15-850: Algorithms for Indexing and Searching

Meeting time: 3:00-5:20pm Wednesdays

Instructors: Blelloch, Lafferty, Miller

Topics

Here is a rough outline with some readings. It will be filled in during the semester.

Week 1: Introduction to IR

Week 2: Introduction to Web Searching (Henzinger Slides)

Week 3: Latent Semantic Indexing [5, 43]

Week 4: Link-based Search Analysis [30, 6, 18]

Week 5: Efficient SVD

Week 6: High-dimensional Nearest Neighbors [27, 20, 31]

Week 7: Cluster Analysis [15]

Week 8: Leave Open

Week 9: Properties of the Web [49]

Week 10 : Removing duplicate documents [12]

Week 11: Web Caching [28]

Week 12: Erasure Codes

Our list of references is augmented from the one provided with the following tutorial:

> A. Broder and M. Henzinger (1998). "Information retrieval on the web: Tools and algorithmic issues," Invited tutorial at Foundations of Computer Science (FOCS). Slides available at `http://www.research.digital.com/SRC/personal/monika/-focs-talk-m78/ppframe.htm`

[1] Proceedings of the Sixth International World Wide Web Conference, Santa Clara, California, April 1997.

[2] WWW7 Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia, April 1998.

[3] P.K. Agarwal, L. Arge, T.M. Murali, K.R. Varadarajan, and J.S. Vitter. I/O-efficient algorithms for contour-line extraction and planar graph blocking (extended abstract). In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 117-126, San Francisco, California, Jan. 1998.

[4] O. Arocena, A.O. Mendelzon, and G.A. Mihaila. Applications of a web query language. In Proceedings of the Sixth International World Wide Web Conference (WWW6), pages 587- 595.

[5] Michael W. Berry and Susan T. Dumais, and Gavin W. O'Brien. Using Linear Algebra for Intelligent Information Retrieval. Published in SIAM Review 37:4 (1995), pp. 573-595.

[6] K. Bharat and A.Z. Broder. A technique for measuring the relative size and overlap of public web search engines. In Proceedings of the Seventh International World Wide Web Conference (WWW7), pages 379-388.

[7] K. Bharat, A.Z. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The connectivity server: Fast access to linkage information on the web. In Proceedings of the Seventh International World Wide Web Conference (WWW7), pages 469-477.

[8] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), pages 111-104, 1998.

[9] S. Brin, J. Davis, and H. Garçia-Molina. Copy detection mechanisms for digital documents. In M.J. Carey and D.A. Schneider, editors, Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, pages 398-409, San Jose, California, May 1995.

[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International World Wide Web Conference (WWW7), pages 107-117.

[11] A.Z. Broder. On the resemblance and containment of documents. In Proceedings of Compression and Complexity of Sequences 1997, pages 21-29. IEEE Computer Society, 1988.

[12] A.Z. Broder. Filtering near-duplicate documents. In Proceedings of FUN 98, 1998. To appear.

[13] A.Z. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher. Min-wise independent permutations. In Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC-98), pages 327-336, New York, May 1998. ACM Press.

[14] A.Z. Broder, M. Charikar, and M. Mitzenmacher. A deran domization using min-wise independent permutations. In Proceedings of Random 98, 1998. To appear.

[15] A.Z. Broder, S.C. Glassman, M.S. Manasse, and G. Zweig. Syntactic clustering of the Web. In Proceedings of the Sixth International World Wide Web Conference (WWW6), pages 391- 404.

[16] J. Carriere and R. Kazman. Webquery: Searching and visualizing the web through connectivity. In Proceedings of the Sixth International World Wide Web Conference (WWW6), pages 701- 711.

[17] S. Chakrabarti, B. Dom, D. Gibson, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Experiments in topic distillation. In ACM-SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web, 1998.

[18] S. Chakrabarti, B. Dom, R.P., S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In Proceedings of the Seventh International World Wide Web Conference (WWW7), pages 65-74.

[19] J. Cho, H. Garçia-Molina, and L. Page. Efficient crawling through URL ordering. In Proceedings of the Seventh International World Wide Web Conference (WWW7), pages 161-172.

[20] Edith Cohen and David D. Lewis, Approximating matrix multiplication for pattern recognition tasks. In Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 682-691, 1997.

[21] E.G. Coffman, Z. Liu, and R.R. Weber. Optimal robot scheduling for web search engines. Technical Report 3317, INRIA, Dec. 1997.

[22] R.B. Doorenbos, O. Etzioni, and D.S. Weld. A scalable comparison-shopping agent for the World-Wide Web. In W.L. Johnson and B. Hayes-Roth, editors, Proceedings of the 1st International Conference on Autonomous Agents, pages 39-48, New York, Feb. 1997. ACM Press.

[23] M. Fang, N. Shivakumar, H. Garçia-Molina, R. Motwani, and J. Ullman. Computing iceberg queries efficiently. In Proceedings of the 1998 International Conference on Very Large Databases (VLDB'98), Aug. 1998. To appear.

[24] W.B. Frakes and R. Baeza-Yates. Information Retrieval: Data Structures and Algorithms. Prentice Hall, 1992.

[25] E. Garfield. Citation analysis as a tool in journal evaluation. Science, 178, 1972.

[26] N. Heintze. Scalable document fingerprinting. In Second USENIX Workshop on Electronic Commerce, pages 191-200, Oakland, California, Nov. 1996.

[27] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC-98), pages 604-613, New York, May 1998. ACM Press.

[28] D. Karger, E. Lehman, T. Leighton, M. Levine, D. Lewin, and R. Panigrahy, "Consistent hashing and random trees: Distributed cachine protocols for relieving hot spots on the World Wide Web," In Proceedings of the 31st Annual ACM Symposium on Theory of Computing (STOC-97).

[29] M.M. Kessler. Bibliographic coupling between scientific papers. American Documentation, 14, 1963.

[30] J. Kleinberg. Authoritative sources in a hyperlinked environment. In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, pages 668-677, January 1998.

[31] E. Kushilevitz, R. Ostrovsky and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. In Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, 1998. pages 614-23.

[32] S. Lawrence and C.L. Giles. Inquirus, the NECI meta search engine. In Proceedings of the Seventh International World Wide Web Conference (WWW7), pages 95-105.

[33] S. Lawrence and C.L. Giles. Searching the World Wide Web. Science, 280(5360):98, 1998.

[34] Y. Maarek, M. Jacovi, M. Shtalhaim, S. Ur, D. Zernik, and I.Z. Ben Shaul. Webcutter: A system for dynamic and tailorable site mapping. In Proceedings of the Sixth International World Wide Web Conference (WWW6), pages 713-722.

[35] U. Manber. Finding similar files in a large file system. In Pro ceedings of the Winter 1994 USENIX Conference, pages 1-10, Berkeley, CA, USA, 1994.

[36] M. Marchiori. The quest for correct information on the web: Hyper search engines. In Proceedings of the Sixth International World Wide Web Conference (WWW6), pages 265-274.

[37] G.A. Mihaila. WebSQL - an SQL-like query language for the world wide Web. Master's thesis, University of Toronto, 1996.

[38] J. Mogul and P. Leach. RFC 2227: Simple hit-metering and usage-limiting for HTTP, Oct. 1997. Status: Proposed standard. URL: ftp://ftp.math.utah.edu/-pub/rfc/rfc2227.txt.

[39] M. Naor and B. Pinkas. Oblivious transfer and polynomial evaluation. Manuscript, 1998.

[40] M. Naor and B. Pinkas. Secure accounting and auditing on the Web. In Proceedings of the Seventh International World Wide Web Conference (WWW7), pages 541-550.

[41] M.H. Nodine, M.T. Goodrich, and J.S. Vitter. Blocking for external graph searching. Algorithmica, 16(2):181-214, Aug. 1996.

[42] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Work in progress. URL: http://google.stanford.edu/-*backrub/pageranksub.ps.

[43] Christos Papadimitriou, Prabhakar Raghavan, Hisao Tamaki and Santosh Vempala. Latent Semantic Indexing: A Probabilistic Analysis. Proc. 17th ACM Symposium on the Principles of Database Systems, Seattle, 1998.

[44] P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear: Extracting usable structures from the web. In Proceedings of the Conference on Human Factors in Computing Systems (CHI 96), pages 118-125, April 1996.

[45] J. Pitkow. In search of: reliable usage data on the WWW. In Proceedings of the Sixth International World Wide Web Conference (WWW6), pages 452-463.

[46] J. Shakes, M. Langheinrich, and O. Etzioni. Dynamic reference sifting: A case study in the homepage domain. In Proceedings of the Sixth International World Wide Web Conference (WWW6), pages 189-199.

[47] N. Shivakumar and H. Garçia-Molina. SCAM: A copy detection mechanism for digital documents. In DL '95 Proceedings, 1995.

[48] N. Shivakumar and H. Garçia-Molina. Finding near-replicas of documents on the web. In Proceedings of Workshop on Web Databases (WebDB'98), March 1998.

[49] C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a very large AltaVista query log. Technical Report 1998-014, Compaq Systems Research Center, Palo Alto, California, 1998.

[50] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. J. Amer. Soc. Info. Sci., 24, 1973.

[51] K. Sparck-Jones and P. Willet, editors. Readings in Information Retrieval. Morgan Kaufmann, 1997.

[52] E. Spertus. Parasite: Mining structural information on the web. In Proceedings of the Sixth International World Wide Web Conference (WWW6), pages 587-595.

[53] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), pages 2-10, 1998.