Bornholm Web Mining Techniques

Finn Årup Nielsen

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

Email: fn@imm.dtu.dk
WWW: http://www.imm.dtu.dk/~fn

OVERVIEW

- Downloading
- Mass-downloading
- Focused crawling
- Counting links
- Size of web-sites
- PostScript/PDF conversion
- Generation of Graphs
- Collaborative efforts

DOWNLOADING

- `telnet`
- `wget` command line program.

  `wget http://www.imm.dtu.dk/cisp/`
- Perl with the LWP-library

  ```
  perl -MLWP::Simple -e 'getprint "http://www.imm.dtu.dk/cisp/"'
  ```
- Parallel download
  - In serial download each download has to wait for previous download to finish.
  - Solution is multiple threads/processes
  - Perl modul for parallel download `ParallelUserAgent` by Marc Langheinrich.

MASS DOWNLOADING

- Crawling. Start on a page and follow the links with depth-first or breath-first searching.
- Focused crawling.
- Search engine results (Lawrence and Giles, 1998). "[...] no engine indexes more than about 16% of the web" (Lawrence and Giles, 1999).
- Scan IP-numbers (Lawrence and Giles, 1999). 4.3 milliard possible servers with 1999 IP version. Problems:
  - Some servers contain the same information.
  - Names might share an IP-number, e.g, in web-hotels. HTTP/1.1 name-based virtual hosting.
- User registration. Let the user submit web-pages.

FOCUSED CRAWLING

- Problem of downloading all web-pages.
- Obtain relevant web-pages with minimum download (Chakrabarti et al., 1999), http://www.cs.berkeley.edu/ soumen/focus/.
- Relevant for niche search engine (McCallum et al., 1999).
- Term based.
  - Let the user select a set of documents.
  - Determine the relevance of a document by its content (terms) — classify!
  - Crawl only the links in relevant documents.
- Adaptive crawlers: Update the classifier for the document relevance with downloaded documents
- Some interesting document might be separated by non-relevant documents, e.g.,
  - Reinforcement learning (Rennie and McCallum, 1999)
  - Context focused crawler (CFC) (Diligenti et al., 2000).

COUNTING LINKS

- Web search engines typically report number of found links.
- Statistics of web-citations.
- Reverse engineer the CGI arguments.
- Counting the number of webpages by restricting to a domain (Almind and Ingwersen, 1997).
- Citation counting with advanced web-searches (Ingwersen, 1998).
  - AltaVista: `domain:` and `link:`
  - AllTheWeb (Fast): "must include" `imm.dtu.dk` "in the link to URL" with domain filter "Only Include" `edu`.
  - Google? inurl, site, link, e.g., `link:www.dtu.dk`
- Web search engine does not cover all the web.

SIZE OF WEB-SITES

- Distribution of web-sites is approximately distributed according to a power-law (Huberman and Adamic, 1999).
- Few site with many web-pages: rich get richer.
- Can affect a learning algorithm since data set is dominated by a single site.
- Out-degree and in-degree power-law distributed (Albert et al., 1999; Faloutsos et al., 1999).
- Note! Steve Lawrence: "Everything is a straight line in a double logarithmic plot".
- Specific subcategories of web-pages, e.g., university homepages, are typically unimodal on a log scale, (Pennock et al., 2002), http://modelingtheweb.com/.

POSTSCRIPT/PDF CONVERSION

- Conversion of PDF and PostScript files to text files.
- Implemented in Google web search engine and ResearchIndex.
- Problems: Two-column layout, kerning, encryption, equations, images, formatting, tables.
- Tools
  - `Prescript` (Miller, 1998). by New Zealand digital libraries. Based on python (should be installed), relative fast.
  - `pstotext`. By Andrew Birrell and Paul McJones from (then) Digital Equipment Corporation. Slow, but convert files better.
  - `ps2ascii`. Distributed with ghostscript distribution, fast. Interlace text in two-column layout (not important for bag-of-words representation), robust (i.e., works!)
  - PS^{2}text. Commercial program.

GENERATION OF GRAPHS

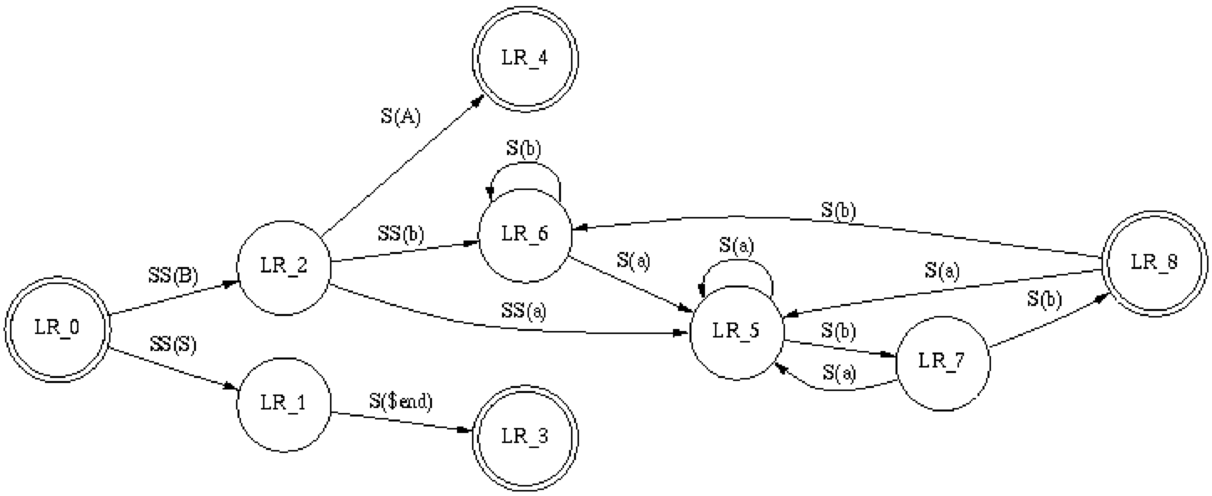
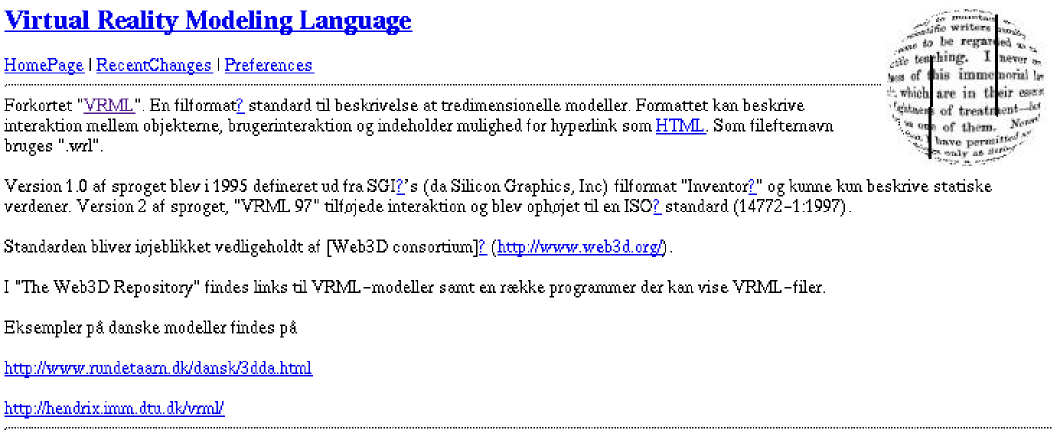


Figure 1: Graphviz example.

- DAG (Gansner et al., 1988)
- GraphViz (Gansner and North, 2000; North, 1992; Koutsofios and North, 1996)
  - Collection of programs to render directed and undirected graphs.
  - WebDot, a CGI program that converts a dot to images. Image maps are possible.
  - Used in PubGene (Jenssen et al., 2001).
- Other examples: Hyperbolic displays, interactive visualization for large hierarchies (Lamping and Rao, 1996), path finder network scaling on author from ResearchIndex (Chen and Paul, 2001).

COLLABORATIVE EFFORTS



Virtual Reality Modeling Language

HomePage | RecentChanges | Preferences

Forkortet "VRML". En filformat? standard til beskrivelse at tredimensionelle modeller. Formattet kan beskrive interaktion mellem objekterne, brugerinteraktion og indeholder mulighed for hyperlink som HTML. Som filefternavn bruges ".wrl".

Version 1.0 af sproget blev i 1995 defineret ud fra SGI?'s (da Silicon Graphics, Inc) filformat "Inventor?" og kunne kun beskrive statiske verdener. Version 2 af sproget, "VRML 97" tilføjede interaktion og blev ophøjet til en ISO? standard (14772-1:1997).

Standarden bliver iøjeblikket vedligeholdt af [Web3D consortium]? (http://www.web3d.org/).

I "The Web3D Repository" findes links til VRML-modeller samt en række programmer der kan vise VRML-filer.

Eksempler på danske modeller findes på

http://www.rundetaarn.dk/dansk/3dda.html

http://hendrix.imm.dtu.dk/vrml/

Figure 2: Wikipedia example.

- Open directory, http://dmoz.org/
  - Collaborative human-edited directory
  - Receives 250 site submissions per hour (August 2000, searchenginewatch.com). Used by Google.
- Wiki, WikiWikiWeb, (Leuf and Cunningham, 2001)
  - Web-page, that is editable by users.
  - Simple markup and hyperlinks (e.g., automatic link by “camel case”)
  - Wikipedia: “A collaborative project to produce a complete encyclopedia from scratch” start in January 2001 and have 28,214 articles (April 2002).
- Open Mind, http://www.openmind.org
  - Let “users” enter knowledge, e.g., common sense.

References

Albert, R., Jeong, H., and Barabasi, A.-L. (1999). diameter of the world-wide web. *Nature*, 401:130–131.

Almind, T. C. and Ingwersen, P. (1997). Informetric analyses on the world wide web: Methodological approaches to “webometrics”. *Journal of Documentation*, 53(4):404–426.

Chakrabarti, S., van den Berg, M., and Domc, B. (1999). Focused crawling: A new approach to topic-specific web resource discovery. In *WWW8*.

Chen, C. and Paul, R. J. (2001). Visualizing a knowledge domain’s intellectual structure. *IEEE Computer*, 34(3):65–71. http://www.brunel.ac.uk/~cssrccc2/papers/ieeecomputer2001.pdf.

Diligenti, M., Coetzee, F. M., Lawrence, S., Giles, C. L., and Gori, M. (2000). Focused crawling using context graphs. In *Proc. Very Large Databases*.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *ACM SIGCOMM’99*.

Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software — Practice and Experience*, 30(11):1203–1234. http://www.research.att.com/sw/tools/graphviz/GN99.pdf. ISSN 00380644, [defkat.dk — bibliotek.dk].

Gansner, E. R., North, S. C., and Vo, K. P. (1988). Dag: A program that draws directed graphs. *Software — Practice and Experience*, 18(11):1047–1062. ftp://ftp.cs.utexas.edu/pub/code2/kleyn/graphs/AttDag/dagdoc.ps. ResearchIndex: http://citeseer.nj.nec.com/gansner89dag.html.

Huberman, B. A. and Adamic, L. A. (1999). Growth dynamics of the world-wide web. *Nature*, 401:131.

Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, 54(2):236–243.

Jenssen, T.-K., Læreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28. PMID: 11326270. http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v28/n1/full/ng0501_21.html. Gene cocitation analysis involving data from over 10 million articles from PubMed citing 13,712 human genes names obtained from HUGO, LocusLink, The Genome Database and GENATLAS. Network diagrams are automatically generated with the GraphViz software.

Koutsofios, E. and North, S. C. (1996). *Drawing graphs with dot*. AT&T Bell Laboratories, Murray Hill, New Jersey.

Lamping, J. and Rao, R. (1996). The hyperbolic browser: A focus + context technique for visualizing large hierarchies. *Journal of Visual Languages and Computing*, 7(1):33–35.

Lawrence, S. and Giles, C. L. (1998). Searching the world wide web. *Science*, 280:98–100.

Lawrence, S. and Giles, C. L. (1999). Accessibility and information on the web. *Nature*, 400:107–109.

Leuf, B. and Cunningham, W. (2001). *The Wiki Way: Collaboration and Sharing on the Internet*. Addison-Wesley. ISBN 020171499X, [defkat.dk — bibliotek.dk — amazon.com — bn.com].

McCallum, A., Nigam, K., Rennie, J., and Seymore, K. (1999). Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*.

Miller, D. J. (1998). *Prescript: Programme Structure and functional Description*. The New Zealand Digital Library, University Waikato, New Zealand. Distributed with Prescript available at http://www.nzdl.org/html/prescript.html.

North, S. C. (1992). *NEATO User’s Guide*. AT&T Bell Laboratories, Murray Hill, NJ.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., and Giles, C. L. (2002). Winners don’t take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211. http://modelingtheweb.com/modelingtheweb.ps.

Rennie, J. and McCallum, A. (1999). Using reinforcement learning to spider the web efficiently. In *Proc. 16th International Conf. on Machine Learning*.