Modern Information Retrieval

the concepts and technology behind search

Second edition

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Modern Information Retrieval

The Concepts and Technology behind Search

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Second edition

Addison-Wesley

Harlow, England • Reading, Massachusetts
Menlo Park, California • New York
Don Mills, Ontario • Amsterdam • Bonn
Sydney • Singapore • Tokyo • Madrid
San Juan • Milan • Mexico City • Seoul • Taipei

References

[1] I. Aalbersberg. Incremental relevance feedback. In *Proc of the Fifteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 11–22, Denmark, 1992.

[2] H. Abdi. Kendall rank correlation. In N. Salkind, editor, *Encyclopedia of Measurement and Statistics*, Thousand Oaks, CA, 2007. Sage.

[3] H. Abdi. The Kendall rank correlation coefficient. Technical report, Univ. of Texas at Dallas, 2007.

[4] K. Aberer, F. Klemm, M. Rajman, and J. Wu. An architecture for peer-to-peer information retrieval. In *Workshop on Peer-to-Peer Information Retrieval*, Sheffield, UK, July 2004.

[5] S. Abiteboul. Querying semi-structured data. In F. N. Afrati and P. Kolaitis, editors, *Int. Conf. on Database Theory (ICDT)*, number 1186 in LNCS, pages 1–18, Delphi, Greece, 1997. Springer-Verlag.

[6] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *Proceedings of the twelfth international conference on World Wide Web*, pages 280–290, Budapest, Hungary, 2003. ACM Press.

[7] M. Abolhassani and N. Fuhr. Applying the divergence from randomness approach for content-only search in xml documents. *Lecture Notes in Computer Science*, 2997:409–420, 2004.

[8] M. Abrams, editor. *World Wide Web: Beyond the Basics*. Prentice Hall, 1998.

[9] M. Abrol, N. Latarche, U. Mahadevan, J. Mao, R. Mukherjee, P. Raghavan, M. Tourn, J. Wang, and G. Zhang. Navigating large-scale semi-structured data in business portals. In *Proceedings of the 27th VLDB Conference*, pages 663–666, Roma, Italy, 2001. http://www.vldb.org/conf/2001/P663.pdf.

[10] A. Adams and A. Blandford. Digital libraries' support for the user's 'information journey'. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 160–169, Denver, Colorado, 2005.

[11] E. Adar. User 4xxxxx9: Anonymizing query logs. In *Query Log Analysis: Social and Technological Challenges, Workshop in WWW'07*, 2007.

[12] J. Adiego, G. Navarro, and P. de la Fuente. Scm: Structural contexts model for improving compression in semistructured text databases. In *Proc. 10th International Symposium on String Processing and Information Retrieval (SPIRE 2003)*, LNCS

2857, pages 153–167. Springer, 2003. Extended version appeared in *Information Processing and Management* 43(3), May 2007, pp. 769–790.

[13] J. Adiego, G. Navarro, and P. de la Fuente. Lempel-Ziv compression of structured text. In *Proc. 14th IEEE Data Compression Conference (DCC'04)*, pages 112–121, 2004. Extended version appeared in *JASIST* 58(4), 2007, pp. 461–478.

[14] M. Adler and M. Mitzenmacher. Towards compressing Web graphs. In *Data Compression Conference*, pages 203–212, 2001.

[15] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, 2005.

[16] D. Agarwal and S. Merugu. Predictive discrete latent-factor models for large-scale dyadic data. In *KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 26–35, New York, NY, USA, 2007. ACM.

[17] E. Agichtein, E. Brill, and S. Dumais. Improving Web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA, 2006. ACM Press.

[18] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting Web search result preferences. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–10. ACM Press, 2006.

[19] A. Agogino and J. Ghosh. Increasing pagerank through reinforcement learning. In *Proceedings of Intelligent Engineering Systems Through Artificial Neural Networks*, volume 12, pages 27–32, St. Louis, Missouri, USA, November 2002. ASME Press.

[20] M. Agosti and A. F. Smeaton, editors. *Information retrieval and hypertext*. Kluwer Academic Publishers, Boston/London/Dordrecht, 1996.

[21] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. Bocca, M. Jarke, and C. Zaniolo, editors, *20th Int Conference on Very Large Databases*, pages 487–499. Morgan Kaufmann Publishers, 1994.

[22] A. Aho and M. Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, June 1975.

[23] AIR Workshops: Adversarial Web Retrieval. http://airweb.cse.lehigh.edu/, 2005.

[24] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 773–774, 2007.

[25] S. M. Alessi and S. R. Trollip. *Multimedia for learning: methods and development*. Allyn and Bacon, 2001. 580 pages.

[26] S. Ali, M. Consens, G. Kazai, and M. Lalmas. A common basis for the evaluation of structured document retrieval. In *ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany*, 2008. In Press.

[27] W. Alink, R. Bhoedjang, P. Boncz, and A. de Vries. XIRAF - XML-based indexing and querying for digital forensics. *Digital Investigation*, 3(Supplement-1):50–58, 2006.

[28] Alis Technologies. Web languages hit parade, 1997.

[29] J. Allan. Incremental relevance feedback for information filtering. In *Proc of the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 270–278, Zurich, Switzerland, 1996.

[30] J. Allan. Perspectives on information retrieval and speech. In *Information Retrieval Techniques for Speech Applications, from the workshop "Information Retrieval Techniques for Speech Applications," held as part of the 24th Annual International ACM SIGIR Conference*, pages 1–10, London, UK, 2002. Springer-Verlag.

[31] J. Allan. HARD Track overview in TREC 2004 high accuracy retrieval from documents. *Proceedings of the Thirteenth Text REtrieval Conference (TREC'04)*, 2005.

[32] B. L. Allen. *Information Tasks: Toward a User-Centered Approach to Information Systems*. Academic Press, San Diego, CA, 1996.

[33] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[34] O. Alonso and R. Baeza-Yates. A model for visualizing large answers in WWW. In *XVIII Int. Conf. of the Chilean CS Society*, pages 2–7, Antofagasta, Chile, 1998. IEEE CS Press.

[35] O. Alonso, R. Baeza-Yates, and M. Gertz. Effectiveness of temporal snippets. In *WSSP Workshop at the World Wide Web Conference—WWW'09*, 2009.

[36] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *SIGIR Evaluation Workshop*, 2009.

[37] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[38] G. Amati. *Probability Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, Department of Computing Science University of Glasgow, 2003. http://www.dcs.gla.ac.uk/~gianni/selectedPapers.html.

[39] G. Amati and C. van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transaction on Office and Information Systems (TOIS)*, 20(4), 2002.

[40] Amazon. Mechanical Turk, 2009. http://www.mturk.com.

[41] G. M. Amdahl. Validity of the single-processor approach to achieving large scale computing capabilities. In *Proc. AFIPS 1967 Spring Joint Computer Conf.*, volume 30, pages 483–485, Atlantic City, N.J., Apr. 1967.

[42] S. Amer-Yahia, C. Botev, J. Dörre, and J. Shanmugasundaram. Full-Text extensions explained. *IBM Systems Journal*, 45(2):335–352, 2006.

[43] S. Amer-Yahia, C. Botev, and J. Shanmugasundaram. TexQuery: a full-text search extension to XQuery. In *13th international conference on World Wide Web, New York, NY, USA*, pages 583–594, 2004.

[44] S. Amer-Yahia, P. Case, T. Roelleke, J. Shanmugasundaram, and G. Weikum. Report on the DB/IR panel at SIGMOD 2005. *SIGMOD Record*, 34(4):71–74, 2005.

[45] S. Amer-Yahia, S. Cho, and D. Srivastava. Tree Pattern Relaxation. In *Advances in Database Technology - EDBT 2002, 8th International Conference on Extending Database Technology, Prague, Czech Republic*, pages 496–513, 2002.

[46] S. Amer-Yahia, D. Hiemstra, T. Roelleke, D. Srivastava, and G. Weikum. Ranked XML Querying. In S. Amer-Yahia, D. Srivastava, and G. Weikum, editors, *Workshop on Ranked XML Querying*, number 08111 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2008.

[47] S. Amer-Yahia and M. Lalmas. XML search: languages, INEX and scoring. *SIGMOD Record*, 35(4):16–23, 2006.

[48] A. Amir, S. Srinivasan, and A. Efrat. Search the audio, browse the video: A generic paradigm for video collections. *EURASIP Journal on Applied Signal Processing*, 2003(2):209–222, 2003. doi:10.1155/S111086570321012X.

[49] E. Amitay and C. Paris. Automatically summarising Web sites - is there a way around it? In *Proceedings of the 2000 ACM CIKM International Conference on Information and Knowledge Management*, pages 173–179, McLean, VA, USA, November 2000.

[50] C. Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, New York, revised edition, 2008.

[51] T. Anderson, A. Hussam, B. Plummer, and N. Jacobs. Pie charts for visualizing query term frequency in search results. *Proceedings of the Fifth International Conference on Asian Digital Library*, pages 440–451, 2002.

[52] K. Andrews. Visualising cyberspace: Information visualization in the Harmony Internet browser. In *Proceedings '95 Information Visualization*, pages 97–104, Atlanta, Oct. 1995.

[53] K. Andrews, C. Gütl, J. Moser, V. Sabol, and W. Lackner. Search Result Visualisation with xFIND. *Proceedings of User Interfaces to Data Intensive Systems (UIDIS 2001)*, pages 50–58, 2001.

[54] V. N. Anh, O. de Kretser, and A. Moffat. Vector-space ranking with effective early termination. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 35–42, New Orleans, LA, Sept. 2001. ACM Press, New York.

[55] V. N. Anh and A. Moffat. Impact transformation: Effective and efficient Web retrieval. In M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, and K. Järvelin, editors, *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–10, Tampere, Finland, Aug. 2002. ACM Press, New York.

[56] V. N. Anh and A. Moffat. Simplified similarity scoring using term ranks. In *Proc. SIGIR 2005*, pages 226–233, 2005.

[57] V. N. Anh and A. Moffat. Pruned query evaluation using pre-computed impacts. In *SIGIR'06: Proceedings of the 29th International ACM SIGIR conference on Research and Development in Information Retrieval*, Seattle, WA, USA, 2006.

[58] V. N. Anh and A. Moffat. Pruning strategies for mixed-mode querying. In *CIKM'06: Proceedings of the 15th ACM International conference on Information and Knowledge Management*, Arlington, Virginia, USA, 2006.

[59] P. Anick. Using Terminological Feedback for Web Search Refinement: A Log-Based Study. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'03)*, pages 88–95, 2003.

[60] P. Anick, J. Brennan, R. Flynn, D. Hanssen, B. Alvey, and J. Robbins. A direct manipulation interface for Boolean information retrieval via natural language query. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'90)*, pages 135–150, Brussels, Belgium, 1990.

[61] P. Anick and R. Kantamneni. A longitudinal study of real-time search assistance adoption. In *Proceedings of the 31st Annual International ACM SIGIR Conference*

on Research and development in information retrieval (SIGIR'08), pages 701–702, New York, NY, USA, 2008. ACM.

[62] ANSI/NISO Standards. Z39.50-information retrieval: Application service definition and protocol specification. Technical report, International Standard Maintenance Agency, Washington, USA, 1995. See http://lcweb.loc.gov/z3950/agency.

[63] A. Apostolico and Z. Galil, editors. *Combinatorial Algorithms on Words*. Springer-Verlag, New York, 1985.

[64] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, 2001.

[65] M. Araújo, G. Navarro, and N. Ziviani. Large text searching allowing errors. In *Proc. WSP'97*, pages 2–20. Carleton University Press, 1997.

[66] Y. Aridor, D. Carmel, Y. Maarek, A. Soffer, and R. Lempel. Knowledge encapsulation for focused search from pervasive devices. In *WWW'01: Proceedings of the 10th international conference on World Wide Web*, pages 754–764, New York, NY, USA, 2001. ACM.

[67] F. Arman, R. Depommier, A. Hsu, and M.-Y. Chiu. Content-based browsing of video sequences. In *Proceedings of ACM Multimedia*, pages 97–103, 1994.

[68] W. Y. Arms. Implementing Policies for Access Management. *D-Lib Magazine*, February 1998. http://www.dlib.org/dlib/february98/arms/02arms.html.

[69] W. Y. Arms. *Digital Libraries*. MIT Press, 2000.

[70] S. Arnold. How Google's Internet search is transforming application software. In *The Google Legacy*, pages 169–188. Infonortics, 2005.

[71] G. Arocena and A. Mendelzon. WebOQL: Restructuring documents, databases and Webs. In *Int. Conf. on Data Engineering*, pages 24–33, Orlando Florida, 1998.

[72] G. O. Arocena, A. O. Mendelzon, and G. A. Mihaila. Applications of a Web query language. In *Proc. 6th. Int'l. WWW Conf.*, Apr. 1997.

[73] P. Arvola, M. Junkkari, and J. Kekäläinen. Generalized contextualization method for XML information retrieval. In *ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany*, pages 20–27, 2005.

[74] E. Ashoori, M. Lalmas, and T. Tsikrika. Examining topic shifts in content-oriented XML retrieval. *International Journal on Digital Libraries*, 8(1):39–60, 2007.

[75] Ask MyStuff. urlhttp://about.ask.com/en/docs/mystuff/tour.shtml.

[76] Ask.com. Advanced search tips. http://help.ask.com/en/docs/about/adv_search_tips.shtml.

[77] J. A. Aslam and V. Pavlu. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *Advances in Information Retrieval: 28th European Conference on IR Research*, pages 198–209, 2007.

[78] R. Attar and A. Fraenkel. Local feedback in full-text retrieval systems. *Journal of the ACM*, 24(3):397–417, 1977.

[79] G. Attardi and M. Ciaramita. Tree revision learning for dependency parsing. In C. L. Sidner, T. Schultz, M. Stone, and C. Zhai, editors, *HLT-NAACL*, pages 388–395, Rochester, NY, USA, April 2007. The Association for Computational Linguistics.

[80] A. Aula. Enhancing the readability of search result summaries. *Proceedings of HCI 2004*, pages 6–10, 2004.

[81] A. Aula. *Studying user strategies and characteristics for developing Web search interfaces.* PhD thesis, University of Tampere, Finland, Ph.D. Dissertation, Dissertations in Interactive Technology, Number 3., 2005.

[82] S. Axelrod, V. Goel, R. A. Gopinath, P. Olsen, and K. Visweswariah. Subspace constrained Gaussian mixture models for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(6):1144–1160, November 2005.

[83] C. Badue, R. Baeza-Yates, B. A. Ribeiro-Neto, A. Ziviani, and N. Ziviani. Analyzing imbalance among homogeneous index servers in a Web search system. *Information Processing & Management*, 43(3), 2007.

[84] C. S. Badue, J. Almeida, V. Almeida, R. Baeza-Yates, B. A. Ribeiro-Neto, A. Ziviani, and N. Ziviani. Capacity planning for vertical search engines. Submitted for publication, 2009.

[85] C. S. Badue, R. Baeza-Yates, B. A. Ribeiro-Neto, A. Ziviani, and N. Ziviani. Modeling performance-driven workload characterization of Web search systems. In P. S. Yu, V. J. Tsotras, E. A. Fox, and B. Liu, editors, *CIKM*, pages 842–843, Arlington, Virginia, USA, November 2006. ACM.

[86] R. Baeza-Yates. Challenges in the interaction of information retrieval and natural language processing. In A. F. Gelbukh, editor, *5th Int. Conf. on Computational Linguistics and Intelligent Text Processing*, volume 2945 of *Lecture Notes in Computer Science*, pages 445–456, Seoul, South Korea, November 2004. Springer.

[87] R. Baeza-Yates. A fast set intersection algorithm for sorted sequences. In S. C. Sahinalp, S. Muthukrishnan, and U. Dogrusöz, editors, *CPM*, volume 3109 of *Lecture Notes in Computer Science*, pages 400–408, Istanbul,Turkey, 2004. Springer.

[88] R. Baeza-Yates. Query usage mining in search engines. *Web Mining: Applications and Techniques, Anthony Scime, editor. Idea Group*, 2004.

[89] R. Baeza-Yates. Applications of Web query mining. *European Conference on Information Retrieval (ECIR'05), D. Losada, J. Fernández-Luna (editors), Springer LNCS 3408*, pages 7–22, 2005.

[90] R. Baeza-Yates. Graphs from search engine queries. In J. van Leeuwen, G. F. Italiano, W. van der Hoek, C. Meinel, H. Sack, and F. Plasil, editors, *SOFSEM: Theory and Practice of Computer Science*, volume 4362 of *Lecture Notes in Computer Science*, pages 1–8, Harrachov, Czech Republic, January 2007. Springer.

[91] R. Baeza-Yates, P. Boldi, and C. Castillo. Generalizing PageRank: Damping functions for link-based ranking algorithms. In *Proceedings of SIGIR*, Seattle, Washington, USA, August 2006. ACM Press.

[92] R. Baeza-Yates, P. Boldi, and C. Castillo. Generic damping functions for propagating importance in link-based ranking. *Internet Mathematics*, 3(4):445–478, 2006.

[93] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The intention behind Web queries. In F. Crestani, P. Ferragina, and M. Sanderson, editors, *Proceedings of String Processing and Information Retrieval (SPIRE)*, volume 4209 of *Lecture Notes in Computer Science*, pages 98–109. Springer, 2006.

[94] R. Baeza-Yates and C. Castillo. Relating Web characteristics with link based web page ranking. In *Proceedings of String Processing and Information Retrieval SPIRE*, pages 21–32, Laguna San Rafael, Chile, 2001. IEEE CS Press.

[95] R. Baeza-Yates and C. Castillo. Balancing volume, quality and freshness in Web crawling. In *Soft Computing Systems - Design, Management and Applications*, pages 565–572, Santiago, Chile, 2002. IOS Press Amsterdam.

[96] R. Baeza-Yates and C. Castillo. Crawling the infinite web: five levels are enough. *Journal of Web Engineering*, 6:49–72, 2007.

[97] R. Baeza-Yates, C. Castillo, and E. Efthimiadis. Characterization of national Web domains. *ACM TOIT*, 7(2), 2007.

[98] R. Baeza-Yates, C. Castillo, and F. S. Jean. Web dynamics, structure and page quality. In M. Levene and A. Poulovassilis, editors, *Web Dynamics*, pages 93–109. Springer, 2004.

[99] R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, and F. Silvestri. Challenges on distributed Web retrieval. In *Proceedings of ICDE 2007*, pages 6–20. IEEE, 2007.

[100] R. Baeza-Yates, C. Castillo, M. Marín, and A. Rodríguez. Crawling a country: Better strategies than breadth-first for Web page ordering. In *Proceedings of the 14th international conference on World Wide Web*, pages 864–872, Chiba, Japan, 2005. ACM Press.

[101] R. Baeza-Yates, M. Ciaramita, P. Mika, and H. Zaragoza. Towards semantic search. In E. Kapetanios, V. Sugumaran, and M. Spiliopoulou, editors, *Natural Language and Information Systems, 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008*, volume 5039 of *Lecture Notes in Computer Science*, pages 4–11, London, UK, June 2008. Springer.

[102] R. Baeza-Yates, W. Cunto, U. Manber, and S. Wu. Proximity matching using fixed-queries trees. In *Proc. of Combinatorial Pattern Matching*, number 807 in LNCS, pages 198–212. Springer-Verlag, 1994.

[103] R. Baeza-Yates, N. Fuhr, and Y. Maarek. Second edition of the XML and information retrieval workshop held at sigir'2002. *SIGIR Forum*, 36(2):53–57, 2002.

[104] R. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. The Impact of Caching on Search Engines. In *SIGIR'07: Proceedings of the 30th International ACM SIGIR conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 2007.

[105] R. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. Design trade-offs for search engine caching. *TWEB*, 2(4), 2008.

[106] R. Baeza-Yates, A. Gionis, F. Junqueira, V. Plachouras, and L. Telloli. On the feasibility of multi-site Web search engines. In *ACM CIKM 2009*, pages 425–434, Hong Kong, China, November 2009.

[107] R. Baeza-Yates and G. Gonnet. Efficient Text Searching of Regular Expressions. In G. Ausiello, M. Dezani-Ciancaglini, and S. R. D. Rocca, editors, *ICALP'89*, number 372 in LNCS, pages 46–62, Stresa, Italy, 1989. Springer-Verlag.

[108] R. Baeza-Yates and G. Gonnet. Fast text searching for regular expressions or automaton searching on a trie. *J. of the ACM*, 43(6):915–936, Nov 1996.

[109] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query clustering for boosting Web page ranking. *Advances in Web Intelligence, AWIC 2004, Springer LNCS*, 3034:164–175, 2004.

[110] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *Current Trends in Database Technology - EDBT*, volume 3268, pages 588–596. Springer-Verlag GmbH, 2004.

[111] R. Baeza-Yates, C. A. Hurtado, and M. Mendoza. Improving search engines by query clustering. *JASIST*, 58(12):1793–1804, 2007.

[112] R. Baeza-Yates, F. Junqueira, V. Plachouras, and H. F. Witschel. Admission Policies for Caches of Search Engine Results. In *SPIRE'07: Proceedings of the 14th Symposium on String Processing and Information Retrieval*, Santiago, Chile, 2007.

[113] R. Baeza-Yates, A. Moffat, and G. Navarro. Searching large text collections. In J. Abello, P. M. Pardalos, and M. G. C. Resende, editors, *Handbook of Massive Data Sets*, pages 195–244. Kluwer Academic Publishers, 2002.

[114] R. Baeza-Yates, V. Murdock, and C. Hauff. Efficiency trade-offs in two-tier Web search systems. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *SIGIR*, pages 163–170, Boston, MA, USA, 2009. ACM.

[115] R. Baeza-Yates and G. Navarro. Integrating contents and structure in text retrieval. *ACM SIGMOD Record*, 25(1):67–79, Mar. 1996.

[116] R. Baeza-Yates and G. Navarro. Block-addressing indices for approximate text retrieval. In *Proc. of the 6th CIKM Conference*, pages 1–8, Las Vegas, Nevada, 1997.

[117] R. Baeza-Yates and G. Navarro. Faster approximate string matching. *Algorithmica*, 23(2):127–158, 1999.

[118] R. Baeza-Yates, G. Navarro, J. Vegas, and P. de la Fuente. A model and a visual query language for structured text. In B. A. Ribeiro-Neto, editor, *Proc. of the 5th Symposium on String Processing and Information Retrieval*, pages 7–13, Santa Cruz, Bolivia, Sept 1998. IEEE CS Press.

[119] R. Baeza-Yates, N.Fuhr, and Y. Maarek. Introduction to the special issue on XML retrieval. *ACM Transactions on Information Systems*, 24(4):405–406, 2006.

[120] R. Baeza-Yates, A. Pereira, and N. Ziviani. Genealogical trees on the Web: A search engine user perspective. In *WWW'08: Proceedings of the 17th international conference on World Wide Web*, pages 367–376, Beijing, China, 2008.

[121] R. Baeza-Yates and B. Poblete. Dynamics of the Chilean Web structure. In *Proceedings of the 3rd International Workshop on Web Dynamics*, New York, USA, 2004.

[122] R. Baeza-Yates and B. Poblete. A website mining model centered on user queries. In *Semantics, Web and Mining, Joint International Workshops, EWMF 2005 and KDO 2005*, volume 4289 of *Lecture Notes in Computer Science*, pages 1–17, Porto, Portugal, October 2005. Springer.

[123] R. Baeza-Yates and F. Saint-Jean. A three level search engine index based in query log distribution. In M. A. Nascimento, E. S. de Moura, and A. L. Oliveira, editors, *SPIRE*, volume 2857 of *Lecture Notes in Computer Science*, pages 56–65, Manaus, Brazil, October 2003. Springer.

[124] R. Baeza-Yates and A. Salinger. Experimental analysis of a fast intersection algorithm for sorted sequences. In M. P. Consens and G. Navarro, editors, *SPIRE*, volume 3772 of *Lecture Notes in Computer Science*, pages 13–24, Buenos Aires, Argentina, November 2005. Springer.

[125] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In P. Berkhin, R. Caruana, and X. Wu, editors, *KDD*, pages 76–85, San Jose, CA, USA, 2007. ACM.

[126] P. Bailey, N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the trec 2007 enterprise track. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, Gaithersburg, MD, 2007. NIST. TREC Special Publication: SP 500-274. trec.nist.gov/pubs/trec16/papers/ENT.OVERVIEW16.pdf.

[127] P. Bailey, D. Hawking, and B. Matson. Secure search in enterprise webs: Tradeoffs in efficient implementation for document level security. In *Proceedings of CIKM 2006*, 2006. http://es.csiro.au/pubs/cikm127_bailey.pdf.

[128] D. Bainbridge, J. Thompson, and I. H. Witten. Assembling and enriching digital library collections. In *JCDL'03: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 323–334, Houston, Texas, 2003.

[129] J. Baker. UCLA-NSF Social Aspects of Digital Libraries Workshop, January 1996. http://www.gslis.ucla.edu/DL/.

[130] P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. John Wiley & Sons, May 2003.

[131] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *Information Processing and Management*, 45(1):1–19, January 2009. doi:10.1016/j.ipm.2008.06.003.

[132] K. Balog and M. de Rijke. Combining candidate and document models for expert search. In *The Seventeenth Text Retrieval Conference (TREC 2008)*. NIST, 2009. Special Publication.

[133] S. Baluja and M. Covell. Learning 'forgiving' hash functions: Algorithms and large-scale tests. In *International Joint Conference on AI*, January 2007.

[134] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for YouTube: Taking random walks through the view graph. In *WWW'08: Proceeding of the 17th International Conference on World Wide Web*, pages 895–904, New York, NY, USA, 2008. ACM.

[135] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. In *WWW'06: Proceedings of the 15th international conference on World Wide Web*, pages 367–376, New York, NY, USA, 2006. ACM Press.

[136] Z. Bar-Yossef and M. Gurevich. Efficient search engine measurements. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors, *WWW*, pages 401–410, Banff, Canada, 2007. ACM.

[137] Z. Bar-Yossef and M. Gurevich. Mining search engine query logs via suggestion sampling. In *VLDB 2008*, 2008.

[138] Z. Bar-Yossef and M. Gurevich. Estimating the impression rank of Web pages. In *WWW 2009*, 2009.

[139] A.-L. Barabási. *Linked: the New Science of Networks*. Perseus Books Group, May 2002.

[140] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[141] M. Barbaro and T. Zeller. A face is exposed for AOL searcher no. 4417749. New York Times, 2006.

[142] J. Barbay and C. Kenyon. Adaptive intersection and t-threshold problems. In *SODA*, pages 390–399, 2002.

[143] J. Barbay, A. López-Ortiz, T. Lu, and A. Salinger. An experimental investigation of set intersection algorithms for text searching. *Journal of Experimental Algorithms (JEA)*, 14(3):7–24, 2009.

[144] L. Barbosa, F. Junqueira, V. Plachouras, and R. Baeza-Yates. Variability as a measure a search quality, 2009. Submitted.

[145] R. Barbosa. *Query Performance on Distributed Digital Libraries*. CS Department, Federal University of Minas Gerais, Brazil, 1998. Master Thesis. In Portuguese.

[146] P. Barford and M. Crovella. Generating representative Web workloads for network and server performance evaluation. In *ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 151–160, July 1998.

[147] H. Barlow. Unsupervised learning. *Neural Computation*, 1(3):295–311, 1989.

[148] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Special Issue on Text and Images, Journal of Machine Learning Research*, 2002.

[149] J. Baron, D. Lewis, and D. Oard. TREC-2006 Legal Track Overview. In *Proceedings of TREC 2006*, pages 79–98. NIST, 2007. http://trec.nist.gov/pubs/trec15/t15_proceedings.html.

[150] D. Barreau and B. Nardi. Finding and Reminding: File Organization From the Desktop. *ACM SIGCHI Bulletin*, 27(3):39–43, 1995.

[151] L. A. Barroso, J. Dean, and U. Hölzle. Web search for a planet: the Google Cluster Architecture. *IEEE Micro Magazine*, 23(2):22–28, Mar./Apr. 2003.

[152] L. A. Barroso and U. Hölzle. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, volume 6 of *Synthesis Lectures on Computer Architecture*. Morgan Claypool, 2009.

[153] B. T. Bartell, G. W. Cottrell, and R. K. Belew. Latent semantic indexing is an optimal special case of multidimensional scaling. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Information Retrieval Theory, pages 161–167, 1992.

[154] M. Bartsch and G. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 15–18, Octobter 2001.

[155] N. A. Basbanes. Foreword. The Library – An Illustrated History, authored by Stuart A.P. Murray, published by Skyhorse Publishing, 2009.

[156] M. Bates. Information search tactics. *Journal of the American Society for Information Science*, 30(4):205–214, 1979.

[157] M. Bates. The design of browsing and berrypicking techniques for the on-line search interface. *Online Review*, 13(5):407–431, 1989.

[158] M. Bates. Where should the person stop and the information search interfaces start? *Information Processing and Management*, 26(5), 1990.

[159] M. Bates. Improving user access to library catalog and portal information. Task force recommendation 2.3, final report (version 3), Library of Congress Bicentennial Conference on Bibliographic Control for the New Millenium, 2003. http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf.

[160] P. Baudisch, B. Lee, and L. Hanna. Fishnet, a fisheye Web browser with search term popouts: a comparative evaluation with overview and linear view. *Proceedings of the working conference on Advanced Visual Interfaces (AVI'04)*, pages 133–140, 2004.

[161] H. Bay, T. Tuytelaars, and L. J. V. Gool. Surf: Speeded up robust features. In *ECCV (1)*, pages 404–417, 2006.

[162] D. Bearman. Digital Libraries. In B. Cronin, editor, *Annual Review of Information Science and Technology*, volume 41, pages 223–272. American Society for Information Science and Technology, 2007.

[163] M. Beaudouin-Lafon and W. Mackay. Prototyping Tools and Techniques. In *Human-Computer Interaction Handbook*. Lawrence Erlrbaum Associates, 2003.

[164] J. Becker and R. Hayes. *Information storage and retrieval: tools, elements, theories*. Wiley, 1963.

[165] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: An efficient and robust access method for points and rectangles. *ACM SIGMOD*, pages 322–331, May 1990.

[166] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 407–416. ACM, 2000.

[167] C. Beeri and Y. Kornatzky. A logical query language for hypertext systems. In *Proc. of the European Conference on Hypertext*, pages 67–80. Cambridge University Press, 1990.

[168] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized Web query log. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 321–328, Sheffield, United Kingdom, 2004. ACM Press.

[169] S. M. Beitzel, E. C. Jensen, O. Frieder, D. A. Grossman, D. D. Lewis, A. Chowdhury, and A. Kolcz. Automatic Web query classification using labeled and unlabeled training data. In R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, editors, *SIGIR*, pages 581–582, Salvador, Brazil, August 2005. ACM.

[170] R. Belew. *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge University Press, 2000.

[171] N. Belkin, D. Kelly, G. Kim, J. Kim, H. Lee, G. Muresan, M. Tang, X. Yuan, and C. Cool. Query length in interactive information retrieval. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'03)*, pages 205–212, 2003.

[172] R. M. Bell, Y. Koren, and C. Volinsky. Chasing $1,000,000: How we won the Netflix progress prize. *Statistical Computing and Statistical Graphics Newsletter*, 18:4–12, 2007.

[173] T. Bell, J. Cleary, and I. Witten. Data compression using adaptive coding and partial string matching. *IEEE Trans. on Communications*, 32(4):396–402, 1984.

[174] T. Bell, J. Cleary, and I. Witten. *Text Compression*. Prentice-Hall, 1990.

[175] T. C. Bell, A. Moffat, C. Nevill-Manning, I. H. Witten, and J. Zobel. Data compression in full-text retrieval systems. *Journal of the American Society for Information Science*, 44:508–531, 1993.

[176] A. Belussi and C. Faloutsos. Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In *Proc. of VLDB Conf.*, pages 299–310, Zurich, Switzerland, Sept. 1995.

[177] Y. Ben-Aharon, S. Cohen, Y. Grumbach, Y. Kanza, J. Mamou, Y. Sagiv, B. Sznajder, and E. Twito. Searching in an XML corpus using content and structure. In *INEX 2003 Proceedings*, pages 46–52, 2003.

[178] I. Ben-Shaul, M. Herscovici, M. Jacovi, Y. Maarek, D. Pelleg, M. Shtalhaim, V. Soroka, and S. Ur. Adding support for dynamic and focused search with fetuccino. *Computer Networks*, 31(11-16):1653–1665, 1999. Also appeared in the Proceedigns of WWW8.

[179] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Improving Collection Selection With Overlap Awareness in P2P Search Engines. In *SIGIR'05: Proceedings of the 28th International ACM SIGIR conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005.

[180] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. MINERVA: collaborative P2P search. In *VLDB'05: Proceedings of the 31st International conference on Very Large Data Bases*, Trondheim, Norway, 2005.

[181] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. P2P content search: Give the Web back to the people. International Workshop on Peer-to-Peer Systems (IPTPS), February 2006.

[182] Y. Benkler. Coase's penguin, or, Linux and the nature of the firm. *Yale Law Journal*, 112:371–446, 2002.

[183] P. N. Bennett, S. T. Dumais, and E. Horvitz. The combination of text classifiers using reliability indicators. *Inf. Retr.*, 8(1):67–100, 2005.

[184] J. Bentley, D. Sleator, R. Tarjan, and V. Wei. A locally adaptive data compression scheme. *Communications of the ACM*, 29(4):320–330, apr 1986.

[185] S. Berchtold, D. A. Keim, and H.-P. Kriegel. The X-tree : An index structure for high-dimensional data. *VLDB*, pages 28–39, 1996.

[186] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Who's in the picture. In *Proceedings of the Neural Information Processing Society*, 2004.

[187] A. Berger and J. Lafferty. Information retrieval as a statistical translation. In *ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, 1999.

[188] D. Bergmark. Collection synthesis. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 253–262, Portland, OR, 2002.

[189] P. Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2(2):73–120, 2005.

[190] T. Berners-Lee. The World Wide Web Consortium. http://www.w3.org.

[191] T. Berners-Lee. Universal resource identifiers in WWW. RFC 1630. http://www.w3.org/Addressing/rfc1630.txt, 1994.

[192] T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, and A. Secret. The World-Wide Web. *Comm. of the ACM*, 37(8):76–82, 1994.

[193] T. Berners-Lee, R. Fielding, and L. Masinter. Uniform resource identifiers (uri): Generic syntax. RFC 2396. http://www.ietf.org/rfc/rfc2396.txt, 1998.

[194] M. Berry and M. Browne. *Understanding Search Engines – Mathematical Modeling and Text Retrieval*. Siam, 2005.

[195] S. Betsi, M. Lalmas, A. Tombros, and T. Tsikrika. User expectations from XML element retrieval. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA*, pages 611–612, 2006.

[196] K. Bharat. Searchpad: explicit capture of search context to support Web search. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications networking*, pages 493–501, Amsterdam, the Netherlands, the Netherlands, 2000. North-Holland Publishing Co.

[197] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The connectivity server: fast access to linkage information on the Web. In *7th WWW Conf.*, Brisbane, Australia, April 1998.

[198] K. Bharat and A. Z. Broder. A technique for measuring the relative size and overlap of public Web search engines. In *7th WWW Conference*, pages 379–388, Brisbane, Australia, 1998.

[199] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104–111, Melbourne, Australia, August 1998. ACM Press, New York.

[200] M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Trans. Inter. Tech.*, 5(1):92–128, February 2005.

[201] D. Bilal. Children's use of the Yahooligans! Web search engine. I. Cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American Society for Information Science*, 51(7):646–665, 2000.

[202] Bing. Advanced search keywords. http://help.live.com/Help.aspx?market=en-US&project=WL_Searchv1&querytype=topic&query=WL_SEARCH_REF_Keywords.htm.

[203] Bing. Use advanced search. http://help.live.com/help.aspx?mkt=en-us&project=wl_searchv1&querytype=keyword&query=redliub&tmt=&domain=www.bing.com:80.

[204] J. Bing. *Handbook of Legal Information Retrieval*. Elsevier Science Inc., New York, NY, USA, 1984.

[205] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[206] R. Blanco and A. Barreiro. Document identifier reassignment through dimensionality reduction. In *Advances in Information Retrieval: 27th European Conference on IR research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings*, pages 375 – 387, 2005.

[207] A. Blandford, S. Keith, I. Connell, and H. Edwards. Analytical usability evaluation for digital libraries: a case study. In *Proc. of JCDL'04*, pages 27–36, Tucson, AZ, 2004.

[208] A. Blandford, H. Stelmaszewska, and N. Bryan-Kinns. Use of multiple digital libraries: A case study. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 179–188, Roanoke, Virginia, 2001.

[209] D. Blandford and G. Blelloch. Index compression through document reordering. In *Proceedings of the Data Compression Conference (DCC'02)*, pages 342–351, Washington, DC, USA, 2002. IEEE Computer Society.

[210] H. Blanken, T. Grabs, H.-J. Schek, R. Schenkel, and G. Weikum, editors. *Intelligent Search on XML Data, Applications, Languages, Models, Implementations, and Benchmarks*, volume 2818. Springer, 2003.

[211] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–134, New York, NY, USA, 2003. ACM.

[212] J. F. Blinn. What's that deal with the DCT? *Computer Graphics and Applications*, 13:78–83, July 1993.

[213] B. H. Bloom. Space/Time Trade-offs in Hash Coding with Allowable Errors. *Communications of the ACM*, 13(7), 1970.

[214] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[215] J. Blustein, I. Ahmed, and K. Instone. An evaluation of look-ahead breadcrumbs for the WWW. In S. Reich and M. Tzagarakis, editors, *Hypertext*, pages 202–204, Salzburg, Austria, September 2005. ACM.

[216] Budapest open access initiative, 2001. http://www.soros.org/openaccess/.

[217] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 609–618, Napa Valley, California, USA, 2008. ACM.

[218] P. Boldi, F. Bonchi, C. Castillo, and S. Vigna. From "Dango" to "Japanese Cakes": Query reformulation models and patterns. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 183–190, Milano, Italy, 2009. IEEE CS Press.

[219] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler: a scalable fully distributed Web crawler. *Software, Practice and Experience*, 34(8):711–726, 2004.

[220] P. Boldi, M. Santini, and S. Vigna. Do your worst to make the best: Paradoxical effects in pagerank incremental computations. In *Proceedings of the third Workshop on Web Graphs (WAW)*, volume 3243 of *Lecture Notes in Computer Science*, pages 168–180, Rome, Italy, October 2004. Springer.

[221] P. Boldi, M. Santini, and S. Vigna. Pagerank as a function of the damping factor. In *Proceedings of the 14th international conference on World Wide Web*, pages 557–566, Chiba, Japan, 2005. ACM Press.

[222] P. Boldi and S. Vigna. The webgraph framework I: Compression techniques. In S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, editors, *Proceedings of the 13th conference on World Wide Web*, pages 595–602, New York, NY, USA, 2004. ACM Press.

[223] J. Bollen, R. Luce, S. S. Vemulapalli, and W. Xu. Usage analysis for the identification of research trends in digital libraries. *D-Lib Magazine*, 9, 2003.

[224] M. Bolsky and D. Korn. *The New KornShell Command and Programming Language*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1995.

[225] C. Bonacic, C. García, M. Marín, M. Prieto, and F. Tirado. Exploiting Hybrid Parallelism in Web Search Engines. In *14th European International Conference on Parallel Processing (Euro-Par 2008)*, LNCS 5168, pages 414–423, Las Palmas de Gran Canaria, Spain, August 2008.

[226] P. Bonnet and A. Tomasic. Partial answers for unavailable data sources. In *Workshop on Flexible Query-Answering Systems*, pages 43–54, 1998.

[227] A. Bookstein. On the perils of merging Boolean and weighted retrieval systems. *Journal of the American Society for Information Sciences*, 29(3):156–158, 1978.

[228] A. Bookstein. Fuzzy requests: An approach to weighted Boolean searches. *Journal of the American Society for Information Sciences*, 31:240–247, 1980.

[229] A. Bookstein. Implication of Boolean structure for probabilistic retrieval. In *Proc of the Eight Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 11–17, Montreal, Canada, 1985.

[230] J. Boreczky, A. Girgensohn, G. Golovchinsky, and S. Uchihashi. An interactive comic book presentation for exploring video. In *CHI '00: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 185–192, New York, NY, USA, 2000. ACM.

[231] J. Boreczky and L. Rowe. Comparison of video shot boundary detection techniques. In *IS&T SPIE Symposium on Electronic Imaging*, volume 2670, pages 170–179, San Jose, 1996.

[232] C. Borgman. Why are online catalogs *still* hard to use? *Journal of the American Society for Information Science*, 47(7):493–503, 1996.

[233] C. L. Borgman. Social aspects of digital libraries. In *DL'96: Proceedings of the 1st ACM International Conference on Digital Libraries*, D-Lib Working Session 2A, pages 170–171, 1996.

[234] C. L. Borgman. What are digital libraries? competing visions. *Inf. Process. Manage.*, 35(3):227–243, 1999.

[235] C. L. Borgman, G. Leazer, A. J. Gilliland-Swetland, K. Millwood, C. Champeny, J. Finley, and L. J. Smart. How geography professors select materials for classroom lectures: Implications for the design of digital libraries. In *Proc. of JCDL'04*, pages 179–185, Tucson, AZ, 2004.

[236] C. L. Borgman, G. H. Leazer, A. J. Gilliland-Swetland, and R. Gazan. Iterative design and evaluation of a geographic digital library for university students: A case study of the Alexandria Digital Earth Prototype (ADEPT). *LNCS*, 2163:390, 2001.

[237] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.

[238] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *WWW'01: Proceedings of the 10th international conference on World Wide Web*, pages 415–429, New York, NY, USA, 2001. ACM Press.

[239] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Inter. Tech.*, 5(1):231–297, February 2005.

[240] D. Borthakur. *The Hadoop Distributed File System: Architecture and Design*. The Apache Software Foundation, 2007.

[241] J. Bosak. XML, Java, and the future of the Web. Technical report, Sun Microsystems, 1997. http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm.

[242] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *European Conference on Computer Vision*, 2006.

[243] C. P. Bourne and T. B. Hahn. *A history of online information services, 1963-1976*. MIT Press, Cambridge, Mass., 2003.

[244] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz. The Harvest information discovery and access system. In *Proc. 2nd Inter. World Wide Web Conf.*, pages 763–771, Oct. 1994.

[245] J. Boyan, D. Freitag, and T. Joachims. A machine learning architecture for optimizing web search engines. In *AAAI Workshop on Internet Based Information Systems*, pages 1 – 8, August 1996.

[246] R. S. Boyer and J. S. Moore. A fast string searching algorithm. *Communications of the ACM*, 20(10):762–772, 1977.

[247] T. Bozkaya and Z. M. Ozsoyoglu. Distance-based indexing for high-dimensional metric spaces. In *Proc. of ACM SIGMOD Conference*, pages 357–368, Tucson, AZ, USA, 1997.

[248] P. D. Bra and R. Post. Searching for arbitrary information in the WWW: the fish search for Mosaic. In *Proc. of the Second International World Wide Web Conference*, Chicago, Oct. 1994. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/www-fall94.html.

[249] A. Bratko and B. Filipic. Exploiting structural information for semi-structured document categorization. *Information Processing and Management*, 42(3):679–694, 2006.

[250] T. Bray. Measuring the web. In *Fifth International World Wide Web Conference*, Paris, May 1996. http://www5conf.inria.fr/fich_html/papers/P9/Overview.html.

[251] M. Breaks. The eLib Hybrid Library Projects. *Ariadne*, (28), 2001. http://www.ariadne.ac.uk/issue28/hybrid/.

[252] M. Breeding. Musings on the state of the ILS in 2006. *Computers in Libraries*, 26(26):26–29, 2006.

[253] M. Breeding. Making a business case for open source ILS. *Computers in Libraries*, 28(28):36–39, 2008.

[254] M. Breeding. Open source integrated library systems. *Library Technology Reports*, 44(8), 2008.

[255] M. Breeding. Opportunity out of turmoil. *Library Journal*, 133(6):32, 2008.

[256] L. Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996.

[257] B. Brewington and G. Cybenko. Keeping up with the changing web. *IEEE Computer*, 33(5):52–58, May 2000.

[258] B. Brewington, G. Cybenko, R. Stata, K. Bharat, and F. Maghoul. How dynamic is the web? In *Proceedings of the Ninth Conference on World Wide Web*, Amsterdam, Netherlands, May 2000. ACM Press.

[259] Bright-Planet. Deep Web white paper. Available online at brightplanet.com, July 2000.

[260] S. Brin. Near neighbor search in large metric spaces. In *Proc. of VLDB Conf.*, pages 574–584, Zurich, Switzerland, Sept 1995.

[261] S. Brin. Extracting patterns and relations from the World Wide Web. In *Workshop on Web Databases*, Valencia, Spain, March 1998.

[262] S. Brin, J. Davis, and H. Garcia-Molina. Copy detection mechanisms for digital documents. In M. J. Carey and D. A. Schneider, editors, *SIGMOD Conference*, pages 398–409, San Jose, CA, USA, 1995. ACM Press.

[263] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *WWW7: Proceedings of the Seventh International Conference on World Wide Web 7*, pages 107–117. Elsevier Science Publishers B. V., 1998.

[264] T. Brinkhoff, H.-P. Kriegel, R. Schneider, and B. Seeger. Multi-step processing of spatial joins. In *Proc. of ACM SIGMOD*, pages 197–208, Minneapolis, MN, USA, May 1994.

[265] N. Brisaboa, A. Fariña, G. Navarro, and M. Esteller. (s,c)-dense coding: An optimized compression code for natural language text databases. In *Proceedings of the 10th International Symposium on String Processing and Information Retrieval (SPIRE 2003)*, LNCS 2857, pages 122–136. Springer, 2003.

[266] N. Brisaboa, A. Fariña, G. Navarro, and J. Paramá. Efficiently decodable and searchable natural language adaptive compression. In *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*, pages 234–241. ACM Press, 2005.

[267] A. Broder. On the resemblance and containment of documents. In *SEQUENCES: Conf. on Compression and Complexity of Sequences*, pages 21–29, Salerno, Italy, 1997. IEEE Computer Society.

[268] A. Broder. A taxonomy of Web search. *ACM SIGIR Forum*, 36(2):3–10, 2002. http://www.acm.org/sigir/forum/F2002/broder.pdf.

[269] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic clustering of the Web. In *6th Int'l WWW Conference*, pages 391–404, Santa Clara, CA, USA, April 1997.

[270] A. Broder and V. Josifovski. Introduction to computational advertising. Course at Stanford University, http://www.stanford.edu/class/msande239/, Sept-Dec 2009.

[271] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: Experiments and models. In *Proceedings of the Ninth Conference on World Wide Web*, pages 309–320, Amsterdam, Netherlands, May 2000. ACM Press.

[272] A. Z. Broder. The future of Web search: From information retrieval to information supply. In *NGITS*, page 362, 2006.

[273] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *Proceedings of CIKM 2003*, pages 426–434, New York, NY, USA, 2003. ACM Press.

[274] A. Z. Broder and A. C. Ciccolo. Towards the next generation of enterprise search technology. *IBM Syst. J.*, 43(3):451–454, 2004.

[275] A. Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. Search advertising using Web relevance feedback. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1013–1022, New York, NY, USA, 2008. ACM.

[276] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using Web knowledge. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR*, pages 231–238, Amsterdam, The Netherlands, 2007 2007. ACM.

[277] A. Z. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR*, pages 559–566, Amsterdam, The Netherlands, November 2007. ACM.

[278] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. In *Selected papers from the sixth international conference on World Wide Web*, pages 1157–1166, Essex, UK, 1997. Elsevier Science Publishers Ltd.

[279] A. Z. Broder and Y. S. Maarek, editors. *Proceedings of the SIGIR 2006 Workshop on Faceted Search*, Seattle, WA, USA, August 2006.

[280] J. Broglio, J. Callan, W. Croft, and D. Nachbar. Document retrieval and routing using the INQUERY system. In D. Harman, editor, *Overview of the Third Retrieval Conference (TREC-3)*, pages 29–38. NIST Special Publication 500-225, 1995.

[281] A. Broschart, R. Schenkel, M. Theobald, and G. Weikum. TopX @ INEX 2007. In *Focused access to XML documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*, Dagstuhl Castle, Germany, 2008. Selected Papers.

[282] E. W. Brown. *Execution Performance Issues in Full-Text Information Retrieval*. PhD thesis, University of Massachusetts, Amherst, 1996. Available as UMass Comp. Sci. Tech. Rep. TR95-81.

[283] S. Browne, J. Dongarra, J. Horner, P. McMahan, and S. Wells. Technologies for repository interoperation and access control. In *Proceedings of the 3rd ACM International Conference on Digital Libraries*, pages 40–48, 1998.

[284] R. Brunelli, O. Mich, and C. Modena. A survey on video indexing. Technical Report 9612-06, IRST, 1996.

[285] P. Bruza, R. McArthur, and S. Dennis. Interactive Internet Search: Keyword, Directory and Query Reformulation Mechanisms Compared. *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'00)*, pages 280–287, 2000.

[286] P. D. Bruza and T. P. Van Der Weide. Stratified hypermedia structures for information disclosure. *The Computer Journal*, 35(3):208–220, 1992.

[287] G. Buchanan, D. Bainbridge, K. J. Don, and I. H. Witten. A new framework for building digital library collections. In *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries (JCDL05)*, pages 23–31, Denver, CA, USA, 2005.

[288] G. Buchanan, S. Cunningham, A. Blandford, J. Rimmer, and C. Warwick. Information Seeking by Humanities Scholars. *Proceedings of the European Conference on Digital Libraries (ECDL'05)*, 2005.

[289] C. Buckley and G. Salton. Optimization of relevance feedback weights. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum)*, pages 351–357, 1995.

[290] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In *Proc. of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 292–300, Dublin, Ireland, 1994.

[291] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40, 2000.

[292] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, 2004.

[293] N. Budhiraja, K. Marzullo, F. Schneider, and S. Toueg. Optimal primary-backup protocols. In *Proceedings of the International Workshop on Distributed Algorithms (WDAG)*, pages 362–378, Haifa, Israel, November 1992. Springer-Verlag.

[294] P. Buneman. Semistructured data. In *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 117–121, Tucson, Arizona, 1997.

[295] C. J. C. Burges, R. Ragno, and Q. V. Le. Learning to rank with nonsmooth cost functions. In B. Schölkopf, J. C. Platt, T. Hoffman, B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*, pages 193–200. MIT Press, 2006.

[296] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender. Learning to rank using gradient descent. In L. D. Raedt and S. Wrobel, editors, *ICML*, volume 119, pages 89–96, Bonn, Germany, August 2005. ACM Press.

[297] W. Burkhard and R. Keller. Some approaches to best-match file searching. *Communications of the ACM*, 16(4):230–236, Apr. 1973.

[298] F. Burkowski. An algebra for hierarchically organized text-dominated databases. *Information Processing & Management*, 28(3):333–348, 1992.

[299] F. Burkowski. Retrieval activities in a database consisting of heterogeneous collections of structured text. In *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark*, pages 112–125, 1992.

[300] A. W. Burks, H. H. Goldstine, and J. von Neumann. Preliminary discussion of the logical design of an electronic computing instrument. In W. Aspray and A. Burks, editors, *Papers of John von Neumann on Computers and Computer Theory*, pages 97–142. The MIT Press, Cambridge, MA, 1987. (originally appeared in 1946).

[301] M. Burner. Crawling towards eternity - building an archive of the world wide web. *Web Techniques*, 2(5), May 1997.

[302] M. Burrows and D. Wheeler. A block sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation, 1994.

[303] V. Bush. As we may think. *The Atlantic Monthly*, July 1945.

[304] S. Büttcher, C. L. Clarke, and G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010.

[305] D. Byrd. A Scrollbar-based Visualization for Document Navigation. *Proceedings of the Fourth ACM International Conference on Digital Libraries*, 1999.

[306] D. Byrd and R. Podorozhny. Adding Boolean-quality control to best-match searching via an improved user interface. Technical Report IR-210, Computer Science Dept., Univ. of Massachusetts/Amherst, 2000.

[307] J. Byrum, Jr. Recommendations for urgently needed improvement of OPAC and the role of the national bibliographic agency in achieving it. In *World Library and Information Congress: 71th IFLA General Conference and Council, August 14-18, Oslo, Norway*, 2005. http://www.ifla.org/IV/ifla71/papers/124e-Byrum.pdf.

[308] F. Cacheda, V. Carneiro, V. Plachouras, and I. Ounis. Performance analysis of distributed information retrieval architectures using an improved network simulation model. *Information Processing and Management*, 43(1):204–224, 2007.

[309] B. Cahoon and K. McKinley. Performance evaluation of a distributed architecture for information retrieval. In *Proc. 19th Inter. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 110–118, Zurich, Switzerland, Aug. 1996.

[310] N. Caidi and A. Clement. Digital libraries and community networking: the canadian experience. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, page 386, Tucson, Arizona, 2004.

[311] P. Calado, B. A. Ribeiro-Neto, N. Ziviani, E. Moura, and I. Silva. Local versus global link information in the web. *ACM Trans. Inf. Syst.*, 21(1):42–63, January 2003.

[312] P. Calado, M. Cristo, E. S. de Moura, N. Ziviani, B. A. Ribeiro-Neto, and M. A. Gonçalves. Combining link-based and content-based methods for Web document classification. In *CIKM*, pages 394–401, 2003.

[313] P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, B. A. Ribeiro-Neto, and N. Ziviani. Link-based similarity measures for the classification of Web documents. *J. Am. Soc. Inf. Sci. Technol.*, 57(2):208–221, 2006.

[314] P. Calado, A. S. da Silva, A. H. F. Laender, B. A. Ribeiro-Neto, and R. C. Vieira. A Bayesian network approach to searching Web databases through keyword-based queries. *Inf. Process. Manage.*, 40(5):773–790, 2004.

[315] P. Calado, A. S. da Silva, R. C. Vieira, A. H. F. Laender, and B. A. Ribeiro-Neto. Searching Web databases by structuring keyword-based queries. In *CIKM*, pages 26–33, 2002.

[316] P. Calado, M. A. Gonçalves, E. A. Fox, B. A. Ribeiro-Neto, A. H. F. Laender, A. S. da Silva, D. C. Reis, P. A. Roberto, M. V. Vieira, and J. P. Lage. The web-DL environment for building digital libraries from the web. In *JCDL'03: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 346–357, Houston, Texas, 2003.

[317] P. Calado and B. A. Ribeiro-Neto. An information retrieval approach for approximate queries. *IEEE Trans. Knowl. Data Eng.*, 15(1):236–239, 2003.

[318] J. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310. Springer-Verlag New York, Inc., 1994.

[319] J. Callan. Document filtering with inference networks. In *Proceedings of the 19th ACM SIGIR Conference*, pages 262–269, Zurich, Switzerland, August 1996.

[320] J. Callan. Distributed information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval. Recent Research from the Center for Intelligent Information Retrieval*, volume 7 of *The Kluwer International Series on Information Retrieval*, chapter 5, pages 127–150. Kluwer Academic Publishers, Boston/Dordrecht/London, 2000.

[321] J. Callan, M. Connell, and A. Du. Automatic discovery of language models for text databases. In *Proceedings of ACM SIGMOD'99*, pages 479–490, New York, 1999.

[322] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *DEXA*, pages 78–83, 1992.

[323] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of ACM SIGIR'95*, pages 21–28, Seattle, WA, July 1995. ACM Press.

[324] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, 85(25):5468–5471, Dec 2000.

[325] B. B. Cambazoglu, V. Plachouras, and R. Baeza-Yates. Quantifying performance and quality gains in distributed Web search engines. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *SIGIR*, pages 411–418, Boston, MA, USA, July 2009. ACM.

[326] L. J. Camp. DRM: doesn't really mean digital copyright management. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, pages 78–87, Washington, DC, USA, 2002.

[327] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[328] P. Cao and S. Irani. Cost-aware WWW proxy caching algorithms. In *USITS*, 1997.

[329] P. Cao and Z. Wang. Efficient top-K Query Calculation in Distributed Networks. In *PODS'04: Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems*, Paris, France, 2004.

[330] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon. Adapting ranking svm to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193, Seattle, Washington, USA, 2006. ACM Press.

[331] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 129–136, Corvalis, Oregon, 2007. ACM Press.

[332] J. Carbonell and J. Goldstein. The use of MMR, Diversity-based reranking for reordering documents and producing summaries. In *Proceedings of ACM SIGIR'98*, pages 335–336, Melbourne, Australia, 1998.

[333] D. Carmel, Y. Maarek, M. Mandelbrod, Y.Mass, and A. Soffer. Searching XML documents via XML fragments. In *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada*, pages 151–158, 2003.

[334] D. Carmel, Y. S. Maarek, and A. Soffer. XML and Information Retrieval: a SIGIR 2000 Workshop. *SIGIR Forum*, 34(1):31–36, 2000.

[335] D. Carmel and E. Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan Claypool, 2010.

[336] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What makes a query difficult? In *Proc. of the ACM Int'l Conf. on Information Retrieval*, pages 390–397, 2006.

[337] Carnegie and Reuters. Reuters-21578 text categorization collection, 1987. Produced by Carnegie Group Inc. and Reuters Ltd., Reuters-21578 is made available for research purposes only. Data formatting and organization done by David Lewis.

[338] J. S. Carriére and R. Kazman. Webquery: searching and visualizing the Web through connectivity. *Computer Networks and ISDN Systems*, 29(8-13):1257–1267, September 1997.

[339] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *ICML '06: Proceedings of the 23rd International Conference on Machine learning*, pages 161–168, New York, NY, USA, 2006. ACM.

[340] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, April 2008.

[341] Cassandra. http://incubator.apache.org/cassandra/, 2008.

[342] C. Castillo. *Effective Web Crawling*. PhD thesis, University of Chile, 2004.

[343] C. Castillo and R. Baeza-Yates. A new crawling model. In *Poster proceedings of the eleventh conference on World Wide Web*, Honolulu, Hawaii, USA, 2002.

[344] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[345] R. G. G. Cattell and D. K. Barry. *The Object Data Standard: ODMG 3.0*. Morgan Kaufmann, 2000.

[346] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US, 1994.

[347] D. Chakrabarti, R. Kumar, and K. Punera. Quicklink selection for navigational query results. In *Proceedings of the Eighteenth International World Wide Web Conference*, Madrid, Spain, May 2009.

[348] S. Chakrabarti. Recent results in automatic Web resource discovery. *ACM Computing Surveys*, 31(4):17, 1999.

[349] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, August 2002.

[350] S. Chakrabarti. Learning to rank in vector spaces and social networks. *Internet Mathematics*, 4(2-3):267–298, 2007.

[351] 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*, Melbourne, Australia, 1998.

[352] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *7th WWW Conference*, pages 65–74, Brisbane, Australia, April 1998.

[353] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11–16):1623–1640, 1999.

[354] D. Chamberlin, J. Robie, and D. Florescu. Quilt: An XML Query Language for Heterogeneous Data Sources. In *The World Wide Web and Databases, Third International Workshop WebDB 2000, Dallas, Texas, USA, Selected Papers*, pages 1–25, 2000.

[355] C.-C. Chang and C.-J. Lin. LibSVM – A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[356] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A distributed storage system for structured data. In *OSDI 2006*, pages 205–218, 2006.

[357] K. C. Chang, B. He, C. Li, M. Patel, and Z. Zhang. Structured databases on the web: observations and implications. *SIGMOD Rec.*, 33(3):61–70, September 2004.

[358] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[359] E. Chávez, K. Figueroa, and G. Navarro. Effective proximity retrieval by ordering permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(9):1647–1658, 2008.

[360] E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín. Proximity searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, Sept. 2001.

[361] C. Chen. Structuring and visualizing the WWW by generalized similarity analysis. In *8th ACM Conference on Hypertext and Hypermedia*, pages 177–186, Southampton, England, 1997.

[362] H. Chen, H. Jin, J. Wang, L. Chen, Y. Liu, and L. M. Ni. Efficient multi-keyword search over P2P Web. In *WWW'08: Proceeding of the 17th International World Wide Web conference*, Beijing, China, 2008.

[363] M. Chen, M. Hearst, J. Hong, and J. Lin. Cha-Cha: A System for Organizing Intranet Search Results. *Proceedings of the 2nd conference on USENIX Symposium on Internet Technologies and Systems*, pages 11–14, 1999.

[364] M. Chen, A. S. LaPaugh, and J. P. Singh. Predicting category accesses for a user in a structured information space. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 65–72, Tampere, Finland, July 2002. ACM.

[365] P. M. Chen, E. K. Lee, G. A. Gibson, R. H. Katz, and D. A. Paterson. RAID: High-performance, reliable secondary storage. *ACM Comput. Surv.*, 26(2):145–185, June 1994.

[366] S. Chen and J. Goodman. *An Empirical Study of Smoothing Techniques for Language Modeling*. Harvard University, 1998. Tech Report TR-10-98.

[367] J. Cheney. Compressing XML with multiplexed hierarchical PPM models. In *Proc. 11th IEEE Data Compression Conference (DCC'01)*, pages 163–172, 2001.

[368] F. Cheong. *Internet agents spiders, wanderers, brokers and bots*. New Riders, 1996.

[369] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 490–497, Seattle, WA, USA, 2001. ACM.

[370] E. H. Chi, P. Pirolli, and J. Pitkow. The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a web site. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 161–168, The Hague, The Netherlands, 2000. ACM.

[371] Y. Chiaramella, P. Mulhem, and F. Fourel. A model for multimedia information retrieval. Technical report, University of Glasgow, 1996.

[372] T. T. Chinenyanga and N. Kushmerick. Expressive Retrieval from XML Documents. In *24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana*, pages 163–171, 2001.

[373] J. Cho. The evolution of the Web and implications for an incremental crawler. In *Proceedings of 26th International Conference on Very Large Databases (VLDB)*, pages 527–534, Cairo, Egypt, September 2000. Morgan Kaufmann Publishers.

[374] J. Cho and H. Garcia-Molina. Synchronizing a database to improve freshness. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, pages 117–128, Dallas, Texas, USA, 2000.

[375] J. Cho and H. Garcia-Molina. Parallel crawlers. In *Proceedings of the eleventh international conference on World Wide Web*, pages 124–135, Honolulu, Hawaii, USA, 2002. ACM Press.

[376] J. Cho and H. Garcia-Molina. Effective page refresh policies for Web crawlers. *ACM Transactions on Database Systems*, 28(4), 2003.

[377] J. Cho and H. Garcia-Molina. Estimating frequency of change. *ACM Transactions on Internet Technology*, 3(3), 2003.

[378] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. In *Proceedings of the seventh conference on World Wide Web*, Brisbane, Australia, 1998. Elsevier Science.

[379] J. Cho, N. Shivakumar, and H. Garcia-Molina. Finding replicated Web collections. In *ACM SIGMOD*, pages 355–366, 1999.

[380] A. Chowdhury, O. Frieder, D. A. Grossman, and M. C. McCabe. Collection statistics for fast duplicate document detection. *ACM Trans. Inf. Syst.*, 20(2):171–191, 2002.

[381] A. Chowdhury and G. Pass. Operational requirements for scalable search systems. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 435–442, New York, NY, USA, 2003. ACM Press.

[382] G. Chowdhury. *Introduction to Modern Information Retrieval*. Facet Publishing, 2003. 488 pages.

[383] M. G. Christel, D. B. Winkler, and C. R. Taylor. Multimedia abstractions for a digital video library. In *DL '97: Proceedings of the Second ACM International Conference on Digital libraries*, pages 21–29, New York, NY, USA, 1997. ACM.

[384] K. Church and W. Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.

[385] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *Proc. of VLDB Conf.*, pages 426–435, Athens, Greece, Aug. 1997.

[386] Citeseer. http://citeseer.ist.psu.edu/, 1997.

[387] CiteseerX. http://citeseerX.ist.psu.edu/, 2008.

[388] CITIDEL. Computing and Information Technology Interactive Digital Educational Library. http://www.citidel.org, 2004.

[389] C. Clarke. Controlling overlap in content-oriented XML retrieval. In *28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil*, pages 314–321, 2005.

[390] C. Clarke, E. Agichtein, S. Dumais, and R. White. The influence of caption features on clickthrough patterns in Web search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'07)*, pages 135–142. ACM Press New York, NY, USA, 2007.

[391] C. Clarke, G. Cormack, and F. Burkowski. Shortest substring ranking (multitext experiments for TREC-4). In D. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, 1996.

[392] C. L. Clarke, G. V. Cormack, and F. J. Burkowski. An algebra for structured text search and a framework for its implementation. *The Computer Journal*, 38:43–56, 1995.

[393] CLEF 2009. http://clef-campaign.org/2009/2009_agenda.html, 2009.

[394] C. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield, 1962.

[395] C. Cleverdon. The Cranfield tests on index language devices. *ASLIB Proceedings*, 19(6):173–194, 1967.

[396] C. Cleverdon. Optimizing convenient online access to bibliographic databases. In *Document retrieval systems*, pages 32–41. Taylor Graham Publishing, London, UK, UK, 1988.

[397] C. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. Technical report, ASLIB, 1966.

[398] C. Cleverdon and R. Thorne. An experiment with the uniterm system. Technical Report Library Memo 7, RAE, 1954.

[399] C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, 1991.

[400] CliffsNotes. About pride and prejudice — publication history and critical reception, 2009. http://www.cliffsnotes.com/WileyCDA/LitNote/Pride-and-Prejudice-About-Pride-and-Prejudice-Publication-History-and-Critical-Reception.id-147,pageNum-9.html.

[401] CMU. WebKB hypertext collection. http://www.cs.cmu.edu/~webkb/.

[402] CMU. 20 newsgroup, 1999. Originally owned by Tom Mitchell, Computer Science Department, Carnegie Mellon University.

[403] T. A. S. Coelho, P. P. Calado, L. V. Souza, B. A. Ribeiro-Neto, and R. Muntz. Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):408–417, 2004.

[404] E. G. Coffman, Z. Liu, and R. R. Weber. Optimal robot scheduling for Web search engines. *Journal of Scheduling*, 1(1):15–29, 1998.

[405] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv. XSEarch: A Semantic Search Engine for XML. In *29th International Conference on Very Large Data Bases, Berlin, Germany*, pages 45–56, 2003.

[406] W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. *ACM Transaction on Office and Information Systems*, 17(2):141–173, 1999.

[407] W. W. Cohen, R. E. Schapire, and Y. Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270, 1999.

[408] D. E. Comer and D. L. Stevens. *Internetworking with TCP/IP Vol III: Client-Server Programming and Applications*. Prentice-Hall, Inc., Englewood Cliffs, NJ, USA, 1993.

[409] B. Commentz-Walter. A string matching algorithm fast on the average. In *Proc. ICALP'79*, pages 118–132. Springer-Verlag, 1979.

[410] Communications of the ACM. Hypermedia, February 1994. 37(2).

[411] Communications of the ACM. Hypermedia, August 1995. 38(8).

[412] J. Conklin. Hypertext: An introduction and survey. *IEEE Computer*, 20(9):17–41, Sept. 1987.

[413] D. Connolly. Dan Connolly on the Architecture of the Web: Let a Thousand Flowers Bloom. *IEEE Internet Computing*, 2(2):22–31, 1998.

[414] N. E. O. Connor, S. Marlow, N. Murphy, A. Smeaton, P. Browne, S. Deasy, H. Lee, and K. McDonald. Fischlar: An on-line system for indexing and broadcasting television content. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 3:1633 – 1636, 2001.

[415] M. Consens and T. Milo. Algebras for Querying Text Regions. In *Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, San Jose, California*, pages 11–22, 1995.

[416] M. P. Consens and A. O. Mendelzon. The G+/GraphLog visual query system. In H. Garcia-Molina and H. V. Jagadish, editors, *SIGMOD Conference*, page 388, Atlantic City, NJ, USA, May 1990. ACM Press.

[417] Consultative Committee for Space Data Systems. Reference model for an open archival information system (OAIS), 2001. http://public.ccsds.org/publications/archive/650x0b1.pdf.

[418] M. Cooke and D. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communications*, pages 141–177, 2001.

[419] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *ICTAI*, pages 558–567, 1997.

[420] A. Cooper. A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Transactions on the Web (TWeb)*, 2(4), 2008.

[421] B. F. Cooper and H. Garcia-Molina. Sil: a model for analyzing scalable peer-to-peer search networks. *Comput. Netw.*, 50(13):2380–2400, 2006.

[422] W. S. Cooper. The formalism of probability theory in IR: A foundation or an encumbrance? In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Probabilistic Models, pages 242–247, 1994. Triennial ACM-SIGIR Award Paper.

[423] W. S. Cooper, F. C. Gey, and D. P. Dabney. Probabilistic retrieval based on staged logistic regression. In *Proc. of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 198–210, Copenhagen, Denmark, 1992.

[424] J. Cope, N. Craswell, and D. Hawking. Automatic discovery of search interfaces on the web. In *The Fourteenth Australasian Database Conference*, volume 17 of *Conferences in Research and Practice in Information Technology*, Adelaide, Australia, 2003.

[425] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. The MIT Press/McGraw-Hill, Cambridge, MA, 1990.

[426] D. E. Corporation. AltaVista, 1996. `http://altavista.com`.

[427] Corporation for National Research Initiatives, the Handle System, May 1998. http://www.handle.net/.

[428] P. Correia-Saraiva, E. Silva de Moura, N. Ziviani, W. Meira, R. Fonseca, and B. A. Ribeiro-Neto. Rank-preserving two-level caching for scalable search engines. In *Proceedings of the 24th International ACM Conference on Research and Development in Information Retrieval*, September 2001. New Orleans, USA.

[429] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[430] E. Cortez, A. S. da Silva, M. A. Gonçalves, F. Mesquita, and E. S. de Moura. FLUX-CM: Flexible Unsupervised Extraction of Citation Metadata. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 2007.

[431] I. G. Councill, C. L. Giles, H. Han, and E. Manavoglu. Automatic acknowledgement indexing: expanding the semantics of contribution in the citeseer digital library. In *Proceedings of the 3rd International Conference on Knowledge Capture (K-CAP 2005)*, pages 19–26, Banff, Alberta, Canada, 2005.

[432] T. Couto, M. Cristo, M. A. Gonçalves, P. Calado, N. Ziviani, E. Moura, and B. A. Ribeiro-Neto. A comparative study of citations and links in document classification. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 75–84, New York, NY, USA, 2006. ACM Press.

[433] M. Covell and S. Ahmad. Analysis-by-synthesis dissolve detection. In *Proceedings of IEEE International Conference on Image Processing*, Rochester, NY, September 2002.

[434] M. Covell, M. Withgott, and M. Slaney. Mach1: Nonuniform time-scale modification of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Language Processing*, volume 1, pages 349–352, 1998.

[435] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Pichunter: Bayesian relevance feedback for image retrieval. *International Conference on Pattern Recognition*, 13:361–369, 1996.

[436] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[437] N. Craswell, P. Bailey, and D. Hawking. Server selection on the world wide web. In *Proceedings of the 5th ACM Digital Libraries Conference*, pages 37–46, June 2000.

[438] N. Craswell, F. Crimmins, D. Hawking, and A. Moffat. Performance and cost tradeoffs in Web search. In *Proceedings of the 15th Australasian Database Conference*, pages 161–169, Dunedin, New Zealand, January 2004.

[439] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the trec 2005 enterprise track. In *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*, Gaithersburg, MD, 2005. NIST. TREC Special Publication: SP 500-266. trec.nist.gov/pubs/trec14/papers/ENTERPRISE.OVERVIEW.pdf.

[440] N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins. P@noptic expert: Searching for experts not just for documents. In *Poster Proceedings of AusWeb'01*, 2001. /urlausweb.scu.edu.au/aw01/papers/edited/vercoustre/paper.htm.

[441] A. Crauser and P. Ferragina. A theoretical and experimental study on the construction of suffix arrays in external memory. *Algorithmica*, 32(1):1–35, 2002.

[442] A. Crespo and H. Garcia-Molina. Archival storage for digital libraries. In *DL'98: Proceedings of the 3rd ACM International Conference on Digital Libraries*, pages 69–78, 1998.

[443] A. Crespo and H. Garcia-Molina. Semantic overlay networks for P2P systems. Technical Report 2003-75, Stanford University, 2003.

[444] F. Crestani, M. Lalmas, C. J. van Rijsbergen, and I. Campbell. "Is this document relevant? ... probably": A survey of probabilistic models in information retrieval. *ACM Computing Surveys*, 30(4):528–552, Dec. 1998.

[445] M. Cristo, P. Calado, M. de Lourdes da Silveira, I. Silva, R. R. Muntz, and B. A. Ribeiro-Neto. Bayesian belief networks for ir. *Int. J. Approx. Reasoning*, 34(2-3):163–179, 2003.

[446] M. Cristo, P. Calado, E. S. de Moura, N. Ziviani, and B. A. Ribeiro-Neto. Link information as a similarity measure in Web classification. In *SPIRE*, pages 43–55, 2003.

[447] M. Crochemore and W. Rytter. *Text Algorithms*. Oxford University Press, Oxford, UK, 1994.

[448] M. Crochemore and W. Rytter. *Jewels of Stringology — Text algorithms*. World Scientific, 2002. ISBN 981-02-4782-6. 320 pages.

[449] B. Croft, D. Metzler, and T. Strohman. *Search Engines – Information Retrieval in Practice*. Addison Wesley, February 2009.

[450] W. Croft. Experiments with representation in a document retrieval system. *Information Technology: Research and Development*, 2(1):1–21, 1983.

[451] W. Croft, S. Cronen-Townsend, and V. Lavrenko. Relevance Feedback and Personalization: A Language Modeling Perspective. *Delos Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001.

[452] W. Croft and D. Harper. Using probabilistic models of retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, 1979.

[453] W. B. Croft and J. Lafferty, editors. *Language Modeling for Information Retrieval*. Number 13 in the Information Retrieval Series. Kluwer/Springer, 2003.

[454] S. Cronen-Townsend, Y. Zhou, and W. Croft. A Language Modeling Framework for Selective Query Expansion. Technical Report Technical Report IR-338, Center for Intelligent Information Retrieval, University of Massachusetts, 2004.

[455] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306, 2002.

[456] T. Crook, B. Frasca, R. Kohavi, and R. Longbotham. Seven pitfalls to avoid when running controlled experiments on the web. In J. F. E. IV, F. Fogelman-Soulié, P. A. Flach, and M. J. Zaki, editors, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 1105–1114. ACM, 2009.

[457] C. J. Crouch and B. Yang. Experiments in automatic statistical thesaurus construction. In *Proc. ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 77–88, Denmark, 1992.

[458] M. E. Crovella and A. Bestavros. Self-similarity in world wide Web traffic: evidence and possible causes. In *SIGMETRICS '96: Proceedings of the 1996 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, volume 24, pages 160–169, New York, NY, USA, May 1996. ACM Press.

[459] R. Crow. The case for institutional repositories: A SPARC position paper. Technical report, The Scholarly Publishing & Academic Resources Coalition, Washington, D.C., Aug. 2002.

[460] Computer Science Bibliography. http://liinnwww.ira.uka.de/bibliography.

[461] S. Cucerzan and E. Brill. Spelling correction as an iterative process that exploits the collective knowledge of Web users. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 293–300, Barcelona, Spain, July 2004.

[462] F. M. Cuenca-Acuna and T. D. Nguyen. Text-based content search and retrieval in ad-hoc P2P communities. In *Revised Papers from the NETWORKING 2002 Workshops on Web Engineering and Peer-to-Peer Computing*, pages 220–234, London, UK, 2002. Springer-Verlag.

[463] J. Culpepper and A. Moffat. Compact set representation for information retrieval. In *SPIRE 2007*, volume 4726 of *Lecture Notes in Computer Science*, pages 137–148. Springer, 2007.

[464] S. J. Cunningham, N. Reeves, and M. Britland. An ethnographic study of music information seeking: implications for the design of a music digital library. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, Portland, Oregon, 2003.

[465] F. Curbera, M. Duftler, R. Khalaf, W. Nagy, N. Mukhi, and S. Weerawarana. Unraveling the Web services web: An introduction to SOAP, WSDL, and UDDI. *IEEE Distributed Systems Online*, 3(4), 2002.

[466] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin. Fast, flexible filtering with Phlat. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06)*, pages 261–270, 2006.

[467] D. Cutting, J. Pedersen, D. Karger, and J. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'92)*, pages 318–329, Copenhagen, Denmark, 1992.

[468] M. Czerwinski, M. van Dantzich, G. Robertson, and H. Hoffman. The contribution of thumbnail image, mouse-over text and spatial location memory to Web page retrieval in 3D. *Proceedings of Human-Computer Interaction (INTERACT'99)*, 99:163–170, 1999.

[469] A. Czumaj, M. Crochemore, L. Gasieniec, S. Jarominek, T. Lecroq, W. Plandowski, and W. Rytter. Speeding up two string-matching algorithms. *Algorithmica*, 12:247–267, 1994.

[470] L. C. da Rocha, F. Mourão, A. Pereira, M. A. Gonçalves, and W. Meira Jr. Exploiting temporal contexts in text classification. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa, California, USA, October 26-30, 2008*, pages 243–252, 2008.

[471] A. S. da Silva, P. Calado, R. C. Vieira, A. H. F. Laender, and B. A. Ribeiro-Neto. Keyword-based queries over Web databases. In *Effective Databases for Text & Document Management*, pages 74–92. 2003.

[472] A. S. da Silva, E. A. Veloso, P. B. Golgher, A. H. F. Laender, and N. Ziviani. CoBWeb - a crawler for the Brazilian web. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, pages 184–191, Cancun, México, 1999. IEEE CS Press.

[473] R. da Silva Torres, C. B. Medeiros, M. A. Gonçalves, and E. A. Fox. A digital library framework for biodiversity information systems. *International Journal on Digital Libraries*, 6(1):3–17, 2006.

[474] R. Darnell. *HTML 4.0 Unleashed, Professional Reference Edition*. Samms.net Publishing, 1998.

[475] J. R. Davis and C. Lagoze. NCSTRL: Design and deployment of a globally distributed digital library. *Journal of the American Society for Information Science*, 51(3):273–280, 2000.

[476] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28:357–366, August 1980.

[477] B. D. Davison. Topical locality in the web. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, pages 272–279, Athens, Greece, 2000. ACM Press.

[478] Digital Bibliography & Library Project. http://www.informatik.uni-trier.de/~ley/db/.

[479] H. M. de Almeida, M. A. Gonçalves, M. Cristo, and P. Calado. A combined component approach for finding collection-adapted ranking functions based on genetic programming. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 399–406. ACM, 2007.

[480] R. de Freitas Vale, B. A. Ribeiro-Neto, L. R. S. de Lima, A. H. F. Laender, and H. R. Freitas-Junior. Improving text retrieval in medical collections through automatic categorization. In *SPIRE*, pages 197–210, 2003.

[481] M. de Lourdes da Silveira and B. A. Ribeiro-Neto. Concept-based ranking: a case study in the juridical domain. *Inf. Process. Manage.*, 40(5):791–805, 2004.

[482] M. de Lourdes da Silveira, B. A. Ribeiro-Neto, R. de Freitas Vale, and R. T. Assumpção. Vertical searching in juridical digital libraries. In *ECIR*, pages 491–501, 2003.

[483] A. de Vries, G. Kazai, and M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO 2004 Conference on Coupling approaches, coupling media and coupling languages for information retrieval, Vaucluse, France*, pages 463–473, 2004.

[484] J. Dean. Challenges in building large-scale information retrieval systems: invited talk presentation. In R. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, and B. B. Cambazoglu, editors, *WSDM*, page 1, Barcelona, Spain, 2009. ACM.

[485] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of ACM*, 51(1):107–113, 2008. Preliminary version published in OSDI 2004, p. 137-150.

[486] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: Amazon's highly available key-value store. *SIGOPS Oper. Syst. Rev.*, 41(6):205–220, 2007.

[487] K. J. Delaney. Yahoo Contends It Tops Google in Number of Pages Searched. Wall Street Journal, August 15, 2005, page B4.

[488] A. Delgado and R. Baeza-Yates. An analysis of query languages for XML. *UPGRADE. The European Online Magazine for the IT Professional, Special Issue on Information Retrieval and the Web*, III(3), 2002.

[489] DELOS Network of Excellence. Reference Model for Digital Library Management Systems. http://www.delos.info/ReferenceModel.

[490] E. D. Demaine, A. López-Ortiz, and J. I. Munro. Adaptive set intersections, unions, and differences. In *SODA*, pages 743–752, San Francisco, CA, USA, 2000.

[491] E. D. Demaine, A. López-Ortiz, and J. I. Munro. Experiments on adaptive set intersections for text retrieval systems. In A. L. Buchsbaum and J. Snoeyink, editors, *ALENEX*, volume 2153 of *Lecture Notes in Computer Science*, pages 91–104, Washington, DC, USA, January 2001. Springer.

[492] S. Dennis, R. Mcarthur, and P. Bruza. Searching the World Wide Web Made Easy? the Cognitive Load Imposed By Query Refinement Mechanisms. *Proceedings of the Australian Document Computing Conference*, pages 65–71, 1998.

[493] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40(1):64–69, 2006.

[494] A. Deutsch, M. Fernández, D. Florescu, A. Levy, and D. Suciu. XML-QL. In *Query Languages 1998*, 1998.

[495] Z. Dezso, E. Almaas, A. Lukacs, B. Racz, I. Szakadat, and A. Barabasi. Fifteen minutes of fame: the dynamics of information access on the web, May 2005.

[496] F. Diaz and R. Jones. Using temporal profiles of queries for precision prediction. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 18–24, 2004.

[497] M. Diligenti, F. Coetzee, S. Lawrence, L. C. Giles, and M. Gori. Focused crawling using context graphs. In *Proceedings of 26th International Conference on Very Large Databases (VLDB)*, pages 527–534, Cairo, Egypt, September 2000.

[498] S. Dill, R. Kumar, K. S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Trans. Inter. Tech.*, 2(3):205–223, 2002.

[499] J. Dinet, M. Favart, and J. Passerault. Searching for information in an online public access catalogue(OPAC): the impacts of information search expertise on the use of Boolean operators. *Journal of Computer Assisted Learning*, 20(5):338–346, 2004.

[500] S. Ding, J. He, H. Yan, and T. Suel. Using graphics processors for high performance IR query processing. In J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, editors, *WWW*, pages 421–430, Madrid, Spain, 2009. ACM.

[501] A. Divoli, M. Hearst, and M. Wooldridge. Evidence for showing gene/protein name suggestions in bioscience literature search interfaces. *Pacific Symposium on Biocomputing*, 568:79, 2008.

[502] P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita. Using annotations in enterprise search. In *Proceedings of WWW'06*, pages 811–817, New York, NY, USA, 2006. ACM.

[503] M. Dodge. The geography of cyberspace directory. http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/geography_Of_Cyberspace.html, 1997-2004.

[504] P. Dömel. Webmap: A graphical hypertext navigation tool. In *Electronic Proceedings of the Second World Wide Web Conference '94: Mosaic and the Web*, 1994. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/WWW2_Proceedings.html.

[505] S. Dominich. *Mathematical foundations of information retrieval*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

[506] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the Web graph. In *Eigth international workshop on the Web and databases WebDB*, Baltimore, USA, June 2005.

[507] P. Donmez, K. M. Svore, and C. J. Burges. On the local optimality of LambdaRank. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 460–467, Boston, MA, USA, 2009. ACM Press.

[508] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *WWW'07: Proceedings of the 16th international conference on World Wide Web*, pages 581–590, New York, NY, USA, 2007. ACM.

[509] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. C. Mogul. Rate of change and other metrics: a live study of the world wide web. In *USENIX Symposium on Internet Technologies and Systems*, pages 147–158, Monterey, California, USA, December 1997.

[510] C. Doulkeridis, K. Nørvåg, and M. Vazirgiannis. Peer-to-peer similarity search over widely distributed document collections. In *LSDS-IR '08: Proceeding of the 2008 ACM workshop on Large-Scale distributed systems for information retrieval*, pages 35–42, New York, NY, USA, 2008. ACM.

[511] D. Dreilinger. Savvy Search, 1996. http://savvy.cs.colostate.edu:2000/form?beta.

[512] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: learning to identify link spam. In *Proceedings of the 16th European Conference on Machine Learning (ECML)*, volume 3720 of *Lecture Notes in Artificial Intelligence*, pages 233–243, Porto, Portugal, 2005.

[513] D. D'Souza and J. A. Thom. Collection Selection Using n-Term Indexing. In *Proceedings of CODAS*, pages 52–63, 1999.

[514] D. D'Souza, J. A. Thom, and J. Zobel. Collection selection for managed distributed document databases. *Information Processing & Management*, 40, 2004.

[515] D. D'Souza, J. Zobel, and J. Thom. Is CORI effective for collection selection? an exploration of parameters, queries, and data. In *Proceedings of Australian Document Computing Symposium*, pages 41–46, Melbourne, Australia, 2004.

[516] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW'06: Proceedings of the 15th International Conference on World Wide Web*, pages 193–202, New York, NY, USA, 2006. ACM Press.

[517] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2000.

[518] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff I've seen: a system for personal information retrieval and re-use. In *Proceedings of ACM SIGIR '03*, pages 72–79, Toronto, Canada, 2003. ACM.

[519] S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*, pages 277–284, 2001.

[520] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155, 1998.

[521] S. T. Dumais and H. Chen. Hierarchical classification of Web content. In *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 256–263, Athens, GR, 2000.

[522] G. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, editors, *SIGIR*, pages 331–338. ACM, November 2008.

[523] S. Dziadosz and R. Chandrasekar. Do Thumbnail Previews Help Users Make Better Relevance Decisions about Web Search Results. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'02)*, pages 365–366, 2002.

[524] J.-P. Eckmann and E. Moses. Curvature of co-links uncovers hidden thematic layers in the World Wide Web. *PNAS*, 99(9):5825–5829, April 2002.

[525] J. Edwards, K. S. Mccurley, and J. A. Tomlin. An adaptive model for optimizing performance of an incremental Web crawler. In *Proceedings of the Tenth Conference on World Wide Web*, pages 106–113, Hong Kong, May 2001. Elsevier Science.

[526] R. Edwards, S. Clarke, and A. Kellett. Organisations waste 10% of salary bill searching for information. http://www.butlergroup.com/pdf/PressReleases/ESRReportPressRelease.pdf, October 2006.

[527] D. Egan, J. Remde, L. Gómez, T. Landauer, J. Eberhardt, and C. Lochbaum. Formative design evaluation of SuperBook. *Transaction on Information Systems*, 7(1), 1989.

[528] D. Egan, J. Remde, T. Landauer, C. Lochbaum, and L. Gómez. Behavioral evaluation and analysis of a hypertext browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'89)*, pages 205–210, May 1989.

[529] D. Eichmann. The RBSE spider: balancing effective search against Web load. In *Proceedings of the first World Wide Web Conference*, Geneva, Switzerland, May 1994.

[530] S. G. Eick and G. J. Wills. Navigating large networks with hierarchies. In *Proc. of the Conference on Visualization '93*, pages 204–209, San Jose, Oct. 1993.

[531] N. Eiron, K. S. Curley, and J. A. Tomlin. Ranking the Web frontier. In *Proceedings of the 13th international conference on World Wide Web*, pages 309–318, New York, NY, USA, 2004. ACM Press.

[532] P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21:194–203, 1975.

[533] E. Elliott and G. Davenport. Video streamer. *ACM CHI-94: Proceedings on Human Factors in Computing Systems: Celebrating Independence*, pages 65–66, 1994.

[534] Endeca. Endeca. http://www.endeca.com/.

[535] EPrints. Registry of open access repositories (ROAR), 2006. http://roar.eprints.org/.

[536] P. Erdõs and A. Rényi. Random graphs. *Publication of the Mathematical Institute of the Hungarian Acadamy of Science*, 5, 1960.

[537] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. W. Han, and U. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.

[538] J. Exposto, J. Macedo, A. Pina, A. Alves, and J. Rufino. Geographical partition for distributed Web crawling. In *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*, pages 55–60, Bremen, Germany, 2005. ACM Press.

[539] G. Eysenbach and C. Kohler. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *British Medical Journal*, 324(7337):573–577, 2002.

[540] Facebook. http://www.facebook.com/, 2004.

[541] R. Fagin, R. Kumar, K. S. McCurley, J. Novak, D. Sivakumar, J. A. Tomlin, and D. P. Williamson. Searching the workplace web. In *Proceedings of WWW2003*, Budapest, Hungary, May 2003. http://www2003.org/cdrom/papers/refereed/p641/xhtml/p641-mccurley.html.

[542] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. In *PODS'01: Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems*, Santa Barbara, CA, USA, 2001.

[543] R. Fagin and E. L. Wimmers. Incorporating user preferences in multimedia queries. In *Proceedings of the 1997 International Conference on Database Theory*, pages 247–261, 1997.

[544] T. Fagni, R. Perego, F. Silvestri, and S. Orlando. Boosting the performance of Web search engines: Caching and prefetching query results by exploiting historical usage data. *ACM Trans. Inf. Syst.*, 24(1):51–78, 2006.

[545] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *J. of Intelligent Information Systems*, 3(3/4):231–262, July 1994.

[546] C. Faloutsos and R. Chan. Text access methods for optical and large magnetic disks: design and performance comparison. In *Proc. of VLDB'88*, pages 280–293, 1988.

[547] C. Faloutsos and S. Christodoulakis. Description and performance analysis of signature file methods. *ACM TOIS*, 5(3):237–257, 1987.

[548] C. Faloutsos and V. Gaede. Analysis of n-dimensional quadtrees using the Hausdorff fractal dimension. In *Proc. of VLDB Conf.*, pages 40–50, Bombay, India, Sept. 1996.

[549] C. Faloutsos and I. Kamel. Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In *Proc. ACM SIGACT-SIGMOD-SIGART PODS*, pages 4–13, Minneapolis, MN, May 1994.

[550] A. Fariña. *New Compression Codes for Text Databases*. PhD thesis, Computer Science Department, University of A Coruña, A Coruña, Spain, 2005.

[551] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.

[552] R. Feldman and J. Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, Cambridge, MA, USA, December 2006.

[553] S. Feldman and C. Sherman. The high cost of not finding information. White Paper #29127, IDC, April 2003. http://www.idc.com.

[554] T. Ferl and L. Millsap. The knuckle-cracker's dilemma: a transaction log study of OPAC subject searching. *Information Technology and Libraries*, 15(2):113–126, 1996.

[555] M. Fernandez, D. Florescu, A. Levy, and D. Suciu. A query language for a Web-site management system. *SIGMOD Record*, 26(3):4–11, September 1997.

[556] P. Ferragina and G. Manzini. Opportunistic data structures with applications. In *Proc. 41st IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 390–398, 2000.

[557] P. Ferragina and G. Manzini. Indexing compressed texts. *Journal of the ACM*, 52(4):552–581, 2005.

[558] P. Ferragina, G. Manzini, V. Mäkinen, and G. Navarro. Compressed representation of sequences and full-text indexes. *ACM Transactions on Algorithms*, 3(2), 2007. Earlier version in *Proc. SPIRE 2004*.

[559] D. Fetterly, M. Manasse, and M. Najork. On the evolution of clusters of near-duplicate Web pages. *Journal of Web Engineering*, 2(4):228–246, 2004.

[560] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide web. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 170–177, New York, NY, USA, 2005. ACM Press.

[561] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of Web pages. In *Proceedings of the Twelfth Conference on World Wide Web*, Budapest, Hungary, 2003. ACM Press.

[562] S. Few. *Information Dashboard Design: the Effective Visual Communication of Data*. O'Reilly, 2006.

[563] S. Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

[564] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. RFC 2616 - HTTP/1.1, the hypertext transfer protocol. http://w3.org/Protocols/rfc2616/rfc2616.html, 1999.

[565] R. A. Finkel, A. B. Zaslavsky, K. Monostori, and H. W. Schmidt. Signature extraction for overlap detection in documents. In M. J. Oudshoorn, editor, *Twenty-Fifth Australasian Computer Science Conference (ACSC2002)*, volume 4 of *CRPIT*, pages 59–64, Melbourne, Australia, 2002. Australian Computer Society.

[566] K. Fishkin and M. C. Stone. Enhanced dynamic queries via movable filters. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'95)*, volume 1, pages 415–420, 1995.

[567] L. Fitzpatrick and M. Dent. Automatic feedback using past queries: Social searching? *20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 306–313, 1997.

[568] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: the QBIC System. *Computer*, 28(9):23–32, 1995.

[569] D. Florescu, A. Y. Levy, and A. O. Mendelzon. Database techniques for the world-wide web: A survey. *SIGMOD Record*, 27(3):59–74, 1998.

[570] M. J. Flynn. Very high-speed computing systems. In *Proc. IEEE*, volume 54, pages 1901–1909, 1966.

[571] B. M. Fonseca, P. B. Golgher, E. S. De Moura, and N. Ziviani. Using association rules to discovery search engines related queries. In *First Latin American Web Congress (LA-WEB'03)*, November, 2003. Santiago, Chile.

[572] B. M. Fonseca, P. B. Golgher, B. Pôssas, B. A. Ribeiro-Neto, and N. Ziviani. Concept-based interactive query expansion. In *CIKM*, pages 696–703, 2005.

[573] E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21:768–780, 1965.

[574] D. Foskett. Thesaurus. In K. S. Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 111–134. Morgan Kaufmann Publishers, Inc., 1997.

[575] M. Foulonneau, Cole, T. W., Habing, T. G., Shreeves, and S. L. Using collection descriptions to enhance an aggregation of harvested item-level metadata. In *JCDL'05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 32–41, Denver, Colorado, USA, 2005.

[576] F. Fouss, M. Saerens, and J.-M. Renders. Links between kleinberg's hubs and authorities, correspondence analysis, and Markov chains. In *Proceedings of the third IEEE international conference on data mining (ICDM)*, pages 521–524, Melbourne, Florida, USA, November 2003.

[577] C. Fox. Lexical analysis and stoplists. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 102–130. Prentice Hall, 1992.

[578] E. A. Fox, R. K. France, E. Sahle, A. Daoud, and B. E. Cline. Development of a modern OPAC: From REVTOLC to MARIAN. In *Proc. of the 16th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 248–259, 1993.

[579] E. A. Fox and S. Urs. Digital Libraries. In B. Cronin, editor, *Annual Review of Information Science and Technology*, volume 36, Ch. 12, pages 503–589. American Society for Information Science and Technology, 2002.

[580] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve Web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, 2005.

[581] W. Frakes. Stemming algorithms. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 131–160. Prentice Hall, 1992.

[582] W. Frakes and R. Baeza-Yates, editors. *Information Retrieval: Data Structures & Algorithms*. Prentice Hall, 1992.

[583] W. Francis and H. Kucera. *Frequency Analysis of English Usage*. Houghton Mifflin Co., 1982.

[584] A. P. Francisco, R. Baeza-Yates, and A. L. Oliveira. Clique analysis of query log graphs. In A. Amir, A. Turpin, and A. Moffat, editors, *SPIRE*, volume 5280 of *Lecture Notes in Computer Science*, pages 188–199, Melbourne, Australia, November 2008. Springer.

[585] K. Franzen and J. Karlgren. Verbosity and interface design. Technical report, Technical Report T2000, 2000.

[586] K. Fredriksson and S. Grabowski. Practical and optimal string matching. In *Proc. SPIRE 2005*, pages 376–387, 2005.

[587] K. Fredriksson and G. Navarro. Average-optimal single and multiple approximate string matching. *ACM Journal of Experimental Algorithmics (JEA)*, 9(1.4), 2004.

[588] H. R. Freitas-Junior, B. A. Ribeiro-Neto, R. de Freitas Vale, A. H. F. Laender, and L. R. S. de Lima. Categorization-driven cross-language retrieval of medical information. *JASIST*, 57(4):501–510, 2006.

[589] L. Freund and E. G. Toms. Enterprise search behaviour of software engineers. In *Proceedings of ACM SIGIR '06*, pages 645–646, New York, NY, USA, 2006. ACM.

[590] L. Freund, E. G. Toms, and C. L. Clarke. Modeling task-genre relationships for IR in the workplace. In *Proceedings of ACM SIGIR '05*, pages 441–448, New York, NY, USA, 2005. ACM.

[591] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 170–178, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[592] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, pages 148–156, 1996.

[593] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. www.cs.princeton.edu/~schapire/boost.html.

[594] A. Friedlander. D-lib Program: Research in Digital Libraries, May 1998. http://www.dlib.org/.

[595] J. Friedman, R. Kohavi, and Y. Yun. Lazy decision trees. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 717–724. AAAI Press and the MIT Press, Aug. 1996.

[596] Friendster. http://www.friendster.com/, 2002.

[597] N. Fuhr. Models for retrieval with probabilistic indexing. *Information Processing & Management*, 25:55–72, 1989.

[598] N. Fuhr. Optimal polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems*, 7(3):183–204, 1989.

[599] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[600] N. Fuhr. A decision-theoretic approach to database selection in networked ir. *ACM Trans. Inf. Syst.*, 17(3):229–249, 1999.

[601] N. Fuhr and N. Gövert. Retrieval quality vs. effectiveness of specificity-oriented search in XML collections. *Information Retrieval*, 9(1):55–70, 2006.

[602] N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, editors. *INitiative for the Evaluation of XML Retrieval (INEX). Proceedings of the First INEX Workshop. Dagstuhl, INEX 2002*, ERCIM Workshop Proceedings, Dagstuhl, Germany, 2003. ERCIM.

[603] N. Fuhr and K. Großjohann. XIRQL: An XML query language based on information retrieval concepts. *ACM Transaction on Information Systems*, 22(2):313–356, 2004.

[604] N. Fuhr, P. Hansen, M. Mabe, A. Micsik, and I. Sølvberg. Digital libraries: A generic classification and evaluation scheme. *Lecture Notes in Computer Science*, 2163:187–199, 2001.

[605] N. Fuhr, S. Hartmann, G. Knorz, G. Lustig, M. Schwantner, and K. Tzeras. AIR/X – a rule-based multistage indexing system for large subject fields. In A. Lichnerowicz, editor, *Proceedings of RIAO-91, 3rd International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 606–623, Barcelona, ES, 1991. Elsevier Science Publishers, Amsterdam, NL.

[606] N. Fuhr, J. Kamps, M. . Lalmas, S. Malik, and A. Trotman, editors. *Focused Access to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*, Dagstuhl Castle, Germany, 2008. Selected Papers.

[607] N. Fuhr and M. Lalmas. Introduction to the Special Issue on INEX. *Information Retrieval*, 8(4):515–519, 2005.

[608] N. Fuhr and M. Lalmas. Advances in XML retrieval: the INEX initiative. In *Proceedings of the International Workshop on Research Issues in Digital Libraries, IWRIDL 2006*, page 16, Kolkata, India, 2006.

[609] N. Fuhr, M. Lalmas, and S. Malik, editors. *INitiative for the Evaluation of XML Retrieval (INEX). Proceedings of the Second INEX Workshop. Dagstuhl, Germany, December 15–17, 2003*, 2004.

[610] N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors. *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 of *Lecture Notes in Computer Science*. Springer-Verlag, 2006.

[611] N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors. *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*, volume 3493 of *Lecture Notes in Computer Science*, Dagstuhl Castle, Germany, 2005. Springer. Revised Selected Papers.

[612] N. Fuhr, M. Lalmas, and A. Trotman, editors. *Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, volume 4518 of *Lecture Notes in Computer Science*. Springer-Verlag, 2007.

[613] N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovas, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Solvberg. Evaluation of digital libraries. *International Journal of Digital Libraries*, 8(1):21–38, 2007.

[614] G. Furnas, S. Deerwester, S. Dumais, T. Landauer, R. Harshman, L. Streeter, and K. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proc of the Eleventh Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 465–480, 1988.

[615] D. Gabor. Theory of communication, part iii. *The Journal of the Institute of Electrical Engineers*, pages 429–457, 1946.

[616] V. Gaede and O. Günther. Multidimensional access methods. *ACM Computing Surveys*, 30(2):170–231, 1998.

[617] Q. Gan and T. Suel. Improved techniques for result caching in Web search engines. In J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, editors, *WWW*, pages 431–440, Madrid, Spain, 2009. ACM.

[618] H. Garcia-Molina, L. Gravano, and N. Shivakumar. dSCAM: Finding document copies across multiple databases. In *Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems*, pages 68–79, Miami Beach, FL, USA, December 1996. IEEE Computer Society.

[619] N. Gershon, J. LeVasseur, J. Winstead, J. Croall, A. Pernick, and W. Ruh. Visualizing Internet resources. In *Proceedings '95 Information Visualization*, pages 122–128, Atlanta, Oct. 1995.

[620] S. Geva. GPX - Gardens Point XML IR at INEX 2005. In *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, pages 240–253, Dagstuhl Castle, Germany, 2006. Revised Selected Papers.

[621] S. Geva, J. Kamps, and A. Trotman, editors. *Advances in Focused Retrieval, 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008*, volume 5631 of *Lecture Notes in Computer Science*, Dagstuhl Castle, Germany, 2009. Springer. Revised and Selected Papers.

[622] F. C. Gey. Inferring probability of relevance using the method of logistic regression. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Probabilistic Models, pages 222–231, 1994.

[623] S. Ghemawat, H. Gobioff, and S.-T. Leung. The Google file system. In M. L. Scott and L. L. Peterson, editors, *Proceedings of the 19th ACM Symposium on Operating Systems Principles 2003*, pages 29–43, Bolton Landing, NY, USA, October 2003. ACM.

[624] G. Giacinto and F. Roli. Adaptive selection of image classifiers. In *International Conference on Image Analysis and Processing*, pages 38–45, 1997.

[625] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web communities from link topologies. In *9th ACM Conference on Hypertext and Hypermedia*, Pittsburgh, USA, 1998.

[626] H. M. Gladney. Trustworthy 100-year digital objects: Evidence after every witness is dead. *ACM Transactions on Information Systems*, 22(3):406–436, 2004.

[627] H. M. Gladney. Principles for digital preservation. *Communications of the ACM*, 49(2):111–116, 2006.

[628] H. M. Gladney and R. A. Lorie. Trustworthy 100-year digital objects: durable encoding for when it's too late to ask. *ACM Transactions on Information Systems*, 23(3):299–324, 2005.

[629] K. Goel, R. Guha, and O. Hansson. Introducing rich snippets. Google Webmaster Central Blog, May 2009. http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html.

[630] S. Goel, A. Broder, E. Gabrilovich, and B. Pang. Anatomy of the long tail: Ordinary people with extraordinary tastes. In *Third ACM Conference on Web Search and Data Mining (WSDM)*, New York, USA, 2010.

[631] N. Goevert, N. Fuhr, M. Lalmas, and G. Kazai. Evaluating the effectiveness of content-oriented XML retrieval methods. *Journal of Information Retrieval*, 9(6):699–722, 2006.

[632] A. Göker and D. He. Analysing Web search logs to determine session boundaries for user-oriented learning. In P. Brusilovsky, O. Stock, and C. Strapparava, editors, *Adaptive Hypermedia*, volume 1892 of *Lecture Notes in Computer Science*, pages 319–322, Trento, Italy, August 2000. Springer.

[633] C. Goldfarb. *The SGML Handbook*. Oxford University Press, Oxford, 1990.

[634] C. Goldfarb and P. Prescod. *The XML Handbook*. Prentice Hall, Oxford, 1998.

[635] S. W. Golomb. Run-length encodings. *IEEE Transactions on Information Theory*, 12(3):399–401, 1966.

[636] G. H. Golub and C. Greif. Arnoldi-type algorithms for computing stationary distribution vectors, with application to pagerank. Technical Report SCCM-04-15, Stanford University, 2004.

[637] B. Gomes. Search quality, continued. *The Official Google Blog*, Jan 2008. http://googleblog.blogspot.com/2008/08/search{}-quality{}-continued.html.

[638] A. Gómez-Pérez, F. Ortíz-Rodríguez, and B. Villazón-Terrazas. Ontology-based legal information retrieval to improve the information access in e-government. In *WWW'06: Proceedings of the 15th international conference on World Wide Web*, pages 1007–1008, New York, NY, USA, 2006. ACM.

[639] M. A. Gonçalves, M. Luo, R. Shen, M. F. Ali, and E. A. Fox. An XML log standard and tool for digital library logging analysis. *Lecture Notes in Computer Science*, 2458:129–143, 2002.

[640] M. A. Gonçalves and E. A. Fox. 5SL – A language for declarative specification and generation of digital libraries. In *Proc. of the 2nd Joint Conf. on Digital Libraries (JCDL'2002)*, pages 263–272, Portland, Oregon, July 14-18, 2002.

[641] M. A. Gonçalves, E. A. Fox, A. Krowne, P. Calado, A. H. F. Laender, A. S. da Silva, and B. A. Ribeiro-Neto. The effectiveness of automatically structured queries in digital libraries. In *JCDL*, pages 98–107, 2004.

[642] M. A. Gonçalves, E. A. Fox, and L. T. Watson. Towards a digital library theory: a formal digital library ontology. *Int. J. on Digital Libraries*, 8(2):91–114, 2008.

[643] M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Transanctions on Information Systems*, 22(2):270–312, 2004.

[644] M. A. Gonçalves, B. Lagoeiro, L. T. Watson, and E. A. Fox. "What is a good digital library?" - a quality model for digital libraries. *Information Processing & Management*, 43(5), 2007.

[645] M. A. Gonçcalves, G. Panchanathan, U. Ravindranathan, A. Krowne, F. Fox, F. Jagodzinski, and L. Cassel. The XML log standard for digital libraries: analysis, evolution, and deployment. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 312–314, Portland, Oregon, 2003.

[646] G. Gonnet. Examples of PAT applied to the Oxford English Dictionary. Technical Report OED-87-02, UW Centre for the New OED and Text Research, Univ. of Waterloo, 1987.

[647] G. Gonnet, R. Baeza-Yates, and T. Snider. New indices for text: Pat trees and Pat arrays. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, pages 66–82. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.

[648] G. Gonnet and F. Tompa. Mind your grammar: a new approach to modeling text. In *Proc. of the Thirteenth Int. Conf. on Very Large Data Bases*, pages 339–346, Brighton, England, Sept 1987.

[649] G. H. Gonnet and R. Baeza-Yates. *Handbook of Algorithms and Data Structures in Pascal and C*. Addison-Wesley, Wokingham, England, 2nd edition, 1991.

[650] Google. Google introduces personalized search services; site enhancements emphasize efficiency. Google Press Center, http://www.google.com/press/pressrel/enhancements.html, March 2004.

[651] Google blog, 2008. http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

[652] Google. Advanced search tips. http://www.google.com/support/websearch/bin/answer.py?hl=en\&answer=136861, 2009.

[653] Google notebook. http://www.google.com/notebook, 2009.

[654] Google. Search features. http://www.google.com/help/features.html, 2009.

[655] Google. Search sponsored links. http://www.google.com/sponsoredlinks, 2009.

[656] Google Books. http://books.google.com/, 2009.

[657] Google Scholar. http://scholar.google.com/, 2004.

[658] D. Gorton. Animated Images for a Multimedia Database. Master's thesis, Dept. of Computer Science, Univ. of Maryland, College Park, May 1989.

[659] N. Gövert, M. Abolhassani, N.Fuhr, and K. Großjohann. Content-oriented XML retrieval with HyRex. In *First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*, pages 26–32, Dagstuhl, Germany, 2002.

[660] N. Gövert and G. Kazai. Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In *First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*, pages 1–17, Dagstuhl, Germany, 2002.

[661] J. Graham. The reader's helper: a personalized document reading environment. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'99)*, pages 481–488, 1999.

[662] M. Granitzer, W. Kienreich, V. Sabol, K. Andrews, and W. Klieber. Evaluating a System for Interactive Exploration of Large, Hierarchically Structured Document Repositories. *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, pages 127–133, 2004.

[663] L. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'04)*, pages 478–479, 2004.

[664] L. Gravano, C.-C. K. Chang, and H. García-Molina. STARTS: Stanford proposal for Internet meta-searching. In *Proc. ACM SIGMOD Inter. Conf. on Management of Data*, pages 207–218, Tucson, AZ, May 1997.

[665] L. Gravano, K. Chang, H. Garcia-Molina, C. Lagoze, and A. Paepcke. STARTS - Stanford protocol proposal for internet retrieval and search. http://infolab.stanford.edu/~gravano/starts.html, January 1997. accessed 15 Jun 2009.

[666] L. Gravano and H. Garcia-Molina. Generalizing GlOSS to vector-space databases and broker hierarchies. In U. Dayal, P. M. D. Gray, and S. Nishio, editors, *VLDB*, pages 78–89, Zurich, Switzerland, September 1995. Morgan Kaufmann.

[667] L. Gravano, H. García-Molina, and A. Tomasic. The effectiveness of GlOSS for the text-database discovery problem. In *Proc. ACM SIGMOD Inter. Conf. on Management of Data*, pages 126–137, Minneapolis, MN, May 1994.

[668] M. Gray. Web characterization studies, 1993.

[669] M. Gray. `www-talk` mailing list, June 1993.

[670] M. Gray. Web growth, 1996.

[671] S. Green. Automated link generation: can we do better than term repetition. In *7th WWW Conference*, Brisbane, Australia, 1998.

[672] S. L. Greene, S. Devlin, P. Cannata, and L. Gómez. No ifs, ands, or ors: A study of database querying. *International Journal of Man [sic] -Machine Studies*, 32(3):303–326, 1990.

[673] G. Grefenstette. Comparing two language identification schemes. In *Proceedings of the 3rd international conference on Statistical Analysis of Textual Data (JADT 1995)*, 1995.

[674] G. Grefenstette. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, Boston, USA, 1998.

[675] G. Grefenstette. Upcoming industrial needs for search. In *Proceedings of ECIR'09*, volume 5478 of *Lecture Notes in Computer Science*, page 3. Springer, 2009.

[676] G. Grefenstette and J. Nioche. Estimation of English and non-English language use on the WWW. In *Proceedings of Content-Based Multimedia Information Access (RIAO)*, pages 237–246, Paris, France, 2000.

[677] G. Griffin, A. D. Holub, and P. Perona. The Caltech-256. Technical report, Caltech, 2006.

[678] D. Griffiths. A pragmatic approach to Spearman's rank correlation coefficient. *Teaching Statistics*, 2:10–13, 1980.

[679] G. Grinstein, T. O'Connell, S. Laskowski, C. Plaisant, J. Scholtz, and M. Whiting. VAST 2006 Contest–A Tale of Alderwood. *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST'06)*, pages 215–216, 2006.

[680] R. Grossi, A. Gupta, and J. S. Vitter. High-order entropy-compressed text indexes. In *SODA*, pages 841–850, 2003.

[681] R. Grossi and J. Vitter. Compressed suffix arrays and suffix trees with applications to text indexing and string matching. *SIAM Journal on Computing*, 35(2):378–407, 2006. Preliminary version in Proc. 32nd ACM Symposium on Theory of Computing (STOC), pp. 397–406.

[682] D. A. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics*. Kluwer Academic Publishers, 1998.

[683] V. Guduvada, V. Raghavan, W. Grosky, and R. Kasanagottu. Information retrieval on the world wide web. *IEEE Internet Computing*, Oct-Nov:58–68, 1997.

[684] J. Guiver and E. Snelson. Learning to rank with softrank and gaussian processes. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 259–266, New York, NY, USA, 2008. ACM Press.

[685] A. Gulli and A. Signorini. The indexable Web is more than 11.5 billion pages. In *Poster proceedings of the 14th international conference on World Wide Web*, pages 902–903, Chiba, Japan, 2005. ACM Press.

[686] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRANK: Ranked Keyword Search over XML Documents. In *SIGMOD Conference*, pages 16–27, 2003.

[687] V. Gupta and R. H. Campbell. Internet search engine freshness by Web server help. In *Proceedings of the Symposium on Internet Applications (SAINT)*, pages 113–119, San Diego, California, USA, 2001.

[688] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proc. ACM SIGMOD*, pages 47–57, Boston, Mass, June 1984.

[689] J. Gwertzman and M. Seltzer. World-wide Web cache consistency. In *Proceedings of the 1996 Usenix Technical Conference*, San Diego, California, USA, January 1996.

[690] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[691] Hadoop. http://hadoop.apache.org/, 2007.

[692] C. D. Hafner. Representation of knowledge in a legal information retrieval system. In *SIGIR '80: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 139–153, Kent, UK, UK, 1981. Butterworth & Co.

[693] D. Haines and W. B. Croft. Relevance feedback and inference networks. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Inference Networks, pages 2–11, 1993.

[694] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2003*, pages 37–48. IEEE Computer Society, 2003.

[695] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *JCDL'04: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 296–305, Tucson, Arizona, USA, 2004.

[696] U. Hanani, B. Shapira, and P. Shoval. Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction (UMUAI)*, 11(3):203–259, 2001.

[697] P. Hansen and K. Järvelin. The information seeking and retrieval process at the swedish patent- and registration office. In *Proceedings of the ACM SIGIR'2000 workshop on Patent Retrieval*, July 2000. http://www.sics.se/humle/projects/pir/patent.html.

[698] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybertinetics*, 3(6):610–621, 1973.

[699] E. Hargittai. Classifying and Coding Online Actions. *Social Science Computer Review*, 22(2):210–227, 2004.

[700] D. Harman. Ranking algorithms. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 363–392. Prentice Hall, 1992.

[701] D. Harman. Relevance feedback and other query modification techniques. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 241–263. Prentice Hall, 1992.

[702] D. Harman. Overview of the first text retrieval conference (TREC-1). In D. Harman, editor, *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 1–20. NIST Special Publication 500-207, 1993.

[703] D. Harman. Overview of the second text retrieval conference (TREC-2). In D. Harman, editor, *Proceedings of the Second Text REtrieval Conference (TREC-2)*. NIST Special Publication, 1994.

[704] D. Harman. Overview of the third text retrieval conference (TREC-3). In D. Harman, editor, *Proceedings of the Third Text REtrieval Conference (TREC-3)*. NIST Special Publication, 1995.

[705] D. Harman, E. Fox, R. Baeza-Yates, and W. Lee. Inverted files. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Algorithms and Data Structures*, chapter 3, pages 28–43. Prentice-Hall, Englewood Cliffs, NJ, USA, 1992.

[706] S. Harnad. The self-archiving initiative. *Nature*, 410, 2001.

[707] D. Harper. *Relevance Feedback in Document Retrieval Systems: An Evaluation of Probabilistic Strategies*. PhD thesis, Jesus College, Cambridge, England, 1980.

[708] D. Harper and C. van Rijsbergen. An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34(3):189–216, 1978.

[709] S. P. Harter and C. A. Hert. Evaluation of information retrieval systems: Approaches, issues, and methods. *Review of Information Science and Technology (ARIST)*, 32:3–94, 1997.

[710] S. Harum. Digital Library Initiative, January 1998. http://dli.grainger.uiuc.edu/national.htm.

[711] B. G. Haskell, A. Puri, and A. N. Netravali. *Digital Video: An introduction to MPEG-2*. Kindle Book, 12 1996.

[712] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Machine Learning*. Springer, 2008. 2nd Edition.

[713] A. Hatter and E. Trapasso. Managers say the majority of information obtained for their work is useless, accenture survey finds. accenture.tekgroup.com/article_display.cfm?article_id=4484, January 2007.

[714] C. Hauff and L. Azzopardi. When is query performance prediction effective? In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *SIGIR*, pages 829–830, Boston, MA, USA, 2009. ACM.

[715] C. Hauff, L. Azzopardi, and D. Hiemstra. The combination and evaluation of query performance prediction methods. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, 2009.

[716] C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K.-S. Choi, and A. Chowdhury, editors, *CIKM*, pages 1419–1420, Napa Valley, California, USA, 2008. ACM.

[717] C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K.-S. Choi, and A. Chowdhury, editors, *CIKM*, pages 439–448, Napa Valley, CA, USA, 2008. ACM.

[718] T. Haveliwala. Efficient computation of pagerank. Technical report, Stanford University, 1999.

[719] D. Hawking. Challenges in enterprise search. In *Proceedings of the Australasian Databases Conference ADC2004*, pages 15–26, Dunedin, New Zealand, January 2004. Australian Computer Society. Invited paper: http://es.csiro.au/pubs/hawking_adc04keynote.pdf.

[720] D. Hawking, N. Craswell, F. Crimmins, and T. Upstill. How valuable is external link evidence when searching enterprise webs? In *Proceedings of the Australasian Database Conference ADC2004*, pages 77–84, January 2004. http://es.csiro.au/pubs/hawking_adc04.pdf.

[721] D. Hawking, T. Rowlands, and M. Adcock. Improving rankings in small-scale Web search using click-implied descriptions. *Australian Journal of Intelligent Information Processing Systems. ADCS 2006 special issue.*, 9(2):17–24, December 2006. http://es.csiro.au/pubs/hawking-rowlands-adcock-adcs2006.pdf.

[722] D. Hawking, T. Rowlands, and P. Thomas. C-test: Supporting novelty and diversity in testfiles for search evaluation. In *Proceedings of the SIGIR workshop on redundancy, diversity and interdependent document relevance*, 2009. http://david-hawking.net/pubs/hawking-rowlands-thomas09.pdf.

[723] D. Hawking, E. Voorhees, N.Craswell, and P. Bailey. Overview of the TREC-8 Web Track. In *The Eighth Text REtrieval Conference (TREC-8), Gaithersburg, Maryland, National Institute of Standards and Technology (NIST)*, 1999.

[724] D. Hawking and J. Zobel. Does topic metadata help with Web search? *JASIST*, 58(5):613–628, 2007. Preprint at http://es.csiro.au/pubs/hawking_zobel_jasist.pdf.

[725] Hbase. http://hadoop.apache.org/hbase/, 2008.

[726] HDFS Architecture. http://hadoop.apache.org/common/docs/current/hdfs_design.html, 2008.

[727] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *The Eleventh Symposium on String Processing and Information Retrieval (SPIRE)*, pages 43–54, 2004.

[728] D. He, A. Göker, and D. J. Harper. Combining evidence for automatic Web session identification. *Information Processing & Management*, 38(5):727–742, 2002.

[729] J. He, M. Larson, and M. de Rijke. Using coherence-based measures to predict query difficulty. In *Advances in Information Retrieval: 29th European Conference on IR Research*, pages 689–694, 2008.

[730] H. Heaps. *Information Retrieval - Computational and Theoretical Aspects*. Academic Press, 1978.

[731] M. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics (ACL'94)*, pages 9–16, Las Cruces, NM, June 1994.

[732] M. Hearst. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'95)*, Denver, CO, May 1995.

[733] M. Hearst. Improving full-text precision using simple query constraints. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'96)*, Las Vegas, NV, 1996.

[734] M. Hearst. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64, 1997.

[735] M. Hearst. *Search User Interfaces*. Cambridge University Press, 2009.

[736] M. Hearst, A. Divoli, H. Guturu, A. Ksikes, P. Nakov, M. Wooldridge, and J. Ye. BioText Search Engine: beyond abstract search. *Bioinformatics*, 23(16):2196, 2007.

[737] M. Hearst, J. English, R. Sinha, K. Swearingen, and K.-P. Yee. Finding the flow in Web site search. *Communications of the ACM*, 45(9), September 2002.

[738] M. A. Hearst. Clustering versus faceted categories for information exploration. *Commun. ACM*, 49(4):59–61, 2006.

[739] M. A. Hearst. Design recommendations for hierarchical faceted search interfaces. In A. Z. Broder and Y. S. Maarek, editors, *Proceedings of the SIGIR 2006 Workshop on Faceted Search*, pages 26–30, August 2006.

[740] J. Heer and E. H. Chi. Separating the swarm: categorization methods for user sessions on the web. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 243–250, New York, NY, USA, 2002. ACM.

[741] T. Heimonen and N. Jhaveri. Visualizing Query Occurrence in Search Result Lists. *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'05)*, pages 877–882, 2005.

[742] N. Heintze. Scalable document fingerprinting. In *1996 USENIX Workshop on Electronic Commerce*, November 1996.

[743] J. M. Hellerstein, J. F. Naughton, and A. Pfeffer. Generalized search trees for database systems. In *Proc. of VLDB Conf.*, pages 562–573, Zurich, Switzerland, Sept 1995.

[744] M. Hemmje, C. Kunkel, and A. Willett. LyberWorld – a visualization user interface supporting fulltext retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'94)*, pages 249–259, Dublin, Ireland, July 1994.

[745] M. Henzinger. Web information retrieval - an algorithmic perspective. In *European Symposium on Algorithms*, pages 1–8, 2000. http://citeseer.nj.nec.com/571448.html.

[746] M. Henzinger. Hyperlink analysis for the web. *IEEE Internet Computing*, 5(1):45–50, 2001.

[747] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in Web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[748] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In Smola, Bartlett, Schoelkopf, and Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 2000.

[749] W. Hersh. *Information Retrieval – A Health and Biomedical Perspective*. Springer, 2009. Third Edition.

[750] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–201. Springer-Verlag New York, Inc., 1994.

[751] W. R. Hersh. Improving health care through information. *Jama*, 288:1955–1958, 2002.

[752] W. R. Hersh. Health care information technology—progress and barriers. *Jama*, 292:2273–2274, 2004.

[753] W. R. Hersh and D. H. Hickam. How well do physicians use electronic information retrieval systems? a framework for investigation and systematic review. *Jama*, 280:1347–1352, 1998.

[754] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur. The shark-search algorithm. An application: tailored Web site mapping. In *Proceedings of the seventh conference on World Wide Web*, pages 317–326, Brisbane, Australia, April 1998. Elsevier Science.

[755] M. Hertzum and E. Frokjaer. Browsing and querying in online documentation: A study of user interfaces and the interaction process. *ACM Transactions on Computer-Human Interaction (ToCHI)*, 3(2):136–161, 1996.

[756] M. Hertzum and A. M. Pejtersen. The information-seeking practices of engineers: searching for documents as well as for people. *Information Processing and Management*, 36:761–778, 2000.

[757] E. v. Herwijnen. *Practical SGML*. Kluwer Academic Publishers, second edition edition, 1994.

[758] E. Hetzler and A. Turner. Analysis Experiences Using Information Visualization. *IEEE Computer Graphics and Applications*, 24(5):22–26, 2004.

[759] A. Heydon and M. Najork. Mercator: A scalable, extensible Web crawler. *World Wide Web Conference*, 2(4):219–229, April 1999.

[760] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the Second European Conference on Research and Advance Technology for Digital Libraries (ECDL)*, pages 569–584, 1998.

[761] D. Hiemstra and R. Baeza-Yates. Structured text retrieval models. In *Encyclopedia of Database Systems*. Springer, 2009.

[762] D. Hiemstra and W. Kraaij. Twenty-one at TREC-7: Ad-hoc and cross-language track. In *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, pages 227–238, 1999.

[763] C. Hildreth. OPAC research: laying the groundwork for future OPAC design. *The Online Catalogue: Development and Directions*, pages 1–24, 1989.

[764] C. Hildreth. Online catalog design models: Are we moving in the right direction? Report Submitted to the Council on Library and Information Resources, 1995, updated 2000. http://myweb.cwpost.liu.edu/childret/clr-opac.html.

[765] L. L. Hill, G. Janee, R. Dolin, J. Frew, and M. Larsgaard. Collection metadata solutions for digital library applications. *JASIS*, 50(13):1169–1181, 1999.

[766] W. Hill, J. Hollan, D. Wroblewski, and T. McCandless. Edit wear and read wear. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'92)*, 92:3–9, 1992.

[767] R. Himmeroder, G. Lausen, B. Ludascher, and C. Schlepphorst. On a declarative semantics for Web queries. In *Int. Conf. on Deductive and Object-Oriented Database (DOOD)*, pages 386–398, Singapore, December 1997.

[768] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. Webbase: a repository of Web pages. *Computer Networks (Amsterdam, Netherlands: 1999)*, 33(1–6):277–293, 2000.

[769] D. S. Hirschberg and D. A. Lelewer. Efficient decoding of prefix codes. *Communications of the ACM*, 33(4), 1990.

[770] O. Hoeber and X. D. Yang. A comparative user study of Web search interfaces: HotMap, Concept Highlighter, and Google. *IEEE/WIC/ACM International Conference on Web Intelligence*, 2006.

[771] C. Hölscher and G. Strube. Web search behavior of Internet experts and newbies. *Computer Networks*, 33(1-6):337–346, 2000.

[772] J. Hopcroft and J. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison Wesley, Reading, Mass., 1979.

[773] K. Hornbæk and E. Frøkjær. Do Thematic Maps Improve Information Retrieval. *Human-Computer Interaction (INTERACT'99)*, pages 179–186, 1999.

[774] R. N. Horspool. Practical fast searching in strings. *Software Practice and Experience*, 10:501–506, 1980.

[775] R. N. Horspool and G. V. Cormack. Constructing word-based text compression algorithms. In *Proc. of IEEE Second Data Compression Conference*, pages 62–81, 1992.

[776] E. Hörster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval CIVR 07*, July 2007.

[777] M. Hosseini. A study on performance volatility in information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (doctoral consortium)*, page 853, Boston, MA, USA, 2009.

[778] A. J. M. Houtsma. Pitch and timbre: Definition, meaning and use. *Journal of New Music Research*, 26:104–115, 1997.

[779] P. Howarth and S. Rüger. Robust texture features for still-image retrieval. *IEE Proceedings on Vision, Image and Signal*, 152(6):868–874, 2005.

[780] D. Howe and H. Nissenbaum. Trackmenot: Resisting surveillance in Web search. In I. Kerr, C. Lucock, and V. Steeves, editors, *On the Identity Trail: Privacy, Anonymity and Identity in a Networked Society*. Oxford: Oxford University Press, 2009.

[781] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, Mar. 2002.

[782] W. Hsu, L. Kennedy, and S.-F. Chang. Reranking methods for visual search. *Multimedia, IEEE*, 14(3):14–22, July-Sept. 2007.

[783] ht://Dig. http://www.htdig.org/, 2007.

[784] Y. Hu, H. Li, Y. Cao, D. Meyerzon, and Q. Zheng. Automatic extraction of titles from general documents using machine learning. In *JCDL'05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, Tools & techniques: supporting classification, pages 145–154, 2005.

[785] F. Huang, S. Watt, D. Harper, and M. Clark. Compact representations in XML retrieval. In *Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, pages 64–72, Dagstuhl Castle, Germany, 2006. Revised and Selected Papers.

[786] J. Huang, S. Ertekin, Y. Song, H. Zha, and C. L. Giles. Efficient multiclass boosting classification with active learning. In *SDM*. SIAM, 2007. http://www.siam.org/meetings/proceedings/2007/datamining/papers/027Huang.pdf.

[787] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 762–768, 1997.

[788] X. Huang, A. Acero, and H.-W. Hon. *Spoken language processing*. Prentice Hall PTR, 2000.

[789] X. Huang, F. Peng, A. An, and D. Schuurmans. Dynamic Web log session identification with statistical language models. *JASIST*, 55(14):1290–1303, 2004.

[790] Z. Huang, W. Chung, T.-H. Ong, and H. Chen. A graph-based recommender system for digital library. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 65–73, Portland, Oregon, 2002.

[791] Z. Huang, X. Li, and H. Chen. Link prediction approach to collaborative filtering. In *Proceedings ACM/IEEE Joint Conference on Digital Libraries, JCDL 2005*, pages 141–142, Denver, CA, USA, 2005.

[792] B. A. Huberman. *The Laws of the Web: Patterns in the Ecology of Information*. The MIT Press, October 2001.

[793] B. A. Huberman and L. A. Adamic. Evolutionary dynamics of the World Wide Web. *Condensed Matter*, January 1999.

[794] B. A. Huberman and L. A. Adamic. Growth dynamics of the world-wide web. *Nature*, 399, 1999.

[795] D. Huffman. A method for the construction of minimum-redundancy codes. *Proc. of the I.R.E.*, 40(9):1090–1101, 1952.

[796] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338, 1993.

[797] D. A. Hull. Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 282–291, Dublin, Ireland, July 1994.

[798] J. Hunter and L. Armstrong. A comparison of schemas for video metadata representation. In *WWW'99: Proceedings of the eighth international conference on World Wide Web*, pages 1431–1451, New York, NY, USA, 1999. Elsevier North-Holland, Inc.

[799] T. Hybrid Library. The HyLIFE Hybrid Library Toolkit, 2001. http://hylife.unn.ac.uk/toolkit/.

[800] Hypertable. http://www.hypertable.org/, 2009.

[801] IAB. Interactive advertising bureau, September 2006. http://www.iab.net/news/pr_2006_05_30.asp.

[802] R. Iannella. Digital rights management (DRM) architectures. *D-Lib Magazine*, 7, 2001.

[803] Ibiblio. Internet Pioneers: Tim Berners-Lee. Ibiblio, the Public's Library and Digital Archive, http://www.ibiblio.org/pioneers/lee.html, September, 2006.

[804] IBM OmniFind Yahoo! Edition. http://omnifind.ibm.yahoo.net/, 2007.

[805] IDC. The Enterprise Workplace: How It Will Change the Way We Work. IDC Report 32919, February 2005.

[806] E. Ide. New experiments in relevance feedback. In G. Salton, editor, *The SMART Retrieval System*, pages 337–354. Prentice Hall, 1971.

[807] IEEE Standards Committee on Optical Disk and Multimedia Platforms (SCODMP). IEEE SFQL. Technical report, IEEE, Washington, USA, 1992.

[808] IEEE Computer, February 1999. 32(2).

[809] Indri. http://www.lemurproject.org/indri/, 2007.

[810] P. Ingwersen and K. Jarvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, 2005.

[811] Inktomi. http://www.inktomi.com, 1998.

[812] Internet Systems Consortium. Internet domain survey. http://ftp.isc.org/www/survey/reports/current/, 2009.

[813] Information Processing & Management, March 1999. 35(3).

[814] K. Itakura and C. A. Clarke. University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks. In *Focused Access to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*, pages 417–425, Dagstuhl Castle, Germany, 2007. Selected Papers.

[815] IWS. Internet world stats, January 2010. http://www.internetworldstats.com/top20.htm.

[816] P. Jaccard. étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Socit Vaudoise des Sciences Naturelles*, 37:547579, 1901.

[817] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[818] S. C. Jane Hunter. Implementing Preservation Strategies for Complex Multimedia Objects. In *Proc. 7th European Conf. Research and Advanced Technology for Digital Libraries, ECDL 2003*, pages 473–486, Trodheim, Norway, August 17-22, 2003.

[819] B. Jansen, A. Spink, and S. Koshman. Web searcher interaction with the Dogpile.com metasearch engine. *Journal of the American Society for Information Science and Technology*, 58(5):744–755, 2007.

[820] B. Jansen, A. Spink, and J. Pedersen. A Temporal Comparison of AltaVista Web Searching. *Journal of the American Society for Information Science and Technology*, 56(6):559–570, 2005.

[821] B. J. Jansen. *Understanding User-Web Interactions via Web Analytics*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan Claypool, 2009.

[822] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of Web search engine queries. In *Proc. of the 16th international conference on World Wide Web*, pages 1149–1150. ACM Press, 2007.

[823] B. J. Jansen and A. Spink. An analysis of Web searching by European AlltheWeb.com users. *Information Processing and Management: an International Journal*, 41(2):361–381, 2005.

[824] B. J. Jansen and A. Spink. How are we searching the World Wide Web? a comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248–263, 2006.

[825] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on Web search engines. *JASIST*, 58(6):862–871, 2007.

[826] B. P. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *ACM SIGIR Forum*, 32(1):5–17, Spring 1998.

[827] N. Jardine and C. Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

[828] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR International Conference on Information Retrieval*, pages 41–48, 2000.

[829] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[830] N. Jayant, J. Johnston, and R. Safranek. Signal compression based on models of human perception. *Proceedings of the IEEE*, 81(10), October 1993.

[831] G. Jeh and J. Widom. Scaling personalized Web search. In *WWW'03: Proceedings of the 12th international conference on World Wide Web*, pages 271–279, New York, NY, USA, 2003. ACM.

[832] C. Jenkins, C. Corritore, and S. Wiedenbeck. Patterns of information seeking on the Web: a qualitative study of domain expertise and Web expertise. *IT & Society*, 1(3):64–89, 2003.

[833] B.-S. Jeong and E. Omiecinski. Inverted file partitioning schemes in multiple disk systems. *IEEE Trans. Par. and Dist. Syst.*, 6(2):142–153, Feb. 1995.

[834] S. Ji, G. Li, C. Li, and J. Feng. Efficient interactive fuzzy keyword search. In *WWW'09: Proceedings of the 18th international conference on World wide web*, pages 371–380, New York, NY, USA, 2009. ACM.

[835] Y. Jing and S. Baluja. PageRank for product image search. In *WWW'08: Proceeding of the 17th International Conference on World Wide Web*, pages 307–316, New York, NY, USA, 2008. ACM.

[836] J.Nielsen. *Hypertext and Hypermedia*. Academic Press, 1990.

[837] T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *International Conference on Machine Learning (ICML)*, pages 143–151, 1997.

[838] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

[839] T. Joachims. SVMLight – Support Vector Machine, 1999. http://svmlight.joachims.org/.

[840] T. Joachims. SVMPerf – Support Vector Machine, 1999. http://svmlight.joachims.org/svm_perf.html.

[841] T. Joachims. Optimizing search engines using clickthrough data. In D. Hand, D. Keim, and R. Ng, editors, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-02)*, pages 132–142, Edmonton, Alberta, Canada, July 2002.

[842] T. Joachims. Evaluating retrieval performance using clickthrough data. In J. Franke, G. Nakhaeizadeh, and I. Renz, editors, *Text Mining*, pages 79–96. Physica/Springer Verlag, 2003.

[843] T. Joachims. Training linear SVMs in linear time. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226, 2006.

[844] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2005. ACM.

[845] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems*, 25(2), 2007.

[846] T. Joachims and F. Radlinski. Search engines that learn from implicit feedback. *IEEE Computer*, 40(8):34–40, August 2007.

[847] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K.-S. Choi, and A. Chowdhury, editors, *CIKM*, pages 699–708, Napa Valley, CA, USA, November 2008. ACM.

[848] R. Jones, R. Kumar, B. Pang, and A. Tomkins. "I know what you did last summer": query logs and user privacy. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 909–914, New York, NY, USA, 2007. ACM Press.

[849] R. Jones, R. Kumar, B. Pang, and A. Tomkins. Vanity fair: privacy in query log bundles. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 853–862, New York, NY, USA, 2008. ACM Press.

[850] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW'06: Proceedings of the 15th international conference on World Wide Web*, pages 387–396, New York, NY, USA, 2006. ACM Press.

[851] S. Jones. Graphical query specification and dynamic result previews for a digital library. In *Proceedings of the 11th annual ACM symposium on User Interface Software and Technology (UIST'98)*, pages 143–151, San Francisco, USA, November 1998.

[852] W. Jones, H. Bruce, and S. Dumais. Keeping Found Things Found on the Web. *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM'01)*, pages 119–126, 2001.

[853] W. Jones, S. Dumais, and H. Bruce. Once Found, What Then? A Study of "Keeping" Behaviors in the Personal Use of Web Information. *Proceedings of the American Society for Information Science and Technology*, 39(1):391–402, 2002.

[854] D. Jonker, W. Wright, D. Schroh, P. Proulx, and B. Cort. Information Triage with TRIST. *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[855] W.-K. Joo and S. H. Myaeng. Improving retrieval effectiveness with hyperlink information. In *Proceedings of International Workshop on Information Retrieval with Asian Languages (IRAL)*, Singapore, October 1998.

[856] T. Joyce and R. Needham. The thesaurus approach to information retrieval. In K. S. Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 15–20. Morgan Kaufmann Publishers, Inc., 1997.

[857] F. Junqueira and K. Marzullo. Coterie availability in sites. In *Proceedings of the International Conference on Distributed Computing (DISC)*, number 3724 in LNCS, pages 3–17, Krakow, Poland, September 2005. Springer Verlag.

[858] S. Kaasten, S. Greenberg, and C. Edwards. How People Recognise Previously Seen Web Pages from Titles, URLs and Thumbnails. *People and Computers*, pages 247–266, 2002.

[859] B. Kahle. Archiving the Internet. http://www.alexa.com/~brewster/essays/sciam_article.html, 1997.

[860] B. Kahle and A. Medlar. An information server for corporate users: Wide Area Information Servers. *ConneXions - the Interoperability Report*, 5(11):2–9, 1991. ftp://think.com/wais/wais-corporate-paper.text.

[861] M. Kaisser, M. Hearst, and J. Lowe. Improving Search Results Quality by Customizing Summary Lengths. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technlogies (ACL-HLT'08)*, 2008.

[862] M. Käki. Findex: Search Result Categories Help Users When Document Ranking Fails. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'05)*, pages 131–140, 2005.

[863] J. Kalbach. *Designing Web navigation*. O'Reilly, 2007.

[864] T. Kalt. A new probabilistic model of text classification and retrieval. Technical Report IR-78, CIIR, Univ. of Massachussets at Amherst, 1996. http://ciir.cs.umass.edu/~kalt/kaltTr96.pdf.

[865] I. Kamel and C. Faloutsos. Hilbert R-tree: An improved R-tree using fractals. In *Proc. of VLDB Conference*, pages 500–509, Santiago, Chile, Sept 1994.

[866] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. Length normalization in XML retrieval. In *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK*, pages 80–87, 2004.

[867] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. INEX 2007 Evaluation Metrics. In *Focused access to XML documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*, Dagstuhl Castle, Germany, 2008. Selected Papers.

[868] J. Kamps and B. Sigurbjörnsson. What do users think of an XML element retrieval system? In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 of *Lecture Notes in Computer Science*, pages 411–421, 2006.

[869] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the Web for computing pagerank, 2003.

[870] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the twelfth international conference on World Wide Web*, pages 261–270. ACM Press, 2003.

[871] B. Kang and R. Wilensky. Toward a model of self-administering data. In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 322–330, Roanoke, Virginia, 2001.

[872] I.-H. Kang and G. Kim. Query type classification for Web document retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71, New York, NY, USA, 2003. ACM Press.

[873] T. Kanungo and D. Orr. Predicting the readability of short Web summaries. In *ACM WSDM '09: 2nd ACM International Conference on Web Search and Data Mining*, 2009.

[874] J. Kapms. Indexing units. In *Encyclopedia of Database Systems*. Springer, 2009.

[875] D. Karger, E. Lehman, T. Leighton, R. Panigrahy, M. Levine, and D. Lewin. Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the world wide web. In *STOC '97: Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 654–663, El Paso, TX, USA, 1997. ACM.

[876] J. Kärkkäinen and P. Sanders. Simple linear work suffix array construction. In *Proc. ICALP'03*, pages 943–955, 2003.

[877] R. Karp and M. Rabin. Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development*, 31(2):249–260, Mar. 1987.

[878] M. Kaszkiel and J. Zobel. Passage retrieval revisited. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM Press, 1997.

[879] N. Katayama and S. Satoh. The SR-tree: An index structure for high-dimensional nearest neighbor queries. In *Proc. of ACM SIGMOD*, pages 369–380, Tucson, AZ, 1997.

[880] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.

[881] H. Kautz, B. Selman, and M. Shah. The hidden Web. *AI Magazine*, 18(2):27–36, 1997.

[882] G. Kazai. Choosing an Ideal Recall-Base for the Evaluation of the Focused Task: Sensitivity Analysis of the XCG Evaluation Measures. In *Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, pages 35–44, Dagstuhl Castle, Germany, 2007. Revised and Selected Papers.

[883] G. Kazai. INitiative for the Evaluation of XML retrieval (INEX). In *Encyclopedia of Database Systems*. Springer, 2009.

[884] G. Kazai and A. Doucet. Overview of the INEX 2007 Book Search Track (BookSearch '07). *SIGIR Forum*, 42(1):2–15, 2008.

[885] G. Kazai and M. Lalmas. eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval. *ACM Transactions on Information Systems*, 24(4):503–542, 2006.

[886] G. Kazai, M. Lalmas, and A. de Vries. The overlap problem in content-oriented XML retrieval evaluation. In *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK*, pages 72–79, 2004.

[887] G. Kazai, M. Lalmas, and J. Reid. Construction of a test collection for the focused retrieval of structured documents. In *Advances in Information Retrieval, 25th European Conference on IR Research, ECIR 2003, Pisa, Italy*, pages 88–103, 2003.

[888] G. Kazai, N. Milic-Frayling, and J. Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *ACM SIGIR International Conference on Information Retrieval*, 2009.

[889] G. Kazai and T. Rölleke. A Scalable Architecture for XML Retrieval. In *First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*, pages 49–56, Dagstuhl, Germany, 2002.

[890] Y. Ke, L. Deng, W. Ng, and D.-L. Lee. Web dynamics and their ramifications for the development of Web search engines. *Computer Networks*, 50(10):1430–1447, July 2006.

[891] KEA: Keyphrase extraction algorithm. http://www.nzdl.org/Kea/, 2009.

[892] J. Kekäläinen. Binary and graded relevance in IR evaluations: comparison of the effects on ranking of IR systems. *Information Processing & Management*, 41(5):1019–1033, 2005.

[893] J. Kekäläinen, P. Arvola, and M. Junkkari. Contextualization. In *Encyclopedia of Database Systems*. Springer, 2009.

[894] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224, 2009.

[895] D. Kelly, V. Dollu, and X. Fu. The Loquacious User: A Document-Independent Source of Terms for Query Expansion. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'05)*, pages 457–464, 2005.

[896] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[897] K. Kelly. Scan This Book!, May 2006. New York Times Magazine.

[898] M. G. Kendall. *Rank Correlation Methods*. Hafner Publishing Company, 1955.

[899] R. Kengeri, C. D. Seals, H. D. Harley, H. P. Reddy, and E. A. Fox. Usability study of digital libraries: ACM, IEEE-CS, NCSTRL, NDLTD. *Int. J. on Digital Libraries*, 2(2-3):157–169, 1999.

[900] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How Flickr helps us make sense of the world: Context and content in community-contributed media collections. In *MULTIMEDIA '07: Proceedings of the 15th International Conference on Multimedia*, pages 631–640, New York, NY, USA, 2007. ACM.

[901] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev. To search or to label?: Predicting the performance of search-based automatic image classifiers. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia IR*, pages 249–258, New York, NY, USA, 2006. ACM.

[902] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *WWW'08: Proceeding of the 17th International Conference on World Wide Web*, pages 297–306, New York, NY, USA, 2008. ACM.

[903] A. Kent, M. M. Berry, F. U. Luehrs Jr., and J. W. Perry. Operational criteria for designing information retrieval systems. *American Documentation*, 6(2), 1955.

[904] R. Khare and A. Rifkin. XML: A door to automated Web applications. *IEEE Internet Computing*, 1(4):78–86, 1977.

[905] R. Khare and A. Rifkin. The origin of (document) species. *Computer Networks and ISDN Systems*, 30(1-7), 1998. WWW7 Conference, Brisbane, Australia, available at http://decweb.ethz.ch/WWW7/00/.

[906] P. Kilpeläinen and H. Mannila. Retrieval from hierarchical texts by partial patterns. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA, USA, 1993*, pages 214–222, 1993.

[907] B. J. Kim, A. Trusina, P. Minnhagen, and K. Sneppen. Self organized scale-free networks from merging and regeneration, Mar 2004.

[908] D. W. King and C. Tenopir. Evolving journal costs: Implications for publishers, libraries, and readers. *Learned Publishing*, 12:251–258, Oct. 1999.

[909] S. T. Kirsch. Document retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents. US Patent 5,659,732, Aug. 1997.

[910] A. Kleiboemer, M. Lazear, and J. Pedersen. Tailoring a retrieval system for naive users. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR '96)*, Las Vegas, NV, 1996.

[911] J. Kleinberg. Authoritative sources in a hyperlinked environment. *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 46(5):604–632, 1998. http://www.cs.cornell.edu/home/kleinber/auth.pdf.

[912] D. Knuth, J. Morris, and V. Pratt. Fast pattern matching in strings. *SIAM J. Comput*, 6(2):323–350, June 1977.

[913] D. E. Knuth. *The Art of Computer Programming*, volume 3: Searching and Sorting. Addison-Wesley, 1973.

[914] T. Kochtanek and J. Matthews. *Library Information Systems: From Library Automation to Distributed Information*. Libraries Unlimited, 2002.

[915] T. R. Kochtanek and K. K. Hein. Delphi study of digital libraries. *Information Processing & Management*, 35(3):245–254, 1999.

[916] I. Kodratoff and J. Carbonell. *Machine Learning: An Artificial Intelligence Approach, Vol. III*. Kaufman Publishers Inc., 1990.

[917] W. Koehler. A longitudinal study of Web pages continued: a consideration of document persistence. *Information Research*, 9(2), January 2004.

[918] J. Koenemann and N. J. Belkin. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 205–212, 1996.

[919] R. Kohavi, R. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'07)*, pages 959–967. ACM Press New York, NY, USA, 2007.

[920] R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In P. Berkhin, R. Caruana, and X. Wu, editors, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 959–967. ACM, 2007.

[921] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.*, 18(1):140–181, 2009.

[922] R. Kohavi, L. Mason, R. Parekh, and Z. Zheng. Lessons and Challenges from Mining Retail E-Commerce Data. *Machine Learning*, 57(1):83–113, 2004.

[923] R. Kohavi, D. Sommerfield, and J. Dougherty. Data mining using MLC++, a machine learning library in C++. In *ICTAI '96: Proceedings of the 8th International Conference on Tools with Artificial Intelligence*, page 234, Washington, DC, USA, 1996. IEEE Computer Society.

[924] A. Kolcz and A. Chowdhury. Hardening fingerprinting by context. In *CEAS 2007 - The Fourth Conference on Email and Anti-Spam*, Mountain View, CA, USA, 2007.

[925] A. Kolcz and A. Chowdhury. Lexicon randomization for near-duplicate detection with I-Match. *The Journal of Supercomputing*, 45(3):255–276, 2008.

[926] A. Kolcz, A. Chowdhury, and J. Alspector. The impact of feature selection on signature-driven spam detection. In *CEAS 2004 - First Conference on Email and Anti-Spam*, Mountain View, CA, USA, July 2004.

[927] A. Kolcz, A. Chowdhury, and J. Alspector. Improved robustness of signature-based near-replica detection via lexicon randomization. In W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, editors, *KDD*, pages 605–610, Seattle, WA, USA, August 2004. ACM.

[928] D. Konopnicki and O. Shmueli. W3QS: A query system for the World Wide Web. In *Proc. of VLDB'95*, pages 54–65, Zurich, Switzerland, Sept. 1995.

[929] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In *WWW 2009*, pages 171–180, 2009.

[930] J. Korpela. Lurching Toward Babel: HTML, CSS, and XML. *IEEE Computer*, 31(7):103–106, 1998.

[931] R. Korphage. *Information Storage and Retrieval*. John Wiley & Sons, Inc., 1997.

[932] J. Korst and V. Pronk. *Multimedia Storage and Retrieval: An Algorithmic Approach*. John Wiley & Sons, 2005.

[933] M. Koster. Guidelines for robots writers. http://www.robotstxt.org/wc/guidelines.html, 1993.

[934] M. Koster. Robots in the web: threat or treat ? *ConneXions*, 9(4), April 1995.

[935] M. Koster. A standard for robot exclusion. http://www.robotstxt.org/wc/exclusion.html, 1996.

[936] N. Koudas, C. Faloutsos, and I. Kamel. Declustering spatial databases on a multi-computer architecture. *EDBT Conf. Proc.*, pages 592–614, Mar. 1996.

[937] G. Kowalski and M. T. Maybury. *Information Storage and Retrieval Systems: Theory and Implementation*. Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[938] D. Kraft and D. Buel. Fuzzy sets and generalized Boolean retrieval systems. *International Journal of Man-Machine Studies*, 19:45–56, 1983.

[939] R. Krishnan. Google notebook blog. http://googlenotebookblog.blogspot.com/2009/01/stopping-development-on-google-notebook.html, Jan 2009.

[940] A. Krowne. Planetmath. http://planetmath.org/.

[941] A. Krumpholz and D. Hawking. InexBib - retrieving XML elements based on external evidence. *Australian Journal of Intelligent Information Processing Systems. ADCS 2006 special issue.*, 9(2):72–79, December 2006. http://es.csiro.au/pubs/krumpholz-hawking-adcs2006.pdf.

[942] U. Kruschwitz. *Intelligent Document Retrieval: Exploiting Markup Structure (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[943] J. Kubica, A. Moore, D. Cohn, and J. Schneider. cgraph: A fast graph-based method for link analysis and queries, 2003.

[944] K. Kukich. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–440, Dec. 1992.

[945] B. Kules and B. Shneiderman. Users can change their Web search tactics: Design guidelines for categorized overviews. *Information Processing and Management*, 44(2):463–484, 2008.

[946] R. Kumar, J. Novak, B. Pang, and A. Tomkins. On anonymizing query logs via token-based hashing. In *WWW'07: Proceedings of the 16th international conference on World Wide Web*, pages 629–638, New York, NY, USA, 2007. ACM Press.

[947] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks*, 31(11–16):1481–1493, 1999.

[948] M. Kuniavsky. *Observing the User Experience: A Practitioner's Guide to User Research*. Morgan Kaufmann, 2003.

[949] K. Kwok. A neural network for probabilistic information retrieval. In *Proc. ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 21–30, 1989.

[950] K. Kwok. Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information Systems*, 8(4):363–386, October 1990.

[951] K. Kwok. A network approach to probabilistic information retrieval. *ACM Transactions on Information Systems*, 13(3):324–353, July 1995.

[952] K. Kwok, L. Papadopolous, and Y. Kwan. Retrieval experiments with a large collection using pircs. In *Proc of the First TExt Retrieval Conference (TREC-1)*, USA, 1993. Special Publication 500-267, National Institute of Standards and Technology (NIST).

[953] K. L. Kwok. An attempt to identify weakest and strongest queries. In *Proceedings of the 28th Annual Conference on Research and Development in Information Retrieval (SIGIR)*, 2005.

[954] A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, and B. A. Ribeiro-Neto. Learning to advertise. In *Proceedings of the 29th ACM Int. Conference on Information Retrieval, ACM SIGIR*, pages 549–556, 2006.

[955] A. H. F. Laender, B. A. Ribeiro-Neto, and A. S. da Silva. DEByE - data extraction by example. *Data and Knowledge Engineering*, 40(2):121–154, 2002.

[956] B. Lagoeiro, M. A. Gonçalves, and A. H. F. Laender. 5SQual - A Quality Assessment Tool for Digital Libraries. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, page (accepted for publication), 2007.

[957] C. Lagoze. Networked Computer Science Technical Reference Library. http://www.ncstrl.org.

[958] C. Lagoze. The Warwick framework: A container architecture for diverse sets of metadata. *D-Lib Magazine*, 2(7), July 1996.

[959] C. Lagoze, W. Arms, S. Gan, D. Hillmann, C. Ingram, D. Krafft, R. Marisa, J. Phipps, J. Saylor, C. Terrizzi, W. Hoehn, D. Millman, J. Allan, S. Guzman-Lara, and T. Kalt. Core services in the architecture of the national science digital library (NSDL). In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 201–209, Portland, Oregon, 2002.

[960] C. Lagoze, W. Y. Arms, S. Gan, D. Hillmann, C. Ingram, D. B. Krafft, R. J. Marisa, J. Phipps, J. Saylor, C. Terrizzi, W. Hoehn, D. Millman, J. Allan, S. Guzman-Lara, and T. Kalt. Core services in the architecture of the national science digital library (nsdl). In *ACM/IEEE Joint Conference on Digital Libraries (JCDL 2002)*, pages 201–209, Portland, Oregon, 2002.

[961] C. Lagoze, D. Fielding, and S. Payette. Making global digital libraries work: Collection services, connectivity regions, and collection views. In I. Witten, R. Akscyn, and F. M. Shipman, editors, *Proc. of the 3rd ACM Conf. on Digital Libraries (DL-98)*, pages 134–143, jun 1998.

[962] C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: an architecture for complex objects and their relationships. *Int. J. on Digital Libraries*, 6(2):124–138, 2006.

[963] C. Lagoze and H. van de Sompel. The Open Archives Initiative. In *Proc. of the 1st Joint Conf. on Digital Libraries (JCDL'2001)*, pages 54–62, Roanoke, Virginia, June 24-28, 2001.

[964] L. V. S. Lakshmanan, F. Sadri, and I. N. Subramanian. A declarative language for querying and restructuring the Web. In *Proc. of 6th. International Workshop on Research Issues in Data Engineering, RIDE '96*, New Orleans, Feb. 1996.

[965] M. Lalmas. *XML Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan Claypool, 2009.

[966] M. Lalmas and V. Murdock, editors. *ACM SIGIR Workshop on Aggregated Search*, Singapore, 2008.

[967] M. Lalmas and I. Ruthven. Representing and retrieving structured documents using the dempster-shafer theory of evidence: Modelling and evaluation. *Journal of Documentation*, 54(5):529–565, 1998.

[968] M. Lalmas and A. Tombros. Evaluating XML Retrieval Effectiveness at INEX. *SIGIR Forum*, 41(1):40–57, 2007.

[969] Lam, Wai and Lai, Kwok-Yin. A meta-learning approach for text categorization. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 303–309, New Orleans, Lousiana, 2001.

[970] B. LaMacchia. The Internet fish construction kit. In *6th. Int'l. WWW Conference*, Santa Clara, CA, USA, Apr. 1997.

[971] L. Lamport. Paxos made simple. *ACM SIGACT News*, 32(4):51–58, December 2001.

[972] F. Lancaster. *Indexing and Abstracting in Theory and Practice*. University of Illinois, 3rd edition, 2003.

[973] G. Landau and U. Vishkin. Fast string matching with k differences. *Journal of Computer Systems Science*, 37:63–78, 1988.

[974] T. Landauer, D. Egan, J. Remde, M. Lesk, C. Lochbaum, and D. Ketchum. Enhancing the usability of text through computer delivery and formative evaluation: the superbook project. In C. McKnight, A. Dillon, and J. Richardson, editors, *Hypertext: A Psychological Perspective*, pages 71–136. Ellis Horwood, 1993.

[975] P. Langley. *Elements of Machine Learning*. Morgan Kaufmann, 1996.

[976] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, July 2006.

[977] A. Large, L. Tedd, and R. Hartley, editors. *Information Seeking in The Online Age: Principles and Practice*. Bowker-Saur, London, UK, 1999.

[978] L. S. Larkey, M. E. Connell, and J. Callan. Collection selection and results merging with topically organized u.s. patents and trec data. In *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*, pages 282–289, New York, NY, USA, 2000. ACM Press.

[979] L. S. Larkey and W. B. Croft. Combining classifiers in text categorization. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pages 289–297, Zürich, CH, 1996. ACM Press, New York, US.

[980] R. R. Larson. Cheshire II at INEX'03: Component and Algorithm Fusion for XML Retrieval. In *INEX 2003 Proceedings*, pages 38–45, 2003.

[981] N. Larsson and A. Moffat. Offline dictionary-based compression. *Proceedings of the IEEE*, 88(11):1722–1732, 2000.

[982] O. Lassila. Web metadata: A matter of semantics. *IEEE Internet Computing*, 2(4):30–37, 1998.

[983] O. Lassila and R. Swick. World Wide Web Consortium - RDF. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, 1999.

[984] E. Lau and D. Goh. In search of query patterns: a case study of a university OPAC. *Information Processing and Management*, 42(5):1316–1329, 2006.

[985] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, 2001.

[986] S. Lawrence and C. L. Giles. Context and page analysis for improved Web search. *IEEE Internet Computing*, 2(4):38–46, 1998.

[987] S. Lawrence and C. L. Giles. Inquirus, the NECI meta search engine. In *7th WWW Conference*, pages 95–105, Brisbane, Australia, 1998.

[988] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *Computer*, 32(6):67–71, 1999.

[989] S. Lawrence and L. C. Giles. Accessibility of information on the web. *Intelligence*, 11(1):32–39, 2000.

[990] A. Lazonder, H. Biemans, and I. Wopereis. Differences Between Novice and Experienced Users in Searching Information on the World Wide Web. *Journal of the American Society for Information Science*, 51(6):576–581, 2000.

[991] D. Lea. *Concurrent Programming in Java: Design Principles and Patterns*. The Java Series. Addison-Wesley, Reading, MA, 1997.

[992] C. P. Lee, G. H. Golub, and S. A. Zenios. A fast two-stage algorithm for computing pagerank and its extensions. Technical report, Stanford University, 2004.

[993] J. Lee and P. Kantor. A study of probabilistic information retrieval systems in the case of inconsistent expert judgements. *Journal of the American Society for Information Sciences*, 42(3):166–172, 1991.

[994] J. Lee, W. Kim, and Y. Lee. Ranking documents in thesaurus-based Boolean retrieval systems. *Information Processing & Management (IP&M)*, 30(1):79–91, 1993.

[995] J. H. Lee. Properties of extended Boolean models in information retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Statistical Models, pages 182–190, 1994.

[996] J. H. Lee, W. Y. Kim, M. H. Kim, and Y. J. Lee. On the evaluation of Boolean operators in the extended Boolean retrieval framework. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Mathematical Models, pages 291–297, 1993.

[997] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in Web search. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 391–400, New York, NY, USA, 2005. ACM Press.

[998] Y.-B. Lee and S. H. Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of SIGIR-02, 25th ACM International Conference on Research and Development in Information Retrieval*, pages 145–150, Tampere, FI, 2002.

[999] W. G. LeFurgy. Building preservation partnerships: the library of congress national digital information infrastructure and preservation program. *Library Trends*, 54(1), 2005. http://www.digitalpreservation.gov/library/pdf/building.pdf.

[1000] W. G. LeFurgy, M. Hedstrom, T. A. Pardo, and T. O. Walters. Preserving information long-term: digital archiving. In L. M. L. Delcambre and G. Giuliano, editors, *Proceedings of the 2005 National Conference on Digital Government Research, DG.O 2005, Atlanta, Georgia, USA, May 15-18, 2005*, page 15. Digital Government Research Center, 2005.

[1001] J. Lehman. Building a taxonomy. Technical Report 2, New Idea Engineering - NIE Enterprise Search, june 2003. http://www.ideaeng.com/pub/entsrch/issue02/article02.html.

[1002] R. Lempel and S. Moran. Predictive caching and prefetching of query results in search engines. In *WWW'03: Proceedings of the 12th international conference on World Wide Web*, pages 19–28, New York, NY, USA, 2003. ACM Press.

[1003] Lemur toolkit. http://www.lemurproject.org/, 2007.

[1004] M. Lesk. Word-word associations in document retrieval systems. *American Documentation*, 20(1):8–36, 1969.

[1005] M. Lesk. *Practical Digital Libraries; Books, Bytes, & Bucks*. Morgan Kaufman, 1997.

[1006] M. Lesk. *Understanding Digital Libraries*. Morgan Kaufmann, 2nd edition, 2005.

[1007] J. Leskovec, S. Dumais, and E. Horvitz. Web Projections: Learning from Contextual Subgraphs of the Web. In *Int'l Conference of the World Wide Web*, 2007.

[1008] O. Levard. Google peut-il vous traiter d'arnaqueur? *LCI*, July 21 2009. http://tf1.lci.fr/infos/economie/entreprises/0,4490540,00.html (in French).

[1009] O. Levard. Les suggestions très limites de google. *LCI*, July 16 2009. http://tf1.lci.fr/infos/high-tech/0,4227853,00-les-suggestions-tres-limites-de-google-.html (in French).

[1010] M. Levene and A. Poulovassilis. *Web Dynamics*. Springer, 2004.

[1011] M. Levene and A. Poulovassilis. Special issue on Web dynamics. *Computer Networks*, 50(10):1425–1429, 2006.

[1012] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Phys. Dokl*, 6:126–136, 1966.

[1013] D. M. Levy. Heroic measures: reflections on the possibility and purpose of digital preservation. In *DL'98: Proceedings of the 3rd ACM International Conference on Digital Libraries*, pages 152–161, Pittsburgh, PA, 1998.

[1014] D. D. Lewis. Naive (Bayes) at forty: the independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 4–15, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[1015] D. D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In *Symposium on Document Analysis and Information Retrieval*, pages 81–93, Las Vegas, Nevada, Apr. 1994.

[1016] D. D. Lewis, Y. Yang, T. G. Rose, G. Dietterich, F. Li, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[1017] LexisNexis. Data Centres. http://www.lexisnexis.com/presscenter/mediakit/datacenter.asp.

[1018] M. Li, M. Zhu, Y. Zhang, and M. Zhou. Exploring Distributional Similarity Based Models for Query Spelling Correction. In *Annual Meeting-Association for Computational Linguistics (ACL'06)*, 2006.

[1019] P. Li, C. J. C. Burges, and Q. Wu. Mcrank: Learning to rank using multiple classification and gradient boosting. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*. MIT Press, 2007.

[1020] W.-S. Li, J. Shim, K. Candan, and Y. Hara. WebDB: A Web query system and its modeling, language, and implementation. In *Proc. of Advances in Digital Libraries*, Santa Barbara, CA, USA, April 1998.

[1021] Y. Li. Toward a qualitative search engine. *IEEE Internet Computing*, 2(4):24–29, July 1998.

[1022] Y. H. Li and A. K. Jain. Classification of text documents. *Comput. J*, 41(8):537–546, 1998.

[1023] Library of Congress, Z39.50 Maintenance Agency, June 1998. http://lcweb.loc.gov/z3950/agency/.

[1024] Library of Congress. The national digital information infrastructure and preservation program, 2006. http://www.digitalpreservation.gov/.

[1025] Library of Congress, Metadata Encoding and Transmission Standard (METS), May 2006. http://www.loc.gov/standards/mets/.

[1026] Library of Congress, Network Development and MARC Standards Office. Metadata object description schema (MODS) version 3.1, July 2005.

[1027] H. Lie and B. Bos. *Cascading Style Sheets: Designing for the Web*. Addison-Wesley, 1997.

[1028] R. Lienhart. Comparison of automatic shot boundary detection algorithms. *SPIE Image and Video Processing VII*, pages 3656–3729, January 1999.

[1029] R. Lienhart. Reliable transition detection in videos: a survey and practitioner' s guide. *International Journal of Image and Graphics*, 1(3):469–486, 2001.

[1030] R. Lienhart, S. Pfeiffer, and W. Effelsberg. The MoCA Workbench: Support for creativity in movie content analysis. In *ICMCS*, pages 314–321, 1996.

[1031] L. Lim, M. Wang, S. Padmanabhan, J. S. Vitter, and R. Agarwal. Characterizing Web document change. In *Proceedings of the Second International Conference on Advances in Web-Age Information Management*, volume 2118 of *Lecture Notes in Computer Science*, pages 133–144, London, UK, July 2001. Springer.

[1032] L. R. S. Lima, A. H. F. Laender, and B. A. A. Ribeiro-Neto. A hierarchical approach to the automatic categorization of medical documents. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pages 132–139, New York, NY, USA, 1998. ACM.

[1033] J. Lin, M. DiCuccio, V. Grigoryan, and W. Wilbur. Navigating information spaces: A case study of related article search in PubMed. *Information Processing and Management*, 2008.

[1034] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. Karger. What Makes a Good Answer? the Role of Context in Question Answering. *Proceedings of Human-Computer Interaction (INTERACT'03)*, 2003.

[1035] K.-I. Lin, H. Jagadish, and C. Faloutsos. The TV-tree - an index structure for high-dimensional data. *VLDB Journal*, 3:517–542, Oct. 1994.

[1036] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[1037] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, 1st ed. 2007. corr. 2nd printing edition, January 2009.

[1038] F. Liu and R. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(7):722–733, July 1996.

[1039] T.-Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *LR4IR Workshop, in conjunction with SIGIR 2007*, 2007.

[1040] T.-Y. Liu, Y. Yang, H. Wan, H.-J. Zeng, Z. Chen, and W.-Y. Ma. Support vector machines classification with a very large-scale taxonomy. *SIGKDD Explorations*, 7(1):36–43, 2005.

[1041] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193, New York, NY, USA, 2004. ACM Press.

[1042] X. Liu and W. B. Croft. Statistical language modeling for information retrieval. In B. Cronin, editor, *Annual Review of Information Science and Technology*, volume 39. ASIS&T, 2005. Chapter 1.

[1043] Y. Liu, M. Zhang, and R. Cen. Data cleansing for Web information retrieval using query independent features. *Journal of the American Society for Information Science and Technology*, 58(12):001–015, 2007.

[1044] M.-L. Lo and C. V. Ravishankar. Spatial joins using seeded trees. In *Proc. of ACM SIGMOD*, pages 209–220, Minneapolis, MN, USA, May 1994.

[1045] Load monitor project. http://sourceforge.net/projects/monitor, 2007.

[1046] X. Long and T. Suel. Optimized query execution in large search engines with global page ordering. In *Proceedings of VLDB 2003*, pages 129–140, 2003.

[1047] X. Long and T. Suel. Three-Level Caching for Efficient Query Processing in Large Web Search Engines. In *WWW'05: Proceedings of the 14th International World Wide Web conference*, Chiba, Japan, 2005.

[1048] B. T. Loo, R. Huebsch, J. M. Hellerstein, S. Shenker, and I. Stoica. Enhancing P2P File-Sharing with an Internet-Scale Query Processor. In *VLDB'04: Proceedings of the 30th International conference on Very Large Data Bases*, Toronto, Canada, 2004.

[1049] R. A. Lorie. Long term preservation of digital information. In *Proceedings of the 1st ACM/IEEE Joint Conference on Digital Libraries*, pages 346–352, Roanakoe, Virginia, 2001.

[1050] R. Losee and A. Bookstein. Integrating Boolean queries in conjunctive normal form with probabilistic retrieval models. *Information Processing & Management (IP&M)*, 24(3):315–321, 1988.

[1051] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[1052] R. Lowry. *Concepts and Applications of Inferential Statistics*. Vassar College, Poughkeepsie, NY, USA, 2008. http://faculty.vassar.edu/lowry/webtext.html.

[1053] J. Lu and J. Callan. Content-Based Retrieval in Hybrid Peer-to-Peer Networks. In *Proc. of ACM Int'l Conf. on Information and Knowledge Management*, pages 199–206, 2003.

[1054] J. Lu and J. Callan. Federated Search of Text-Based Digital Libraries in Hierarchical Peer-to-Peer Networks. In *ECIR'05: Proceedings of the 27th European conference on IR Research*, Santiago de Compostela, Spain, 2005.

[1055] J. Lu and J. Callan. User Modeling for Full-Text Federated Search in Peer-to-Peer Networks. In *SIGIR'06: Proceedings of the 29th International ACM SIGIR conference on Research and Development in Information Retrieval*, Seattle, WA, USA, 2006.

[1056] Q. Lu, P. Cao, E. Cohen, K. Li, and S. Shenker. Search and replication in unstructured peer-to-peer networks. In *ICS '02: Proceedings of the 16th international conference on Supercomputing*, pages 84–95, New York, NY, USA, 2002. ACM.

[1057] W. Lu, S. Robertson, and A. MacFarlane. Field-Weighted XML Retrieval Based on BM25. In *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, pages 161–171, Dagstuhl Castle, Germany, 2005. Revised Selected Papers.

[1058] Y. Lu, C. Hu, X. Zhu, H. Zhang, and Q. Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 31–37, 2000.

[1059] Z. Lu, K. S. McKinley, and B. Cahoon. The hardware/software balancing act for information retrieval on symmetric multiprocessors. Technical Report TR98-25, Dept. of Comp. Sci., Univ. of Mass., Amherst, MA, 1998.

[1060] C. Lucchese, S. Orlando, R. Perego, and F. Silvestri. Mining query logs to optimize index partitioning in parallel Web search engines. In *Proceedings of the 2nd international conference on Scalable information systems*, Suzhou, China, 2007.

[1061] Lucene. http://jakarta.apache.org/lucene/, 2007.

[1062] H. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957.

[1063] H. Luhn. *Keyword-in-context Index for Technical Literature (KWIC Index)*. International Business Machines Corp., Advanced Systems Development Division, 1959.

[1064] T.-Y. Lui. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[1065] R. P. Luk, H. V. Leong, T. Dillon, A. S. Chan, W. B. Croft, and J. Allan. A survey in indexing and searching XML documents. *Journal of the American Society for Information Science and Technology*, 53(6):415–437, 2002.

[1066] C. A. Lynch and J. K. Lippincott. Institutional Repository Deployment in the United States as of Early 2005. *D-Lib Magazine*, 11, 2005. http://www.dlib.org/dlib/september05/lynch/09lynch.html.

[1067] Y. Maarek, M. Jacovi, M. Shtalhaim, S. Ur, D. Zernik, and I. Ben-Shaul. Webcutter: a system for dynamic and tailorable site mapping. In *Selected papers from the sixth international conference on World Wide Web*, pages 1269–1279, Essex, UK, 1997. Elsevier Science Publishers Ltd.

[1068] Y. Maarek and D. Zernick. Proceedings of the WWW6 Workshop on Site Mapping, April 1997.

[1069] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Sept. 2003.

[1070] I. Macleod. Storage and retrieval of structured documents. *Information Processing & Management*, 26(2):197–208, 1990.

[1071] I. MacLeod. A query language for retrieving information from hierarchic text structures. *The Computer Journal*, 34(3):254–264, 1991.

[1072] I. A. Macleod, T. P. Martin, B. Nordin, and J. R. Phillips. Strategies for building distributed information retrieval systems. *Inf. Process. & Mgmnt.*, 23(6):511–528, 1987.

[1073] V. Mäkinen and G. Navarro. Succinct suffix arrays based on run-length encoding. *Nordic Journal of Computing*, 12(1):40–66, 2005.

[1074] Managing Gigabytes. http://www.cs.mu.oz.au/mg/, 2007.

[1075] U. Manber and G. Myers. Suffix arrays: a new method for on-line string searches. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pages 319–327, San Francisco, USA, 1990.

[1076] U. Manber, A. Patel, and J. Robison. The business of personalization: Experience with personalization of Yahoo! *Commun. ACM*, 43(8):35–39, 2000.

[1077] U. Manber, M. Smith, and B. Gopal. WebGlimpse: combining browsing and searching. In *Proc. of USENIX Technical Conference*, pages 195–206, Anaheim, USA, Jan 1997.

[1078] U. Manber and S. Wu. GLIMPSE: A tool to search through entire file systems. In *Proceedings of the Winter 1994 USENIX Conference: January 17–21, 1994, San Francisco, California, USA*, pages 23–32, Berkeley, CA, USA, Winter 1994.

[1079] P. Maniatis, M. Roussopoulos, T. J. Giuli, D. S. H. Rosenthal, and M. Baker. The LOCKSS peer-to-peer digital preservation system. *ACM Transactions on Computer Systems*, 23(1):2–50, 2005.

[1080] B. Manjunat, P. Salembier, and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley and Sons, 2002.

[1081] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[1082] G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1995.

[1083] G. Marchionini. Evaluating digital libraries: A longitudinal and multifaceted view. *Library Trends*, 49(2), 2000.

[1084] G. Marchionini. A briefing on the evolution and status of the open video digital library. *Int. J. on Digital Libraries*, 4(1):36–38, 2004.

[1085] G. Marchionini. Exploratory Search: From Finding To Understanding. *Communications of the Acm*, 49(4):41–49, 2006.

[1086] M. Marchiori. The quest for correct information of the Web: hyper search engines. In *Proc. of the sixth international conference on the Web*, pages 265–274, Santa Clara, CA, USA, April 1997.

[1087] M. Marín, C. Bonacic, V. G. Costa, and C. Gómez. A Search Engine Accepting On-Line Updates. In *13th European International Conference on Parallel Processing (Euro-Par 2007)*, LNCS 4641, pages 348–357, Rennes, France, 2007.

[1088] M. Marín and V. G. Costa. High-performance Distributed Inverted Files. In *ACM 16th Conference on Information and Knowledge Management (CIKM 2007)*, pages 935–938, Lisbon, Portugal, Nov. 2007.

[1089] M. Marín and V. G. Costa. (Sync|Async)$^+$ MPI Search Engines. In *14th European PVM/MPI Meeting*, LNCS 4757, pages 117–124, Paris, 2007.

[1090] M. Marín and C. Gómez. Load Balancing Distributed Inverted Files. In *9th ACM International Workshop on Web Information and Data Management (WIDM 2007)*, pages 57–64, Lisbon, Portugal, November 2007.

[1091] E. P. Markatos. On caching search engine query results. *Computer Communications*, 24(2):137–143, 2001.

[1092] J. Markwell and D. W. Brooks. Link-rot limits the usefulness of Web-based educational materials in biochemistry and molecular biology. *Biochem. Mol. Biol. Educ.*, 31:69–72, 2003.

[1093] M. Maron and J. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of ACM*, 7(3):216–244, 1960.

[1094] C. C. Marshall. Making metadata: A study of metadata creation for a mixed physical-digital collection. In *DL'98: Proceedings of the 3rd ACM International Conference on Digital Libraries*, pages 162–171, 1998.

[1095] B. Masand, G. Linoff, and D. Waltz. Classifying news stories using memory based reasoning. In N. Belkin, P. Ingwersen, and A. M. Pejtersen, editors, *Proceedings of the 15th Annual International Conference on Reasearch and Development in Information Retrieval*, SIGIR Forum, pages 59–65, New York, NY, USA, June 1992. ACM Press.

[1096] M. Masnick. Two separate rulings in france split over whether google's suggestion algorithm can be libelous. *Techdirt*, July 24 2009. http://www.techdirt.com/articles/20090724/0407145647.shtml.

[1097] Y. Mass and M. Mandelbrod. Retrieving the most relevant XML Components. In *INEX 2003 Proceedings*, pages 53–58, 2003.

[1098] Y. Mass and M. Mandelbrod. Component Ranking and Automatic Query Refinement for XML Retrieval. In *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*, pages 73–84, Dagstuhl Castle, Germany, 2005. Revised Selected Papers.

[1099] Y. Mass and M. Mandelbrod. Using the INEX Environment as a Test Bed for Various User Models for XML Retrieval. In *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, pages 187–195, Dagstuhl Castle, Germany, 2006. Revised Selected Papers.

[1100] M. Massey and W. Bender. Salient stills: process and practice. *IBM Syst. J.*, 35(3-4):557–573, 1996.

[1101] M. T. Maybury. *Intelligent Multimedia Information Retrieval*. MIT Press, 1997.

[1102] M. Mayer. Universal search: the best answer is still the best answer. *The Official Google Blog*, May 2007. http://googleblog.blogspot.com/2007/05/universal-search-best-answer-is-still.html.

[1103] O. A. Mcbryan. GENVL and WWWW: Tools for taming the web. In *Proceedings of the first World Wide Web Conference*, Geneva, Switzerland, May 1994.

[1104] A. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow, 1996.

[1105] A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press, 1998.

[1106] J. McCarthy. Artificial intelligence, logic and formalizing common sense. In R. Thomason, editor, *Philosophical Logic and Artificial Intelligence*. Kluver Academic, 1989.

[1107] J. McCarthy. *Formalizing Common Sense: Papers by John McCarthy*. Ablex Publishing Corporation, 1990.

[1108] E. McCreight. A space-economical suffix tree construction algorithm. *Journal of the ACM*, 23(2):262–272, 1976.

[1109] K. McKeown, J. Hirschberg, M. Galley, and S. Maskey. From text to speech summarization. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages v/997–v1000, March 2005.

[1110] F. McMartin and Y. Terada. Digital library services for authors of learning materials. In *Proc. of JCDL'02*, pages 117–118, Portland, OR, 2002.

[1111] F. Mcsherry. A uniform approach to accelerated pagerank computation. In *WWW'05: Proceedings of the 14th international conference on World Wide Web*, pages 575–582, New York, NY, USA, 2005. ACM Press.

[1112] C. T. Meadow, B. R. Boyce, D. H. Kraft, and C. L. Barry. *Text Information Retrieval Systems, Third Edition*. Academic Press, Inc., Orlando, FL, USA, 2007.

[1113] M. Mealling and R. Denenberg. Uniform resource identifiers (URIs), URLs, and uniform resource names (URNs): Clarifications and recommendations. Internet informational RFC 3305, Aug. 2002.

[1114] R. Meddis and M. J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification, and II: Phase sensitivity. *J. Acoust. Soc. Am.*, 89:2866–2894, 1991.

[1115] MEDLINE. NLM—United States National Library of Medicine, 2009.

[1116] C. Meilhac and C. Nastar. Relevance feedback and category search in image databases. In *IEEE International Conference on Multimedia Computing*, 1999.

[1117] S. Melnik, S. Raghavan, B. Yang, and H. Garcia-Molina. Building a distributed full-text index for the web. *ACM Trans. Inf. Syst.*, 19(3):217–241, 2001.

[1118] D. A. Menascé and V. A. Almeida. *Capacity Planning for Web Performance: Metrics, Models, and Methods*. Prentice Hall, 1998.

[1119] F. Menczer. Lexical and semantic clustering by Web links. *Journal of the American Society for Information Science and Technology*, 55(14):1261–1269, August 2004.

[1120] A. Mendelzon, G. Mihaila, and T. Milo. Querying the World Wide Web. *International Journal on Digital Libraries*, 1(1):54–67, April 1997.

[1121] D. Merrill. http://www.youtube.com/watch?v=syKY8CrHkck\#t=22m11sattimestamp22m11s.

[1122] R. Michalski, J. Carbonell, and T. Mitchell. *Machine Learning: An Artificial Intelligence Approach, Vol. I.* Kaufman Publishers Inc., 1983.

[1123] R. Michalski, J. Carbonell, and T. Mitchell. *Machine Learning: An Artificial Intelligence Approach, Vol. II.* Kaufman Publishers Inc., 1986.

[1124] S. Michel, M. Bender, N. Ntarmos, P. Triantafillou, G. Weikum, and C. Zimmer. Discovering and Exploiting Keyword and Attribute-Value Co-occurrences to Improve P2P Routing Indices. In *CIKM'06: Proceedings of the 15th ACM International conference on Information and Knowledge Management*, Arlington, Virginia, USA, 2006.

[1125] S. Michel, P. Triantafillou, and G. Weikum. KLEE: a framework for distributed top-k query algorithms. In *VLDB'05: Proceedings of the 31st International conference on Very Large Data Bases*, Trondheim, Norway, 2005.

[1126] S. Michel, P. Triantafillou, and G. Weikum. MINERVA∞: A Scalable Efficient Peer-to-Peer Search Engine. In *Middleware'05: Proceedings of the 6th International Middleware conference*, Grenoble, France, 2005.

[1127] V. Mihajlovic, G. Ramírez, T. Westerveld, D. Hiemstra, H. E. Blok, and A. de Vries. TIJAH Scratches INEX 2005: Vague Element Selection, Image Search, Overlap, and Relevance Feedback. In *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, pages 72–87, Dagstuhl Castle, Germany, 2006. Revised Selected Papers.

[1128] P. Mika. *Social Networks and the Semantic Web*, volume 5 of *Semantic Web and Beyond*. Springer-Verlag, Berlin-Heidelberg, 2007.

[1129] P. Mika. Microsearch: An interface for semantic search. In *Proceedings of the SemSearch 2008 Workshop on Semantic Search at the 5th European Semantic Web Conference*, Tenerife, Spain, June 2008. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-334/.

[1130] N. Milic-Frayling, R. Jones, K. Rodden, G. Smyth, A. Blackwell, and R. Sommerer. Smartback: supporting users in back navigation. *Proceedings of the 13th International Conference on World Wide Web (WWW'04)*, pages 63–71, 2004.

[1131] D. R. Millen, J. Feinberg, and B. Kerr. Dogear: Social bookmarking in the enterprise. In *Proceedings of CHI '06*, pages 111–120, New York, NY, USA, 2006. ACM Press.

[1132] D. Miller, T. Leek, and R. Schwartz. A hidden Markov model information retrieval system. In *ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 214–221, 1999.

[1133] G. Miller, E. Newman, and E. Friedman. Length-frequency statistics for written english. *Information and Control*, 1:370–389, 1958.

[1134] R. Miller. Websphinx, a personal, customizable Web crawler. http://www-2.cs.cmu.edu/~rcm/websphinx, 2004.

[1135] R. Miller and K. Bharat. Sphinx: A framework for creating personal, site-specific Web crawlers. In *Proceedings of the seventh conference on World Wide Web*, Brisbane, Australia, April 1998. Elsevier Science.

[1136] M. Mills, J. Cohen, and Y. Y. Wong. A magnifier tool for video data. In *CHI '92: Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 93–98, New York, NY, USA, 1992. ACM.

[1137] J. Minker, G. Wilson, and B. Zimmerman. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8(6):329–348, 1972.

[1138] T. Minohara and R. Watanabe. Queries on structure in hypertext. In *Foundations of Data Organization and Algorithms, FODO '93*, pages 394–411. Springer, 1993.

[1139] R. Mitchell, D. Day, and L. Hirschman. Fishing for information on the Internet. In *Proceedings '95 Information Visualization*, pages 105–111, Atlanta, USA, Oct. 1995.

[1140] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[1141] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In B. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. of 21st Annual International Conference on Research and Development in Information Retrieval, SIGIR 98*, pages 206–214, Melbourne, Australia, 1998.

[1142] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[1143] S. Miyamoto, T. Miyake, and K. Nakayama. Generation of a pseudothesaurus for information retrieval based on cooccurrences and fuzzy set operations. *IEEE Transactions on Systems and Man Cybernetics*, 13(1):62–70, 1983.

[1144] S. Miyamoto and K. Nakayama. Fuzzy information retrieval based on a fuzzy pseudothesaurus. *IEEE Transactions on Systems and Man Cybernetics*, 16(2):278–282, 1986.

[1145] S. Mizzaro. Relevance: the whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

[1146] W. E. Moen. *The development of ANSI/NISO Z39.50: A case study in standards evolution*. PhD thesis, Syracuse University, 1998.

[1147] A. Moffat. Word-based text compression. *Software Practice and Experience*, 19(2):185–198, 1989.

[1148] A. Moffat. *Compression and Coding Algorithms*. Kluwer, 2002.

[1149] A. Moffat and T. Bell. In situ generation of compressed inverted files. *Journal of the American Society for Information Science*, 46(7):537–550, 1995.

[1150] A. Moffat and R. Wan. Re-Store: A system for compressing, browsing, and searching large documents. In *Proc. 8th International Symposium on String Processing and Information Retrieval*, pages 162–174, 2001.

[1151] A. Moffat, W. Webber, and J. Zobel. Load balancing for term-distributed parallel retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 348–355, New York, NY, USA, 2006. ACM Press.

[1152] A. Moffat and J. Zobel. Information retrieval systems for large document collections. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 85–94, Gaithersburg, MD, USA, 1995. Dept. of Commerce, National Institute of Standards and Technology. Special Publication 500-226.

[1153] A. Moffat and J. Zobel. What does it mean to "measure performance"? In X. Zhou, S. Su, M. P. Papazoglou, M. E. Owlowska, and K. Jeffrey, editors, *Proc. Fifth International Conf. on Web Informations Systems*, pages 1–12, Brisbane, Australia, Nov. 2004. LNCS 3306, Springer.

[1154] K. Monostori, R. A. Finkel, A. B. Zaslavsky, G. Hodász, and M. Pataki. Comparison of overlap detection techniques. In P. M. A. Sloot, C. J. K. Tan, J. Dongarra, and A. G. Hoekstra, editors, *International Conference on Computational Science (I)*, volume 2329 of *Lecture Notes in Computer Science*, pages 51–60, Amsterdam, The Netherlands, 2002. Springer.

[1155] K. Monostori, A. B. Zaslavsky, and H. W. Schmidt. Efficiency of data structures for detecting overlaps in digital documents. In *24th Australasian Computer Science Conference (ACSC 2001)*, pages 140–147, Gold Coast, QU, Australia, February 2001. IEEE Computer Society.

[1156] M. R. Morris and J. Teevan. *Collaborative Web Search: Who, What, Where, When, and Why*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan Claypool, 2009.

[1157] P. Morville and L. Rosenfeld. *Information Architecture for the World Wide Web: Designing Large-Scale Web Sites*. O'Reilly Media, 3rd edition, December 2006.

[1158] E. Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates. Fast and flexible word searching on compressed text. *ACM Transactions on Information Systems (TOIS)*, 18(2):113–139, 2000.

[1159] M. A. Moura. Personal communication at the Information Sciences School, UFMG, Brazil, 2004.

[1160] F. Mourão, L. C. da Rocha, R. B. Araújo, T. Couto, M. A. Gonçalves, and W. Meira Jr. Understanding temporal aspects in document classification. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, pages 159–170, 2008.

[1161] S. Mukherjea and J. Foley. Visualizing the World Wide Web with the Navigational View Builder. *Computer Networks and ISDN Systems*, 27:1075–1087, 1995.

[1162] R. Mukherjee and J. Mao. Enterprise search: Tough stuff. *Queue*, 2(2):36–46, 2004.

[1163] H. Müller, W. Müller, D. Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters*, 22(5):593–601, 2001.

[1164] S. A. Murray. *The Library – An Illustrated History*. Skyhorse Publishing, 2009.

[1165] S.-H. Myaeng, D.-H. Jang, M.-S. Kim, and Z.-C. Zhoo. A flexible model for retrieval of SGML documents. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia*, pages 138–145, 1998.

[1166] G. Myers. A fast bit-vector algorithm for approximate string matching based on dynamic progamming. *Journal of the ACM*, 46(3):395–415, 1999.

[1167] J. L. Myers and A. D. Well. *Research Design and Statistical Analysis*. Lawrence Erlbaum, 2003. Second Edition, 508 pages.

[1168] MySpace. http://www.myspace.com/, 2003.

[1169] M. Najork and J. L. Wiener. Breadth-first crawling yields high-quality pages. In *Proceedings of the Tenth Conference on World Wide Web*, pages 114–118, Hong Kong, May 2001. Elsevier Science.

[1170] R. Nallapati. Discriminative models for information retrieval. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *SIGIR*, pages 64–71, Sheffield, UK, July 2004. ACM Press.

[1171] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at TRECVID. In *MULTIMEDIA '04: Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 660–667, New York, NY, USA, 2004. ACM Press.

[1172] P. Nardiello, F. Sebastiani, and A. Sperduti. Discretizing continuous attributes in adaboost for text categorization. In *Proceedings of the 25th European Conference on Advances in Information Retrieval*, pages 320–334, 2003.

[1173] A. Nation. Visualizing websites using a hierarchical table of contents browser: Webtoc. In *Proceedings of the Third Conference on Human Factors and the Web*, Denver, CO, 1997.

[1174] National Library of Medicine (NLM). UMLS - Unified Medical Language System. http://www.nlm.nih.gov/research/umls/about_umls.html, September, 2006.

[1175] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001.

[1176] G. Navarro. NR-grep: a fast and flexible pattern matching tool. *Software Practice and Experience (SPE)*, 31:1265–1312, 2001.

[1177] G. Navarro. Indexing text using the Ziv-Lempel trie. *Journal of Discrete Algorithms*, 2(1):87–114, 2004.

[1178] G. Navarro and R. Baeza-Yates. A language for queries on structure and contents of textual databases. In *18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA*, pages 93–101, 1995.

[1179] G. Navarro and R. Baeza-Yates. Proximal nodes: A model to query document databases by content and structure. *ACM Transactions on Information Systems*, 15(4):400–435, 1997.

[1180] G. Navarro and R. Baeza-Yates. Very fast and simple approximate string matching. *Information Processing Letters*, 72:65–70, 1999.

[1181] G. Navarro and R. Baeza-Yates. A hybrid indexing method for approximate string matching. *Journal of Discrete Algorithms (JDA)*, 1(1):205–239, 2000.

[1182] G. Navarro, R. Baeza-Yates, E. Barbosa, N. Ziviani, and W. Cunto. Binary searching with non-uniform costs and its application to text retrieval. *Algorithmica*, 27:145–169, 2000.

[1183] G. Navarro, J. Kitajima, B. A. Ribeiro-Neto, and N. Ziviani. Distributed generation of suffix arrays. In A. Apostolico and J. Hein, editors, *Proc. of Combinatorial Pattern Matching*, number 1264 in LNCS, pages 102–115, Aarhus, Denmark, 1997. Springer-Verlag.

[1184] G. Navarro and V. Mäkinen. Compressed full-text indexes. *ACM Comput. Surv.*, 39(1), 2007.

[1185] G. Navarro, E. Moura, M. Neubert, N. Ziviani, and R. Baeza-Yates. Adding compression to block addressing inverted indexes. *Information Retrieval*, 3(1):49–77, 2000.

[1186] G. Navarro and M. Raffinot. Fast and flexible string matching by combining bit-parallelism and suffix automata. *ACM Journal of Experimental Algorithmics (JEA)*, 5(4), 2000.

[1187] G. Navarro and M. Raffinot. *Flexible Pattern Matching in Strings – Practical online search algorithms for texts and biological sequences*. Cambridge University Press, 2002. ISBN 0-521-81307-7. 280 pages.

[1188] G. Navarro and M. Raffinot. New techniques for regular expression searching. *Algorithmica*, 41(2):89–116, 2005.

[1189] G. Navarro and J. Tarhio. LZgrep: A Boyer-Moore string matching tool for Ziv-Lempel compressed text. *Software Practice and Experience (SPE)*, 35(12):1107–1130, 2005.

[1190] NDLTD. Networked Digital Library of Theses and Dissertations. http://www.ndltd.org, 2004.

[1191] M. Needleman. The shibboleth authentication/authorization system. *Serials Review*, 30(3):252–253, 2004.

[1192] M. Nelson. Data compression with the burrows-wheeler transform. *Dr. Dobb's Journal*, Sept. 1996.

[1193] M. L. Nelson and K. Maly. Buckets: smart objects for digital libraries. *Communications of the ACM*, 44(5):60–62, 2001.

[1194] M. L. Nelson, K. Maly, M. Zubair, and S. N. T. Shen. Soda: Smart objects, dumb archives. In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)*, pages 453–464, 1999.

[1195] J. Nesbit. The accuracy of approximate string matching algorithms. *J. of Computer-Based Instruction*, 13(3):80–83, 1986.

[1196] Netcraft. Web Server Survey. http://news.netcraft.com/, 2010.

[1197] M. L. Neufeld and M. Cornog. Database history: from dinosaurs to compact discs. *Journal of the American Society for Information Science*, 37(4):183–190, 1986.

[1198] T. Neumann, M. Bender, S. Michel, and G. Weikum. A reproducible benchmark for P2P retrieval. In *Proc. First Int. Workshop on Performance and Evaluation of Data Management Systems, ExpDB*, 2006.

[1199] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, December 2005.

[1200] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Link analysis, eigenvectors and stability. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 903–910, Seattle, Washington, USA, 2001.

[1201] H. Ng, W. Goh, and K. Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 67–73, 1997.

[1202] D. Ngu and X. Wu. SiteHelper: a localized agent that helps incremental exploration of the World Wide Web. In *6th. Int'l. WWW Conference*, Santa Clara, CA, USA, Apr. 1997.

[1203] L. T. Nguyen, W. G. Yee, and O. Frieder. Adaptive distributed indexing for structured peer-to-peer networks. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1241–1250, New York, NY, USA, 2008. ACM.

[1204] J. Nielsen. *Usability Engineering*. Academic Press, 1993.

[1205] J. Nielsen. Statistics for traffic referred by search engines and navigation directories to USEIT. http://www.useit.com/about/searchreferrals.html, 2004.

[1206] J. Nielsen. When Search Engines Become Answer Engines, 2004. http://www.useit.com/alertbox/20040816.html.

[1207] K. Nigam, A. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[1208] G. Noether. Why Kendall Tau? Technical report, RSSCSE, 2008. http://rsscse.org.uk/ts/bts/noether/text.html.

[1209] G. Notess. Search Engines Showdown: The User's Guide to Search Engines. http://www.searchengineshowdown.com/, 1998.

[1210] H. Nottelmann and N. Fuhr. Evaluating different methods of estimating retrieval quality for resource selection. In *Proc. of the ACM Int'l Conf. on Information Retrieval*, pages 290–297, 2003.

[1211] H. Nottelmann and N. Fuhr. Combining CORI and the decision-theoretic approach for advanced resource selection. In *ECIR*, Sunderalnd, UK, 2004.

[1212] NTCIR—NII Test Collection for IR Project, 2009.

[1213] NTCIR-7 PATMT—Patent Translation Test Collection, 2009.

[1214] A. Ntoulas and J. Cho. Pruning Policies for Two-Tiered Inverted Index with Correctness Guarantee. In *SIGIR'07: Proceedings of the 30th International ACM SIGIR conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 2007.

[1215] A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the Web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, pages 1–12, 2004.

[1216] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam Web pages through content analysis. In *Proceedings of the World Wide Web conference*, pages 83–92, Edinburgh, Scotland, May 2006.

[1217] Nutch. http://lucene.apache.org/nutch/, 2007.

[1218] OCLC. Web Services and SRW/U, 2006. http://www.oclc.org/research/projects/webservices/default.htm.

[1219] V. L. O'Day and R. Jeffries. Orienteering in an information landscape: how information seekers get from here to there. In *Proceedings of the INTERCHI Conference on Human Factors in Computing Systems (CHI'93)*, Amsterdam, April 1993. IOS Press.

[1220] Open directory project: http://www.dmoz.org/, 2009.

[1221] C. of the ACM. Digital Libraries, April 1995. 38(4).

[1222] C. of the ACM. Digital Libraries: Global Scope, Unlimited Access, April 1998. 41(4).

[1223] C. of the ACM. Digital Libraries, May 2001. 44(5).

[1224] Y. Ogawa, T. Morita, and K. Kobayashi. A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy Sets and Systems*, 39:163–179, 1991.

[1225] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proceedings of ACM SIGIR '03*, pages 143–150, New York, NY, USA, 2003. ACM Press.

[1226] P. Ogilvie and M. Lalmas. Investigating the exhaustivity dimension in content-oriented XML element retrieval evaluation. In *ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA*, pages 84–93, 2006.

[1227] R. A. O'Keefe and A. Trotman. The simplest query language that could possibly work. In *INEX 2003 Proceedings*, pages 167–174, 2003.

[1228] K. Olsen, R. Korfhage, K. Sochats, M. Spring, and J. Williams. Visualization of a Document Collection with Implicit and Explicit Links-The Vibe System. *Scandinavian Journal of Information Systems*, 5:79–95, 1993.

[1229] C. Olston and M. Najork. Web crawling. *Foundations and Trends in Information Retrieval*, 4(3):172–246, 2009.

[1230] E. O'Neill, B. Lavoie, and P. McClain. OCLC Web characterization project (position paper). In *Web Characterization Workshop*, Boston, USA, Nov 1998. http://www.w3.org/1998/11/05/WC-workshop/Papers/oneill.htm.

[1231] Open linking data project: http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData, 2007.

[1232] N. Orio. Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, 1(1):1–90, 2006.

[1233] B. O'Riordan, K. Curran, and D. Woods. Investigating text input methods for mobile phones. *Journal of Computer Science*, 1(2):189–199, 2005.

[1234] Orkut. http://www.orkut.com/, 2004.

[1235] S. Orlando, R. Perego, and F. Silvestri. Design of a Parallel and Distributed WEB Search Engine. In *Proceedings of Parallel Computing (ParCo) 2001 conference*, pages 197–204. Imperial College Press, September 2001.

[1236] M. Özsu and L. Liu, editors. *Encyclopedia of Database Systems*. Springer, 2009.

[1237] T. Paek, S. Dumais, and R. Logan. WaveLens: A new view onto Internet search results. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)*, pages 727–734, 2004.

[1238] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[1239] G. Panagopoulos and C. Faloutsos. Bit-sliced signature files for very large text databases on a parallel machine architecture. In *Proc. 4th Inter. Conf. on Extending Database Technology (EDBT)*, number 779 in LNCS, pages 379–392, London, 1994. Springer-Verlag.

[1240] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2006.

[1241] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing*, pages 79–86, Philadelphia, US, 2002. Association for Computational Linguistics, Morristown, US.

[1242] G. Pant, K. Tsioutsiouliklis, J. Johnson, and C. L. Giles. Panorama: extending digital libraries with topical crawlers. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 142–150, Tucson, Arizona, 2004.

[1243] A. Papadopoulos and Y. Manolopoulos. Performance of nearest neighbor queries in R-trees. In F. N. Afrati and P. Kolaitis, editors, *Proc. of 6th Int. Conf. on Database Theory*, number 1186 in LNCS, pages 394–408, Delphi, Greece, Jan 1997.

[1244] J. Park and J. Kim. Effects of contextual navigation aids on browsing diverse web systems. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 257–264, New York, NY, USA, 2000. ACM.

[1245] N. Paskin. The DOI Handbook. Edition 4.2.0. International DOI Foundation (IDF). http://www.doi.org/hb.html, 1994.

[1246] A. Patterson. Why writing your own search engine is hard. *ACM Queue*, April 2004.

[1247] E. Patterson, E. Roth, and D. Woods. Predicting Vulnerabilities in Computer-Supported Inferential Analysis under Data Overload. *Cognition, Technology & Work*, 3(4):224–237, 2001.

[1248] V. Paxson. End-to-end routing behavior in the Internet. *ACM SICOMM Computer Communication Review*, 35(5):43–56, October 2006.

[1249] S. Payette and T. Staples. The Mellon Fedora Project. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'02)*, pages 406–421, Rome, Italy, 2002.

[1250] G. W. Paynter. Developing practical automatic metadata assignment and evaluation tools for Internet resources. In M. Marlino, T. Sumner, and F. M. S. III, editors, *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2005, Denver, CA, USA, June 7-11, 2005, Proceedings*, pages 291–300. ACM, 2005.

[1251] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1988.

[1252] J. Pehcevski and B. Piwowarski. Evaluation metrics for structured text retrieval. In *Encyclopedia of Database Systems*. Springer, 2009.

[1253] J. Pehcevski and J. A. Thom. Hixeval: Highlighting XML retrieval evaluation. In *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, pages 43–57, Dagstuhl Castle, Germany, 2006. Revised Selected Papers.

[1254] D. A. Pereira, B. A. Ribeiro-Neto, N. Ziviani, and A. H. F. Laender. Using Web information for creating publication venue authority files. In *JCDL*, pages 295–304, 2008.

[1255] D. A. Pereira, B. A. Ribeiro-Neto, N. Ziviani, A. H. F. Laender, M. A. Gonçalves, and A. A. Ferreira. Using Web information for author name disambiguation. In *JCDL*, pages 49–58, 2009.

[1256] A. Perkins. The classification of search engine spam. Available online at http://www.silverdisc.co.uk/articles/spam-classification/, September 2001.

[1257] M. Persin. Document filtering for fast ranking. In *Proc. of the 17th ACM SIGIR Conference*. Springer Verlag, 1994.

[1258] M. Persin, J. Zobel, and R. Sacks-Davis. Filtered document retrieval with frequency-sorted indexes. *Journal of the American Society for Information Science*, 47(10):749–764, Oct. 1996.

[1259] S. Perugini, M. A. Gonçalves, and E. A. Fox. Recommender systems research: A connection-centric survey. *Journal of Intelligent Information Systems*, 23(2):107–143, 2004.

[1260] S. Perugini, K. McDevitt, R. Richardson, M. Perez-Quinones, R. Shen, N. Ramakrishnan, C. Williams, and E. A. Fox. Enhancing Usability in CITIDEL: Multimodal, Multilingual, and Interactive Visualization Interfaces. In *Proc. of the 4th Joint Conf. on Digital Libraries (JCDL'2004)*, pages 315–324, Tucson, Arizona, June 7-11, 2004.

[1261] N. Pharo and A. Trotman. The use case track at INEX 2006. *SIGIR Forum*, 41(1):64–66, 2007.

[1262] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos. Web usage mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction*, 13(4):311–372, 2003.

[1263] D. Pimienta. Languages, culture, and Internet (in French). http://funredes.org/, March 1998.

[1264] K. Pinel-Sauvagnat. Propagation-based structured text retrieval. In *Encyclopedia of Database Systems*. Springer, 2009.

[1265] B. Pinkerton. `comp.infosystems.announce` newsgroup, June 1994.

[1266] B. Pinkerton. Finding what people want: Experiences with the WebCrawler. In *Proceedings of the first World Wide Web Conference*, Geneva, Switzerland, May 1994.

[1267] S. Piontek and K. Garlock. Creating a World Wide Web resource collection. *Internet Research: Electronic Networking Applications and Policy*, 6(4):20–26, 1996.

[1268] P. Pirolli. Computational models of information scent-following in a very large browsable text collection. In *CHI*, pages 3–10, 1997.

[1269] P. Pirolli. *Information Foraging Theory*. Oxford University Press, 2007.

[1270] P. Pirolli and S. Card. Information foraging models of browsers for very large document spaces. In *Advanced Visual Interfaces*, L'Aquila, Italy, May 1998.

[1271] P. Pirolli and S. Card. Information foraging. *Psychological Review*, 106(4):643–675, 1999.

[1272] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of the 2005 International Conference on Intelligence Analysis*, McClean, VA, May 2005.

[1273] P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear: Extracting usable structures from the Web. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 118–125, Zurich, Switzerland, May 1996. ACM Press.

[1274] J. Pitkow and K. Bharat. WebViz: A tools for World Wide Web access log analysis. In *Proc. of the First International World Wide Web Conference*, Geneva, Switzerland, May 1994. http://www1.cern.ch/PapersWWW94/pitkow-webvis.ps.

[1275] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *Commun. ACM*, 45(9):50–55, 2002.

[1276] B. Piwowarski and G. Dupret. Evaluation in (XML) information retrieval: expected precision-recall with user modelling (EPRUM). In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA*, pages 260–267, 2006.

[1277] B. Piwowarski, G. Dupret, and R. Jones. Mining user Web search activity with layered Bayesian networks or how to capture a click in its context. In R. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, and B. B. Cambazoglu, editors, *WSDM*, pages 162–171, Barcelona, Spain, February 2009. ACM.

[1278] B. Piwowarski and M. Lalmas. Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *12th ACM international conference on Information and knowledge management, Washington, DC, USA*, pages 361–370, 2004.

[1279] B. Piwowarski, A. Trotman, and M. Lalmas. Sound and complete relevance assessments for XML retrieval. *ACM Transactions in Information Systems*, 27(1), 2008.

[1280] B. Piwowarski and H. Zaragoza. Predictive user click models based on click-through history. In M. J. Silva, A. H. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, editors, *CIKM*, pages 175–182, Lisbon, Portugal, November 2007. ACM.

[1281] C. Plaisant, B. Shneiderman, K. Doan, and T. Bruns. Interface and data architecture for query preview in networked information systems. *ACM Transactions on Information Systems (TOIS)*, 17(3):320–341, 1999.

[1282] B. Poblete and R. Baeza-Yates. Query-sets: using implicit feedback and query patterns to organize Web documents. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, editors, *Proceedings of the 17th International Conference on World Wide Web, WWW 2008*, pages 41–50, Beijing, China, April 2008. ACM Press.

[1283] B. Poblete, M. Spiliopoulou, and R. Baeza-Yates. Website privacy preservation for query log publishing. In *Proceedings of the First SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD'07), Lecture Notes in Computer Science*, volume 4890. Springer, 2008.

[1284] S. Podlipnig and L. Boszormenyi. A survey of Web cache replacement strategies. *ACM Computing Surveys*, 35(4):374–398, 2003.

[1285] I. Podnar, M. Rajman, T. Luu, F. Klemm, and K. Aberer. Scalable Peer-to-Peer Web Retrieval with Highly Discriminative Keys. In *ICDE'07: Proceedings of the 23nd International conference on Data Engineering*, Istambul, Turkey, 2007.

[1286] P.Ogilvie and J. Callan. Hierarchical language models for XML component retrieval. In *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*, pages 224–237, Dagstuhl Castle, Germany, 2005. Revised Selected Papers.

[1287] C. A. Pogue and P. Willet. Use of text signatures for document retrieval in a highly parallel environment. *Parallel Computing*, 4:259–268, 1987.

[1288] A. Pollock and A. Hockley. What's Wrong with Internet Searching. *D-Lib Magazine*, 1997. http://www.dlib.org.

[1289] D. B. Ponceleón and A. Dieberger. Hierarchical brushing in a collection of video data. In *Proceedings of Hawaii International Conference on System Science (HICSS)*, Maui, HI, 2001.

[1290] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, 1998.

[1291] E. Popovici, G. Ménier, and P.-F. Marteau. SIRIUS XML IR System at INEX 2006: Approximate Matching of Structure and Textual Content. In *Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, pages 185–199, Dagstuhl Castle, Germany, 2007. Revised and Selected Papers.

[1292] M. Porter. An algorithm for suffix striping. In K. S. Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 313–316. Morgan Kaufmann Publishers, Inc., 1997.

[1293] B. Pôssas, N. Ziviani, and W. Meira. Enhancing the set-based model using proximity information. In *SPIRE - 9th Int. Symposium on String Processing and Information Retrieval*, pages 104–116, 2002. Lisbon, Portugal.

[1294] B. Pôssas, N. Ziviani, W. Meira, and B. A. Ribeiro-Neto. Set-based model: A new approach to information retrieval. In *25th ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 230–237, 2002. Tampere, Finland.

[1295] B. Pôssas, N. Ziviani, W. Meira, and B. A. Ribeiro-Neto. Set-based vector model: An efficient approach for correlation-based ranking. *ACM Transactions on Information Systems*, 23(4):397–429, 2005.

[1296] B. Pôssas, N. Ziviani, B. A. Ribeiro-Neto, and W. Meira. Processing conjunctive and phrase queries with the set-based model. In *SPIRE - 11th Int. Symposium on String Processing and Information Retrieval*, pages 171–183, 2004. Padova, Italy.

[1297] B. Pôssas, N. Ziviani, B. A. Ribeiro-Neto, and W. Meira Jr. Maximal termsets as a query structuring mechanism. In *CIKM*, pages 287–288, 2005.

[1298] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Automatic recognition of audio-visual speech: Recent progress and challenges. *Proceedings of the IEEE*, 91(9), September 2003.

[1299] A. L. Powell and J. C. French. Growth and server availability of the NCSTRL digital library. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, pages 264–265, San Antonio, Texas, 2000.

[1300] A. L. Powell and J. C. French. Comparing the performance of collection selection algorithms. *ACM Trans. Inf. Syst.*, 21(4):412–456, 2003.

[1301] W. Pratt, M. Hearst, and L. Fagan. A knowledge-based approach to organizing retrieved documents. In *Proceedings of 16th Annual Conference on Artificial Intelligence(AAAI 99)*, Orlando, FL, 1999.

[1302] A. Pretschner and S. Gauch. Ontology based personalized search. In *ICTAI '99: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, page 391, Washington, DC, USA, 1999. IEEE Computer Society.

[1303] P. Proulx, S. Tandon, A. Bodnar, D. Schroh, W. Wright, D. Schroh, R. Harper, and W. Wright. Avian Flu Case Study with nSpace and GeoTime. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST'06)*. IEEE, 2006.

[1304] D. Puppin, F. Silvestri, and D. Laforenza. Query-driven document partitioning and collection selection. In *INFOSCALE 2006: Proceedings of the first International Conference on Scalable Information Systems*, 2006.

[1305] The PURL Team, Persistent Uniform Resource Locator (PURL). http://purl.oclc.org/.

[1306] J. Qin, Y. Zhou, and M. Chau. Building domain-specific Web collections for scientific digital libraries: a meta-search enhanced focused crawling method. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2004, Proceedings*, pages 135–141, Tuscon, AZ, USA, 2004.

[1307] T. Qin, X. D. Zhang, M. F. Tsai, D. S. Wang, T. Y. Liu, and H. Li. Query-level loss functions for information retrieval. *Information Processing & Management*, 44(2):838–855, 2008.

[1308] T. Qin, X.-D. Zhang, D.-S. Wang, T.-Y. Liu, W. Lai, and H. Li. Ranking with multiple hyperplanes. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 279–286, 2007.

[1309] Y. Qiu and H. Frei. Concept based query expansion. In *Proc. ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, PA, USA, 1993.

[1310] Qos project. http://qos.sourceforge.net/, 2007.

[1311] J. Quinlan. Discovering rules by induction from large collections of examples. In *Expert Systems in the Micro Electronic Age*, Edinburgh, UK, 1979. Edinburgh University Press.

[1312] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[1313] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Mateo, CA, 1993.

[1314] T. Radecki. Mathematical model of information retrieval system based on the concept of fuzzy thesaurus. *Information Processing & Management*, 12:313–318, 1976.

[1315] T. Radecki. Mathematical model of time-effective information retrieval system based on the theory of fuzzy sets. *Information Processing & Management*, 13:109–116, 1977.

[1316] T. Radecki. Fuzzy set theoretical approach to document retrieval. *Information Processing & Management*, 15:247–259, 1979.

[1317] T. Radecki. On the inclusiveness of information retrieval systems with documents indexed by weighted descriptors. *Fuzzy Sets and Systems*, 5:159–176, 1981.

[1318] T. Radecki. Trends in research on information retrieval—the potential for improvements in conventional Boolean retrieval systems. *Information Processing & Management*, 24:219–227, 1988.

[1319] F. Radlinski, A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel. Optimizing relevance and revenue in ad search: a query substitution approach. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 403–410, New York, NY, USA, 2008. ACM.

[1320] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 239–248, 2005.

[1321] F. Radlinski and T. Joachims. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In *Conference of the Association for the Advancement of Artificial Intelligence (AAAI)*, pages 1406–1412, 2006.

[1322] F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[1323] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Conference on Information and Knowledge Management (CIKM)*, 2008.

[1324] S. Raghavan and H. Garcia-Molina. Crawling the hidden web. In *Proceedings of the Twenty-seventh International Conference on Very Large Databases (VLDB)*, pages 129–138, Rome, Italy, 2001. Morgan Kaufmann.

[1325] V. Raghavan, P. Bollmann, and G. Jung. Retrieval system evaluation using recall and precision: Problems and answers. In *Proceedings of the 12th ACM SIGIR Conference*, pages 59–68, 1989.

[1326] V. Raghavan, G. Jung, and P. Bollmann. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Office and Information Systems*, 7(3):205–229, 1989.

[1327] V. Raghavan and S. Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Sciences*, 37(5):279–287, 1986.

[1328] V. V. Raghavan, P. Bollmann, and G. S. Jung. Retrieval system evaluation using recall and precision: Problems and answers. In *12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Cambridge, Massachusetts, USA*, pages 59–68, 1989.

[1329] V. V. Raghavan and H. Sever. On the reuse of past optimal queries. *18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 344–350, 1995.

[1330] C. Raiciu, F. Huici, M. Handley, and D. S. Rosenblum. Roar: increasing the flexibility and performance of distributed search. *SIGCOMM Comput. Commun. Rev.*, 39(4):291–302, 2009.

[1331] G. Ramírez. *Structural Features in XML Retrieval*. PhD thesis, University of Amsterdam, 2007.

[1332] G. Ramírez. Processing overlaps. In *Encyclopedia of Database Systems*. Springer, 2009.

[1333] K. H. Randall, R. Stata, J. L. Wiener, and R. G. Wickremesinghe. The link database: Fast access to graphs of the web. In *DCC '02: Proceedings of the Data Compression Conference (DCC '02)*, page 122, Washington, DC, USA, 2002. IEEE Computer Society.

[1334] E. Rasmussen. Clustering algorithms. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 419–442. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.

[1335] Y. Rasolofo, D. Hawking, and J. Savoy. Result merging strategies for a current news metasearcher. *Information Processing and Management*, 39:581–609, 2003. http://david-hawking.net/pubs/rasolofo_ipm03.pdf.

[1336] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker. A scalable content-addressable network. In *SIGCOMM '01: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 161–172, New York, NY, USA, 2001. ACM.

[1337] A. Rauber, G. Widmer, S. Downie, S. Dixon, and D. Bainbridge, editors. *Proceedings International Symposium for Audio Information Retrieval (ISMIR)*. Austrian Computer Society, Vienna, Austria, October 2007.

[1338] J. Ray, R. Dale, R. Moore, V. Reich, W. Underwood, McCray, and A. T. Panel on digital preservation. In *JCDL'02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 365–367, 2002.

[1339] K. Rayner. Eye movements in reading and information processing. *Psychological Bulletin*, 124:372–252, 1998.

[1340] R. Reddy and I. Wladawsky-Berger. Digital Libraries: Universal Access to Human Knowledge - A Report to the President. President's Information Technology Advisory Committee (PITAC), Panel on Digital Libraries. http://www.itrd.gov/pubs/pitac/pitacdl-9feb01.pdf, 2001.

[1341] W. J. Reed. The Pareto, Zipf and Other Power Laws. *Economics Letters*, 74(15-19), 2001.

[1342] H. Reiterer, G. Tullius, and T. Mann. Insyder: a content-based visual-information-seeking system for the web. *International Journal on Digital Libraries*, pages 25–41, Mar 2005.

[1343] P. Resnick and H. R. Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, 1997.

[1344] M. Rettig. Prototyping for Tiny Fingers. *Communications of the Acm*, 37(4), 1994.

[1345] Reuters Corpus Volume 1 (RCV1), 2000. Produced by Reuters Ltd., RCV1 is made available for use in research and development of natural language-processing, information-retrieval or machine learning systems.

[1346] Reuters Corpus Volume 2 (RCV2), 2005. Produced by Reuters Ltd., RCV2 is made available for use in research and development of natural language-processing, information-retrieval or machine learning systems.

[1347] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[1348] P. Reynolds and A. Vahdat. Efficient Peer-to-Peer Keyword Searching. In *Middleware'03: Proceedings of the 4th International Middleware conference*, Rio de Janeiro, Brazil, 2003.

[1349] B. Ribeiro-Neto and R. R. Barbosa. Query performance for tightly coupled distributed digital libraries. In *Proc. 3rd ACM Conference on Digital Libraries*, pages 182–190, Pittsburgh, PA, June 1998. ACM Press, New York.

[1350] B. A. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. de Moura. Impedance coupling in content-targeted advertising. In *SIGIR*, pages 496–503, 2005.

[1351] B. A. Ribeiro-Neto, A. H. Laender, and L. R. de Lima. An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science and Technology*, 52(5):391–401, 2001.

[1352] B. A. Ribeiro-Neto and R. Muntz. A belief network model for IR. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Modeling, pages 253–260, 1996.

[1353] B. A. Ribeiro-Neto and R. R. Muntz. Fuzzy ranking of approximate answers. In *Second Int. Conference on Flexible Query Answering Systems (FQAS)*, pages 41–56, 1996.

[1354] I. E. G. Richardson. *H.264 and MPEG-4 Video Compression: Video Coding for Next-Generation Multimedia*. Wiley and Sons, 2003.

[1355] S. Rieh. Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, 53(2):145–161, 2002.

[1356] S. Y. Rieh and H. I. Xie. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, 42(3):751–768, 2006.

[1357] J. Rissanen and G. G. Langdon. Arithmetic coding. *IBM Journal of Research and Development*, 23:149–162, 1979.

[1358] J. Risson and T. Moors. Survey of research towards robust peer-to-peer networks: search methods. *Comput. Netw.*, 50(17):3485–3521, 2006.

[1359] K. M. Risvik. *Scaling Internet Search Engines: Methods and Analysis*. PhD thesis, Norwegian University of Science and Technology, 2004.

[1360] K. M. Risvik, Y. Aasheim, and M. Lidal. Multi-tier architecture for Web search engines. In *LA-WEB*, pages 132–143, Santiago, Chile, 2003.

[1361] K. M. Risvik and R. Michelsen. Search engines and Web dynamics. *Computer Networks*, 39(3):289–302, June 2002.

[1362] RLG/NARA audit checklist for certifying a trusted digital repository, 2005. http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf.

[1363] P. A. Roberto, R. L. T. Santos, M. A. Gonçalves, and A. H. F. Laender. On RDBMS and workflow support for componentized digital libraries. In *XXI Simpósio Brasileiro de Banco de Dados*, pages 87–101, Florianópolis, Santa Catarina, Brasil, October 2006.

[1364] S. Robertson. The probability ranking principle in IR. *Journal of Documentation*, pages 294–304, 1977.

[1365] S. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Sciences*, 27(3):129–146, 1976.

[1366] S. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In D. K. Harman, editor, *The First Text REtrieval Conference (TREC-1)*, pages 21–30, Gaithersburg, MD, USA, 1993. Dept. of Commerce, National Institute of Standards and Technology.

[1367] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-2. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 21–34, Gaithersburg, MD, USA, 1994. Dept. of Commerce, National Institute of Standards and Technology.

[1368] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 109–128, Gaithersburg, MD, USA, 1995. Dept. of Commerce, National Institute of Standards and Technology.

[1369] S. Robertson and H. Zaragoza. The probabilistic relevance model: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[1370] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49, New York, NY, USA, 2004. ACM.

[1371] S. E. Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, 2004.

[1372] S. E. Robertson and M. M. Hancock-Beaulieu. On the evaluation of IR systems. *Inf. Process. Manage.*, 28(4):457–466, 1992.

[1373] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR*, pages 232–241, 1994.

[1374] S. E. Robertson and S. Walker. On relevance weights with little relevance information. In *ACM SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 16–24, New York, NY, USA, 1997.

[1375] J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall Inc, Englewood Cliffs, New Jersey, 1971.

[1376] K. Rodden, W. Basalaj, D. Sinclair, and K. R. Wood. Does organisation by similarity assist image browsing? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*, pages 190–197, 2001.

[1377] P. Roget. *Roget's II the New Thesaurus*. Houghton Mifflin Company, Boston, USA, 1988.

[1378] H. L. Roitblat. Information retrieval and eDiscovery, 2006. http://www.ediscoveryinstitute.org/pubs/InformationRetrievalandeDiscovery.pdf.

[1379] T. Rölleke, M. Lalmas, G. Kazai, I. Ruthven, and S. Quicker. The accessibility dimension for structured document retrieval. In *Advances in Information Retrieval, 24th BCS-IRSG European Colloquium on IR Research, Glasgow, UK*, pages 284–302, 2002.

[1380] D. Rose, D. Orr, and R. Kantamneni. Summary attributes and perceived search quality. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*, pages 1201–1202. ACM Press New York, NY, USA, 2007.

[1381] D. E. Rose and R. K. Belew. Legal information retrieval a hybrid approach. In *ICAIL '89: Proceedings of the 2nd international conference on Artificial intelligence and law*, pages 138–146, New York, NY, USA, 1989. ACM.

[1382] D. E. Rose and D. Levinson. Understanding user goals in Web search. In *Proc. of the 14th international conference on World Wide Web*, pages 13–19. ACM Press, 2004.

[1383] L. Rosenfeld and M. Hurst. *Search Analytics: Conversations with your customers*. Rosenfeld Media, 2010. http://www.rosenfeldmedia.com/books/searchanalytics/.

[1384] R. Rosenfeld. Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, 88(8), 2000.

[1385] S. Ross. *Introduction to probability models*. Harcourt Academic Press, 2000.

[1386] S. Ross and M. Hedstrom. Preservation research and sustainable digital libraries. *Int. J. on Digital Libraries*, 5(4):317–324, 2005.

[1387] J. Rothenberg. *Using Emulation to Preserve Digital Documents*. Koninklijke Bibliotheek, The Netherlands, Aug. 21 2000.

[1388] N. Roussopoulos, S. Kelley, and F. Vincent. Nearest neighbor queries. In *Proc. of ACM-SIGMOD*, pages 71–79, San Jose, CA, May 1995.

[1389] T. Rowlands, D. Hawking, and R. Sankaranarayana. Workload sampling for enterprise search evaluation. In *Proceedings of ACM SIGIR 2007*, pages 887–888, July 2007. Poster paper. http://es.csiro.au/pubs/rowlandsHS07.pdf.

[1390] J. Rowley. The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, 20(2):108–119, 1994.

[1391] A. I. T. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In *Middleware '01: Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms Heidelberg*, pages 329–350, London, UK, 2001. Springer-Verlag.

[1392] S. Rüger. *Multimedia Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan Claypool, 2009.

[1393] Y. Rui and T. Huang. A novel relevance feedback technique in image retrieval. In *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*, pages 67–70, 1999.

[1394] Y. Rui, T. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in mars. In *Proceedings. of the IEEE International Conference on Image Processing*, volume 2, pages 815–818, 1997.

[1395] Y. Rui, T. Huang, and S. Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, 1998.

[1396] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[1397] M. Ruiz and P. Srinivazan. Hierarchical neural networks for text categorization. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 281–282, 1999.

[1398] RuleQuest Research. C5.0 data mining tool, 2008. http://www.rulequest.com/.

[1399] S. Russel and P. Norvig. *Artificial Intelligence – A Modern Approach*. Prentice-Hall, 2002.

[1400] D. Russell, M. Slaney, Y. Qu, and M. Houston. Being literate with large document collections: Observational studies and cost structure tradeoffs. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, 2006.

[1401] D. Russell, M. Stefik, P. Pirolli, and S. Card. The cost structure of sensemaking. In *Proceedings of the INTERCHI Conference on Human Factors in Computing Systems (CHI'93)*, Conceptual Analysis of Users and Activity, pages 269–276, 1993.

[1402] I. Ruthven and M. Lalmas. A Survey on the Use of Relevance Feedback for Information Access Systems. *The Knowledge Engineering Review*, 18(02):95–145, 2003.

[1403] W. Sachs. An approach to associative retrieval through the theory of fuzzy sets. *Journal of the American Society for Information Sciences*, pages 85–87, 1976.

[1404] K. Sadakane. Compressed text databases with efficient query algorithms based on the compressed suffix array. In *Proc. 11th International Symposium on Algorithms and Computation (ISAAC)*, LNCS v. 1969, pages 410–421, 2000.

[1405] K. Sadakane. New text indexing functionalities of the compressed suffix arrays. *Journal of Algorithms*, 48(2):294–313, 2003.

[1406] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. *World Wide Web Conference*, pages 377–386, 2006.

[1407] Y. Saito and M. Shapiro. Optimistic replication. *ACM Computing Surveys*, 37(1):42–81, March 2005.

[1408] G. Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall Inc, Englewood Cliffs, New Jersey, 1971.

[1409] G. Salton and C. Buckley. Parallel text search methods. *Commun. ACM*, 31(2):202–215, Feb. 1988.

[1410] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[1411] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[1412] G. Salton, E. A. Fox, and H. Wu. Extended Boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, Nov. 1983.

[1413] G. Salton and M. E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, Jan. 1968.

[1414] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York, 1983.

[1415] G. Salton, A. Singhal, C. Buckley, and M. Mitra. Automatic text decomposition using text segments and text themes. In *Proceedings of the the seventh ACM conference on Hypertext*, pages 53–65, 1996.

[1416] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[1417] G. Salton, C. Yang, and C. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Sciences*, 26(1):33–44, 1975.

[1418] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351–372, 1973.

[1419] H. Samet. *Foundations of Multidimensional and Metric Data Structures*. Computer Graphics and Geometric Modeling. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[1420] P. Sanders and F. Transier. Intersection in integer inverted indices. In *ALENEX'07*, pages 71–83, 2007.

[1421] T. Sanders. Personal communication, 1993.

[1422] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[1423] R. L. T. Santos, P. A. Roberto, M. A. Gonçalves, and A. H. F. Laender. Design, implementation, and evaluation of a wizard tool for setting up component-based digital libraries. In *Research and Advanced Technology for Digital Libraries, 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, 2006, Proceedings*, volume 4172, pages 135–146, 2006.

[1424] T. Saracevic. Evaluation of evaluation in information retrieval. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 138–146, 1995.

[1425] T. Saracevic. Digital library evaluation: Toward evolution of concepts. *Library Trends*, 49(2):350–369, 2000.

[1426] K. Sauvagnat, M. Boughanem, and C. Chrisment. Answering content and structure-based queries on XML documents using relevance propagation. *Information Systems*, 31(7):621–635, 2006.

[1427] K. Sauvagnat, L. Hlaoua, and M. Boughanem. XFIRM at INEX 2005: Ad-Hoc and Relevance Feedback Tracks. In *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, pages 88–103, Dagstuhl Castle, Germany, 2006. Revised Selected Papers.

[1428] H. Sawhney, R. Kumar, G. Gendel, J. Bergen, D. Dixon, and V. Paragano. VideoBrushTM: experiences with consumer video mosaicing. In *Fourth IEEE Workshop on Applications of Computer Vision, 1998. WACV '98*, pages 56–62, Oct 1998.

[1429] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[1430] R. E. Schapire. The boosting approach to machine learning: An overview, December 2002. www.cs.princeton.edu/~schapire/boost.html.

[1431] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[1432] R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[1433] R. E. Schapire, Y. Singer, and A. Singhal. Boosting and Rocchio applied to text filtering. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 215–223, Melbourne, Australia, 1998.

[1434] B. R. Schatz. Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science*, 275:327–335, January 1997.

[1435] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *ICASSP '97: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)-Volume 2*, page 1331, Washington, DC, USA, 1997. IEEE Computer Society.

[1436] R. Schenkel and M. Theobald. Structural Feedback for Keyword-Based XML Retrieval. In *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK*, pages 326–337, 2006.

[1437] R. Schenkel and M. Theobald. Integrated DB&IR. In *Encyclopedia of Database Systems*. Springer, 2009.

[1438] T. Schlieder and H. Meuss. Querying and ranking XML documents. *JASIST*, 53(6):489–503, 2002.

[1439] B. Schlkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[1440] F. Schneider. Implementing fault-tolerant services using the state machine approach: A tutorial. *ACM Computing Surveys*, 22(4):299–319, December 1990.

[1441] K. Schneider. How OPACs suck, part 1: Relevance rank (or the lack of it), 2006. Available at http://www.techsource.ala.org/blog/2006/03/how-opacs-suck-part-1-relevance-rank-or-the-lack-of-it.html.

[1442] U. Schonfeld, Z. Bar-Yossef, and I. Keidar. Do not crawl in the DUST: different URLs with similar text. In *WWW'06: Proceedings of the 15th international conference on World Wide Web*, pages 1015–1016, New York, NY, USA, 2006. ACM Press.

[1443] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W.H. Freeman and Company, New York, 1991.

[1444] H. Schutze, D. Hull, and J. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Proc. of the 18th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 229–237, Seattle, WA, 1995.

[1445] E. S. Schwartz and B. Kallick. Generating a canonical prefix encoding. *Communications of the ACM*, 7:166–169, 1964.

[1446] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[1447] F. Sebastiani, A. Sperduti, and N. Valdambrini. An improved boosting algorithm and its application to text categorization. In *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*, pages 78–85, New York, NY, USA, 2000. ACM.

[1448] E. Selberg and O. Etzioni. Multi-service search and comparison using the MetaCrawler. In *Proc. of the Fourth International World Wide Web Conference*, Boston, Dec. 1995. http://www.w3.org/pub/Conferences/WWW4/Papers/169.

[1449] P. Sellers. The theory and computation of evolutionary distances: pattern recognition. *Journal of Algorithms*, 1:359–373, 1980.

[1450] P. Serdyukov, R. Aly, and D. Hiemstra. University of twente at the trec 2008 enterprise track: Using the global web as an expertise evidence source. In *Proceedings of TREC-2008*, 2009. http://trec.nist.gov/pubs/trec17/papers/utwente.ent.rev.pdf.

[1451] M. Á. Serrano, A. G. Maguitman, M. Boguñá, S. Fortunato, and A. Vespignani. Decoding the structure of the www: facts versus sampling biases, 2005.

[1452] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:398–403, 1948.

[1453] J. Shapiro, V. G. Voiskunskii, and V. J. Frants. *Automated Information Retrieval : Theory and Text-Only Methods*. Academic Press, 1997.

[1454] W. Shaw, J. Wood, R. Wood, and H. Tibbo. The cystic fibrosis database: Content and research opportunities. *Library and Informatin Science Research*, 13:347–366, 1991.

[1455] W. Shaw Jr., R. Burgin, and P. Howell. Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing & Management*, 33(1):1–14, 1997.

[1456] W. Shaw Jr., R. Burgin, and P. Howell. Performance standards and evaluations in IR test collections: Vector-space and other retrieval models. *Information Processing & Management*, 33(1):15–36, 1997.

[1457] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Q^2C@UST: our winning solution to query classification in KDDCUP 2005. *SIGKDD Explorations*, 7(2):100–110, 2005.

[1458] R. Shen. *Applying the 5S Framework to Integrating Digital Libraries*. PhD thesis, Virginia Tech, 2005.

[1459] R. Shen, M. A. Gonçalves, W. Fan, and E. A. Fox. Requirements gathering and modeling of domain-specific digital libraries with the 5S framework: An archaeological case study with ETANA. In *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries*, volume 3652, pages 1–12, Vienna, Austria, 2005.

[1460] R. Shen, N. S. Vemuri, W. Fan, and E. A. Fox. What is a successful digital library? In *Research and Advanced Technology for Digital Libraries, 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, 2006, Proceedings*, pages 208–219, 2006.

[1461] R. Shen, N. S. Vemuri, W. Fan, and E. A. Fox. Integration of complex archeology digital libraries: An ETANA-DL experience. *Inf. Syst.*, 33(7-8):699–723, 2008.

[1462] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, 2005.

[1463] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[1464] S. Shi, G. Yang, D. Wang, J. Yu, S. Qu, and M. Chen. Making Peer-to-Peer Keyword Searching Feasible Using Multi-level Partitioning. In *IPTPS'04: Proceedings of the 3rd International workshop on Peer-to-Peer Systems*, La Jolla, CA, USA, 2004.

[1465] N. Shivakumar and H. Garcia-Molina. SCAM: A copy detection mechanism for digital documents. In *Digital Libraries*, 1995.

[1466] N. Shivakumar and H. Garcia-Molina. Building a scalable and accurate copy detection mechanism. In *Proceedings of the 1st ACM International Conference on Digital Libraries*, pages 160–168, Bethesda, MD, USA, March 1996. ACM.

[1467] N. Shivakumar and H. Garcia-Molina. Finding near-replicas of documents and servers on the web. In P. Atzeni, A. O. Mendelzon, and G. Mecca, editors, *WebDB'98*, volume 1590 of *Lecture Notes in Computer Science*, pages 204–212, Valencia, Spain, March 1999. Springer.

[1468] V. Shkapenyuk and T. Suel. Design and implementation of a high-performance distributed Web crawler. In *Proceedings of the 18th International Conference on Data Engineering (ICDE)*, San Jose, California, February 2002. IEEE CS Press.

[1469] D. Shkarin. PPM: One step to practicality. In *Proc. 12th IEEE Data Compression Conference (DCC'02)*, page 202, 2002.

[1470] B. Shneiderman and G. Kearsley. *Hypertext Hands-On! An Introduction to a New Way of Organizing and Accessing Information*. Addison-Wesley Publishing Co., Reading, MA, 1989. includes two disks.

[1471] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. *Proceedings of the 2006 conference Advanced Visual Interfaces (AVI'04), Workshop on Beyond time and errors: novel evaluation methods for information visualization*, pages 1–7, 2006.

[1472] B. Shneiderman, C. Plaisant, M. Cohen, and S. Jacobs. *Designing the user interface: strategies for effective human-computer interaction, 5/E*. Addison Wesley, 2009.

[1473] M. Shokouhi, J. Zobel, F. Scholer, and S. Tahaghoghi. Capturing collection size for distributed non-cooperative retrieval. In *Proceedings of the Annual ACM SIGIR Conference*, Seattle, WA, USA, August 2006. ACM Press.

[1474] M. Shokouhi, J. Zobel, S. M. Tahaghoghi, and F. Scholer. Using query logs to establish vocabularies in distributed information retrieval. *Information Processing and Management*, 43(1), January 2007.

[1475] L. Si, R. Jin, J. Callan, and P. Olgilvie. A language modeling framework for resource selection and results mergin g. In *Proceedings of the Conference on Information Knowledge Management (CIKM)*, 2002.

[1476] B. Sigurbjornsson, J. Kamps, and M. de Rijke. An element-based approach to XML retrieval. In *Proceedings INEX 2003 Workshop*, pages 19–26, 2004.

[1477] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW'08: Proceeding of the 17th International Conference on World Wide Web*, pages 327–336, New York, NY, USA, 2008. ACM.

[1478] I. Silva, B. A. Ribeiro-Neto, P. Calado, E. S. de Moura, and N. Ziviani. Link-based and content-based evidential information in a belief network model. In *SIGIR*, pages 96–103, 2000.

[1479] C. Silverstein, M. Henzinger, M. Hannes, and M. Moricz. Analysis of a very large alta vista query log. In *SIGIR Forum*, pages 6–12, 1999. 33(3).

[1480] F. Silvestri. Sorting out the document identifier assignment problem. In *ECIR*, pages 101–112, 2007.

[1481] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174, 2009.

[1482] F. Silvestri, S. Orlando, and R. Perego. Assigning identifiers to documents to enhance the clustering property of fulltext indexes. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*, pages 305–312, New York, NY, USA, 2004. ACM Press.

[1483] Sindice: The semantic Web index, 2008. `http://sindice.com`.

[1484] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, Zurich, Switzerland, 1996.

[1485] A. Singhal, M. Mitra, and C. Buckley. Learning routing queries in a query zone. In *SIGIR*, pages 25–32. ACM Press, 1997.

[1486] G. Skobeltsyn and K. Aberer. Distributed Cache Table: Efficient Query-Driven Processing of Multi-Term Queries in P2P Networks. In *P2PIR'06: Proceedings of the workshop on Information Retrieval in Peer-to-Peer Networks*, Arlington, VA, USA, 2006.

[1487] G. Skobeltsyn, F. Junqueira, V. Plachouras, and R. Baeza-Yates. ResIn: A Combination of Result Caching and Index Pruning for High-performance Web Search Engines. In *SIGIR'08: Proceedings of the 31st International ACM SIGIR conference on Research and Development in Information Retrieval*, Singapore, 2008.

[1488] G. Skobeltsyn, T. Luu, I. Podnar Žarko, M. Rajman, and K. Aberer. Web Text Retrieval with a P2P Query-Driven Index. In *SIGIR'07: Proceedings of the 30th International ACM SIGIR conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 2007.

[1489] M. Slaney. Mixtures of probability experts for audio retrieval and indexing. In *Proc. 2002 IEEE International Conference on Multimedia and Expo*, volume 1, pages 345–348, 2002.

[1490] M. Slaney and M. Casey. Locality-sensitive hashing for finding nearest neighbors. *IEEE Signal Processing Magazine*, 25(2):128–131, March 2008.

[1491] M. Slaney, D. P. W. Ellis, M. Sandler, M. Goto, and M. Goodwin. *Special Issue on Music Information Retrieval, IEEE Transactions on Audio, Speech and Signal Processing*, volume 16. IEEE, February, 2008.

[1492] M. Slaney and G. McRoberts. BabyEars: A recognition system for affective vocalizations. *Speech Communication*, 39:367–384, 2003.

[1493] M. Slaney, D. Ponceleón, and J. Kaufman. Multimedia edges: Finding hierarchy in all dimensions. In *Proceedings of 9th ACM International Conference on Multimedia*, October 2001.

[1494] M. Slaney and W. White. Similarity based on rating data. In *Proceedings on the International Society of Music-Information Retrieval*, September 2007.

[1495] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[1496] J. Smith, M. Campbell, M. Naphade, A. Natsev, and J. Tesic. Learning and classification of semantic concepts in broadcast video. Technical report, IBM, 2004.

[1497] M. A. Smith and T. Kanade. Video skimming for quick browsing based on audio and image characterization. Technical Report CMU-CS-95-186 School of Computer Science Tech Report, Carnegie Mellon University, 1995.

[1498] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2009.

[1499] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP—Conference on Empirical Methods on Natural Language Processing*, 2008.

[1500] I. Soboroff. Dynamic test collections: measuring search effectiveness on the live web. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 276–283, New York, NY, USA, 2006. ACM.

[1501] D. Soergel. *Indexing Languages and Thesauri: Construction and Maintenance*. Melville Publishing Co., Los Angeles, CA, 1974.

[1502] F. Song and B. Croft. A general language model for information retrieval. In *ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 279–280, 1999.

[1503] R. Song, Z. Luo, J.-R. Wen, Y. Yu, and H.-W. Hon. Identifying ambiguous queries in Web search. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors, *WWW*, pages 1169–1170, Banff, Alberta, Canada, May 2007. ACM.

[1504] K. Sparck Jones. A statistical interpretation of term specificity and its application to retrieval. *Journal of Documentation*, 28(1):11–20, 1972.

[1505] K. Sparck Jones. Index term weighting. *Information Storage and Retrieval*, 9(11):619–633, 1973.

[1506] K. Sparck Jones. Experiments in relevance weighting of search terms. *Information Processing & Management*, 15(13):133–144, 1979.

[1507] K. Sparck Jones. Search term relevance weighting given little relevance information. *Journal of Documentation*, 35(1):30–48, 1979.

[1508] K. Sparck Jones. The Cranfield Tests. In K. S. Jones, editor, *Information Retrieval Experiment*, pages 256–284. Butterworth, 1981.

[1509] K. Sparck Jones and E. O. Barber. What makes an automatic keyword classification effective. *J. of the American Society for Information Sciences*, 22(3):166–175, 1971.

[1510] K. Sparck Jones and P. Willet. *Readings in Information Retrieval*. Morgan Kaufmann Publishers, Inc., 1997.

[1511] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[1512] E. Spertus. ParaSite: Mining structural information on the Web. In *6th Int'l WWW Conference*, Santa Clara, CA, USA, April 1997.

[1513] M. Spiliopoulou and L. Faulstich. WUM - A tool for WWW utilization analysis. In *Workshop on Web Databases*, pages 109–115, Valencia, Spain, March 1998.

[1514] D. Spinellis. The decay and failures of Web references. *Communications of the ACM*, 46(1):71–77, January 2003.

[1515] A. Spink and C. Cole, editors. *New Directions in Cognitive Information Retrieval*, volume 29 of *Information Retrieval*. Springer, Netherlands, 2005.

[1516] A. Spink, H. Greisdorf, and J. Bateman. From Highly Relevant to Not Relevant: Examining Different Regions of Relevance. *Information Processing and Management*, 34(5):599–621, 1998.

[1517] A. Spink and B. J. Jansen. *Web Search: Public Searching of the Web*. Information Science and Knowledge Management. Springer, 2004.

[1518] A. Spink, B. J. Jansen, C. Blakely, and S. Koshman. A study of results overlap and uniqueness among major Web search engines. *Information Processing & Management*, 42(5):1379–1391, September 2006.

[1519] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109, 2002.

[1520] A. Spink, S. Ozmutlu, H. C. Ozmutlu, and B. J. Jansen. U.S. versus European Web searching trends. *SIGIR Forum*, 26(2), 2002.

[1521] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[1522] A. Spink and M. Zimmer. *Web Search: Multidisciplinary Perspectives*. Information Science and Knowledge Management. Springer, 2008.

[1523] J. Spool. *Web Site Usability: A Designer's Guide*. Morgan Kaufmann, 1998.

[1524] J. Spool. Usability beyond common sense, 2002. http://www.bcs-hci.org.uk/talks/Spool/UIE-BeyondCommonSense.pdf.

[1525] S. H. Srinivasan and M. Slaney. A bipartite graph model for associating images and text. In *IJCAI-2007 Workshop on Multimodal Information Retrieval*, 2007.

[1526] P. Srinivasdan. Thesaurus construction. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 161–218. Prentice Hall, 1992.

[1527] R. M. Stallman. Emacs the extensible, customizable self-documenting display editor. *SIGPLAN Not.*, 16(6):147–156, 1981.

[1528] C. Stanfill. Partitioned posting files: A parallel inverted file structure for information retrieval. In *Proc. 13th Inter. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 413–428, Brussels, Belgium, 1990.

[1529] C. Stanfill. Parallel information retrieval algorithms. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval Data Structures & Algorithms*, pages 459–497. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.

[1530] C. Stanfill and B. Kahle. Parallel free-text search on the Connection Machine system. *Commun. ACM*, 29(12):1229–1239, Dec. 1986.

[1531] C. Stanfill, R. Thau, and D. Waltz. A parallel indexed algorithm for information retrieval. In *Proc. 12th Inter. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 88–97, Cambridge, USA, June 1989.

[1532] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 1999.

[1533] J. G. Steiner, C. Neuman, and J. I. Schiller. Kerberos: An authentication service for open network systems. In *Winter 1988 USENIX Conference*, pages 191–201, Dallas, TX, 1988. USENIX Association.

[1534] R. Steinmetz and K. Nahrstedt. *Multimedia – Computing, Communications and Applications*. Prentice Hall, 1996. 854 pages.

[1535] D. Stenmark. Method for intranet search engine evaluations. In *Proceedings of IRIS22*, Department of CS/IS, University of Jyväskylä, Finland, August 1999. http://w3.informatik.gu.se/~dixi/publ/method.pdf.

[1536] E. Stoica, M. Hearst, and M. Richardson. Automating Creation of Hierarchical Faceted Metadata Structures. In *Human Language Technologies: the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 244–251, 2007.

[1537] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for Internet applications. In *SIGCOMM '01: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 149–160, New York, NY, USA, 2001. ACM.

[1538] T. Strzalkowski, editor. *Natural Language Information Retrieval*. Kluwer Academic Publishers, 1999.

[1539] A.-J. Su, D. Choffnes, A. Kuzmanovic, and F. Bustamante. Drafting behind Akamai (travelocity-based detouring). In *Proceedings of the ACM SIGCOMM Conference*, pages 435–446, Pisa, Italy, September 2006.

[1540] Q. Su, D. Pavlov, J. Chow, and W. Baker. Internet-scale collection of human-reviewed data. In *WWW'07: Proc. of the International World Wide Web Conference*, 2007.

[1541] T. Suel, C. Mathur, J.-W. Wu, J. Zhang, A. Delis, M. Kharrazi, X. Long, and K. Shanmugasundaram. ODISSEA: A Peer-to-Peer Architecture for Scalable Web Search and Information Retrieval. In *WebDB'03: Proceedings of the International workshop on Web and Databases*, San Diego, CA, USA, 2003.

[1542] H. Suleman, A. Atkins, M. A. Gonçalves, R. K. France, E. A. Fox, V. Chachra, and M. Crowder. Networked digital library of theses and dissertations: Bridging the gaps for global access - part 1. *D-Lib Magazine*, 7(8), 2001.

[1543] D. Sullivan. Search Engine Watch. http://www.searchenginewatch.com, 1997.

[1544] T. Sumner, M. Khoo, M. Recker, and M. Marlino. Understanding educator perceptions of 'quality' in digital libraries. In *Proc. of JCDL'03*, pages 269–279, 2003.

[1545] D. Sunday. A very fast substring search algorithm. *Communications of the ACM*, 33(8):132–142, Aug. 1990.

[1546] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Random House, 2004.

[1547] A. Sutcliffe and M. Ennis. Towards a cognitive theory of information retrieval. *Interacting with Computers*, 10:321–351, 1998.

[1548] R. Swan and J. Allan. Aspect Windows, 3-D Visualizations, and Indirect Comparisons of Information Retrieval Systems. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'98)*, pages 173–181, 1998.

[1549] Swish++. http://homepage.mac.com/pauljlucas/software/swish/, 2007.

[1550] Swish-e. http://www.swish-e.org/, 2007.

[1551] D. Tabatabai and B. Shore. How experts and novices search the Web. *Library & Information Science Research*, 27(2):222–248, 2005.

[1552] J. Tague-Sutcliffe. Measuring the informativeness of a retrieval process. In *Proc of the Fifteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 23–36, Denmark, 1992.

[1553] V. Tahani. A fuzzy model of document retrieval systems. *Information Processing & Management*, 12:177–187, 1976.

[1554] J. I. Tait, editor. *Charting a New Course: Natural Language Processing and Information Retrieval. Essays in Honour of Karen Spärck Jones*. Springer, 2005.

[1555] H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. *IEEE Transactions on System, Man and Cybernatic*, 6, 1978.

[1556] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 718–723, New York, NY, USA, 2006. ACM.

[1557] P. N. Tan and V. Kumar. Discovery of Web robots session based on their navigational patterns. *Data Mining and Knowledge discovery*, 6(1):9–35, 2002.

[1558] C. Tang, Z. Xu, and S. Dwarkadas. Peer-to-peer information retrieval using self-organizing semantic overlay networks. In *SIGCOMM '03: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 175–186, New York, NY, USA, 2003. ACM.

[1559] C. Tang, Z. Xu, and M. Mahalingam. pSearch: Information retrieval in structured overlays. *SIGCOMM Comput. Commun. Rev.*, 33(1):89–94, 2003.

[1560] Y. Taniguchi. An intuitive and efficient access interface to real-time incoming video based on automatic indexing. In *Proc. ACM Multimedia*, pages 25–33, November 1995.

[1561] Y. Taniguchi, A. Akutsu, and Y. Tonomura. PanoramaExcerpts: Extracting and packing panoramas for video browsing. In *MULTIMEDIA '97: Proceedings of the Fifth ACM International Conference on Multimedia*, pages 427–436, New York, NY, USA, Nov 1997. ACM.

[1562] R. Tansley, M. Bass, D. Stuve, M. Branschofsky, D. Chudnov, G. McClellan, and M. Smith. DSpace: An institutional digital repository system. In *Proc. of the 3rd Joint Conference on Digital Libraries*, pages 87–97, Houston, Texas, 2003.

[1563] T. Tao and C. Zhai. An exploration of proximity measures in information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'07)*, pages 295–302. ACM Press, 2007.

[1564] C. M. Taskiran, Z. Pizlo, A. Amir, D. B. Ponceleón, and E. J. Delp. Automated video program summarization using speech transcripts. *IEEE Transactions in Multimedia*, 8(4):775–791, 2006.

[1565] S. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. A simple conceptual model for the internet topology. In *Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE*, volume 3, pages 1667–1671, 2001.

[1566] M. Taylor, J. Guiver, S. Robertson, and T. Minka. Softrank: optimizing non-smooth rank metrics. In *WSDM '08: Proceedings of the international conference on Web search and Web data mining*, pages 77–86, Palo Alto, California, USA, 2008. ACM Press.

[1567] E. S. Team. Eprints services, 2006. http://www.eprints.org/services/.

[1568] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information Re-retrieval: Repeat Queries in Yahoo's Logs. In *SIGIR'07: Proceedings of the 30th International ACM SIGIR conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 2007.

[1569] J. Teevan, C. Alvarado, M. Ackerman, and D. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'04)*, pages 415–422, 2004.

[1570] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of ACM SIGIR '05*, pages 449–456, New York, NY, USA, 2005. ACM.

[1571] J. Teevan, S. T. Dumais, and E. Horvitz. Characterizing the value of personalizing search. In *Proceedings of ACM SIGIR '07*, pages 757–758, New York, NY, USA, 2007. ACM.

[1572] TEI. A gentle introduction to SGML. Technical report, Text Encoding Initiative, 1996. http://www.sil.org/sgml/gentle.html.

[1573] L. Teodosio and W. Bender. Salient stills. *ACM Trans. Multimedia Comput. Commun. Appl.*, 1(1):16–36, 2005.

[1574] Terrier. http://ir.dcs.gla.ac.uk/terrier/, 2007.

[1575] M. Thelwall. *Link Analysis: An Information Science Approach*. Academic Press, December 2004.

[1576] M. Thelwall and D. Wilkinson. Graph structure in three national academic webs: Power laws with anomalies. *Journal of the American Society for Information Science and Technology*, 54(8):706–712, 2003.

[1577] A. Theobald and G. Weikum. The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking. In *EDBT*, pages 477–495, 2002.

[1578] M. Theobald, H. Bast, D. Majumdar, R., and G. Weikum. TopX: efficient and versatile top- query processing for semistructured data. *VLDB Journal*, 17(1):81–115, 2008.

[1579] M. Theobald, R. Schenkel, and G. Weikum. TopX and XXL at INEX 2005. In *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, pages 282–295, Dagstuhl Castle, Germany, 2006. Revised Selected Papers.

[1580] P. Thomas. *Server characterisation and selection for personal metasearch*. PhD thesis, Australian National University, 2008. http://es.csiro.au/pubs/thomas_thesis.pdf.

[1581] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *ACM Int Conference on Information and Knowledge Management (CIKM)*, pages 94–101, 2006.

[1582] P. Thomas and D. Hawking. Evaluating sampling methods for uncooperative collections. In *Proceedings of ACM SIGIR 2007*, pages 503–510, July 2007. http://david-hawking.net/pubs/fp347-thomas.pdf.

[1583] P. Thomas and D. Hawking. Experiences evaluating personal metasearch. In *Proceedings of IIiX*, London, 2008. http://es.csiro.au/pubs/thomas_iiix08.pdf.

[1584] K. Thompson. Regular expression search algorithm. *Communications of ACM*, 11:419–422, 1968.

[1585] K. M. Ting and I. H. Witten. Stacked generalizations: When does it work? In *IJCAI (2)*, pages 866–873, 1997.

[1586] H. Tirri. Search in vain: Challenges for Internet search. *Computer*, 36(1):115–116, 2003.

[1587] TodoCL, 2000. http://www.todocl.com.

[1588] A. Tomasic and H. García-Molina. Caching and database scaling in distributed shared-nothing information retrieval systems. In *Proc. of the ACM SIGMOD Inter. Conf. on Management of Data*, pages 129–138, Washington, D.C., USA, May 1993.

[1589] A. Tomasic and H. Garcia-Molina. Performance of inverted indices in shared-nothing distributed text document information retrieval systems. In *Proceedings of the second international conference on Parallel and distributed information systems*, pages 8–17, San Diego, California, United States, 1993. IEEE Computer Society Press.

[1590] A. Tomasic and H. García-Molina. Performance issues in distributed shared-nothing information retrieval systems. *Inf. Process. & Mgmnt.*, 32(6):647–665, 1996.

[1591] A. Tombros, B. Larsen, and S. Malik. The interactive track at INEX 2004. In *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*, pages 410–423, Dagstuhl Castle, Germany, 2005. Revised Selected Papers.

[1592] A. Tombros, S. Malik, and B. Larsen. Report on the INEX 2004 interactive track. *SIGIR Forum*, 39(1):43–49, 2005.

[1593] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'98)*, pages 2–10, 1998.

[1594] The TREC NIST site, 2008. http://trec.nist.gov.

[1595] A. Trotman. Learning to rank. *Information Retrieval*, 8(3):359–381, 2005.

[1596] A. Trotman. Narrowed Extended XPath I (NEXI). In *Encyclopedia of Database Systems*. Springer, 2009.

[1597] A. Trotman. Processing structural constraints. In *Encyclopedia of Database Systems*. Springer, 2009.

[1598] A. Trotman and S. Geva. Report on the SIGIR 2006 workshop on XML element retrieval methodology. *SIGIR Forum*, 40(2):42–48, 2006.

[1599] A. Trotman, S. Geva, and J. Kamps. Report on the SIGIR 2007 workshop on focused retrieval. *SIGIR Forum*, 41(2):97–103, 2007.

[1600] A. Trotman and M. Lalmas. Report on the INEX 2005 workshop on element retrieval methodology. *SIGIR Forum*, 39(2):46–51, 2005.

[1601] A. Trotman and M. Lalmas. Why structural hints in queries do not help XML retrieval. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA*, pages 711–712, 2006.

[1602] A. Trotman and B. Sigurbjornsson. Narrowed Extended XPath I (NEXI). In *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*, pages 16–40, Dagstuhl Castle, Germany, 2005. Revised Selected Papers.

[1603] M.-F. Tsai, T.-Y. Liu, T. Qin, H.-H. Chen, and W.-Y. Ma. Frank: a ranking method with fidelity loss. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 383–390, Amsterdam, The Netherlands, 2007. ACM Press.

[1604] T. Tsikrika. Aggregation-based Structured Text Retrieval. In *Encyclopedia of Database Systems*. Springer, 2009.

[1605] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Chelshire, CT, 1983.

[1606] E. Tufte. Beautiful Evidence. *Information Design Journal*, 15(2):188–191, 2007.

[1607] D. Tunkelang. *Faceted Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan Claypool, 2009.

[1608] M. Turk. A Random Walk through Eigenspace. *IEICE Transactions on Information and Systems*, Vol. E84-D(12):1586–1595, December 2001.

[1609] H. Turtle and W. B. Croft. Inference networks for document retrieval. In *Proceedings of the Thirteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Information Retrieval Models (1), pages 1–24, 1990.

[1610] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.

[1611] H. R. Turtle. *Inference Networks for Document Retrieval*. PhD thesis, University of Massachusetts at Amherst, Department of Computer Science, February 1991.

[1612] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280, 2008.

[1613] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky. Video manga: generating semantically meaningful video summaries. In *MULTIMEDIA '99: Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, pages 383–392, New York, NY, USA, 1999. ACM.

[1614] E. Ukkonen. Finding approximate patterns in strings. *Journal of Algorithms*, 6:132–137, 1985.

[1615] E. Ukkonen. Approximate string matching over suffix trees. In A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, editors, *Proc. of Combinatorial Pattern Matching*, number 684 in LNCS, pages 228–242, Padova, Italy, 1993. Springer-Verlag.

[1616] E. Ukkonen. Constructing suffix trees on-line in linear time. *Algorithmica*, 14(3):249–260, Sep 1995.

[1617] Unicode Consortium, Unicode. http://www.unicode.org/.

[1618] University of California Libraries. Bibliographic Services Task Force. Rethinking how we provide bibliographic services for the university of california. Final report, University of California, December 2005. Available at http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf.

[1619] T. Upstill, N. Craswell, and D. Hawking. Query-independent evidence in home page finding. *ACM Transactions on Information Systems (TOIS)*, 21(3):286–313, 2003. http://es.csiro.au/pubs/upstill_tois03.pdf.

[1620] P. Vakkari. Relevance and Contributing Information Types of Searched Documents in Task Performance. *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'00)*, pages 2–9, 2000.

[1621] H. van de Sompel, J. A. Young, and T. B. Hickey. Using the OAI-PMH ... differently. *D-Lib Magazine*, 9(7/8), July/Aug. 2003.

[1622] T. P. van der Weide, T. W. C. Huibers, and P. van Bommel. The incremental searcher satisfaction model for information retrieval. *The Computer Journal*, 41(5):311–318, 1998.

[1623] A. van Deursen, P. Klint, and J. Visser. Domain-specific languages: An annotated bibliography. *ACM SIGPLAN Notices*, 35(6):26–36, June 2000.

[1624] C. van Rijsbergen. *Information Retrieval*. Butterwords, 1979.

[1625] C. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, August 2004.

[1626] R. van Zwol. B^3-SDR and Effective Use of Structural Hints. In *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, pages 146–160, Dagstuhl Castle, Germany, 2006. Revised Selected Papers.

[1627] R. van Zwol, J. Baas, H. van Oostendorp, and F. Wiering. Bricks: the Building Blocks to Tackle Query Formulation in Structured Document Retrieval. In *Advances in Information Retrieval, 28th European Conference on IR Research*, pages 314–325, London, UK, 2006.

[1628] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, NY, 1998.

[1629] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proceedings of the IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 2007.

[1630] A. Veerasamy and N. Belkin. Evaluation of a tool for visualization of information retrieval results. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'96)*, pages 85–92, 1996.

[1631] L. Veiga e Silva, M. A. Gonçalves, and A. H. F. Laender. Evaluating a digital library self-archiving service: the bdbcomp user case study. *Information Processing & Management*, 43(4), 2007.

[1632] A. Veloso, H. M. de Almeida, M. A. Gonçalves, and W. Meira Jr. Learning to rank at query-time using association rules. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–274, Singapore, July 2008.

[1633] A. Veloso, W. Meira Jr., M. Cristo, M. A. Gonçalves, and M. J. Zaki. Multi-evidence, multi-criteria, lazy associative document classification. In P. S. Yu, V. J. Tsotras, E. A. Fox, and B. Liu, editors, *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006*, pages 218–227. ACM Press, 2006.

[1634] A. Veloso, W. Meira Jr., M. A. Gonçalves, and M. J. Zaki. Multi-label lazy associative classification. In *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007, Proceedings*, pages 605–612. Springer, 2007.

[1635] N. S. Vemuri, R. da Silva Torres, R. Shen, M. A. Gonçalves, W. Fan, and E. A. Fox. A content-based image retrieval service for archaeology collections. In *Research and Advanced Technology for Digital Libraries, 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, 2006, Proceedings*, pages 438–440. Springer, 2006.

[1636] J. Verhoeff, W. Goffmann, and J. Belzer. Inefficiency of the use of Boolean functions for information retrieval systems. *Communications of the ACM*, 4(12):557–558, 594, Dec. 1961.

[1637] F. Viégas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Many Eyes: A Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics*, pages 1121–1128, 2007.

[1638] M. V. Vieira, B. M. Fonseca, R. Damazio, P. B. Golgher, D. C. Reis, and B. A. Ribeiro-Neto. Efficient search ranking in social networks. In *CIKM*, pages 563–572, 2007.

[1639] R. C. Vieira, P. Calado, A. S. da Silva, A. H. F. Laender, and B. A. Ribeiro-Neto. Structuring keyword-based queries for Web databases. In *JCDL*, pages 94–95, 2002.

[1640] C. L. Viles and J. C. French. Dissemination of collection wide information in a distributed information retrieval system. In *Proc. 18th Inter. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 12–20, Seattle, WA, USA, July 1995.

[1641] V. Vinay, I. J. Cox, N. Milic-Frayling, and K. Wood. On ranking the effectiveness of searches. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 398–404, Seattle, Washington, USA, 2006.

[1642] J.-N. Vittaut and P. Gallinari. Machine learning ranking for structured information retrieval. In *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, 2006,*, pages 338–349, 2006.

[1643] Vivísimo, 1996. http://www.vivisimo.com.

[1644] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326, New York, NY, USA, 2004. ACM Press.

[1645] E. Voorhees. *The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval*. PhD thesis, Cornell University, 1986.

[1646] E. Voorhees. The TREC-8 Question Answering track report. In *TREC-8: Proceedings of the Eighth Text Retrieval Conference*, pages 77–82, 2000.

[1647] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.

[1648] E. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 143–170. Springer Verlag / Heidelberg, 2002. Lecture Notes in Computer Science.

[1649] E. Voorhees. Overview of the TREC 2006. In *Proc. of the Fifteenth Text REtrieval Conference*. NIST Special Publication, Gaithersburg, MD, USA, 2006.

[1650] E. Voorhees. Overview of TREC 2007. In *16th Text Retrieval Conference (TREC)*, 2007.

[1651] E. Voorhees and D. Harman. Overview of the sixth text retrieval conference (TREC-6). In E. Voorhees and D. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication, 1997.

[1652] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird. The collection fusion problem. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 95–104, Gaithersburg, MD, USA, 1995. Dept. of Commerce, National Institute of Standards and Technology. Special Publication 500-226.

[1653] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird. Learning collection fusion strategies. In *Proceedings of ACM SIGIR'95*, pages 172–179, 1995.

[1654] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, Mass., USA, 2005.

[1655] W3C. Extensible markup language (XML) 1.0. Technical report, WWW Consortium (W3C), 1998. http://www.w3.org/TR/1998/REC-xml-19980210.

[1656] W3C. HTML 4.0 specification. Technical report, WWW Consortium (W3C), 1998. http://www.w3.org/TR/1998/REC-html40-19980424/.

[1657] W3C. XML linking language (XLink). Technical report, WWW Consortium (W3C), 1998. http://www.w3.org/TR/1998/WD-xlink-19980303.

[1658] W3C. XSL requirements summary. Technical report, WWW Consortium (W3C), 1998. http://www.w3.org/TR/1998/WD-XSLReq-19980511.

[1659] W3C. XML Schema. http://www.w3.org/XML/Schema, 2001.

[1660] W3C. Resource Description Framework (RDF), 2004. http://www.w3.org/RDF.

[1661] W3C. SPARQL, 2008. http://www.w3.org/TR/rdf-sparql-query/.

[1662] A. Waern. User involvement in automatic filtering: An experimental study. *User Modeling and User-Adapted Interaction (UMUAI)*, 14(2–3):201–237, 2001.

[1663] R. Wan. *Browsing and Searching Compressed Documents*. PhD thesis, Department of Computer Science and Software Engineering, University of Melbourne, Melbourne, Australia, 2003.

[1664] A. Wang. An industrial strength audio search algorithm. In *ISMIR*, 2003.

[1665] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, September 2006.

[1666] S. Warner. Eprints and the open archives initiative. *CoRR*, cs.DL/0307008, 2003.

[1667] S. Wartick. Boolean operations. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 264–292. Prentice Hall, 1992.

[1668] D. J. Waters. What are digital libraries. *CLIR issues*, (4), July/August 1998. http://www.clir.org/pubs/issues/issues04.html#dlf.

[1669] M. Wattenberg and B. Fernanda. The Word Tree, an Interactive Visual Concordance. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1221–1228, 2008.

[1670] M. Wattenberg and J. Kriss. Designing for Social Data Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):549–557, 2006.

[1671] A. Waugh, R. Wilkinson, B. Hills, and J. Dell'oro. Preserving digital information forever. In *DL'00: Proceedings of the 5th ACM International Conference on Digital Libraries*, pages 175–184, San Antonio, Texas, 2000.

[1672] J. Weatherley, T. Sumner, M. Khoo, M. Wright, and M. Hoffmann. Partnership reviewing: a cooperative approach for peer review of complex educational resources. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 106–114, Portland, Oregon, 2002.

[1673] W. Webber, A. Moffat, J. Zobel, and R. Baeza-Yates. A Pipelined Architecture for Distributed Text Query Evaluation. *Information Retrieval*, 10(3), 2007.

[1674] Webglimpse. http://www.webglimpse.net/, 2007.

[1675] S. Weibel and E. Miller. Dublin Core Metadata, 1997. http://purl.org/metadata/dublin_core.

[1676] A. Weigend, E. Wiener, and J. Pedersen. Exploiting hierarchy in text categorization. *Information Retrieval*, 1(3):193–216, 1999.

[1677] K. Weinberger, M. Slaney, and R. van Zwol. Resolving tag ambiguity. In *MULTIMEDIA '08: Proceedings of the 16th International Conference on Multimedia*, New York, NY, USA, 2008. ACM.

[1678] P. Weiner. Linear pattern matching algorithms. In *Proc. IEEE Symp. on Switching and Automata Theory*, pages 1–11, 1973.

[1679] R. Weiss, B. Vélez, M. Sheldon, C. Nemprempre, P. Szilagyi, and D. Gifford. HyPursuit: A hierarchical network engine that exploits content-link hypertext clustering. In *7th ACM Conference on Hypertext and Hypermedia*, pages 180–193, Washington, D.C., USA, 1996.

[1680] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *ESANN*, pages 219–224, 1999.

[1681] R. White, M. Bilenko, and S. Cucerzan. Studying the Use of Popular Destinations to Enhance Web Search Interaction. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'07)*, 2007.

[1682] R. White, J. Jose, and I. Ruthven. A task-oriented study on the influencing effects of query-biased summarisation in Web searching. *Information Processing and Management*, 39(5):707–733, 2003.

[1683] R. White, J. Jose, and I. Ruthven. Using Top-Ranking Sentences for Web Search Result Presentation. *Proceedings of the 12th International Conference on World Wide Web (WWW'03)*, 2003.

[1684] R. White and G. Marchionini. Examining the effectiveness of real-time query expansion. *Information Processing and Management*, 43(3), 2007.

[1685] R. White and D. Morris. Investigating the querying and browsing behavior of advanced search engine users. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'07)*, 2007.

[1686] R. W. White and R. A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan Claypool, 2009.

[1687] R. W. White, I. Ruthven, and J. M. Jose. A study of factors affecting the utility of implicit relevance feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 35–42, New York, NY, USA, 2005. ACM.

[1688] M. Whiting and N. Cramer. WebTheme: Understanding Web Information Through Visual Analytics. In *Proceedings of the First International Semantic Web Conference (ISWC'02)*, pages 460–468. Springer-Verlag London, UK, 2002.

[1689] Wikipedia. http://www.wikipedia.org/, 2001.

[1690] Wikipedia, the Free Encyclopedia. The mother of all demos, September 2006. http://en.wikipedia.org/wiki/The_Mother_of_All_Demos.

[1691] Wikipedia, the Free Encyclopedia. Ted Nelson, September 2006. http://en.wikipedia.org/wiki/Ted_Nelson.

[1692] Wikipedia, the Free Encyclopedia. Information retrieval, 2009. http://en.wikipedia.org/wiki/Information_retrieval.

[1693] Wikipedia, the Free Encyclopedia. Pride & Prejudice (film 2005), 2009. http://en.wikipedia.org/wiki/Pride_&_Prejudice_2005_film.

[1694] Wikipedia, the Free Encyclopedia. Pride and Prejudice, 2009. http://en.wikipedia.org/wiki/Pride_and_Prejudice.

[1695] B. M. Wildemuth, G. Marchionini, M. Yang, G. Geisler, T. Wilkens, A. Hughes, and R. Gruss. How fast is too fast? Evaluating fast forward surrogates for digital video. In *JCDL*, pages 221–230, 2003.

[1696] R. Wilkinson. Effective retrieval of structured documents. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 311–317. Springer-Verlag New York, Inc., 1994.

[1697] R. Wilkinson and P. Hingston. Using the cosine measure in a neural network for document retrieval. In *Proc. ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 202–210, 1991.

[1698] J. Williams. *Bots and other Internet beasts*. Prentice Hall, 1996.

[1699] W. Willinger and V. Paxson. Where mathematics meets the Internet. *Notices of the AMS*, 45(8):961–970, 1998.

[1700] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, and A. Schur. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'95)*, pages 51–58. IEEE Computer Society Press, 1995.

[1701] I. Witten, D. Bainbridge, G. Paynter, and S. Boddie. The greenstone plugin architecture. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 285–286, Portland, Oregon, 2002.

[1702] I. Witten, H. Boddie, J. Stefan, D. Bainbridge, and R. J. McNab. Greenstone: A comprehensive open-source digital library software system. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, pages 113–121, San Antonio, Texas, 2000.

[1703] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, New York, 1994.

[1704] I. H. Witten and D. Bainbridge. *How to Build a Digital Library*. Morgan Kaufmann, 2003.

[1705] I. H. Witten, D. Bainbridge, and S. J. Boddie. Power to the people: End-user building of digital library collections. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, Tools for Constructing and Using Digital Libraries, pages 94–103, Roanoke, VA, 2001.

[1706] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, October 1999.

[1707] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edition, 2005.

[1708] I. H. Witten, M. Gori, and T. Numerico. *Web Dragons: Inside the Myths of Search Engine Technology*. Morgan Kaufmann, 2006.

[1709] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes - Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, Inc, San Francisco, CA, second edition, 1999.

[1710] I. H. Witten, R. M. Neal, and J. G. Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540, 1987.

[1711] K. Wittenburg and E. Sigman. Integration of browsing, searching, and filtering in an applet for Web information access. In *Proc. of the ACM Conference on Human Factors in Computing Systems, Late Breaking Track*, Atlanta, GA, USA, 1997. http://www1.acm.org:82/sigs/sigchi/chi97/proceedings/short-talk/kw.htm.

[1712] J. L. Wolf, M. S. Squillante, P. S. Yu, J. Sethuraman, and L. Ozsen. Optimal crawling strategies for Web search engines. In *WWW'02: Proceedings of the 11th international conference on World Wide Web*, pages 136–147, New York, NY, USA, 2002. ACM Press.

[1713] D. Wolfram. A query-level examination of end user searching behaviour on the excite search engine. In *Proceedings of the 28th Annual Conference Canadian Association for Information Science*, 2000.

[1714] D. Wolfram, P. Wang, and J. Zhang. Identifying Web search session patterns using cluster analysis: A comparison of three search environments. *JASIST*, 60(5):896–910, 2009.

[1715] A. Wolman, G. M. Voelker, N. Sharma, N. Cardwell, A. Karlin, and H. Levy. On the scale and performance of cooperative Web proxy caching. *ACM Operating Systems Review*, 34(5):16–31, December 1999.

[1716] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[1717] S. Wong, W. Ziarko, V. Raghavan, and P. Wong. On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems*, 12(2):299–321, 1987.

[1718] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized vector space model in information retrieval. In *Proc. Eighth ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 18–25, New York, 1985.

[1719] A. Woodruff, P. Aoki, E. Brewer, P. Gauthier, and L. Rowe. An investigation of documents from the World Wide Web. In *5th WWW Conf.*, Paris, France, 1996.

[1720] A. Woodruff, A. Faulring, R. Rosenholtz, J. Morrison, and P. Pirolli. Using thumbnails to search the Web. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*, pages 198–205, 2001.

[1721] K. Woods, W. P. Kegelmeyer, and K. W. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transsactions on Pattern Analysis and Machine Intellegence*, 19(4):405–410, 1997.

[1722] H. Wu and G. Salton. The estimation of term relevance weights using relevance feedback. *Journal of Documentation*, 37(4):194–214, 1981.

[1723] S. Wu and U. Manber. Agrep – a fast approximate pattern-matching tool. In *Proc. of USENIX Technical Conference*, pages 153–162, 1992.

[1724] S. Wu and U. Manber. Fast text searching allowing errors. *Communications of the ACM*, 35(10):83–91, Oct. 1992.

[1725] Xapian code library. http://www.xapian.org/, 2007.

[1726] F. Xia, T. Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: theory and algorithm. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 1192–1199, New York, NY, USA, 2008. ACM Press.

[1727] Y. Xie and D. R. O'Hallaron. Locality in search engine queries and its implications for caching. In *INFOCOM*, 2002.

[1728] L. Xiong and E. Agichtein. Towards privacy preserving query log publishing and analysis. In *Query Log Analysis Workshop, in conjunction with International Conference on World Wide Web (WWW)*, 2007.

[1729] Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, and T. S. Huang. *A Unified Framework for Video Summarization, Browsing and Retrieval: With Applications to Consumer and Surveillance Video*. Elsevier, Amsterdam, 2006.

[1730] J. Xu and J. P. Callan. Effective retrieval with distributed collections. In *SIGIR*, pages 112–120, Melbourne, Australia, August 1998. ACM.

[1731] J. Xu and B. Croft. Cluster-based Language Models for Distributed Retrieval. In *SIGIR'99: Proceedings of the 22nd International ACM SIGIR conference on Research and Development in Information Retrieval*, Berkeley, CA, USA, 1999.

[1732] J. Xu and W. Croft. Query expansion using local and global document analysis. In *Proc. ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Zurich, Switzerland, 1996.

[1733] J. Xu and H. Li. Adarank: a boosting algorithm for information retrieval. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 391–398, New York, NY, USA, 2007. ACM Press.

[1734] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing Web search using web click-through data. In *Proceedings of ACM CIKM '04*, pages 118–126, 2004.

[1735] Yahoo. Searchmonkey, 2008. http://developer.yahoo.com/searchmonkey/.

[1736] Yahoo! Search tips. http://help.yahoo.com/l/us/yahoo/search/basics/basics-04.html, 2009.

[1737] Yahoo! directory: http://search.yahoo.com/dir, 2009.

[1738] H. Yan, S. Ding, and T. Suel. Compressing term positions in Web indexes. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *SIGIR*, pages 147–154, Boston, MA, USA, July 2009. ACM.

[1739] H. Yan, S. Ding, and T. Suel. Inverted index compression and query processing with optimized document ordering. In J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, editors, *WWW*, pages 401–410, Madrid, Spain, April 2009. ACM.

[1740] B. Yang and G. Jeh. Retroactive answering of search queries. In *WWW'06: Proceedings of the 15th international conference on World Wide Web*, pages 457–466, New York, NY, USA, 2006. ACM.

[1741] Y. Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 13–22, Dublin, Ireland, July 1994.

[1742] Y. Yang, T. Ault, and T. Pierce. Combining multiple learning strategies for effective cross validation. In *Proc. 17th International Conf. on Machine Learning*, pages 1167–1174. Morgan Kaufmann, San Francisco, CA, 2000.

[1743] Y. Yang and X. Liu. A re-examination of text categorization methods. In M. A. Hearst, F. Gey, and R. Tong, editors, *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pages 42–49, Berkeley, US, 1999.

[1744] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of the 14th International Conference on Machine Learning – ICML-97*, pages 412–420, Nashville, TN, 1997.

[1745] D. Z. Yazti and M. D. Dikaiakos. Design and implementation of a distributed crawler and filtering processor. In *Proceedings of the fifth Next Generation Information Technologies and Systems (NGITS)*, volume 2382 of *Lecture Notes in Computer Science*, pages 58–74, Caesarea, Israel, June 2002. Springer.

[1746] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'03)*, pages 401–408. ACM New York, NY, USA, 2003.

[1747] B.-L. Yeo and M. M. Yeung. Retrieving and visualizing video. *Commun. ACM*, 40(12):43–52, 1997.

[1748] B.-L. Yeo and M. M. Yeung. Classification, simplification and dynamic visualization of scene transition graphs for video browsing. *IS&T/SPIE Electronic Imaging '98: Storage and Retrieval for Image and Video Databases VI*, pages 60–70, 1998.

[1749] W. Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on Web pages. In L. Carr, D. D. Roure, A. Iyengar, C. A. Goble, and M. Dahlin, editors, *WWW*, pages 213–222, Edinburgh, Scotland, UK, 2006. ACM.

[1750] O. Yilmazel, Finneran, C. M., Liddy, and E. D. Metaextract: an NLP system to automatically assign metadata. In *JCDL'04: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 241–242, 2004.

[1751] O. Yitzhak, N. Golbandi, N. Harel, R. Lempel, A. Neumann, S. Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev. Beyond basic faceted search. In *WSDM '08: Proceedings of the international conference on Web search and Web data mining*, pages 33–44. ACM, 2008.

[1752] E. Yom-Tov, D. Carmel, A. Darlow, D. Pelleg, S. Errera-Yaakov, and S. Fine. Juru at TREC 2005: Query prediction in the terabyte and the robust tracks. In *TREC 2005*, 2005.

[1753] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 512–519, 2005.

[1754] Z. B. Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic transit gloria telae: towards an understanding of the web's decay. In *Proceedings of the 13th conference on World Wide Web*, New York, NY, USA, May 2004. ACM Press.

[1755] D. Young and B. Shneiderman. A graphical filter/flow model for Boolean queries: An implementation and experiment. *Journal of the American Society for Information Science*, 44(6):327–339, July 1993.

[1756] C. T. Yu and G. Salton. Precision weighting–an effective automatic indexing method. *Journal of the ACM*, 23(1):76–88, Jan. 1976.

[1757] H. Yu and M. Young. The impact of Web search engines on subject searching in OPAC. *Information Technology and Libraries*, 23(4):168–180, Dec 2004.

[1758] S. Yu, D. Cai, J. Wen, and W. Ma. Improving pseudo-relevance feedback in Web information retrieval using web page segmentation. In *Proceedings of the 12th international conference on World Wide Web*, pages 11–18, 2003.

[1759] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 271–278, 2007.

[1760] B. Yuwono and D. L. Lee. Search and ranking algorithms for locating resources on the World Wide Web. In *Proceedings of the twelfth International Conference on Data Engineering (ICDE)*, pages 164–171, Washington, DC, USA, February 1996. IEEE CS Press.

[1761] B. Yuwono and D. L. Lee. Server ranking for distributed text retrieval systems on the internet. In *Proceedings of the 5th International Conference on Data Systems for Advanced Applications*, 1997.

[1762] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *MULTIMEDIA '95: Proceedings of the Third ACM International Conference on Multimedia*, pages 189–200, New York, NY, USA, 1995. ACM Press.

[1763] L. Zadeh. Fuzzy sets. In D. Dubois, H. Prade, and R. Yager, editors, *Readings in Fuzzy Sets for Intelligent Systems*. Morgan Kaufmann, 1993.

[1764] O. Zamir and O. Etzioni. Grouper: A Dynamic Clustering Interface to Web Search Results. *Proceedings of the 8th International Conference on World Wide Web (WWW'99)*, 31(11-16):1361–1374, 1999.

[1765] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on Wikipedia. In M. J. Silva, A. H. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, editors, *CIKM*, pages 1015–1018. ACM, November 2007.

[1766] D. Zeinalipour-Yazti, V. Kalogeraki, and D. Gunopulos. Information retrieval techniques for peer-to-peer networks. *Computing in Science and Engg.*, 6(4):20–26, 2004.

[1767] D. Zeinalipour-Yazti, V. Kalogeraki, and D. Gunopulos. Exploiting locality for scalable information retrieval in peer-to-peer networks. *Inf. Syst.*, 30(4):277–298, 2005.

[1768] H. J. Zeng, Q. C. He, Z. Chen, W. Y. Ma, and J. Ma. Learning to cluster Web search results. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 210–217, Sheffield, United Kingdom, 2004. ACM Press.

[1769] Zettair. http://www.seg.rmit.edu.au/zettair/, 2007.

[1770] C. Zhai. *Statistical Language Models for Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan Claypool, 2008.

[1771] C. Zhai. Statistical language models for information retrieval: A critical review. *Foundations and Trends in Information Retrieval*, 2(3):137–213, 2009.

[1772] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM TOIS*, 22:179–214, 2004.

[1773] J. Zhang, X. Long, and T. Suel. Performance of compressed inverted list caching in search engines. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, editors, *WWW*, pages 387–396, Beijing, China, April 2008. ACM.

[1774] J. Zhang and T. Suel. Efficient Query Evaluation on Large Textual Collections in a Peer-to-Peer Environment. In *P2P'05: Proceedings of the 5th International conference on Peer-to-Peer Computing*, Konstanz, Germany, 2005.

[1775] X. Zhang, F. Junqueira, M. Hiltunen, K. Marzullo, and R. Schlichting. Replicating non-deterministic services on grid environments. In *Proceedings of the IEEE International Symposium on High Performance Distributed Computing (HPDC)*, Haifa, Israel, November 2006.

[1776] Y. Zhang. A Comparison on Open Source Search Engine Software. Technical report, School of Information Sciences and Technology, the Pennsylvania State University, April 2002.

[1777] Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *WWW'06: Proceedings of the 15th international conference on World Wide Web*, pages 1039–1040, New York, NY, USA, 2006. ACM.

[1778] Z. Zhang and R. Zhang, editors. *Multimedia Data Mining: A Systematic Introduction to Concepts and Theory*. Data Mining and Knowledge Discovery. Chapman & Hall/CRC, 2008.

[1779] B. Y. Zhao, J. D. Kubiatowicz, and A. D. Joseph. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Technical report, University of California at Berkeley, Berkeley, CA, USA, 2001.

[1780] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Advances in Information Retrieval: 29th European Conference on IR Research*, pages 52–64, 2008.

[1781] E. Zheleva and L. Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *WWW 2009*, 2009.

[1782] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun. A general boosting method and its application to learning ranking functions for web search. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*, Vancouver, BC, Canada, December 2007. MIT Press.

[1783] B. Zhou and J. H. L. Hansen. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. In *ICSLP-2000: International Conference on Spoken Language Processing*, pages 714–717, Beijing, China, October 2000.

[1784] X. Zhou and T. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6), 2003.

[1785] Y. Zhou and W. B. Croft. Query performance prediction in Web search environments. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 543–550, 2007.

[1786] Q. Zhu, M. A. Gonçalves, R. Shen, L. Cassell, and E. A. Fox. Visual semantic modeling of digital libraries. In *Proc. 7th European Conf. Research and Advanced Technology for Digital Libraries, ECDL*, number 2769 in LNCS, Trondheim, Norway, Aug. 2003. Springer.

[1787] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[1788] Y. Zhu. Enhancing search performance on gnutella-like P2P systems. *IEEE Trans. Parallel Distrib. Syst.*, 17(12):1482–1495, 2006. Senior Member-Hu, Yiming.

[1789] Y. Zhu and Y. Hu. Efficient semantic search on dht overlays. *J. Parallel Distrib. Comput.*, 67(5):604–616, 2007.

[1790] Z. Zhuang, R. Wagle, and C. L. Giles. What's there and what's not?: focused crawling for missing documents in digital libraries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 301–310, Denver, Colorado, 2005.

[1791] S. T. Ziliak and D. N. McCloskey. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives (Economics, Cognition, and Society)*. Univ. of Michigan, 2008.

[1792] C. Zimmer, C. Tryfonopoulos, and G. Weikum. Exploiting correlated keywords to improve approximate information filtering. In *SIGIR'08: Proceedings of the 31st International ACM SIGIR conference on Research and Development in Information Retrieval*, Singapore, 2008.

[1793] G. Zipf. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, 1932. Cambridge, MA, USA.

[1794] G. K. Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA, USA, 1949.

[1795] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

[1796] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

[1797] J. Zobel. Collection selection via lexicon inspection. In *Proceedings of the Second Australian Document Computing Symposium*, 1997.

[1798] J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Computing Surveys*, 38(2):1–56, 2006.

[1799] J. Zobel, A. Moffat, and K. Ramamohanarao. Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems*, 23(4):453–490, 1998.

[1800] A. Y. Zomaya, editor. *Parallel and Distributed Computing Handbook*. McGraw-Hill, New York, 1996.