTURAL revolution

- Everythingh happened in less than 15 years.
- What was the dream of they guys, Bill Gates, Steve Jobs, Robert Kahn, Tim Berners-Lee, Sergei Brin, Larry Page, Mark Zuckerberg?
- **Knowledge is power:** make the knowledge accessible to all, it is a dream of democracy (no more intermediaries between people and knowledge)
- After (almost) 15 more years, do you think the web is as hoped?
- Who are the “big players”?

# 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



# 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)

Rome 57° F



Enter search term...



[MAIL](#) [NEWS](#) [JOBS](#) [YELLOWPAGES](#) [TRIPOD](#) [DOMAINS](#) [CHAT](#) [WEATHER](#) [WHOWHERE?](#) [ANGELFIRE](#)



# Google

- Verb “google” has become synonymous with “searching for information on the web”.
- Continuously rises the bar on search quality
- Has been the most popular search engine in the last few years.
- Is the most innovative and dynamic.

The screenshot shows a Google search interface. The search bar contains the text "what is the average incubation time of COVID-19?". Below the search bar, there are navigation links: "All", "Images", "News", "Shopping", "Videos", "More", "Settings", and "Tools". The search results show "About 4,890,000 results (0.43 seconds)". Under the heading "Top stories", there are two featured articles. The first article is titled "Distancing can save lives as COVID-19 incubation, recovery period varies" by Greeley Tribune, dated 1 day ago, with a small image of a person wearing a mask. The second article is titled "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a ..." by The Lancet, dated 13 hours ago, with a small image of a line graph. Below the top stories, there is a link "→ More for what is the average incubation time of COVID-19?". At the bottom, there is a breadcrumb trail "www.jwatch.org › 2020/03/13 › covid-19-incubation-period-update" followed by the title "COVID-19 Incubation Period: An Update - NEJM Journal Watch". The date "Mar 13, 2020" is shown, followed by the text "In the resulting models, estimated median incubation time (IT) of COVID-19 was 5.1 days; mean IT was 5.5 days. For 97.5% of infected persons, ...".


Google


what is the average incubation time of COVID-19?

All Images News Shopping Videos More Settings Tools

About 4,890,000 results (0.43 seconds)

Top stories

 Distancing can save lives as COVID-19 incubation, recovery period varies  
Greeley Tribune · 1 day ago

 The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a ...  
The Lancet · 13 hours ago

→ More for what is the average incubation time of COVID-19?

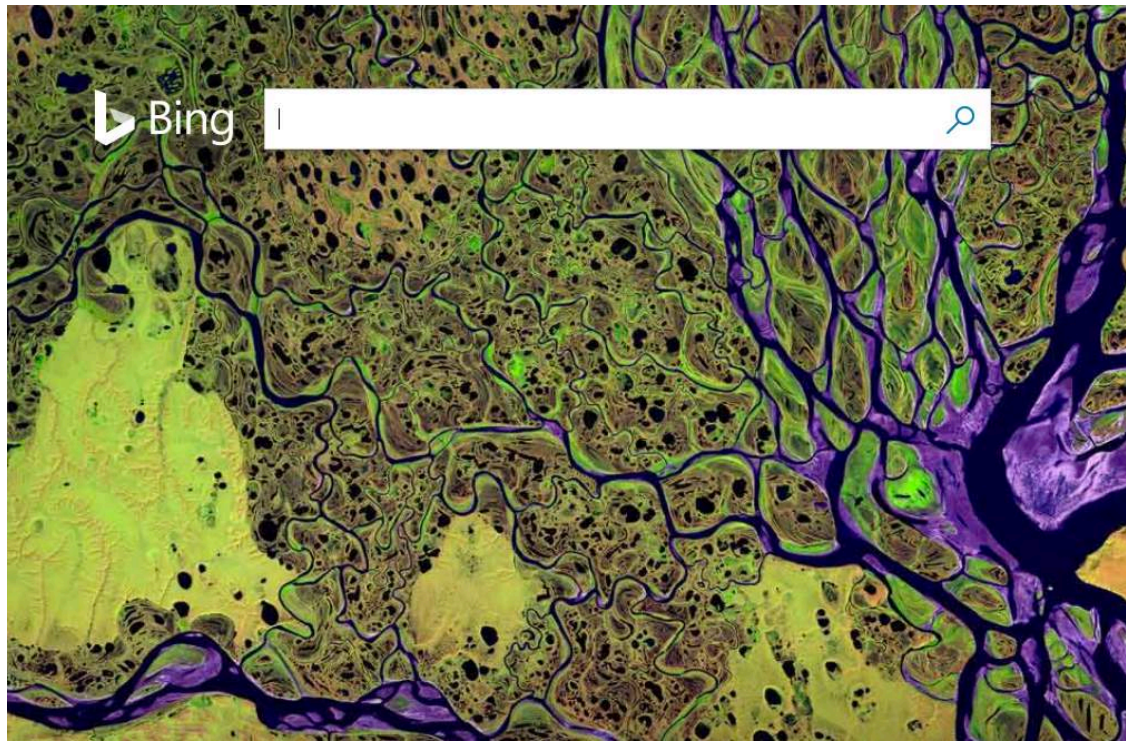
www.jwatch.org › 2020/03/13 › covid-19-incubation-period-update

COVID-19 Incubation Period: An Update - NEJM Journal Watch

Mar 13, 2020 - In the resulting models, estimated median incubation time (IT) of COVID-19 was 5.1 days; mean IT was 5.5 days. For 97.5% of infected persons, ...

# 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)

- Specialised in natural language question answering.
- Search driven by [Teoma](#).



what is the average incubation time of COVID-19?

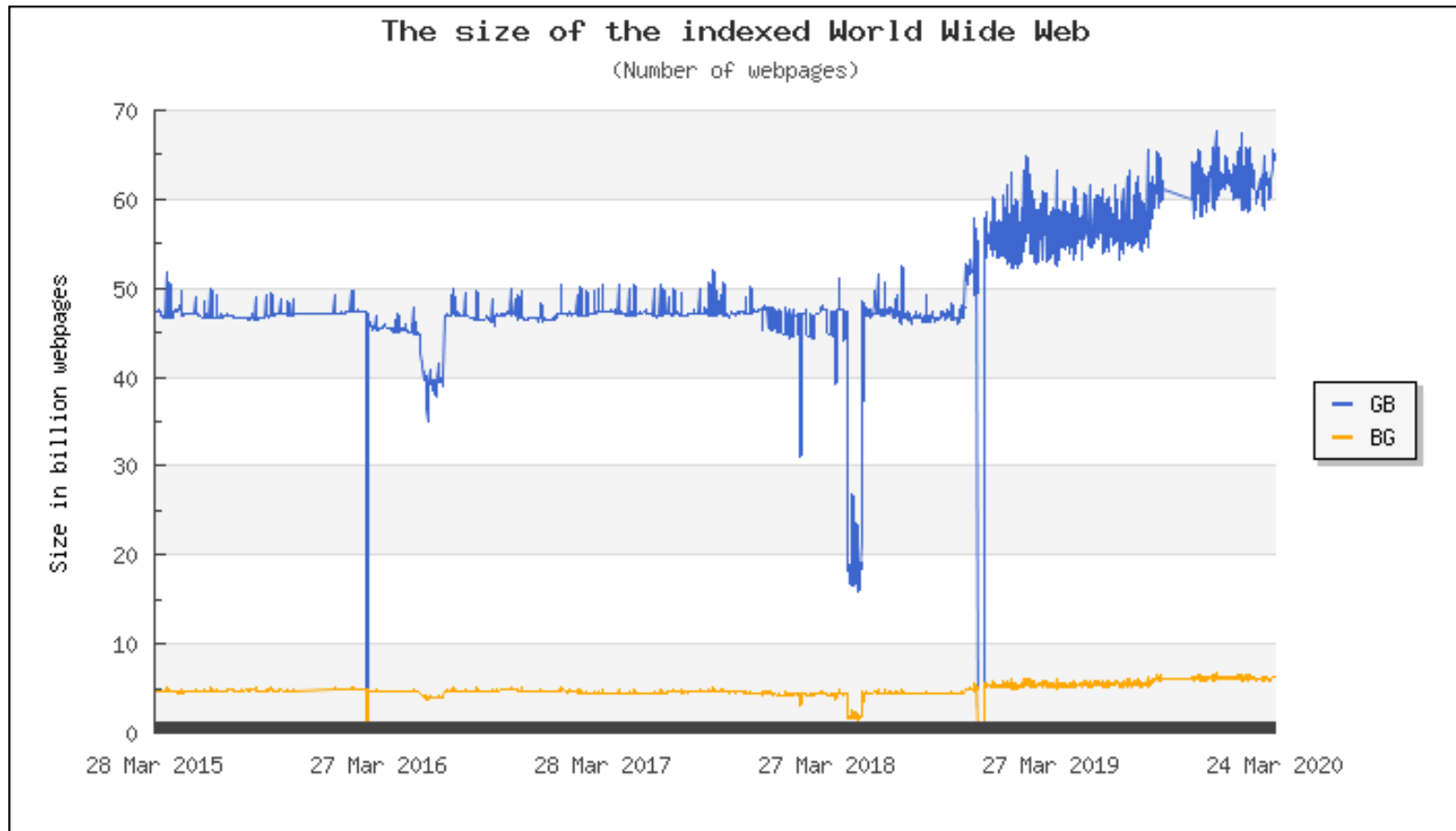


# Web Challenges for IR

- **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.

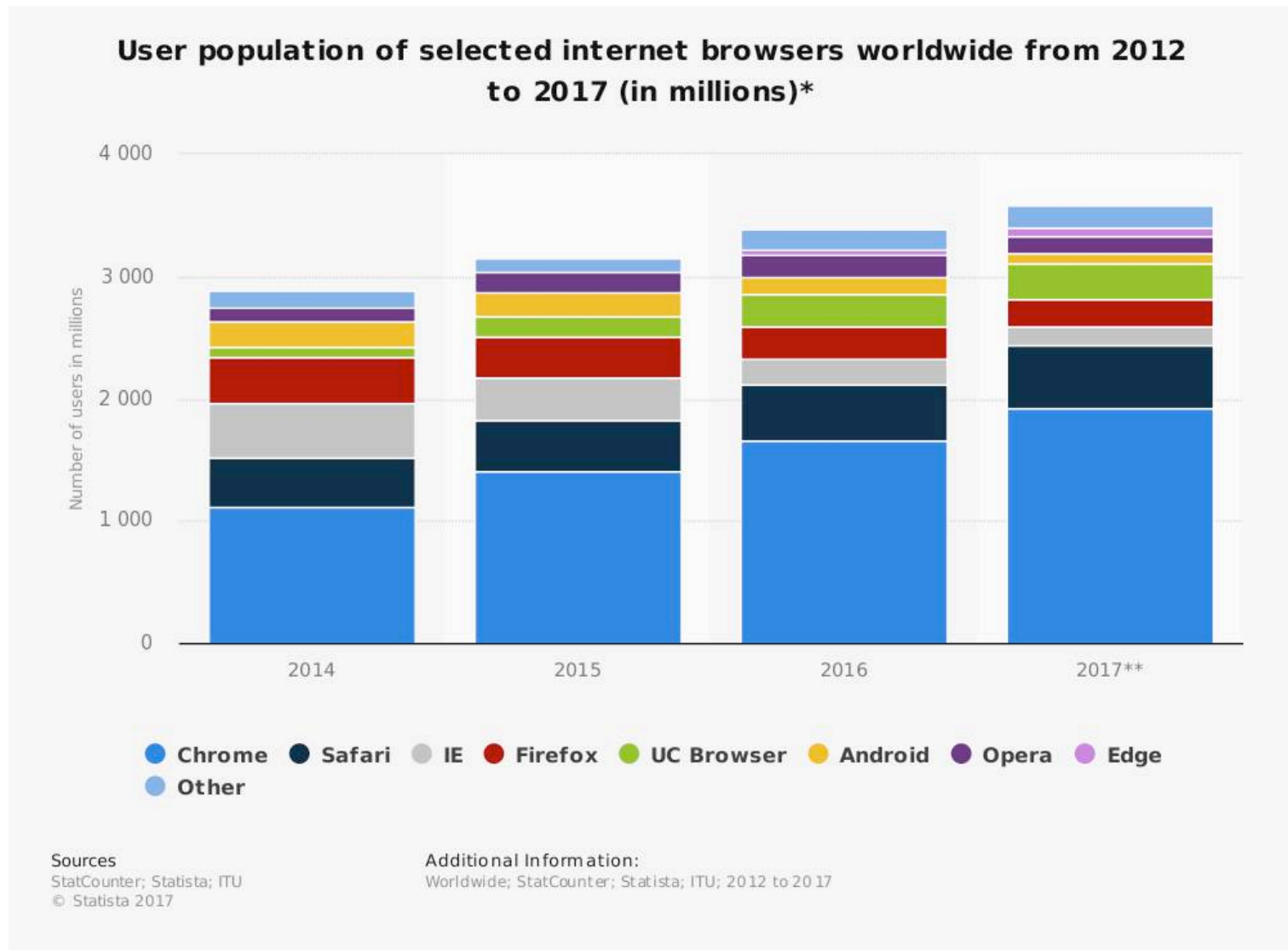
Big, How much Big?

# Indexed pages

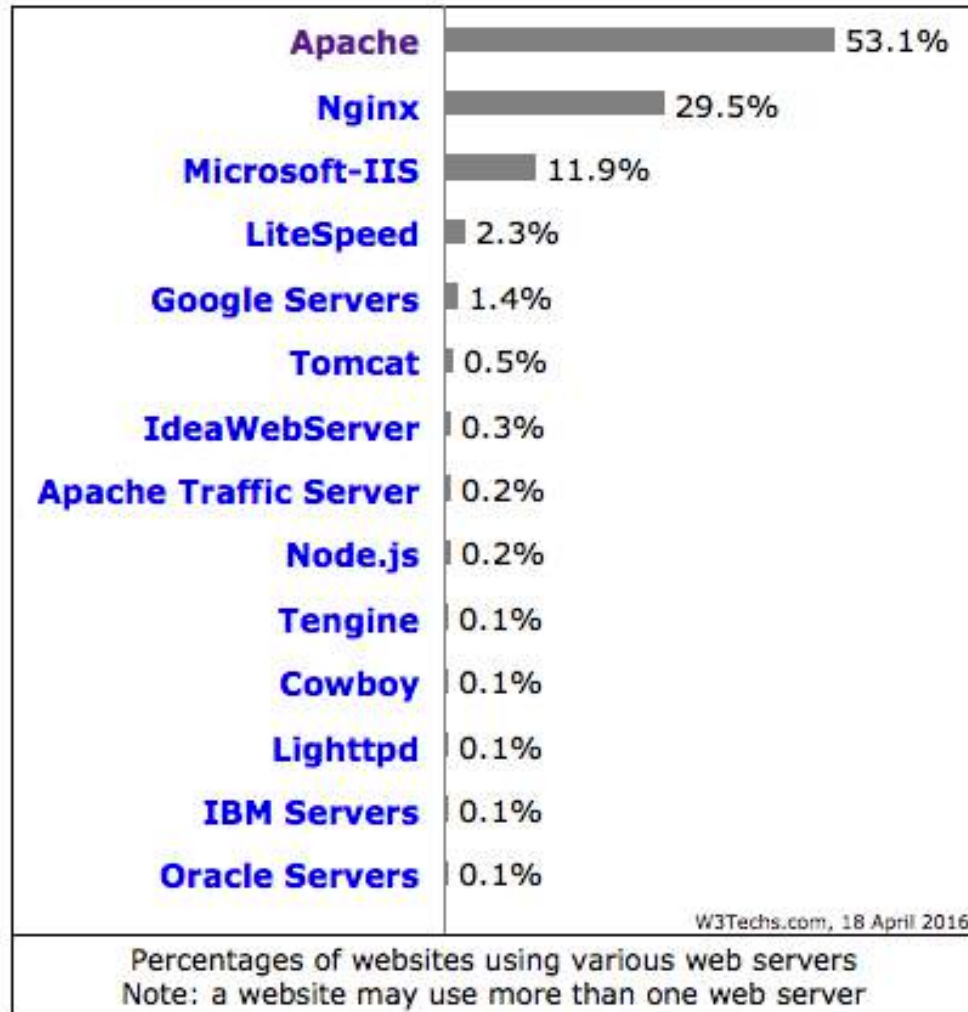


—Note periodic massive deindexing – latest was reported to be due to a bug

# Market share per Web Browsers



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



# Relevant statistics on the web

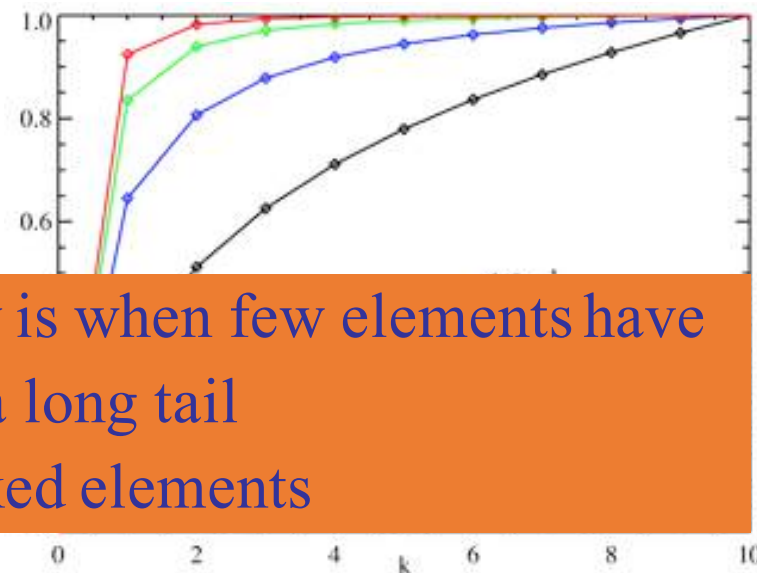
- How many connections per node (indegree, outdegree)?
- How many pages per web site?
- Update frequency per site
- Number of click-through per web site
- Zipfian law dominates!

# Zipfian law

- 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 (= value of some parameter for an element in the population);
- $s$  be the value of the exponent characterizing the distribution

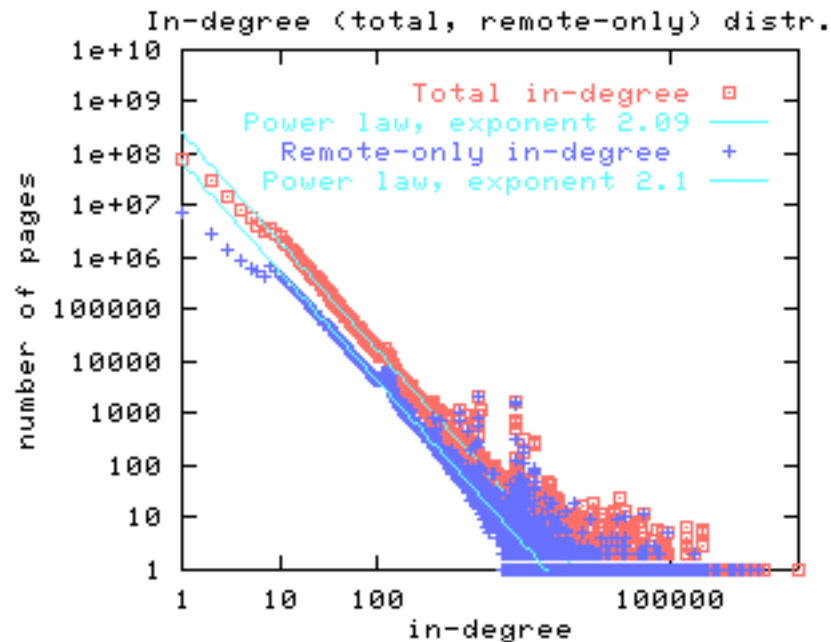




# Zipf's Law dominates on the Web

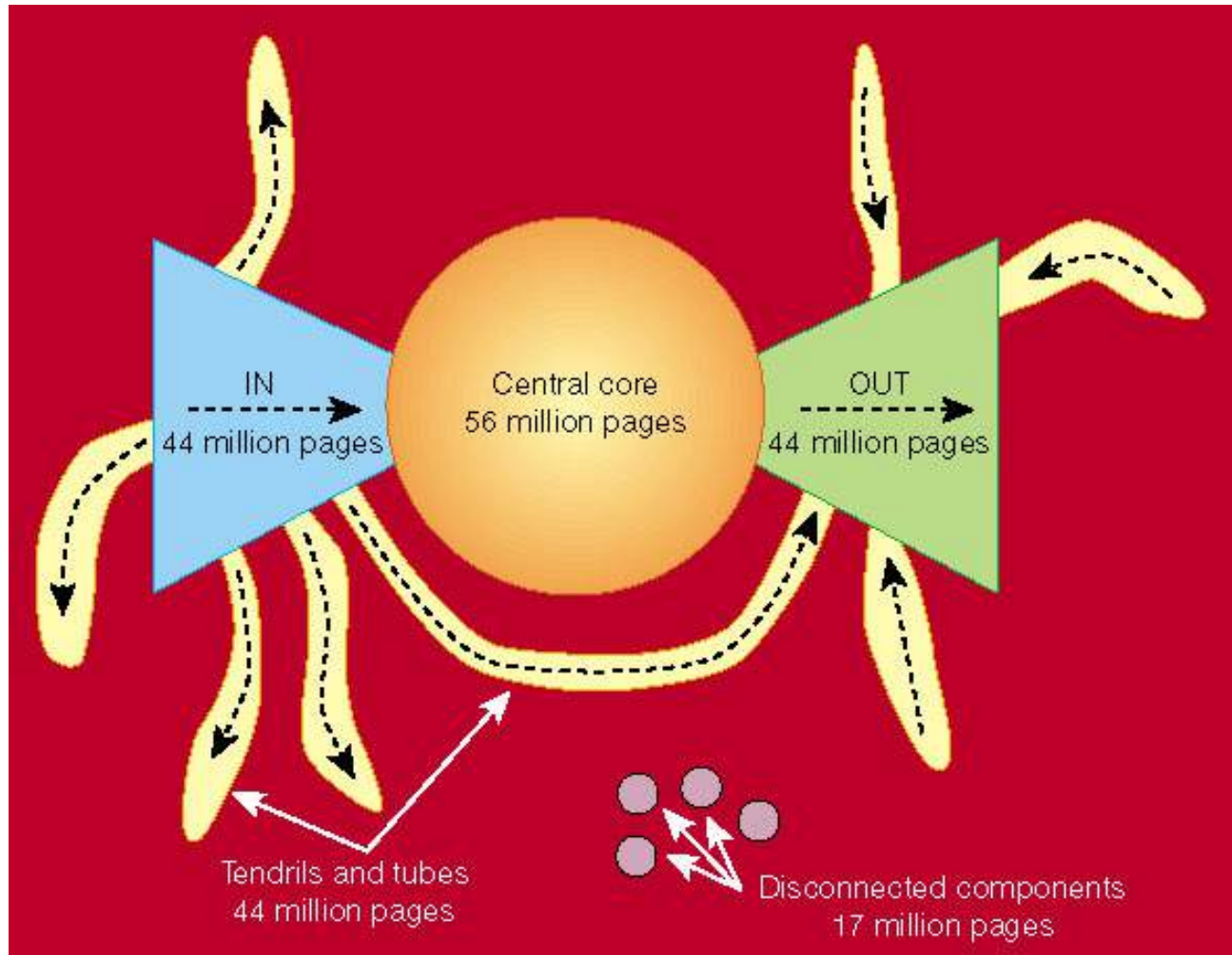
- Number of in-links/out-links to/from a page has a Zipfian distribution (frequency of pages with outdegree= $k$  as a % of the full population  $N$ ).
- Length of web pages has a Zipfian distribution (frequency of pages with length= $k$  as a % of the full population  $N$ ).
- Number of hits to a web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

# 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>

# Crawling, Ranking on the www



—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

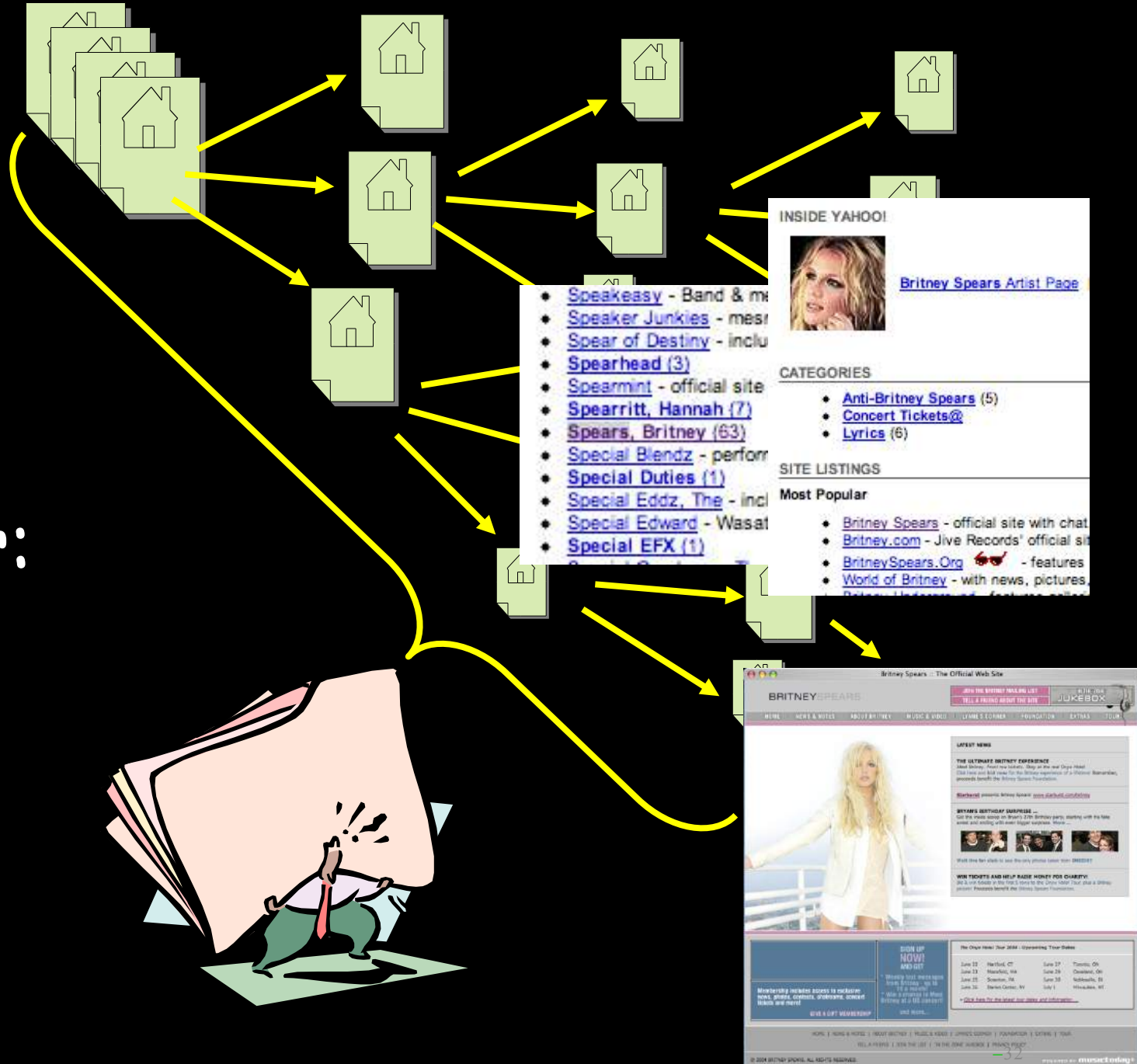
A **Web crawler** is a software application which systematically browses the Web, for the purpose of Web indexing.

# 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





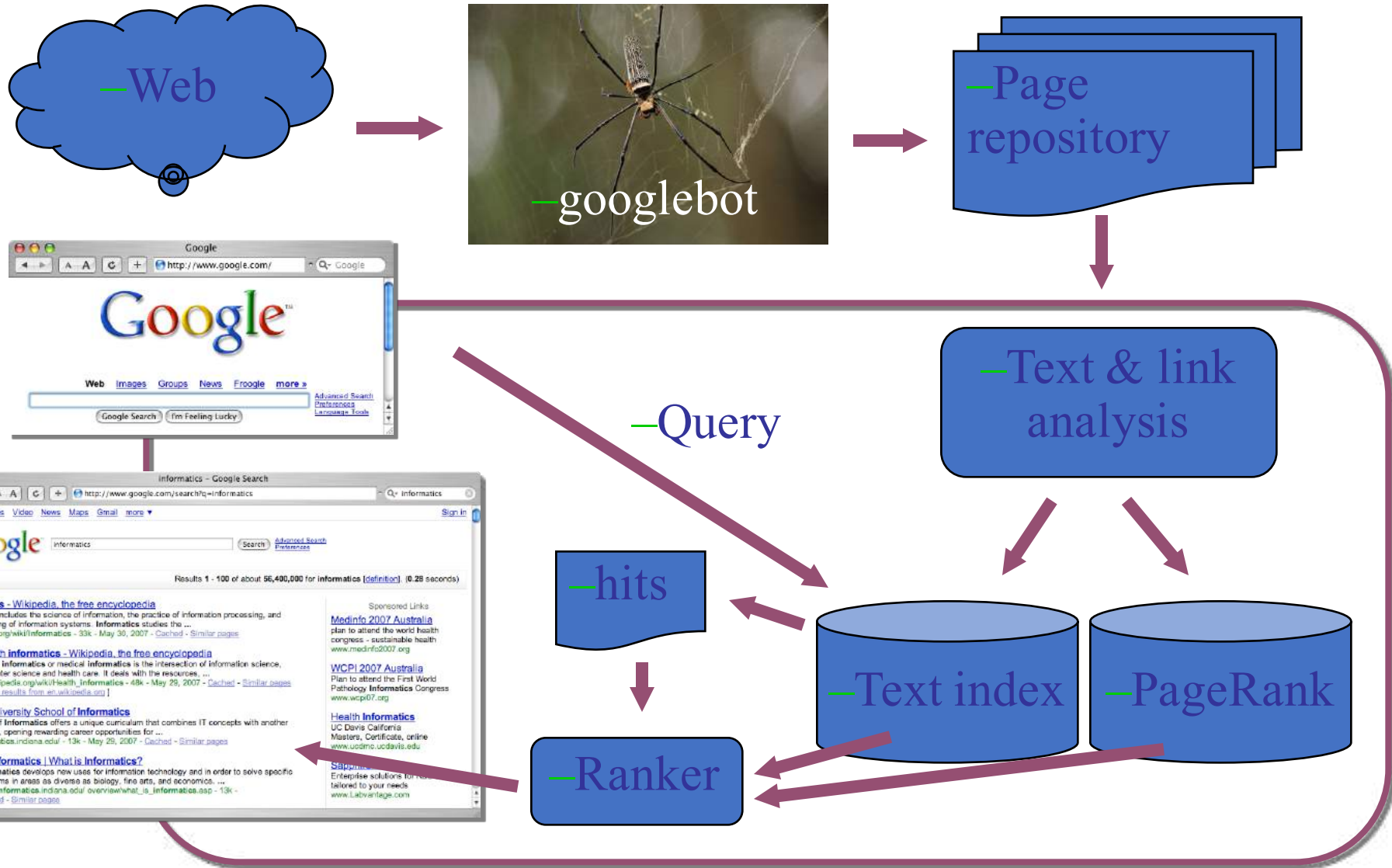
# Many names

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

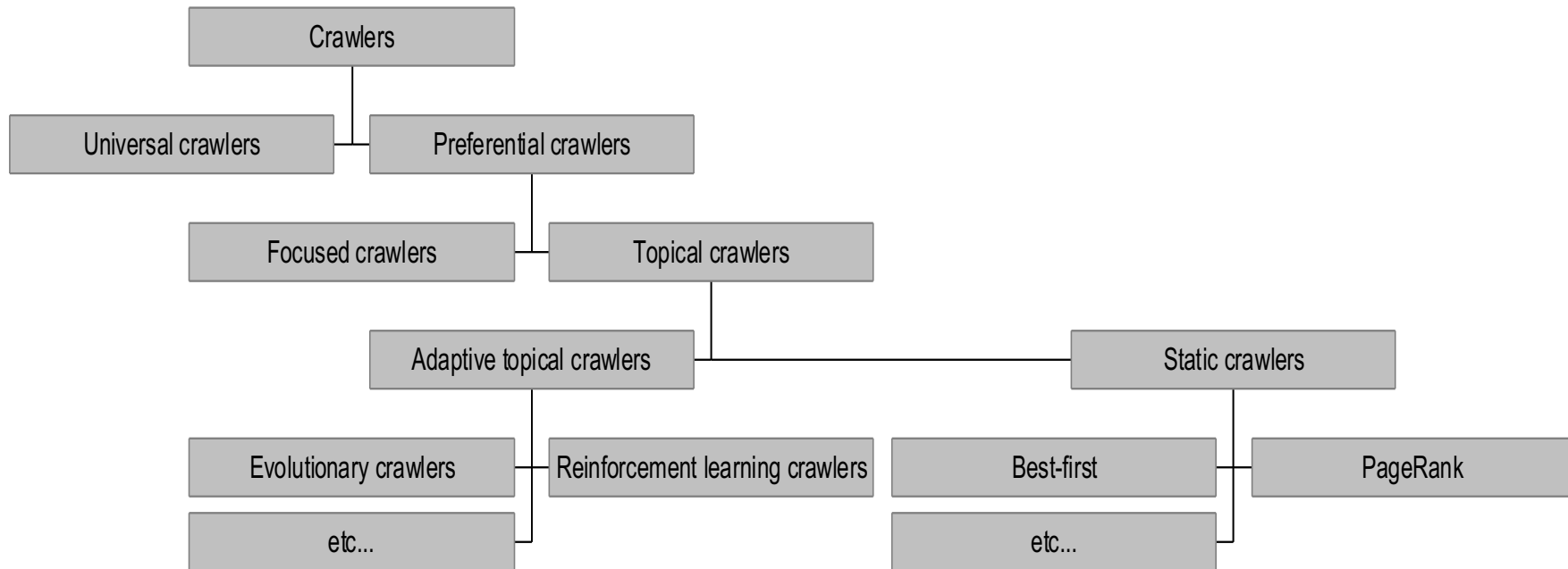
# Crawlers vs Browsers vs Scrapers

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

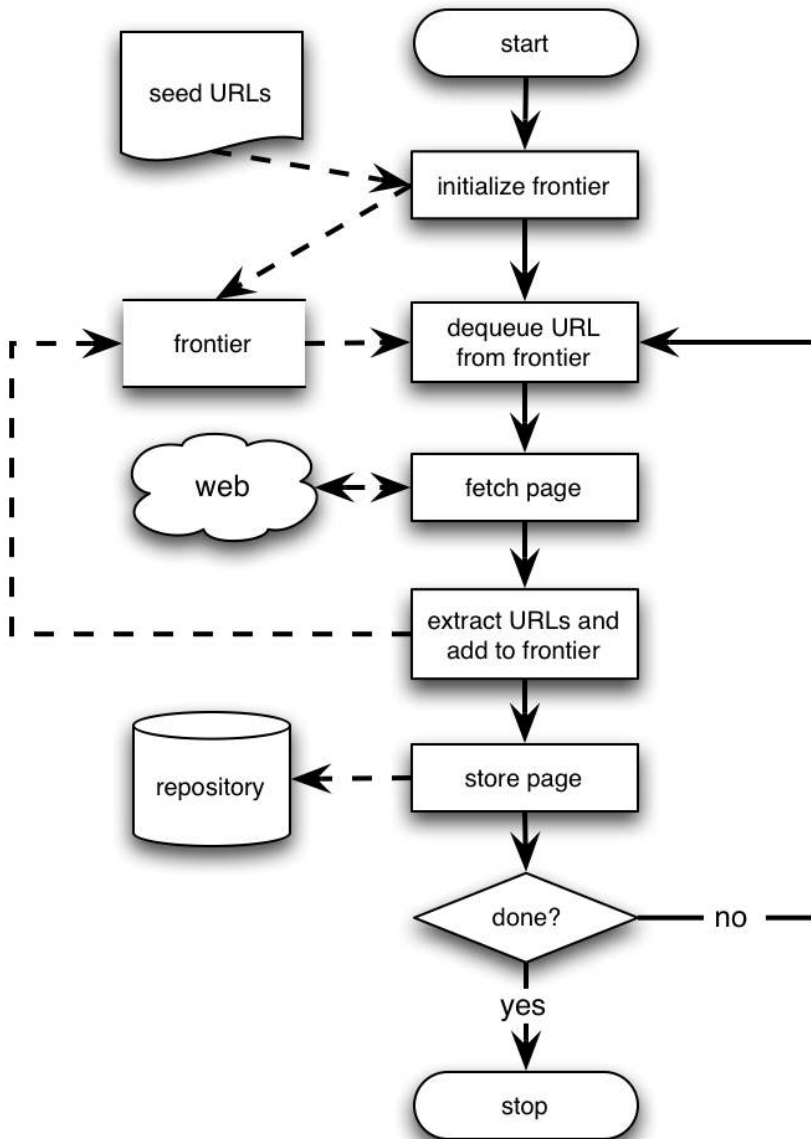
# A crawler within a search engine (e.g. Google)



# Types of crawlers



- **Universal**: support universal web search engines – only content is considered (eg, cosin-similarity)
- **Preferential**: Selective bias toward some pages, eg. most “relevant”/topical, closest to seeds, most popular/largest, PageRank, highest rate/amount of change, etc..
- Universal **really do not exist anymore..** (not on the web) All are preferential (too many pages on the web to avoid preferential criteria)

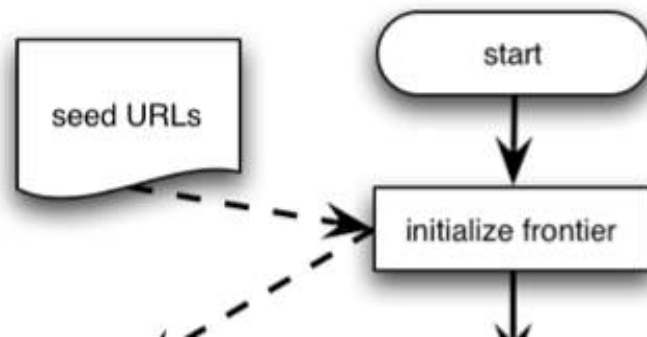


## Basic crawlers

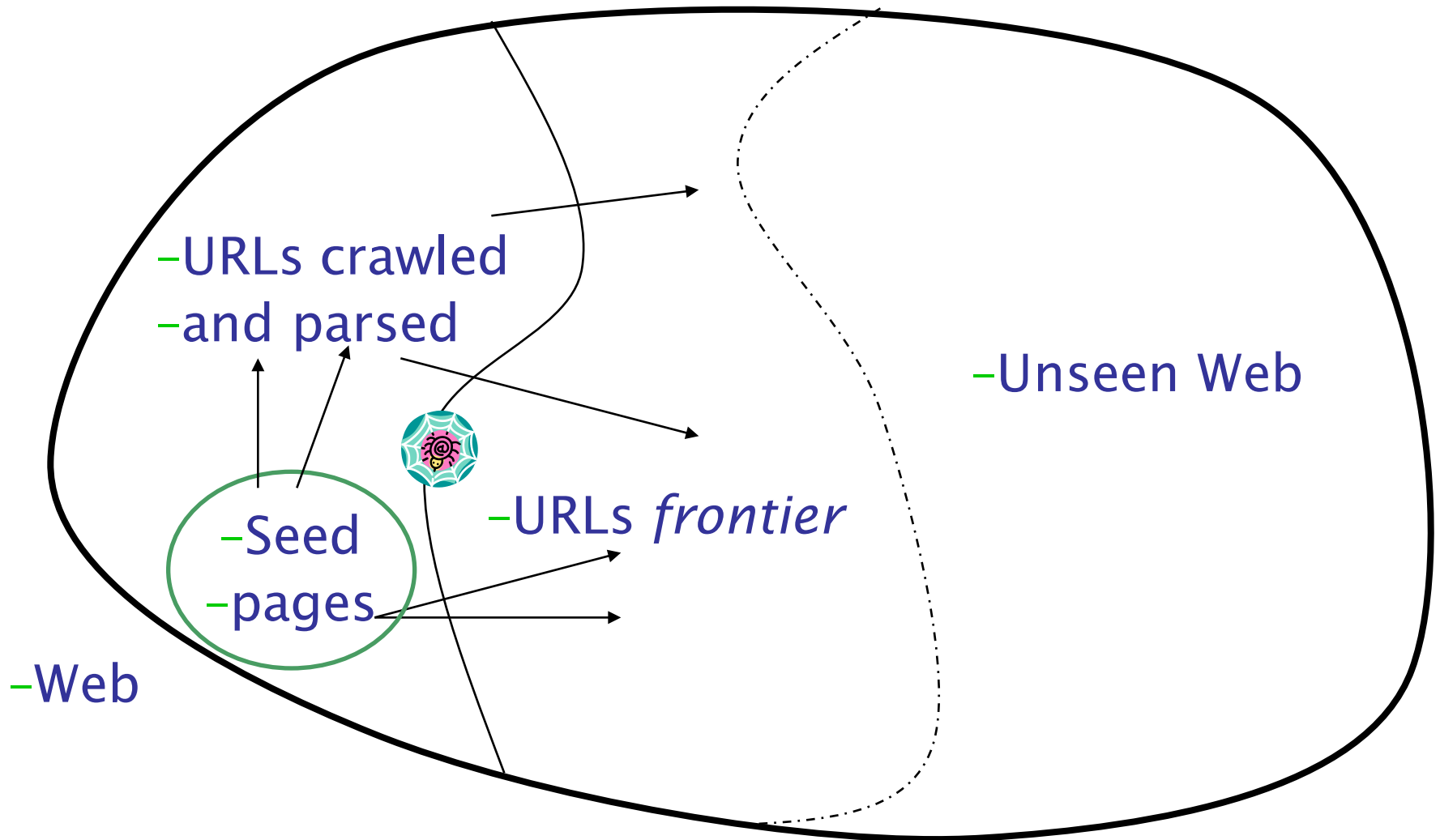
- This is a **sequential universal** crawler (all pages are the same)
- **Seeds** can be any list of starting URLs
- Order of page visits is determined by **frontier** data structure
- **Stop** criterion can be anything

# URL frontier

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



# Basic crawler



# Preferential: 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
- Focused Crawler uses *text classifier* to decide whether a page is “on topic” before indexing it



# Preferential: Topical Crawlers

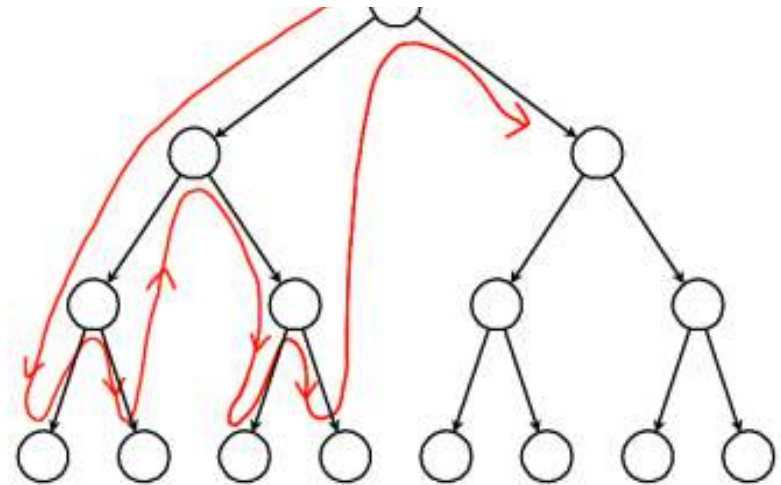
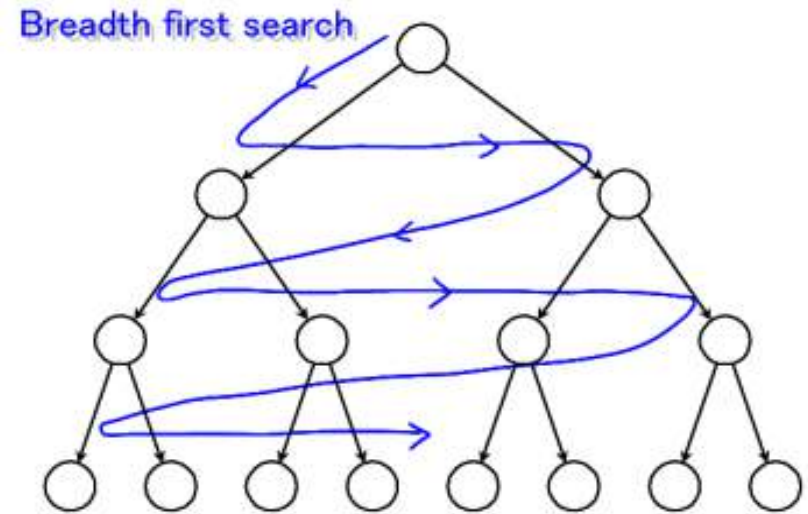
- Topical crawlers are a way to address the **scalability limitations of universal search engines**, by distributing the crawling process across users, queries, or even client computers.
- Strategies are used to assign preferences to pages, **other than similarity with query**.
- Machine learning techniques are employed for link analysis, so that a page's linkage to other pages, together with its content, could be used to estimate its relevance (ex. Page rank by Google)
- They can be STATIC if the strategy to rank is static ADAPTIVE if the strategy depends, e.g., on user past behaviour, etc.
- Will see examples of static and adaptive ranking next

# 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

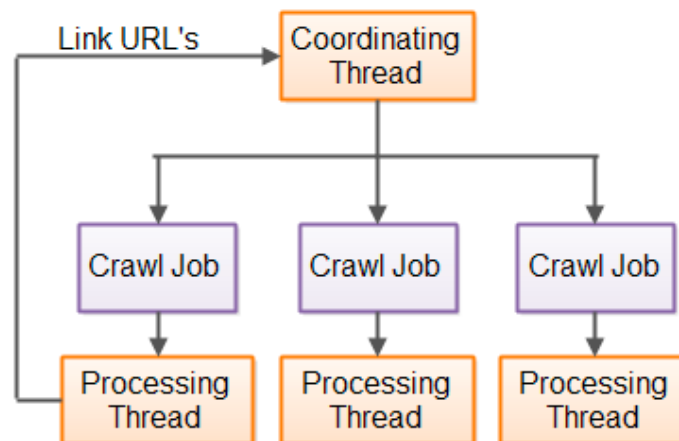
# 1. Web Graph visiting policies

- 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 (2)

- 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

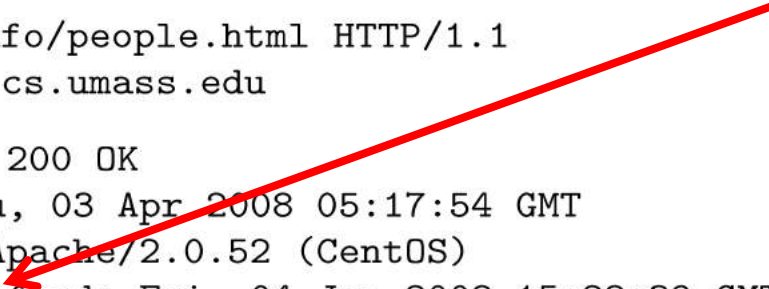
- Web pages are constantly being added, deleted, and modified
- Web crawler must continually **revisit pages** it has already crawled to see if they have changed in order to maintain the *freshness* of the document collection
  - *stale* copies no longer reflect the real contents of the web pages

## 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

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  
Server response: 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:

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

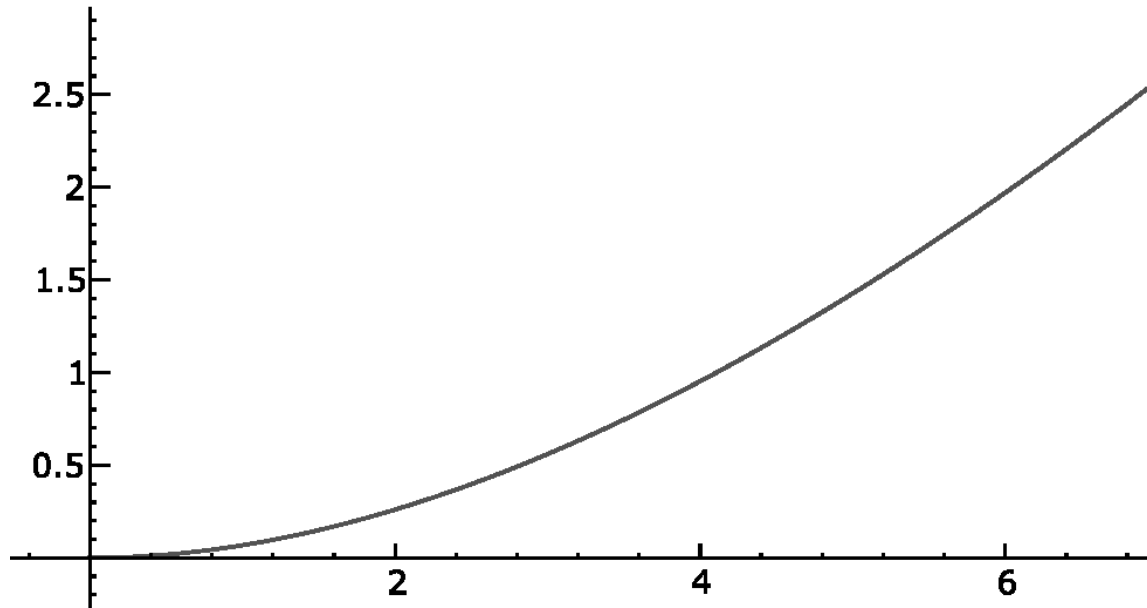
- Web page updates follow the Poisson distribution on average
  - time until the next update is governed by an exponential (Poisson) distribution :

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

$\lambda$  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)

—The vast majority of the Internet lies in the Deep Web, sometimes referred to as the Invisible Web. The actual size of the Deep Web is impossible to measure, but many experts estimate it is about 500 times the size of the web as we know it.

# Deep Web

Not necessarily something illegal (mostly not):

- Data that needs to be accessed by a search interface
- Results of database queries
- Subscription-only information and other password-protected data
- Pages that are not linked to by any other page
- Technically limited content, such as that requiring [CAPTCHA](#) technology
- Text content that exists outside of conventional

*http://* or *https://* protocols



# 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



## 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](#).

Little is known about Googlebot (the Google's crawler). Google crawls the web at varying depths and on several different schedules. It is believed that there are comprehensive crawls (the **deep crawls**), which occur on a per month basis, and then there are intermediate or random crawls (**fresh crawls**), which occur more often, but don't go as deep or index as much. Some refer to the constant motion of the indexes as "*the Google Dance*", but the fact is that new information is ALWAYS being indexed.

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 more refined but “content-based” ranking methods of the BERT flavor)
- 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/>



[About](#) [Become an Editor](#) [Suggest a Site](#) [Help](#) [Login](#)



welcome to our archive of [dmoz.org](#).

Visit [resource-zone](#) to stay in touch with the community.

[#OrganizeTheWeb](#)

+



## 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

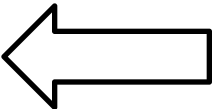
Arts, School Time, Teen Life...





# Games

▼ Subcategories 31



- |  |   |   |   |
|--|---|---|---|
| <ul style="list-style-type: none"><li>Board Games</li><li>Card Games</li><li>CCGs</li><li>Coin-Op</li><li>Dice</li></ul> | <ul style="list-style-type: none"><li>Gambling</li><li>Hand Games</li><li>Hand-Eye Coordination</li><li>Miniatures</li><li>Online</li></ul> | <ul style="list-style-type: none"><li>Paper and Pencil</li><li>Party Games</li><li>Play-By-Mail</li><li>Puzzles</li><li>Roleplaying</li></ul> | <ul style="list-style-type: none"><li>Tile Games</li><li>Trading Card Games</li><li>Video Games</li><li>Yard, Deck, and Table Games</li></ul> |
| <ul style="list-style-type: none"><li>Addiction</li><li>Collecting</li><li>Consumer Information</li></ul>                | <ul style="list-style-type: none"><li>Conventions</li><li>Developers and Publishers</li><li>Game Studies</li></ul>                          | <ul style="list-style-type: none"><li>History</li><li>Play Groups</li><li>Resources</li></ul>   | <ul style="list-style-type: none"><li>Shopping</li><li>Web Hosting</li><li>Women in Gaming</li></ul>  |

▼ 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.

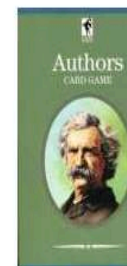
 **Go Fish**  
Rules, information and links.

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



## Go Fish, Authors, Happy Families, Quartet

- [Introduction](#)
- [Go Fish](#)
- [Variations](#)
- [Australian Fish](#)
- [Omben / Minuman](#) (Indonesia)
- [Authors](#)
- [Happy Families](#)
- [Pai Hong](#) (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 thought 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 **Happy 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 whether these games were based on an earlier game played with standard cards, or whether the adaptation to use a standard pack came later.

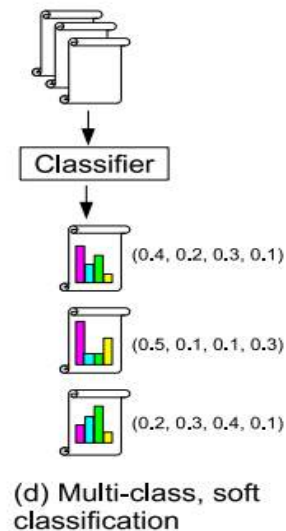
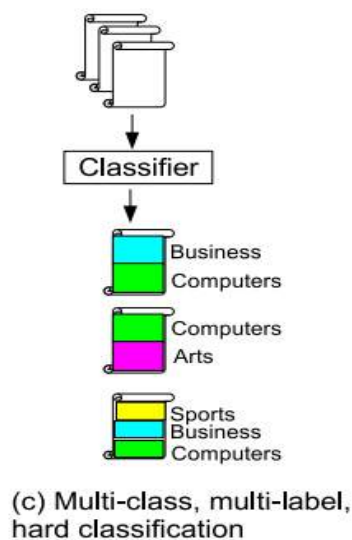
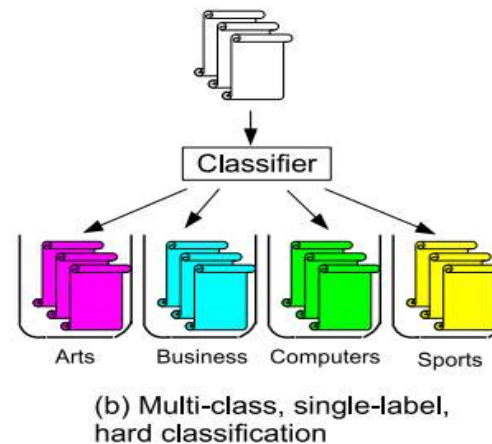
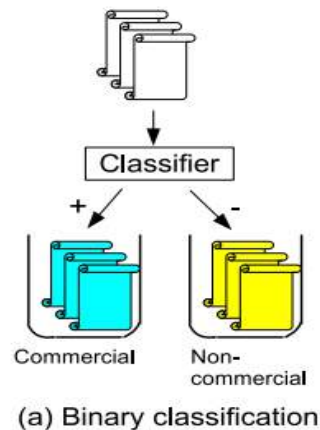
### Go Fish

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 and few others page categorization is “openly” used only by *focused search engines (eg Ebay, Amazon, MESH Medical Subject Index,..)*
- 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



—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 *machine learning* systems on a labeled set of examples (*supervised learning*).
- All algorithms that apply to documents also apply to web pages



# Feature selection in automated web page classification

- 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)
  - Neighbourhood features (selected features in the pages “pointing” at that page)

# Features: On-page

- Textual content and tags

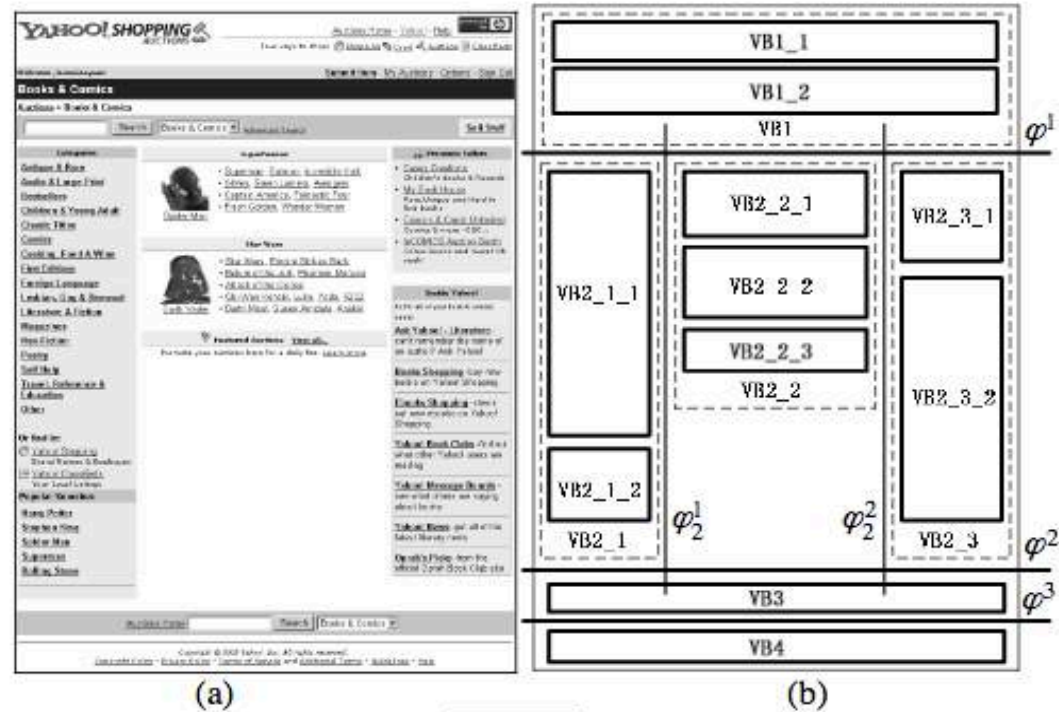
- N-gram feature (n-gram= sequence of n consecutive words)
  - Also called n-words, e.g. “New York” is a biword.
  - In Yahoo!, they used **5-grams** features.
- HTML tags or DOM (document object model)
  - Title, Headings, Metadata and Main text
    - Assigned each of them an arbitrary weight.
    -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

# Features: On-page

- **Visual analysis**

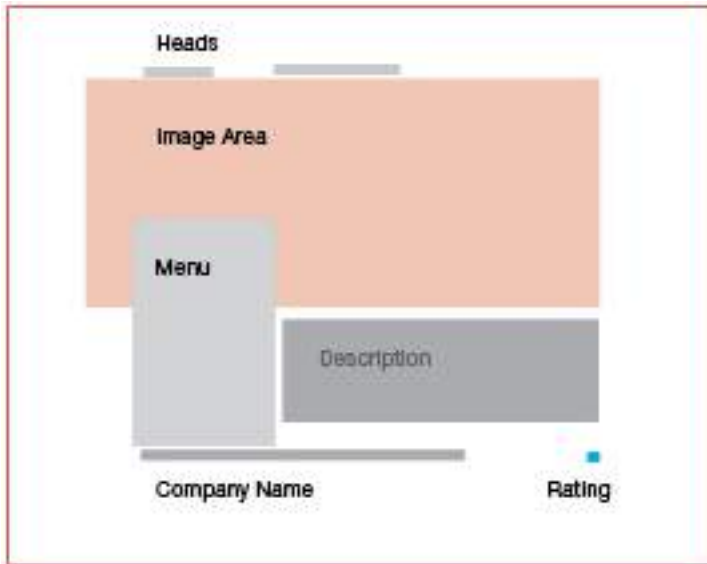
- Each webpage has two representations
  1. Text represented in HTML
  2. The visual representation rendered by a web browser
- visual information is useful as well
  - Each webpage is represented as a hierarchical “Visual adjacency multi graph.”
  - In the graph each node represents an HTML object and 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**
  - Parsing templates is crucial especially for scrapers

# Visual graph representation

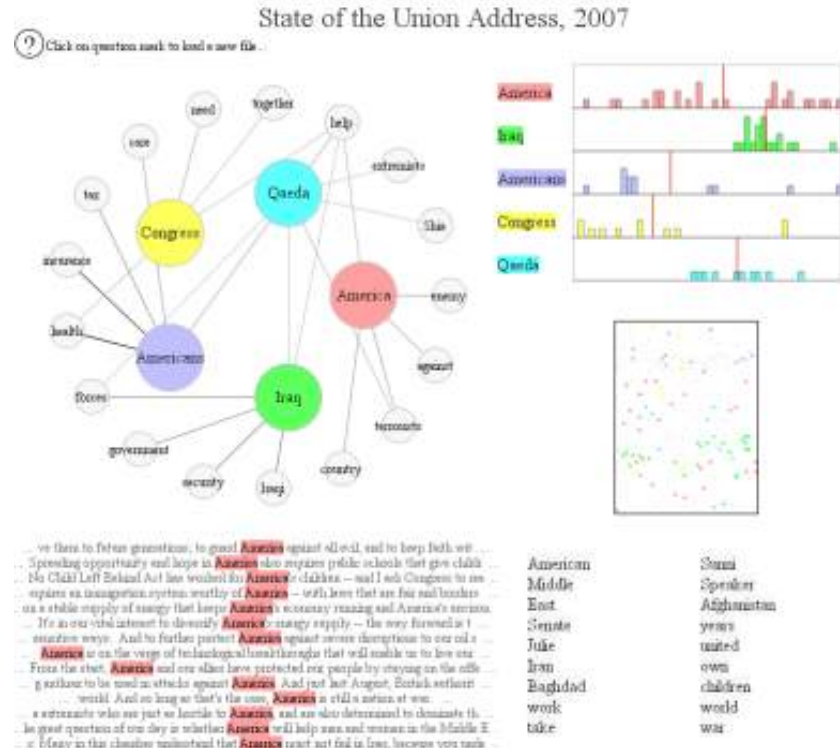
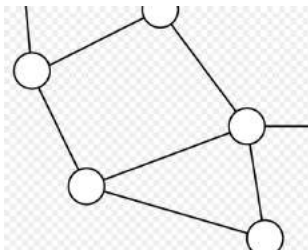


# Visual analysis

heuristic rules can be applied to recognize multiple logical areas, which correspond to different meaningful parts of the page.



## Layout graph



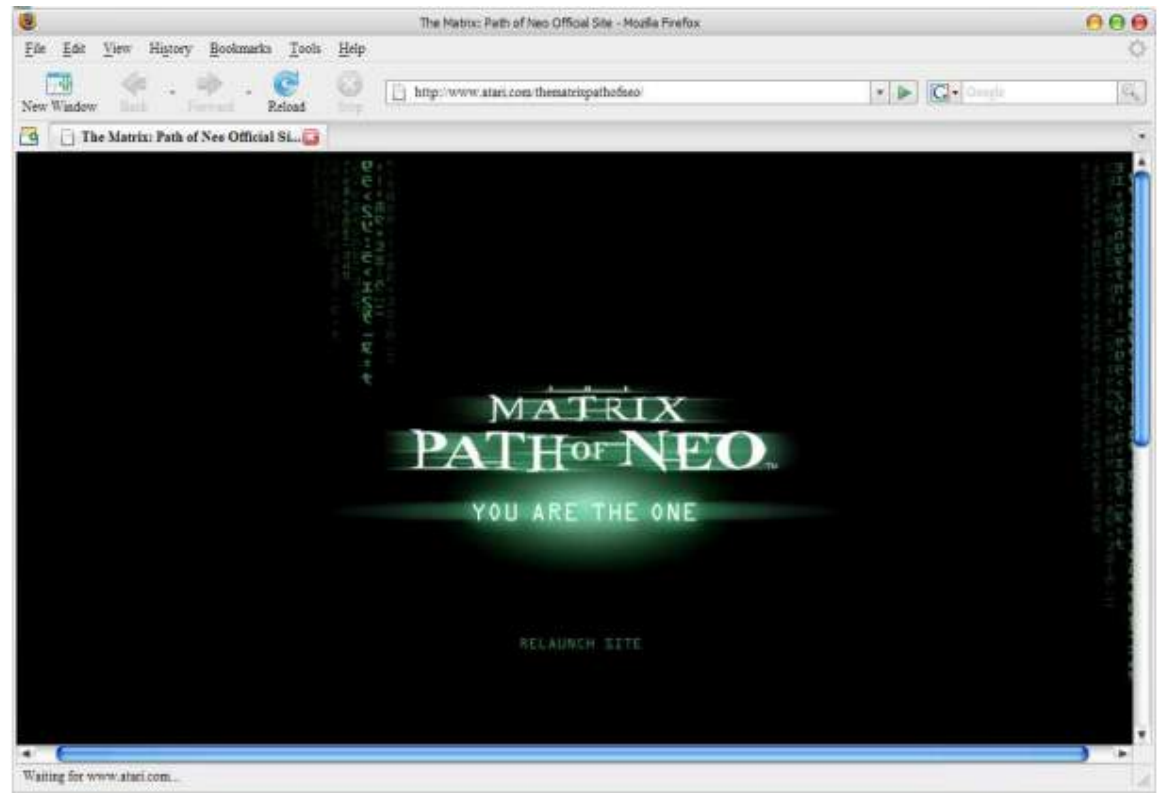
## Content graph

Features: Neighbors Features

# Features: Neighbors Features

- **Motivation**

- Often in-page features are missing or unrecognizable



# Features: Neighbors features

- **Underlying Assumptions**

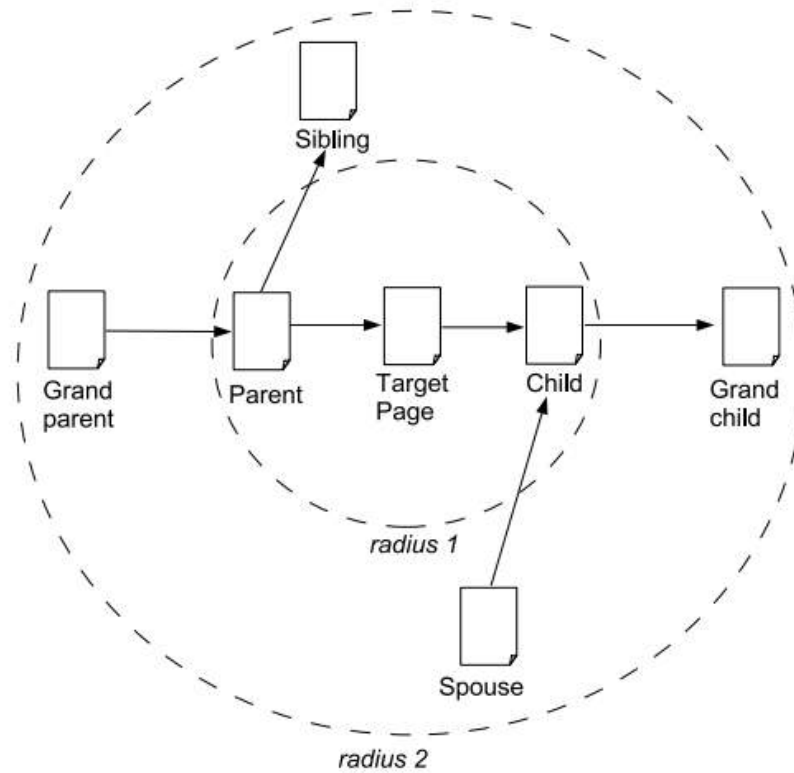
- When exploring the features of neighbors, some assumptions are implicitly made (like e.g. homophily: pages point at similar ones).
- The presence of many “sports” pages in the neighborhood of *Page-a* increases the probability of *Page-a* being in “Sport”.
- linked pages are more likely to have terms 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 within 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

- **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.

Coming Next..

- Ranking with Link Analysis (Page Rank, HITS)