DM79 Algorithms for Web Indexing and Searching
Fall 2004
Exam Curriculum

The curriculum is

- *The Anatomy of a Large-Scale Hypertextual Web Search Engine.* Sergey Brin and Lawrence Page. Proceedings of the 7th International WWW conference, 1998.
- Lecture slides on the I/O-model and on external sorting (except the lower bound part for external sorting).
- *High-Performance Web Crawling.* Marc Najork and Allan Heydon. Compaq SRC Research Report 173, 2001. Also appeared in *Handbook of Massive Data Sets*, Kluwer, 2001.
- *Breadth-First Search Crawling Yields High-Quality Pages.* Marc Najork and Janet L. Wiener. In Proceedings of the Tenth Internal World Wide Web Conference, pages 114-118, May 2001.
- Pages 239-241 in *HTTP: The Definitive Guide.* David Gourley, Brian Totty. O'Reilly, 2002. ISBN: 1-56592-509-2.
- All except Sections 4.2-3 of: *Building a Distributed Full-Text Index for the Web.* Sergey Melnik, Sriram Raghavan, Beverly Yang, and Hector Garcia-Molina. Stanford Technical Report 2000-55. Short version of paper appeared in the proceedings of the 10th International WWW conference, May 2-5, 2001, Hong Kong.
- Pages 109-122, 145-151, 156-161, 169-170, and 175-179 of: *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd edition. Ian H. Witten, Alistair Moffat, and Timothy C. Bell. Morgan Kaufmann Publishing, San Francisco, 1999. ISBN 1-55860-570-3.
- *Efficient Computation of PageRank.* Taher H. Haveliwala. Stanford Technical Report 1999-31.
- Pages 31-45 and 53-58 in *Understanding Search Engines: Mathematical Modeling and Text Retrieval.* M.W. Berry and M. Browne. SIAM Book Series: Software, Environments, and Tools, (June 1999), ISBN: 0-89871-437-0.
- Pages 24-30, 74-81, 99-106, 118-119, and 207-208 in *Modern Information Retrieval.* Ricardo Baeza-Yates, Berthier Ribiero-Neto. Addison Wesley Higher Education, 1999. ISBN: 020139829X.

- Pages 566-576 in *Data Structures and Algorithms in Java.* Michael T. Goodrich and Roberto Tamassia. Wiley, 1998. ISBN: 0-471-19308-9.
- Pages 94-107, 116-119, and 149-155 in *Algorithms on Strings, Trees, and Sequences.* Dan Gusfield. Cambridge University Press, 1997. ISBN: 0521585198.
- *Fast Algorithms for Sorting and Searching Strings.* Jon Bentley and Robert Sedgewick. Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms. New Orleans, January, 1997. Pages 360-369.
- Sections 1, 2, 4, and 5 in *Spectral Analysis of Data.* Yossi Azar, Amos Fiat, Anna R. Karlin, Frank McSherry, and Jared Saia. Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC), pages 619–626, 2001.
- *A Survey of Eigenvector Methods of Web Information Retrieval.* Amy N. Langville and Carl D. Meyer, 2004. To appear in SIAM Review.
- *Graph structure in the web* . Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. In proceedings of Ninth International World Wide Web Conference (WWW9), 2000.
- Non-proof parts of *Stochastic models for the Web graph* . R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Proceedings of the 41th IEEE Symp. on Foundations of Computer Science. November 2000, pp. 57-65.
- *Trawling the Web for emerging cyber-communities.* Ravi Kumar, Prabhakar Raghavan, Rajagopalan Rajagopalan, and Andrew Tomkins. Computer Networks, 31 (11-16), pages 1481-1493, 1999. Also in Proceedings of the 8th WWW conference.
- *On the resemblance and containment of documents.* Andrei Broder. In Compression and Complexity of Sequences (SEQUENCES'97), pages 21-29. IEEE Computer Society, 1998. (Section 4 can be read lightly).
- *Web Search for a Planet: The Google Cluster Architecture.* Luiz André Barroso, Jeffrey Dean, and Urs Hölzle. IEEE Micro, 23(2), pp. 22-28, 2003.
- *Algorithms, games, and the Internet.* Christos Papadimitriou. Proceedings of the 33rd STOC, pp. 749-753, ACM Press, 2001.