**Monday’s Lecture**

History of Internet development

|  |  |
| --- | --- |
| 1961 | First packet switching networks |
| 1969 | advanced research projects agency network (ARPANET) initiated by the United States Department of Defence |
| 1970 | Mark I network (first UK-based network; main figure Donald Davies (1924–2000) from Wales) |
| 1976 | X.25 transport protocol for packet-switching networks |
| 1980 | USENET based on UUCP |
| 1982 | TCP/IP protocol suite, formally introducing the Internet |
| 1982 | SMTP (simple mail transfer protocol) |
| 1983 | DNS (domain name service) |
| 1991 | Gopher (application layer protocol, menu-based approach for document distribution) |
| 1991 | WWW (world wide web) and HTTP (hypertext transfer protocol) |

History of Google

|  |  |
| --- | --- |
| 1995 | Sergej Brin and Larry Page meet at Stanford |
| 1996 | Brin and Page collaborate on research about web search, ranking hypertext and dynamic data mining |
| 1996 | Larry Page sets up BackRub, a web crawler at Stanford |
| 1998 | Brin and Page publish ‘The anatomy of a large-scale hypertextual Web search engine’ at WWW7: Proceedings of the Seventh International Conference on World Wide Web introducing the name Google for the web search, with an architecture aiming at 100 000 000 web pages |
| 08/1998 | Andy Bechtolsheim writes $100 000 cheque for Google, Inc |
| 09/1998 | Google, Inc. is registered |
| 1999 | Brin and Page try to sell Google to Excite for $1 000 000 |
| 1999 | Brin and Page try to sell Google to Excite for $750 000 |

Inputs in a web search (such as a word or several, a phrase, or a search algorithm) gives the expected result of matching web pages which are sorted according to relevance. A search algorithm helps google pin point where you want to search e.g. “Exam time table” *site:aber.ac.uk*”.

Web search will get web pages fast. To do this the web search fast must imply that it cannot access web pages for search. It cannot locally store index information required and web crawling should deliver an index that supports different kinds of searches efficiently.

To exploit the existing structural information and estimate the relevance of a web page we can think of two possible things.

* Would a webpage with more links to it be more relevant i.e. relevance by counting the number of incoming links?
* Is a page more relevant if more relevant pages link to it?

PageRank

PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank is a way of measuring the importance of website pages. According to Google PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.

Algorithm in English:

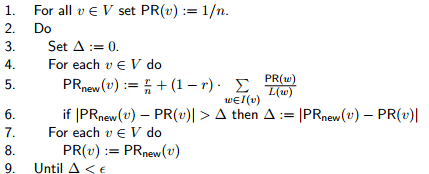
1. Set Page Rank value for all pages to 1/(number of pages) initially.

2. Work in rounds in the following way:

4.-5. Compute the new Page Rank value for v as r/(number of pages) plus, for each page with a link to v, (1 − r) times that page’s Page Rank value divided by the number of different links leaving it.

6. Keep track of the greatest change in Page Rank values.

9. Stop when this difference decreases below ε.



Page Rank problem solving

Page Rank is very useful way beyond ranking web pages. Some examples are:

* Ranking ‘who to follow’ on Twitter
* Support debugging of complex systems
* Traffic prediction
* Book rankings

Problem - *“Page Rank is relatively expensive to compute”*

Solution - The page rank should be recalculated less frequently

Problem – *“Page Rank value is sensitive to manipulations when someone sets up large number of web pages with links to push up Page Rank value of some target page”*

Solution - Punish ‘link farms’ (Google has done this since 2011) and/or use other metrics to determine ranking (done by Google in several secret updates).