10. The World Wide Web (WWW)

= a *hypertext system* which is accessible via *internet*

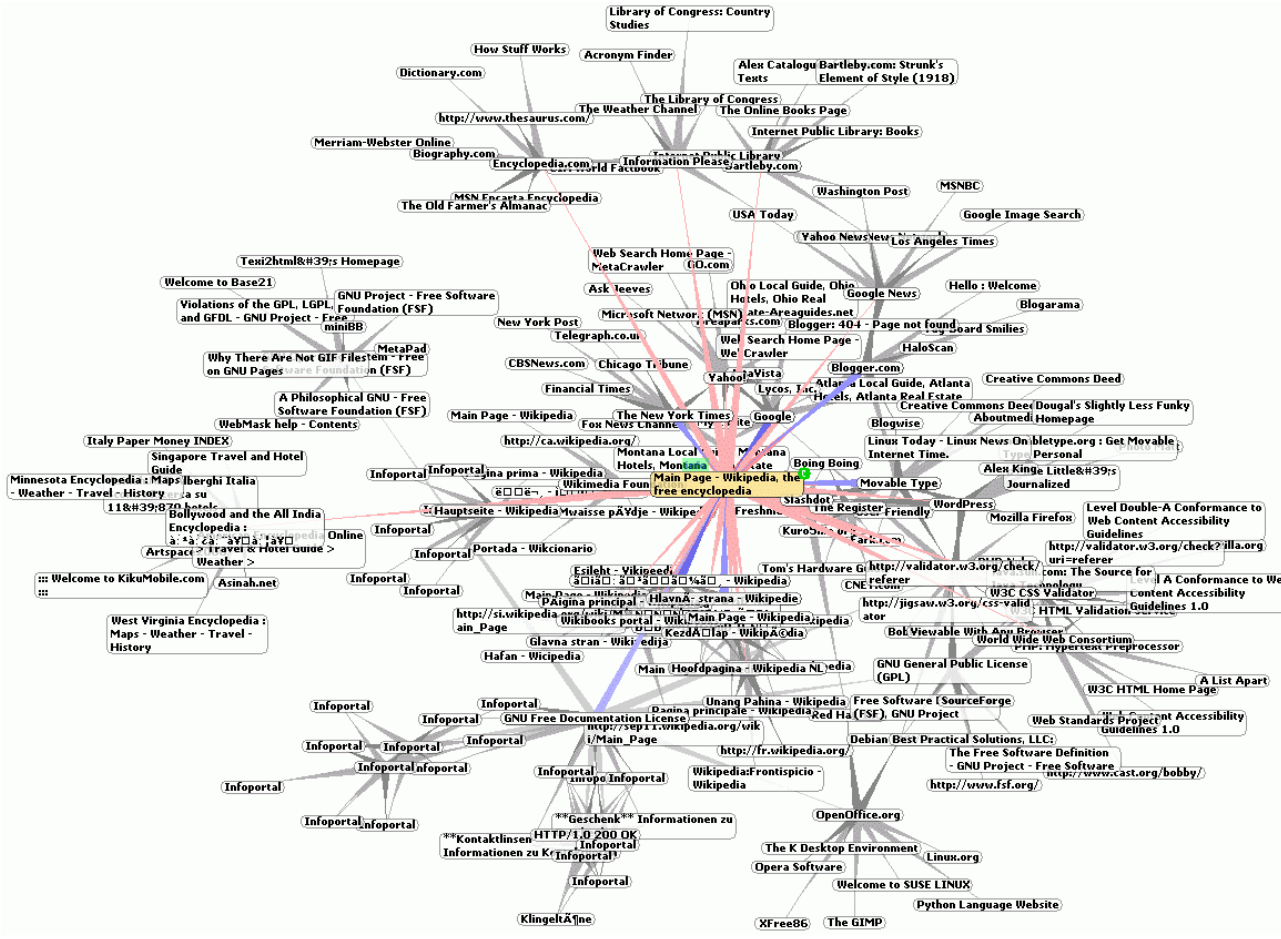
(WWW is only one sort of using the internet – others are e-mail, ftp, telnet, internet telephone ...)

Hypertext: Pages of text containing *hyperlinks* (short: *links*) referring to other pages

Hypertext is viewed using a program called a web browser which retrieves pieces of information, called "documents" or "web pages", from web servers and displays them, typically on a computer monitor. One can then follow hyperlinks on each page to other documents or even send information back to the server to interact with it. The act of following hyperlinks is often called "*surfing*" or "*browsing*" the web. Web pages are often arranged in collections of related material called "web sites."
(from **`www.wikipedia.org`**, the open www encyclopedia)

The link structure of the web forms a very large graph –

the following is a very small subgraph of it:



The Web can be seen as a sort of database –
but very different from relational databases:

- highly distributed, decentralized;
- based on the hypertext model instead of the entity-relationship model;
- with only very weak standards to restrict form and content of the pages;
- very large (in 2001: more than 550 billion documents on the Web)
- without a universal query language.

(*Search engines* try to compensate the last item; see below.)

History of the WWW:

- Origin: a project at CERN (Geneva) in 1989
- *Tim Berners-Lee* and *Robert Cailliau*
- their system: ENQUIRE, realized core ideas of the Web in order to enable access to library information that was scattered on several different computers at CERN
- proposal for the WWW: published by Berners-Lee on November 12, 1990
- *first web page* on November 13 on a NeXT workstation
- Christmas 1990: Berners-Lee built the first web browser and the first web server
- idea of hypertext was older (Vannevar Bush 1945)
- August 6, 1991: summary of the WWW project posted in a newsgroup in the internet
- April 30, 1993: CERN annouced that the WWW would be free to anyone
- 1993: Browser Mosaic (forerunner of Netscape) popularizes the WWW

The three core standards of the Web:

• Uniform Resource Locator (URL): specifies how each page of information is given a unique address at which it can be found (e.g., `http://en.wikipedia.org/wiki/World_Wide_Web`)

• Hypertext Transfer Protocol (HTTP): specifies how the browser and server send the information to each other

• Hypertext Markup Language (HTML): a webpage description language used to encode the information so that it can be displayed on a variety of devices and under different operating systems.

Later extensions:

- Cascading Style Sheets (CSS): define the appearance of elements of a web page, separating appearance and content
- XML: more general language than HTML, designed to enable a better separation of appearance and content; also applicable to other sorts of information
- ECMAScript (also called JavaScript or JScript): a programming language with commands for the browser, enables embedding of programmes (scripts) into web pages. Thus web pages can be changed dynamically.
- Hypertext Transfer Protocol Secure (HTTPS): Extension of HTTP where the protocol SSL is evoked to encrypt the complete data transfer
- Java applets can be embedded in web pages and run on the computer of the Web user

The *World Wide Web Consortium* (*W3C*) develops and maintains some of these standards (HTML, CSS) in order to enable computers to effectively store and communicate different kinds of information.

Problems with the Web:

- *highly decentralized*, no control of the content

→ there is a lot of false and misleading information, hate campaigns, promotion of sexual exploitation and of other crimes...

- *highly dynamic*: *Web pages change all the time!*

Links point to nowhere when the target page was removed...

→ when you give a Web address in the References section of a scientific paper or in your thesis, you should add the *date* when you visited that page!

Archive of (a part of) the Web:

http://www.archive.org

→ lost Web references can (in some cases) be reconstructed if the date is known

- *highly chaotic*: no global index or table of content is available; search for a certain content is complicated and time consuming

→ development of specialized search engines,
the most well-known one: *Google*
(**http://www.google.com**)

How does a search engine work?

- First component: a web crawler, visiting all accessible web pages worldwide, one after the other, following the hyperlinks

but: when you look for a certain keyword, this process would take much too long!
→

- second component: a large database, containing keywords and web addresses where these keywords were already found

the web crawler is working in the background and does only actualize the database
when you invoke Google, you search in Google's database, not in the Web!

→ not all Web pages can be found, because not all are in the database

Usually, you get many, many, many Web pages containing a given keyword (often millions...) →
first remedy: make more intelligent queries
e.g., combining several keywords by "and", or looking for phrases instead of keywords (use quotation marks)
– Google provides such facilities under "extended search"

still there are often too many results

→ priorisation of the found web pages necessary

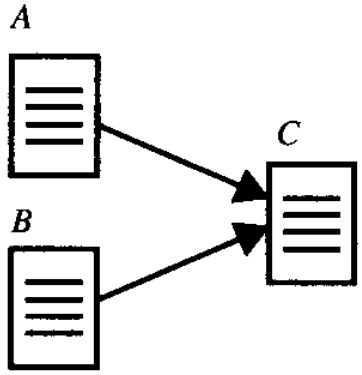
- third component of the search engine (and best capital of the Google company): a *ranking algorithm* for search results

Basic principles of Google ranking of web pages
(Attention: the exact algorithm is changing continuously and is not published)

"Importance" of a web page:
recursively defined, using the hyperlink structure of the Web

The importance of a page is the larger,
the more important pages refer to it!

More precisely:
Let *FLinks*(*A*) be the set of all outgoing links (forward links) of a page *A* and *BLinks*(*A*) the set of all incoming links (backward links) of *A*



$$FLinks(A) = \{C\}$$
$$FLinks(B) = \{C\}$$
$$BLinks(C) = \{A, B\}$$

- *A* has high page rank if the sum of the page ranks of its incoming links is high,
- a page *B* distributes its importance in equal parts to all pages which are referred by it:

$$PageRank(A) = \frac{1}{c} \sum_{B \in BLinks(A)} \frac{PageRank(B)}{|FLinks(B)|}.$$

(*c* = normalisation factor)

Iterative determination of the page rank:

- initially, an arbitrary mapping of values to all web pages is done (typically, the *constant value* 1 is used),
- *iterate the calculation* using the above formula for all pages, until the values remain stable,
- they *converge* against the Eigenvectors of the adjacency matrix of the graph consisting of the web pages (nodes) and their links (edges). (Cf. chapter 8.4.)

Additionally, the Google page rank utilizes:

- *proximity* of the given key words to each other (in the text),
- the *anchor texts* of the links: these are the texts which can be clicked upon. A page *A* gets higher importance when the anchor texts of links referring to *A* contain the keywords, too.

the underlying technology for the WWW:
the **Internet** (short for "Interconnected Networks")

predecessor (end of the 60s): ARPANET (U.S. military project)
was later used to connect universities and research labs

Internet today: A worldwide network of computer networks

- Computers in this network communicate using the standardized *TCP/IP protocol* (Transmission Control Protocol / Internet Protocol: Rules governing the communication)
- Transmission of the information in small portions
- For identification, each computer in the net has a unique number, the *IP address*
- IP address: 32 bit integer; for better comprehensibility usually split in 4 bytes (these 4 bytes are often written as decimal integers, separated by dots: e.g., 194.77.124.35)
  $\rightarrow$ more than 4 billion addresses
- to get identifiers which can better be memorized: *Domain Name System* (DNS)
  – system of (textual) names, association between names and IP addresses
- hierarchy: Domains, subdomains, sub-subdomains..., e.g., **`www-gs.informatik.tu-cottbus.de`** (from right to left!)
- *Top-level domains*: Country abbreviations and some others ("generics"): .de, .fr, .com, .edu, .gov ...
- Lowest level: host name of a single computer (here: www-gs, Web server of the graphics systems chair)

- domain name corresponds to IP address
- transformation of domain names into IP addresses and vice versa: Task of special computers, so-called *nameservers*
- this transformation takes place any time when you click on a hyperlink on a web page!
- each nameserver is responsible for a certain part of the hierarchical name space