

Bibliography

We use the following abbreviated journal and conference names in the bibliography:

CACM Communications of the Association for Computing Machinery.

IP&M Information Processing and Management.

IR Information Retrieval.

JACM Journal of the Association for Computing Machinery.

JASIS Journal of the American Society for Information Science.

JASIST Journal of the American Society for Information Science and Technology.

JMLR Journal of Machine Learning Research.

TOIS ACM Transactions on Information Systems.

Proc. ACL Proceedings of the Annual Meeting of the Association for Computational Linguistics. Available from: <http://www.aclweb.org/anthology-index/>

Proc. CIKM Proceedings of the ACM CIKM Conference on Information and Knowledge Management. ACM Press.

Proc. ECIR Proceedings of the European Conference on Information Retrieval.

Proc. ECML Proceedings of the European Conference on Machine Learning.

Proc. ICML Proceedings of the International Conference on Machine Learning.

Proc. IJCAI Proceedings of the International Joint Conference on Artificial Intelligence.

Proc. INEX Proceedings of the Initiative for the Evaluation of XML Retrieval.

Proc. KDD Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Proc. NIPS Proceedings of the Neural Information Processing Systems Conference.

Proc. PODS Proceedings of the ACM Conference on Principles of Database Systems.

Proc. SDAIR Proceedings of the Annual Symposium on Document Analysis and Information Retrieval.

- Proc. SIGIR* Proceedings of the Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. Available from: <http://www.sigir.org/proceedings/Proc-Browse.html>
- Proc. SPIRE* Proceedings of the Symposium on String Processing and Information Retrieval.
- Proc. TREC* Proceedings of the Text Retrieval Conference.
- Proc. UAI* Proceedings of the Conference on Uncertainty in Artificial Intelligence.
- Proc. VLDB* Proceedings of the Very Large Data Bases Conference.
- Proc. WWW* Proceedings of the International World Wide Web Conference.
- Aberer, Karl. 2001. P-Grid: A self-organizing access structure for P2P information systems. In *Proc. International Conference on Cooperative Information Systems*, pp. 179–194. Springer. [xxxiv](#), [521](#)
- Aizerman, Mark A., Emmanuel M. Braverman, and Lev I. Rozonoér. 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control* 25:821–837. [347](#), [521](#), [522](#), [532](#)
- Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE Transactions on automatic control* 19(6):716–723. [373](#), [521](#)
- Allan, James. 2005. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proc. TREC*. [174](#), [521](#)
- Allan, James, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proc. SIGIR*, pp. 37–45. ACM Press. DOI: [doi.acm.org/10.1145/290941.290954](https://doi.org/10.1145/290941.290954). [399](#), [521](#), [528](#), [530](#)
- Allwein, Erin L., Robert E. Schapire, and Yoram Singer. 2000. Reducing multiclass to binary: A unifying approach for margin classifiers. *JMLR* 1:113–141. URL: www.jmlr.org/papers/volume1/allwein00a/allwein00a.pdf. [315](#), [521](#), [532](#), [533](#)
- Alonso, Omar, Sandeepan Banerjee, and Mark Drake. 2006. GIO: A semantic web application using the information grid framework. In *Proc. WWW*, pp. 857–858. ACM Press. DOI: [doi.acm.org/10.1145/1135777.1135913](https://doi.org/10.1145/1135777.1135913). [373](#), [521](#), [524](#)
- Altingövde, Ismail Sengör, Engin Demir, Fazli Can, and Özgür Ulusoy. 2008. Incremental cluster-based retrieval using compressed cluster-skipping inverted files. *TOIS*. To appear. [372](#)
- Altingövde, Ismail Sengör, Rifat Ozcan, Huseyin Cagdas Ocalan, Fazli Can, and Özgür Ulusoy. 2007. Large-scale cluster-based retrieval experiments on Turkish texts. In *Proc. SIGIR*, pp. 891–892. ACM Press. [521](#), [523](#), [530](#), [534](#)
- Amer-Yahia, Sihem, Chavdar Botev, Jochen Dörre, and Jayavel Shanmugasundaram. 2006. XQuery full-text extensions explained. *IBM Systems Journal* 45(2):335–352. [217](#), [521](#), [522](#), [524](#), [532](#)
- Amer-Yahia, Sihem, Pat Case, Thomas Rölleke, Jayavel Shanmugasundaram, and Gerhard Weikum. 2005. Report on the DB/IR panel at SIGMOD 2005. *SIGMOD Record* 34(4):71–74. DOI: [doi.acm.org/10.1145/1107499.1107514](https://doi.org/10.1145/1107499.1107514). [217](#), [521](#), [523](#), [532](#), [534](#)

- Amer-Yahia, Sihem, and Mounia Lalmas. 2006. XML search: Languages, INEX and scoring. *SIGMOD Record* 35(4):16–23. DOI: doi.acm.org/10.1145/1228268.1228271. 217, 521, 528
- Anagnostopoulos, Aris, Andrei Z. Broder, and Kunal Punera. 2006. Effective and efficient classification on a search-engine model. In *Proc. CIKM*, pp. 208–217. ACM Press. DOI: doi.acm.org/10.1145/1183614.1183648. 315, 521, 522, 531
- Anderberg, Michael R. 1973. *Cluster analysis for applications*. Academic Press. 372, 521
- Andoni, Alexandr, Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab Mirrokni. 2006. Locality-sensitive hashing using stable distributions. In *Nearest Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press. 314, 521, 524, 526, 529
- Anh, Vo Ngoc, Owen de Kretser, and Alistair Moffat. 2001. Vector-space ranking with effective early termination. In *Proc. SIGIR*, pp. 35–42. ACM Press. 149, 521, 528, 529
- Anh, Vo Ngoc, and Alistair Moffat. 2005. Inverted index compression using word-aligned binary codes. *IR* 8(1):151–166. DOI: dx.doi.org/10.1023/B:INRT.0000048490.99518.5c. 106, 521, 529
- Anh, Vo Ngoc, and Alistair Moffat. 2006a. Improved word-aligned binary compression for text indexing. *IEEE Transactions on Knowledge and Data Engineering* 18(6): 857–861. 106, 521, 529
- Anh, Vo Ngoc, and Alistair Moffat. 2006b. Pruned query evaluation using pre-computed impacts. In *Proc. SIGIR*, pp. 372–379. ACM Press. DOI: doi.acm.org/10.1145/1148170.1148235. 149, 521, 529
- Anh, Vo Ngoc, and Alistair Moffat. 2006c. Structured index organizations for high-throughput text querying. In *Proc. SPIRE*, pp. 304–315. Springer. 149, 521, 529
- Apté, Chidanand, Fred Damerau, and Sholom M. Weiss. 1994. Automated learning of decision rules for text categorization. *TOIS* 12(1):233–251. 286, 521, 524, 534
- Arthur, David, and Sergei Vassilvitskii. 2006. How slow is the k -means method? In *Proc. ACM Symposium on Computational Geometry*, pp. 144–153. 373, 521, 534
- Arvola, Paavo, Marko Junkkari, and Jaana Kekäläinen. 2005. Generalized contextualization method for XML information retrieval. In *Proc. CIKM*, pp. 20–27. 216, 521, 527
- Aslam, Javed A., and Emine Yilmaz. 2005. A geometric interpretation and analysis of R-precision. In *Proc. CIKM*, pp. 664–671. ACM Press. 174, 521, 535
- Ault, Thomas Galen, and Yiming Yang. 2002. Information filtering in TREC-9 and TDT-3: A comparative analysis. *IR* 5(2-3):159–187. 315, 521, 535
- Badue, Claudine Santos, Ricardo A. Baeza-Yates, Berthier Ribeiro-Neto, and Nivio Ziviani. 2001. Distributed query processing using partitioned inverted files. In *Proc. SPIRE*, pp. 10–20. 459, 521, 531, 535
- Baeza-Yates, Ricardo, Paolo Boldi, and Carlos Castillo. 2005. The choice of a damping function for propagating importance in link-based ranking. Technical report, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano. 481, 521, 522, 523

- Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley. xxxiv, 84, 105, 175, 400, 521, 531
- Bahle, Dirk, Hugh E. Williams, and Justin Zobel. 2002. Efficient phrase querying with an auxiliary index. In *Proc. SIGIR*, pp. 215–221. ACM Press. 47, 521, 535
- Baldrige, Jason, and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proc. Empirical Methods in Natural Language Processing*, pp. 9–16. 348, 521, 530
- Ball, G. H. 1965. Data analysis in the social sciences: What about the details? In *Proc. Fall Joint Computer Conference*, pp. 533–560. Spartan Books. 373, 521
- Banko, Michele, and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proc. ACL*. 337, 521, 522
- Bar-Ilan, Judit, and Tatyana Gutman. 2005. How do search engines respond to some non-English queries? *Journal of Information Science* 31(1):13–28. 46, 521, 525
- Bar-Yossef, Ziv, and Maxim Gurevich. 2006. Random sampling from a search engine's index. In *Proc. WWW*, pp. 367–376. ACM Press. DOI: doi.acm.org/10.1145/1135777.1135833. 442, 521, 525
- Barroso, Luiz André, Jeffrey Dean, and Urs Hölzle. 2003. Web search for a planet: The Google cluster architecture. *IEEE Micro* 23(2):22–28. DOI: dx.doi.org/10.1109/MM.2003.1196112. 459, 521, 524, 526
- Bartell, Brian Theodore. 1994. *Optimizing ranking functions: A connectionist approach to adaptive information retrieval*. PhD thesis, University of California at San Diego, La Jolla, CA. 150, 521
- Bartell, Brian T., Garrison W. Cottrell, and Richard K. Belew. 1998. Optimizing similarity using multi-query relevance feedback. *JASIS* 49(8):742–761. 150, 521, 522, 523
- Barzilay, Regina, and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Workshop on Intelligent Scalable Text Summarization*, pp. 10–17. 174, 522, 524
- Bast, Holger, and Debapriyo Majumdar. 2005. Why spectral retrieval works. In *Proc. SIGIR*, pp. 11–18. ACM Press. DOI: doi.acm.org/10.1145/1076034.1076040. 417, 522, 529
- Basu, Sugato, Arindam Banerjee, and Raymond J. Mooney. 2004. Active semi-supervision for pairwise constrained clustering. In *Proc. SIAM International Conference on Data Mining*, pp. 333–344. 373, 521, 522, 530
- Beesley, Kenneth R. 1998. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Languages at Crossroads: Proc. Annual Conference of the American Translators Association*, pp. 47–54. 46, 522
- Beesley, Kenneth R., and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications. 46, 522, 527
- Bennett, Paul N. 2000. Assessing the calibration of naive Bayes' posterior estimates. Technical Report CMU-CS-00-155, School of Computer Science, Carnegie Mellon University. 286, 522

- Berger, Adam, and John Lafferty. 1999. Information retrieval as statistical translation. In *Proc. SIGIR*, pp. 222–229. ACM Press. 251, 252, 522, 528
- Berkhin, Pavel. 2005. A survey on pagerank computing. *Internet Mathematics* 2(1): 73–120. 481, 522
- Berkhin, Pavel. 2006a. Bookmark-coloring algorithm for personalized pagerank computing. *Internet Mathematics* 3(1):41–62. 481, 522
- Berkhin, Pavel. 2006b. A survey of clustering data mining techniques. In Jacob Kogan, Charles Nicholas, and Marc Teboulle (eds.), *Grouping Multidimensional Data: Recent Advances in Clustering*, pp. 25–71. Springer. 372, 522
- Berners-Lee, Tim, Robert Cailliau, Jean-Francois Groff, and Bernd Pollermann. 1992. World-Wide Web: The information universe. *Electronic Networking: Research, Applications and Policy* 1(2):74–82. URL: citeseer.ist.psu.edu/article/berners-lee92worldwide.html. 441, 522, 523, 525, 531
- Berry, Michael, and Paul Young. 1995. Using latent semantic indexing for multilanguage information retrieval. *Computers and the Humanities* 29(6):413–429. 417, 522, 535
- Berry, Michael W., Susan T. Dumais, and Gavin W. O'Brien. 1995. Using linear algebra for intelligent information retrieval. *SIAM Review* 37(4):573–595. 417, 522, 524, 530
- Betsi, Stamatina, Mounia Lalmas, Anastasios Tombros, and Theodora Tsirikla. 2006. User expectations from XML element retrieval. In *Proc. SIGIR*, pp. 611–612. ACM Press. 217, 522, 528, 533, 534
- Bharat, Krishna, and Andrei Broder. 1998. A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems* 30(1-7):379–388. DOI: [dx.doi.org/10.1016/S0169-7552\(98\)00127-5](https://doi.org/10.1016/S0169-7552(98)00127-5). 442, 522
- Bharat, Krishna, Andrei Broder, Monika Henzinger, Puneet Kumar, and Suresh Venkatasubramanian. 1998. The connectivity server: Fast access to linkage information on the web. In *Proc. WWW*, pp. 469–477. 459, 522, 526, 528, 534
- Bharat, Krishna, Andrei Z. Broder, Jeffrey Dean, and Monika Rauch Henzinger. 2000. A comparison of techniques to find mirrored hosts on the WWW. *JASIS* 51(12): 1114–1122. URL: citeseer.ist.psu.edu/bharat99comparison.html. 442, 522, 524, 526
- Bharat, Krishna, and Monika R. Henzinger. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In *Proc. SIGIR*, pp. 104–111. ACM Press. URL: citeseer.ist.psu.edu/bharat98improved.html. 481, 522, 526
- Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer. 315, 522
- Blair, David C., and M. E. Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *CACM* 28(3):289–299. 193, 522, 529
- Blanco, Roi, and Alvaro Barreiro. 2006. TSP and cluster-based solutions to the reassignment of document identifiers. *IR* 9(4):499–517. 106, 521, 522
- Blanco, Roi, and Alvaro Barreiro. 2007. Boosting static pruning of inverted files. In *Proc. SIGIR*. ACM Press. 105, 521, 522

- Blandford, Dan, and Guy Blelloch. 2002. Index compression through document re-ordering. In *Proc. Data Compression Conference*, p. 342. IEEE Computer Society. 106, 522
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR* 3:993–1022. 418, 522, 527, 530
- Boldi, Paolo, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. 2002. Ubi-crawler: A scalable fully distributed web crawler. In *Proc. Australian World Wide Web Conference*. URL: citeseer.ist.psu.edu/article/boldi03ubicrawler.html. 458, 522, 523, 532, 534
- Boldi, Paolo, Massimo Santini, and Sebastiano Vigna. 2005. PageRank as a function of the damping factor. In *Proc. WWW*. URL: citeseer.ist.psu.edu/boldi05pagerank.html. 481, 522, 532, 534
- Boldi, Paolo, and Sebastiano Vigna. 2004a. Codes for the World-Wide Web. *Internet Mathematics* 2(4):405–427. 459, 522, 534
- Boldi, Paolo, and Sebastiano Vigna. 2004b. The WebGraph framework I: Compression techniques. In *Proc. WWW*, pp. 595–601. ACM Press. 459, 522, 534
- Boldi, Paolo, and Sebastiano Vigna. 2005. Compressed perfect embedded skip lists for quick inverted-index lookups. In *Proc. SPIRE*. Springer. 46, 522, 534
- Boley, Daniel. 1998. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery* 2(4):325–344. DOI: [dx.doi.org/10.1023/A:1009740529316](https://doi.org/10.1023/A:1009740529316). 400, 522
- Borodin, Allan, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. 2001. Finding authorities and hubs from link structures on the World Wide Web. In *Proc. WWW*, pp. 415–429. 481, 522, 531, 532, 534
- Bourne, Charles P., and Donald F. Ford. 1961. A study of methods for systematically abbreviating English words and names. *JACM* 8(4):538–552. DOI: [doi.acm.org/10.1145/321088.321094](https://doi.org/10.1145/321088.321094). 65, 522, 525
- Bradley, Paul S., and Usama M. Fayyad. 1998. Refining initial points for K-means clustering. In *Proc. ICML*, pp. 91–99. 373, 522, 524
- Bradley, Paul S., Usama M. Fayyad, and Cory Reina. 1998. Scaling clustering algorithms to large databases. In *Proc. KDD*, pp. 9–15. 374, 522, 524, 531
- Brill, Eric, and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proc. ACL*, pp. 286–293. 65, 522, 530
- Brin, Sergey, and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proc. WWW*, pp. 107–117. 149, 458, 480, 522, 530
- Brisaboa, Nieves R., Antonio Fariña, Gonzalo Navarro, and José R. Paramá. 2007. Lightweight natural language text compression. *IR* 10(1):1–33. 107, 522, 524, 530
- Broder, Andrei. 2002. A taxonomy of web search. *SIGIR Forum* 36(2):3–10. DOI: [doi.acm.org/10.1145/792550.792552](https://doi.org/10.1145/792550.792552). 442, 522
- Broder, Andrei, S. Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the web. *Computer Networks* 33(1):309–320. 441, 522, 528, 529, 531, 533, 534

- Broder, Andrei Z., Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. In *Proc. WWW*, pp. 391–404. 442, 522, 525, 529, 535
- Brown, Eric W. 1995. *Execution Performance Issues in Full-Text Information Retrieval*. PhD thesis, University of Massachusetts, Amherst. 149, 522
- Buckley, Chris, James Allan, and Gerard Salton. 1994a. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In *Proc. TREC*, pp. 45–55. 314, 521, 522, 532
- Buckley, Chris, and Gerard Salton. 1995. Optimization of relevance feedback weights. In *Proc. SIGIR*, pp. 351–357. ACM Press. DOI: doi.acm.org/10.1145/215206.215383. 315, 522, 532
- Buckley, Chris, Gerard Salton, and James Allan. 1994b. The effect of adding relevance information in a relevance feedback environment. In *Proc. SIGIR*, pp. 292–300. ACM Press. 185, 194, 314, 521, 522, 532
- Buckley, Chris, Amit Singhal, and Mandar Mitra. 1995. New retrieval approaches using SMART: TREC 4. In *Proc. TREC*. 187, 522, 529, 533
- Buckley, Chris, and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proc. SIGIR*, pp. 33–40. 173, 174, 522, 534
- Burges, Chris, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proc. ICML*. 348, 522, 524, 525, 526, 528, 531, 532
- Burges, Christopher J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2):121–167. 346, 522
- Burner, Mike. 1997. Crawling towards eternity: Building an archive of the World Wide Web. *Web Techniques Magazine* 2(5). 458, 522
- Burnham, Kenneth P., and David Anderson. 2002. *Model Selection and Multi-Model Inference*. Springer. 373, 521, 523
- Bush, Vannevar. 1945. As we may think. *The Atlantic Monthly*. URL: www.theatlantic.com/doc/194507/bush. 17, 441, 523
- Büttcher, Stefan, and Charles L. A. Clarke. 2005a. Indexing time vs. query time: Trade-offs in dynamic information retrieval systems. In *Proc. CIKM*, pp. 317–318. ACM Press. DOI: doi.acm.org/10.1145/1099554.1099645. 84, 523
- Büttcher, Stefan, and Charles L. A. Clarke. 2005b. A security model for full-text file system search in multi-user environments. In *Proc. FAST*. URL: www.usenix.org/events/fast05/tech/buettcher.html. 84, 523
- Büttcher, Stefan, and Charles L. A. Clarke. 2006. A document-centric approach to static index pruning in text retrieval systems. In *Proc. CIKM*, pp. 182–189. DOI: doi.acm.org/10.1145/1183614.1183644. 105, 523
- Büttcher, Stefan, Charles L. A. Clarke, and Brad Lushman. 2006. Hybrid index maintenance for growing text collections. In *Proc. SIGIR*, pp. 356–363. ACM Press. DOI: doi.acm.org/10.1145/1148170.1148233. 84, 523, 529
- Cacheda, Fidel, Victor Carneiro, Carmen Guerrero, and Ángel Viña. 2003. Optimization of restricted searches in web directories using hybrid data structures. In *Proc. ECIR*, pp. 436–451. 372, 523, 525, 534

- Callan, Jamie. 2000. Distributed information retrieval. In W. Bruce Croft (ed.), *Advances in information retrieval*, pp. 127–150. Kluwer. 84, 523
- Can, Fazli, Ismail Sengör Altingövde, and Engin Demir. 2004. Efficiency and effectiveness of query processing in cluster-based retrieval. *Information Systems* 29(8): 697–717. DOI: [dx.doi.org/10.1016/S0306-4379\(03\)00062-0](https://doi.org/10.1016/S0306-4379(03)00062-0). 372, 521, 523, 524
- Can, Fazli, and Esen A. Ozkarahan. 1990. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. *ACM Trans. Database Syst.* 15(4):483–517. 372, 523, 530
- Cao, Guihong, Jian-Yun Nie, and Jing Bai. 2005. Integrating word relationships into language models. In *Proc. SIGIR*, pp. 298–305. ACM Press. 252, 521, 523, 530
- Cao, Yunbo, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. 2006. Adapting Ranking SVM to document retrieval. In *Proc. SIGIR*. ACM Press. 348, 523, 526, 528, 535
- Carbonell, Jaime, and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. SIGIR*, pp. 335–336. ACM Press. DOI: [doi.acm.org/10.1145/290941.291025](https://doi.org/10.1145/290941.291025). 167, 523, 525
- Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22:249–254. 174, 523
- Carmel, David, Doron Cohen, Ronald Fagin, Eitan Farchi, Michael Herscovici, Yoelle S. Maarek, and Aya Soffer. 2001. Static index pruning for information retrieval systems. In *Proc. SIGIR*, pp. 43–50. ACM Press. DOI: [doi.acm.org/10.1145/383952.383958](https://doi.org/10.1145/383952.383958). 105, 149, 523, 524, 526, 529, 533
- Carmel, David, Yoelle S. Maarek, Matan Mandelbrod, Yosi Mass, and Aya Soffer. 2003. Searching XML documents via XML fragments. In *Proc. SIGIR*, pp. 151–158. ACM Press. DOI: [doi.acm.org/10.1145/860435.860464](https://doi.org/10.1145/860435.860464). 216, 523, 529, 533
- Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proc. ICML*. 347, 523, 530
- Castro, R. M., M. J. Coates, and R. D. Nowak. 2004. Likelihood based hierarchical clustering. *IEEE Transactions in Signal Processing* 52(8):2308–2321. 400, 523, 530
- Cavnar, William B., and John M. Trenkle. 1994. N-gram-based text categorization. In *Proc. SDAIR*, pp. 161–175. 46, 523, 534
- Chakrabarti, Soumen. 2002. *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann. 442, 523
- Chakrabarti, Soumen, Byron Dom, David Gibson, Jon Kleinberg, Prabhakar Raghavan, and Sridhar Rajagopalan. 1998. Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proc. WWW*. URL: cite-seer.ist.psu.edu/chakrabarti98automatic.html. 480, 481, 523, 524, 525, 527, 531
- Chapelle, Olivier, Bernhard Schölkopf, and Alexander Zien (eds.). 2006. *Semi-Supervised Learning*. MIT Press. 347, 501, 508, 523, 535
- Chaudhuri, Surajit, Gautam Das, Vagelis Hristidis, and Gerhard Weikum. 2006. Probabilistic information retrieval approach for ranking of database query results. *ACM Transactions on Database Systems* 31(3):1134–1168. DOI: [doi.acm.org/10.1145/1166074.1166085](https://doi.org/10.1145/1166074.1166085). 217, 523, 524, 526, 534

- Cheeseman, Peter, and John Stutz. 1996. Bayesian classification (AutoClass): Theory and results. In *Advances in Knowledge Discovery and Data Mining*, pp. 153–180. MIT Press. 374, 523, 533
- Chen, Hsin-Hsi, and Chuan-Jie Lin. 2000. A multilingual news summarizer. In *Proc. COLING*, pp. 159–165. 373, 523, 528
- Chen, Pai-Hsuen, Chih-Jen Lin, and Bernhard Schölkopf. 2005. A tutorial on v -support vector machines. *Applied Stochastic Models in Business and Industry* 21: 111–136. 346, 523, 528, 532
- Chiararella, Yves, Philippe Mulhem, and Franck Fourel. 1996. A model for multimedia information retrieval. Technical Report 4-96, University of Glasgow. 216, 523, 525, 530
- Chierichetti, Flavio, Alessandro Panconesi, Prabhakar Raghavan, Mauro Sozio, Alessandro Tiberi, and Eli Upfal. 2007. Finding near neighbors through cluster pruning. In *Proc. PODS*. 149, 523, 530, 531, 533, 534
- Cho, Junghoo, and Hector Garcia-Molina. 2002. Parallel crawlers. In *Proc. WWW*, pp. 124–135. ACM Press. DOI: doi.acm.org/10.1145/511446.511464. 458, 523, 525
- Cho, Junghoo, Hector Garcia-Molina, and Lawrence Page. 1998. Efficient crawling through URL ordering. In *Proc. WWW*, pp. 161–172. 458, 523, 525, 530
- Chu-Carroll, Jennifer, John Prager, Krzysztof Czuba, David Ferrucci, and Pablo Duboue. 2006. Semantic search via XML fragments: A high-precision approach to IR. In *Proc. SIGIR*, pp. 445–452. ACM Press. DOI: doi.acm.org/10.1145/1148170.1148247. 216, 523, 524, 531
- Clarke, Charles L.A., Gordon V. Cormack, and Elizabeth A. Tudhope. 2000. Relevance ranking for one to three term queries. *IP&M* 36:291–311. 149, 523, 534
- Cleverdon, Cyril W. 1991. The significance of the Cranfield tests on index languages. In *Proc. SIGIR*, pp. 3–12. ACM Press. 17, 173, 523
- Coden, Anni R., Eric W. Brown, and Savitha Srinivasan (eds.). 2002. *Information Retrieval Techniques for Speech Applications*. Springer. xxiv, 522, 523, 533
- Cohen, Paul R. 1995. *Empirical methods for artificial intelligence*. MIT Press. 286, 523
- Cohen, William W. 1998. Integration of heterogeneous databases without common domains using queries based on textual similarity. In *Proc. SIGMOD*, pp. 201–212. ACM Press. 217, 523
- Cohen, William W., Robert E. Schapire, and Yoram Singer. 1998. Learning to order things. In *Proc. NIPS*. The MIT Press. URL: cite-seer.ist.psu.edu/article/cohen98learning.html. 150, 523, 532, 533
- Cohen, William W., and Yoram Singer. 1999. Context-sensitive learning methods for text categorization. *TOIS* 17(2):141–173. 339, 523, 533
- Comtet, Louis. 1974. *Advanced Combinatorics*. Reidel. 356, 523
- Cooper, William S., Aitao Chen, and Fredric C. Gey. 1994. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In *Proc. TREC*, pp. 57–66. 150, 523, 525

- Cormen, Thomas H., Charles Eric Leiserson, and Ronald L. Rivest. 1990. *Introduction to Algorithms*. MIT Press. 11, 79, 399, 523, 528, 531
- Cover, Thomas M., and Peter E. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1):21–27. 315, 523, 526
- Cover, Thomas M., and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley. 106, 251, 523, 533
- Crammer, Koby, and Yoram Singer. 2001. On the algorithmic implementation of multiclass kernel-based machines. *JMLR* 2:265–292. 347, 523, 533
- Creecy, Robert H., Brij M. Masand, Stephen J. Smith, and David L. Waltz. 1992. Trading MIPS and memory for knowledge engineering. *CACM* 35(8):48–64. DOI: doi.acm.org/10.1145/135226.135228. 314, 523, 529, 533, 534
- Crestani, Fabio, Mounia Lalmas, Cornelis J. Van Rijsbergen, and Iain Campbell. 1998. Is this document relevant? ... probably: A survey of probabilistic models in information retrieval. *ACM Computing Surveys* 30(4):528–552. DOI: doi.acm.org/10.1145/299917.299920. 235, 523, 528, 531
- Cristianini, Nello, and John Shawe-Taylor. 2000. *Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press. 346, 523, 532
- Croft, W. Bruce. 1978. A file organization for cluster-based retrieval. In *Proc. SIGIR*, pp. 65–82. ACM Press. 372, 523
- Croft, W. Bruce, and David J. Harper. 1979. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation* 35(4):285–295. 133, 227, 523, 526
- Croft, W. Bruce, and John Lafferty (eds.). 2003. *Language Modeling for Information Retrieval*. Springer. 252, 524, 528
- Crouch, Carolyn J. 1988. A cluster-based approach to thesaurus construction. In *Proc. SIGIR*, pp. 309–320. ACM Press. DOI: doi.acm.org/10.1145/62437.62467. 374, 524
- Cucerzan, Silviu, and Eric Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proc. Empirical Methods in Natural Language Processing*. 65, 522, 524
- Cutting, Douglas R., David R. Karger, and Jan O. Pedersen. 1993. Constant interaction-time Scatter/Gather browsing of very large document collections. In *Proc. SIGIR*, pp. 126–134. ACM Press. 399, 524, 527, 530
- Cutting, Douglas R., Jan O. Pedersen, David Karger, and John W. Tukey. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proc. SIGIR*, pp. 318–329. ACM Press. 372, 399, 524, 527, 530, 534
- Damerau, Fred J. 1964. A technique for computer detection and correction of spelling errors. *CACM* 7(3):171–176. DOI: doi.acm.org/10.1145/363958.363994. 65, 524
- Davidson, Ian, and Ashwin Satyanarayana. 2003. Speeding up k-means clustering by bootstrap averaging. In *ICDM 2003 Workshop on Clustering Large Data Sets*. 373, 524, 532

- Day, William H., and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification* 1:1–24. 399, 524
- de Moura, Edleno Silva, Gonzalo Navarro, Nivio Ziviani, and Ricardo Baeza-Yates. 2000. Fast and flexible word searching on compressed text. *TOIS* 18(2):113–139. DOI: doi.acm.org/10.1145/348751.348754. 107, 521, 530, 535
- Dean, Jeffrey, and Sanjay Ghemawat. 2004. MapReduce: Simplified data processing on large clusters. In *Proc. Symposium on Operating System Design and Implementation*. xx, 76, 83, 524, 525
- Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *JASIS* 41(6):391–407. 417, 524, 525, 526, 528
- del Bimbo, Alberto. 1999. *Visual Information Retrieval*. Morgan Kaufmann. xxxiv, 535
- Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* 39: 1–38. 373, 524, 528, 532
- Dhillon, Inderjit S. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. KDD*, pp. 269–274. 374, 400, 524
- Dhillon, Inderjit S., and Dharmendra S. Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine Learning* 42(1/2):143–175. DOI: dx.doi.org/10.1023/A:1007612920971. 373, 524, 529
- Di Eugenio, Barbara, and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics* 30(1):95–101. DOI: dx.doi.org/10.1162/089120104773633402. 174, 524, 525
- Dietterich, Thomas G. 2002. Ensemble learning. In Michael A. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks*, 2nd edition. MIT Press. 347, 524
- Dietterich, Thomas G., and Ghulum Bakiri. 1995. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research* 2: 263–286. 315, 521, 524
- Dom, Byron E. 2002. An information-theoretic external cluster-validity measure. In *Proc. UAI*. 373, 524
- Domingos, Pedro. 2000. A unified bias-variance decomposition for zero-one and squared loss. In *Proc. National Conference on Artificial Intelligence and Proc. Conference Innovative Applications of Artificial Intelligence*, pp. 564–569. AAAI Press / The MIT Press. 315, 524
- Domingos, Pedro, and Michael J. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29(2-3):103–130. URL: citeseer.ist.psu.edu/domingos97optimality.html. 286, 524, 530
- Downie, J. Stephen. 2006. The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine* 12(12). xxxiv, 524
- Duda, Richard O., Peter E. Hart, and David G. Stork. 2000. *Pattern Classification*, 2nd edition. Wiley-Interscience. 286, 372, 524, 526, 533

- Dumais, Susan, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proc. CIKM*, pp. 148–155. ACM Press. DOI: doi.acm.org/10.1145/288627.288651. xvii, 282, 333, 334, 347, 524, 526, 531, 532
- Dumais, Susan T. 1993. Latent semantic indexing (LSI) and TREC-2. In *Proc. TREC*, pp. 105–115. 415, 417, 524
- Dumais, Susan T. 1995. Latent semantic indexing (LSI): TREC-3 report. In *Proc. TREC*, pp. 219–230. 416, 417, 524
- Dumais, Susan T., and Hao Chen. 2000. Hierarchical classification of Web content. In *Proc. SIGIR*, pp. 256–263. ACM Press. 347, 523, 524
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1):61–74. 286, 524
- Dunning, Ted. 1994. Statistical identification of language. Technical Report 94-273, Computing Research Laboratory, New Mexico State University. 46, 524
- Eckart, Carl, and Gale Young. 1936. The approximation of a matrix by another of lower rank. *Psychometrika* 1:211–218. 417, 524, 535
- El-Hamdouchi, Abdelmoula, and Peter Willett. 1986. Hierarchic document classification using Ward's clustering method. In *Proc. SIGIR*, pp. 149–156. ACM Press. DOI: doi.acm.org/10.1145/253168.253200. 399, 524, 534
- Elias, Peter. 1975. Universal code word sets and representations of the integers. *IEEE Transactions on Information Theory* 21(2):194–203. 106, 524
- Eyheramendy, Susana, David Lewis, and David Madigan. 2003. On the Naive Bayes model for text categorization. In *International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics. 286, 524, 528, 529
- Fallows, Deborah. 2004. The internet and daily life. URL: www.pewinternet.org/pdfs/PIP_Internet_and_Daily_Life.pdf. Pew/Internet and American Life Project. xxxi, 524
- Fayyad, Usama M., Cory Reina, and Paul S. Bradley. 1998. Initialization of iterative refinement clustering algorithms. In *Proc. KDD*, pp. 194–198. 374, 522, 524, 531
- Fellbaum, Christiane D. 1998. *WordNet – An Electronic Lexical Database*. MIT Press. 194, 524
- Ferragina, Paolo, and Rossano Venturini. 2007. Compressed permuterm indexes. In *Proc. SIGIR*. ACM Press. 65, 524, 534
- Forman, George. 2004. A pitfall and solution in multi-class feature selection for text classification. In *Proc. ICML*. 286, 525
- Forman, George. 2006. Tackling concept drift by temporal inductive transfer. In *Proc. SIGIR*, pp. 252–259. ACM Press. DOI: doi.acm.org/10.1145/1148170.1148216. 286, 525
- Forman, George, and Ira Cohen. 2004. Learning from little: Comparison of classifiers given little training. In *Proc. PKDD*, pp. 161–172. 336, 523, 525
- Fowlkes, Edward B., and Colin L. Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* 78(383):553–569. URL: www.jstor.org/view/01621459/di985957/98p09261/0. 400, 525, 529

- Fox, Edward A., and Whay C. Lee. 1991. FAST-INV: A fast algorithm for building large inverted files. Technical report, Virginia Polytechnic Institute & State University, Blacksburg, VA, USA. 83, 525, 528
- Fraenkel, Aviezri S., and Shmuel T. Klein. 1985. Novel compression of sparse bit-strings – preliminary report. In *Combinatorial Algorithms on Words, NATO ASI Series Vol F12*, pp. 169–183. Springer. 106, 525, 527
- Frakes, William B., and Ricardo Baeza-Yates (eds.). 1992. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall. 497, 510, 521, 525
- Fraley, Chris, and Adrian E. Raftery. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal* 41(8):578–588. 373, 525, 531
- Friedl, Jeffrey E. F. 2006. *Mastering Regular Expressions*, 3rd edition. O'Reilly. 18, 525
- Friedman, Jerome H. 1997. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* 1(1):55–77. 286, 315, 525
- Friedman, Nir, and Moises Goldszmidt. 1996. Building classifiers using Bayesian networks. In *Proc. National Conference on Artificial Intelligence*, pp. 1277–1284. 231, 525
- Fuhr, Norbert. 1989. Optimum polynomial retrieval functions based on the probability ranking principle. *TOIS* 7(3):183–204. 150, 525
- Fuhr, Norbert. 1992. Probabilistic models in information retrieval. *Computer Journal* 35(3):243–255. 235, 348, 525
- Fuhr, Norbert, Norbert Gövert, Gabriella Kazai, and Mounia Lalmas (eds.). 2003a. *Initiative for the Evaluation of XML Retrieval (INEX)*. *Proc. First INEX Workshop*. ERCIM. 216, 525, 527, 528
- Fuhr, Norbert, and Kai Großjohann. 2004. XIRQL: An XML query language based on information retrieval concepts. *TOIS* 22(2):313–356. URL: doi.acm.org/10.1145/984321.984326. 216, 525
- Fuhr, Norbert, and Mounia Lalmas. 2007. Advances in XML retrieval: The INEX initiative. In *International Workshop on Research Issues in Digital Libraries*. 216, 525, 528
- Fuhr, Norbert, Mounia Lalmas, Saadia Malik, and Gabriella Kazai (eds.). 2006. *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*. Springer. 216, 525, 527, 528, 529
- Fuhr, Norbert, Mounia Lalmas, Saadia Malik, and Zoltán Szilávik (eds.). 2005. *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*. Springer. 216, 509, 516, 525, 528, 529, 533
- Fuhr, Norbert, Mounia Lalmas, and Andrew Trotman (eds.). 2007. *Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*. Springer. 216, 503, 506, 525, 528, 534

- Fuhr, Norbert, Saadia Malik, and Mounia Lalmas (eds.). 2003b. *INEX 2003 Workshop*. URL: inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf. 216, 497, 506, 525, 528, 529
- Fuhr, Norbert, and Ulrich Pfeifer. 1994. Probabilistic information retrieval as a combination of abstraction, inductive learning, and probabilistic assumptions. *TOIS* 12(1):92–115. DOI: doi.acm.org/10.1145/174608.174612. 150, 525, 531
- Fuhr, Norbert, and Thomas Rölleke. 1997. A probabilistic relational algebra for the integration of information retrieval and database systems. *TOIS* 15(1):32–66. DOI: doi.acm.org/10.1145/239041.239045. 217, 525, 532
- Gaertner, Thomas, John W. Lloyd, and Peter A. Flach. 2002. Kernels for structured data. In *Proc. International Conference on Inductive Logic Programming*, pp. 66–83. 347, 524, 525, 529
- Gao, Jianfeng, Mu Li, Chang-Ning Huang, and Andi Wu. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics* 31(4):531–574. 46, 525, 526, 528, 535
- Gao, Jianfeng, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. 2004. Dependence language model for information retrieval. In *Proc. SIGIR*, pp. 170–177. ACM Press. 252, 523, 525, 530, 535
- Garcia, Steven, Hugh E. Williams, and Adam Cannane. 2004. Access-ordered indexes. In *Proc. Australasian Conference on Computer Science*, pp. 7–14. 149, 523, 525, 535
- Garcia-Molina, Hector, Jennifer Widom, and Jeffrey D. Ullman. 1999. *Database System Implementation*. Prentice Hall. 84, 525, 534
- Garfield, Eugene. 1955. Citation indexes to science: A new dimension in documentation through association of ideas. *Science* 122:108–111. 480, 525
- Garfield, Eugene. 1976. The permuted subject index: An autobiographic review. *JASIS* 27(5-6):288–291. 65, 525
- Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Computation* 4(1):1–58. 315, 522, 524, 525
- Geng, Xiubo, Tie-Yan Liu, Tao Qin, and Hang Li. 2007. Feature selection for ranking. In *Proc. SIGIR*, pp. 407–414. ACM Press. 348, 525, 528, 531
- Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication* 12(4). URL: jcmc.indiana.edu/vol12/issue4/gerrand.html. article 8. 30, 525
- Gey, Fredric C. 1994. Inferring probability of relevance using the method of logistic regression. In *Proc. SIGIR*, pp. 222–231. ACM Press. 348, 525
- Ghamrawi, Nadia, and Andrew McCallum. 2005. Collective multi-label classification. In *Proc. CIKM*, pp. 195–200. ACM Press. DOI: doi.acm.org/10.1145/1099554.1099591. 315, 525, 529
- Glover, Eric, David M. Pennock, Steve Lawrence, and Robert Krovetz. 2002a. Inferring hierarchical descriptions. In *Proc. CIKM*, pp. 507–514. ACM Press. DOI: doi.acm.org/10.1145/584792.584876. 400, 525, 528, 531

- Glover, Eric J., Kostas Tsioutsoulis, Steve Lawrence, David M. Pennock, and Gary W. Flake. 2002b. Using web structure for classifying and describing web pages. In *Proc. WWW*, pp. 562–569. ACM Press. DOI: doi.acm.org/10.1145/511446.511520. 400, 524, 525, 528, 531, 534
- Gövert, Norbert, and Gabriella Kazai. 2003. Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In Fuhr et al. (2003b), pp. 1–17. URL: inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf. 216, 525, 527
- Grabs, Torsten, and Hans-Jörg Schek. 2002. Generating vector spaces on-the-fly for flexible XML retrieval. In *XML and Information Retrieval Workshop at SIGIR 2002*. 216, 525, 532
- Greiff, Warren R. 1998. A theory of term weighting based on exploratory data analysis. In *Proc. SIGIR*, pp. 11–19. ACM Press. 227, 525
- Grinstead, Charles M., and J. Laurie Snell. 1997. *Introduction to Probability*, 2nd edition. American Mathematical Society. URL: www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf. 235, 525, 533
- Grossman, David A., and Ophir Frieder. 2004. *Information Retrieval: Algorithms and Heuristics*, 2nd edition. Springer. xxxiv, 84, 217, 525
- Gusfield, Dan. 1997. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press. 65, 525
- Hamerly, Greg, and Charles Elkan. 2003. Learning the k in k -means. In *Proc. NIPS*. URL: books.nips.cc/papers/files/nips16/NIPS2003_AA36.pdf. 373, 524, 525
- Han, Eui-Hong, and George Karypis. 2000. Centroid-based document classification: Analysis and experimental results. In *Proc. PKDD*, pp. 424–431. 314, 526, 527
- Hand, David J. 2006. Classifier technology and the illusion of progress. *Statistical Science* 21:1–14. 286, 526
- Hand, David J., and Keming Yu. 2001. Idiot's Bayes: Not so stupid after all. *International Statistical Review* 69(3):385–398. 286, 526, 535
- Harman, Donna. 1991. How effective is suffixing? *JASIS* 42:7–15. 46, 526
- Harman, Donna. 1992. Relevance feedback revisited. In *Proc. SIGIR*, pp. 1–10. ACM Press. 185, 194, 526
- Harman, Donna, Ricardo Baeza-Yates, Edward Fox, and W. Lee. 1992. Inverted files. In Frakes and Baeza-Yates (1992), pp. 28–43. 83, 521, 525, 526, 528
- Harman, Donna, and Gerald Candela. 1990. Retrieving records from a gigabyte of text on a minicomputer using statistical ranking. *JASIS* 41(8):581–589. 83, 523, 526
- Harold, Elliott Rusty, and Scott W. Means. 2004. *XML in a Nutshell*, 3rd edition. O'Reilly. 216, 526, 529
- Harter, Stephen P. 1998. Variations in relevance assessments and the measurement of retrieval effectiveness. *JASIS* 47:37–49. 174, 526
- Hartigan, J. A., and M. A. Wong. 1979. A K-means clustering algorithm. *Applied Statistics* 28:100–108. 373, 526, 535

- Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. 286, 314, 315, 347, 525, 526, 533
- Hatzivassiloglou, Vasileios, Luis Gravano, and Ankineedu Maganti. 2000. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proc. SIGIR*, pp. 224–231. ACM Press. DOI: doi.acm.org/10.1145/345508.345582. 373, 525, 526, 529
- Haveliwala, Taher. 2003. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering* 15(4): 784–796. URL: citeseer.ist.psu.edu/article/haveliwala03topicsensitive.html. 481, 526
- Haveliwala, Taher H. 2002. Topic-sensitive PageRank. In *Proc. WWW*. URL: citeseer.ist.psu.edu/haveliwala02topicsensitive.html. 481, 526
- Hayes, Philip J., and Steven P. Weinstein. 1990. CONSTRUE/TIS: A system for content-based indexing of a database of news stories. In *Proc. Conference on Innovative Applications of Artificial Intelligence*, pp. 49–66. 335, 526, 534
- Heaps, Harold S. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press. 105, 526
- Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64. 217, 526
- Hearst, Marti A. 2006. Clustering versus faceted categories for information exploration. *CACM* 49(4):59–61. DOI: doi.acm.org/10.1145/1121949.1121983. 372, 526
- Hearst, Marti A., and Jan O. Pedersen. 1996. Reexamining the cluster hypothesis. In *Proc. SIGIR*, pp. 76–84. ACM Press. 372, 526, 530
- Hearst, Marti A., and Christian Plaunt. 1993. Subtopic structuring for full-length document access. In *Proc. SIGIR*, pp. 59–68. ACM Press. DOI: doi.acm.org/10.1145/160688.160695. 217, 526, 531
- Heinz, Steffen, and Justin Zobel. 2003. Efficient single-pass index construction for text databases. *JASIST* 54(8):713–729. DOI: dx.doi.org/10.1002/asi.10268. 83, 526, 535
- Heinz, Steffen, Justin Zobel, and Hugh E. Williams. 2002. Burst tries: A fast, efficient data structure for string keys. *TOIS* 20(2):192–223. DOI: doi.acm.org/10.1145/506309.506312. 84, 526, 535
- Henzinger, Monika R., Allan Heydon, Michael Mitzenmacher, and Marc Najork. 2000. On near-uniform URL sampling. In *Proc. WWW*, pp. 295–308. North-Holland. DOI: dx.doi.org/10.1016/S1389-1286(00)00055-4. 442, 526, 529, 530
- Herbrich, Ralf, Thore Graepel, and Klaus Obermayer. 2000. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pp. 115–132. MIT Press. 348, 525, 526, 530
- Hersh, William, Chris Buckley, T. J. Leone, and David Hickam. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proc. SIGIR*, pp. 192–201. ACM Press. 174, 522, 526, 528
- Hersh, William R., Andrew Turpin, Susan Price, Benjamin Chan, Dale Kraemer, Lynetta Sacherek, and Daniel Olson. 2000a. Do batch and user evaluation give the same results? In *Proc. SIGIR*, pp. 17–24. 175, 523, 526, 528, 530, 531, 532, 534

- Hersh, William R., Andrew Turpin, Susan Price, Dale Kraemer, Daniel Olson, Benjamin Chan, and Lynetta Sacherek. 2001. Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *IP&M* 37(3):383–402. 175, 523, 526, 528, 530, 531, 532, 534
- Hersh, William R., Andrew Turpin, Lynetta Sacherek, Daniel Olson, Susan Price, Benjamin Chan, and Dale Kraemer. 2000b. Further analysis of whether batch and user evaluations give the same results with a question-answering task. In *Proc. TREC*. 175, 523, 526, 528, 530, 531, 532, 534
- Hiemstra, Djoerd. 1998. A linguistically motivated probabilistic model of information retrieval. In *Proc. ECDL*, volume 1513 of *LNCS*, pp. 569–584. 252, 526
- Hiemstra, Djoerd. 2000. A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries* 3(2):131–139. 246, 526
- Hiemstra, Djoerd, and Wessel Kraaij. 2005. A language-modeling approach to TREC. In *Voorhees and Harman (2005)*, pp. 373–395. 252, 526, 528
- Hirai, Jun, Sriram Raghavan, Hector Garcia-Molina, and Andreas Paepcke. 2000. WebBase: A repository of web pages. In *Proc. WWW*, pp. 277–293. 458, 525, 526, 530, 531
- Hofmann, Thomas. 1999a. Probabilistic Latent Semantic Indexing. In *Proc. UAI*. URL: citeseer.ist.psu.edu/hofmann99probabilistic.html. 417, 526
- Hofmann, Thomas. 1999b. Probabilistic Latent Semantic Indexing. In *Proc. SIGIR*, pp. 50–57. ACM Press. URL: citeseer.ist.psu.edu/article/hofmann99probabilistic.html. 417, 526
- Hollink, Vera, Jaap Kamps, Christof Monz, and Maarten de Rijke. 2004. Monolingual document retrieval for European languages. *IR* 7(1):33–52. 46, 526, 527, 530, 531
- Hopcroft, John E., Rajeev Motwani, and Jeffrey D. Ullman. 2000. *Introduction to Automata Theory, Languages, and Computation*, 2nd edition. Addison Wesley. 18, 526, 530, 534
- Huang, Yifen, and Tom M. Mitchell. 2006. Text clustering with extended user feedback. In *Proc. SIGIR*, pp. 413–420. ACM Press. DOI: [doi.acm.org/10.1145/1148170.1148242](https://doi.org/10.1145/1148170.1148242). 374, 526, 529
- Hubert, Lawrence, and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2:193–218. 373, 521, 526
- Hughes, Baden, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. 2006. Reconsidering language identification for written language resources. In *Proc. International Conference on Language Resources and Evaluation*, pp. 485–488. 46, 521, 522, 526, 529, 530
- Hull, David. 1993. Using statistical testing in the evaluation of retrieval performance. In *Proc. SIGIR*, pp. 329–338. ACM Press. 173, 526
- Hull, David. 1996. Stemming algorithms – A case study for detailed evaluation. *JASIS* 47(1):70–84. 46, 526
- Ide, E. 1971. New experiments in relevance feedback. In *Salton (1971b)*, pp. 337–354. 193, 526

- Indyk, Piotr. 2004. Nearest neighbors in high-dimensional spaces. In J. E. Goodman and J. O'Rourke (eds.), *Handbook of Discrete and Computational Geometry*, 2nd edition, pp. 877–892. Chapman and Hall/CRC Press. 314, 526
- Ingwersen, Peter, and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer. xxxiv, 526
- Ittner, David J., David D. Lewis, and David D. Ahn. 1995. Text categorization of low quality images. In *Proc. SDAIR*, pp. 301–315. 314, 521, 526, 528
- Iwayama, Makoto, and Takenobu Tokunaga. 1995. Cluster-based text categorization: A comparison of category search strategies. In *Proc. SIGIR*, pp. 273–280. ACM Press. 314, 526, 533
- Jackson, Peter, and Isabelle Moulinier. 2002. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. John Benjamins. 334, 335, 526, 530
- Jacobs, Paul S., and Lisa F. Rau. 1990. SCISOR: Extracting information from on-line news. *CACM* 33:88–97. 335, 526, 531
- Jain, Anil, M. Narasimha Murty, and Patrick Flynn. 1999. Data clustering: A review. *ACM Computing Surveys* 31(3):264–323. 399, 525, 526, 530
- Jain, Anil K., and Richard C. Dubes. 1988. *Algorithms for Clustering Data*. Prentice Hall. 399, 524, 526
- Jardine, N., and Cornelis Joost van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval* 7:217–240. 372, 527, 531
- Järvelin, Kalervo, and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *TOIS* 20(4):422–446. 174, 527
- Jeh, Glen, and Jennifer Widom. 2003. Scaling personalized web search. In *Proc. WWW*, pp. 271–279. ACM Press. 481, 527, 534
- Jensen, Finn V., and Finn B. Jensen. 2001. *Bayesian Networks and Decision Graphs*. Springer. 234, 527
- Jeong, Byeong-Soo, and Edward Omiecinski. 1995. Inverted file partitioning schemes in multiple disk systems. *IEEE Transactions on Parallel and Distributed Systems* 6(2): 142–153. 458, 527, 530
- Ji, Xiang, and Wei Xu. 2006. Document clustering with prior knowledge. In *Proc. SIGIR*, pp. 405–412. ACM Press. DOI: doi.acm.org/10.1145/1148170.1148241. 374, 527, 535
- Jing, Hongyan. 2000. Sentence reduction for automatic text summarization. In *Proc. Conference on Applied Natural Language Processing*, pp. 310–315. 174, 527
- Joachims, Thorsten. 1997. A probabilistic analysis of the Rocchio algorithm with tfidf for text categorization. In *Proc. ICML*, pp. 143–151. Morgan Kaufmann. 314, 527
- Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proc. ECML*, pp. 137–142. Springer. xvii, 282, 333, 334, 527

- Joachims, Thorsten. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola (eds.), *Advances in Kernel Methods - Support Vector Learning*. MIT Press. 347, 527
- Joachims, Thorsten. 2002a. *Learning to Classify Text Using Support Vector Machines*. Kluwer. xvii, 334, 347, 527
- Joachims, Thorsten. 2002b. Optimizing search engines using clickthrough data. In *Proc. KDD*, pp. 133–142. 175, 185, 348, 527
- Joachims, Thorsten. 2006a. Training linear SVMs in linear time. In *Proc. KDD*, pp. 217–226. ACM Press. DOI: doi.acm.org/10.1145/1150402.1150429. 286, 329, 347, 527
- Joachims, Thorsten. 2006b. Transductive support vector machines. In *Chapelle et al. (2006)*, pp. 105–118. 347, 527
- Joachims, Thorsten, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proc. SIGIR*, pp. 154–161. ACM Press. 175, 185, 525, 526, 527, 530
- Johnson, David, Vishv Malhotra, and Peter Vamplew. 2006. More effective web search using bigrams and trigrams. *Webology* 3(4). URL: www.webology.ir/2006/v3n4/a35.html. Article 35. 47, 527, 529, 534
- Jurafsky, Dan, and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 2nd edition. Prentice Hall. xxxiv, 252, 527, 529
- Käki, Mika. 2005. Findex: Search result categories help users when document ranking fails. In *Proc. SIGCHI*, pp. 131–140. ACM Press. DOI: doi.acm.org/10.1145/1054972.1054991. 372, 400, 528
- Kammenhuber, Nils, Julia Luxenburger, Anja Feldmann, and Gerhard Weikum. 2006. Web search clickstreams. In *Proc. ACM SIGCOMM on Internet Measurement*, pp. 245–250. ACM Press. 47, 524, 527, 529, 534
- Kamps, Jaap, Maarten de Rijke, and Börkur Sigurbjörnsson. 2004. Length normalization in XML retrieval. In *Proc. SIGIR*, pp. 80–87. ACM Press. DOI: doi.acm.org/10.1145/1008992.1009009. 216, 527, 531, 532
- Kamps, Jaap, Maarten Marx, Maarten de Rijke, and Börkur Sigurbjörnsson. 2006. Articulating information needs in XML query languages. *TOIS* 24(4):407–436. DOI: doi.acm.org/10.1145/1185877.1185879. 216, 527, 529, 531, 532
- Kamvar, Sepandar D., Dan Klein, and Christopher D. Manning. 2002. Interpreting and extending classical agglomerative clustering algorithms using a model-based approach. In *Proc. ICML*, pp. 283–290. Morgan Kaufmann. 400, 527, 529
- Kannan, Ravi, Santosh Vempala, and Adrian Vetta. 2000. On clusterings – Good, bad and spectral. In *Proc. Symposium on Foundations of Computer Science*, pp. 367–377. IEEE Computer Society. 400, 527, 534
- Kaszkiel, Marcin, and Justin Zobel. 1997. Passage retrieval revisited. In *Proc. SIGIR*, pp. 178–185. ACM Press. DOI: doi.acm.org/10.1145/258525.258561. 217, 527, 535
- Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding groups in data*. Wiley. 373, 527, 532

- Kazai, Gabriella, and Mounia Lalmas. 2006. eXtended cumulated gain measures for the evaluation of content-oriented XML retrieval. *TOIS* 24(4):503–542. DOI: doi.acm.org/10.1145/1185883. 217, 527, 528
- Kekäläinen, Jaana. 2005. Binary and graded relevance in IR evaluations – Comparison of the effects on ranking of IR systems. *IP&M* 41:1019–1033. 174, 527
- Kekäläinen, Jaana, and Kalervo Järvelin. 2002. Using graded relevance assessments in IR evaluation. *JASIST* 53(13):1120–1129. 174, 527
- Kemeny, John G., and J. Laurie Snell. 1976. *Finite Markov Chains*. Springer. 480, 527, 533
- Kent, Allen, Madeline M. Berry, Fred U. Luehrs, Jr., and J. W. Perry. 1955. Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation* 6(2):93–101. 173, 522, 527, 529, 531
- Kernighan, Mark D., Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. In *Proc. ACL*, pp. 205–210. 65, 523, 525, 527
- King, Benjamin. 1967. Step-wise clustering procedures. *Journal of the American Statistical Association* 69:86–101. 399, 527
- Kishida, Kazuaki, Kuang-Hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, and Sung Hyon Myaeng. 2005. Overview of CLIR task at the fifth NTCIR workshop. In *NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*. National Institute of Informatics. 45, 523, 527, 528, 530
- Klein, Dan, and Christopher D. Manning. 2002. Conditional structure versus conditional estimation in NLP models. In *Proc. Empirical Methods in Natural Language Processing*, pp. 9–16. 336, 527, 529
- Kleinberg, Jon M. 1997. Two algorithms for nearest-neighbor search in high dimensions. In *Proc. ACM Symposium on Theory of Computing*, pp. 599–608. ACM Press. DOI: doi.acm.org/10.1145/258533.258653. 314, 527
- Kleinberg, Jon M. 1999. Authoritative sources in a hyperlinked environment. *JACM* 46(5):604–632. URL: citeseer.ist.psu.edu/article/kleinberg98authoritative.html. 481, 527
- Kleinberg, Jon M. 2002. An impossibility theorem for clustering. In *Proc. NIPS*. 373, 527
- Knuth, Donald E. 1997. *The Art of Computer Programming, Volume 3: Sorting and Searching*, 3rd edition. Addison Wesley. 65, 527
- Ko, Youngjoong, Jinwoo Park, and Jungyun Seo. 2004. Improving text categorization using the importance of sentences. *IP&M* 40(1):65–79. 340, 527, 530, 532
- Koenemann, Jürgen, and Nicholas J. Belkin. 1996. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proc. SIGCHI*, pp. 205–212. ACM Press. DOI: doi.acm.org/10.1145/238386.238487. 194, 522, 527
- Kołcz, Aleksander, Vidya Prabakarmurthi, and Jugal Kalita. 2000. Summarization as feature selection for text categorization. In *Proc. CIKM*, pp. 365–370. ACM Press. 340, 527, 531

- Kołcz, Aleksander, and Wen-Tau Yih. 2007. Raising the baseline for high-precision text classifiers. In *Proc. KDD*. 286, 527, 535
- Koller, Daphne, and Mehran Sahami. 1997. Hierarchically classifying documents using very few words. In *Proc. ICML*, pp. 170–178. 347, 527, 532
- Konheim, Alan G. 1981. *Cryptography: A Primer*. John Wiley & Sons. 46, 527
- Korfhage, Robert R. 1997. *Information Storage and Retrieval*. Wiley. xxxiv, 175, 527
- Kozlov, M. K., S. P. Tarasov, and L. G. Khachiyan. 1979. Polynomial solvability of convex quadratic programming. *Soviet Mathematics Doklady* 20:1108–1111. Translated from original in *Doklady Akademii Nauk SSR*, 228 (1979). 328, 527, 533
- Kraaij, Wessel, and Martijn Spitters. 2003. Language models for topic tracking. In W. B. Croft and J. Lafferty (eds.), *Language Modeling for Information Retrieval*, pp. 95–124. Kluwer. 251, 528, 533
- Kraaij, Wessel, Thijs Westerveld, and Djoerd Hiemstra. 2002. The importance of prior probabilities for entry page search. In *Proc. SIGIR*, pp. 27–34. ACM Press. 252, 526, 528, 534
- Krippendorff, Klaus. 2003. *Content Analysis: An Introduction to its Methodology*. Sage. 174, 528
- Krovetz, Bob. 1995. *Word sense disambiguation for large text databases*. PhD thesis, University of Massachusetts Amherst. 46, 528
- Kukich, Karen. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys* 24(4):377–439. DOI: doi.acm.org/10.1145/146370.146380. 65, 528
- Kumar, Ravi, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. 1999. Trawling the Web for emerging cyber-communities. *Computer Networks* 31(11–16): 1481–1493. URL: citeseer.ist.psu.edu/kumar99trawling.html. 442, 528, 531, 533
- Kumar, S. Ravi, Prabhakar Raghavan, Sridhar Rajagopalan, Dandapani Sivakumar, Andrew Tomkins, and Eli Upfal. 2000. The Web as a graph. In *Proc. PODS*, pp. 1–10. ACM Press. URL: citeseer.ist.psu.edu/article/kumar00web.html. 441, 528, 531, 533, 534
- Kupiec, Julian, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proc. SIGIR*, pp. 68–73. ACM Press. 174, 523, 528, 531
- Kurland, Oren, and Lillian Lee. 2004. Corpus structure, language models, and ad hoc information retrieval. In *Proc. SIGIR*, pp. 194–201. ACM Press. DOI: doi.acm.org/10.1145/1008992.1009027. 372, 528
- Lafferty, John, and Chengxiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proc. SIGIR*, pp. 111–119. ACM Press. 250, 251, 528, 535
- Lafferty, John, and Chengxiang Zhai. 2003. Probabilistic relevance models based on document and query generation. In W. Bruce Croft and John Lafferty (eds.), *Language Modeling for Information Retrieval*. Kluwer. 252, 528, 535
- Lalmas, Mounia, Gabriella Kazai, Jaap Kamps, Jovan Pehcevski, Benjamin Piwowarski, and Stephen E. Robertson. 2007. INEX 2006 evaluation measures. In Fuhr et al. (2007), pp. 20–34. 217, 527, 528, 531

- Lalmas, Mounia, and Anastasios Tombros. 2007. Evaluating XML retrieval effectiveness at INEX. *SIGIR Forum* 41(1):40–57. DOI: doi.acm.org/10.1145/1273221.1273225. 216, 528, 533
- Lance, G. N., and W. T. Williams. 1967. A general theory of classificatory sorting strategies 1. Hierarchical systems. *Computer Journal* 9(4):373–380. 399, 528, 535
- Langville, Amy, and Carl Meyer. 2006. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press. 481, 528, 529
- Larsen, Bjornar, and Chinatsu Aone. 1999. Fast and effective text mining using linear-time document clustering. In *Proc. KDD*, pp. 16–22. ACM Press. DOI: doi.acm.org/10.1145/312129.312186. 399, 400, 521, 528
- Larson, Ray R. 2005. A fusion approach to XML structured document retrieval. *IR* 8(4):601–629. DOI: dx.doi.org/10.1007/s10791-005-0749-0. 216, 528
- Lavrenko, Victor, and W. Bruce Croft. 2001. Relevance-based language models. In *Proc. SIGIR*, pp. 120–127. ACM Press. 250, 524, 528
- Lawrence, Steve, and C. Lee Giles. 1998. Searching the World Wide Web. *Science* 280(5360):98–100. URL: citeseer.ist.psu.edu/lawrence98searching.html. 442, 525, 528
- Lawrence, Steve, and C. Lee Giles. 1999. Accessibility of information on the web. *Nature* 500:107–109. 442, 525, 528
- Lee, Whay C., and Edward A. Fox. 1988. Experimental comparison of schemes for interpreting Boolean queries. Technical Report TR-88-27, Computer Science, Virginia Polytechnic Institute and State University. 17, 525, 528
- Lempel, Ronny, and Shlomo Moran. 2000. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks* 33(1–6):387–401. URL: citeseer.ist.psu.edu/lempel00stochastic.html. 481, 528, 530
- Lesk, Michael. 1988. Grab – Inverted indexes with low storage overhead. *Computing Systems* 1:207–220. 83, 528
- Lesk, Michael. 2004. *Understanding Digital Libraries*, 2nd edition. Morgan Kaufmann. xxiv, 528
- Lester, Nicholas, Alistair Moffat, and Justin Zobel. 2005. Fast on-line index construction by geometric partitioning. In *Proc. CIKM*, pp. 776–783. ACM Press. DOI: doi.acm.org/10.1145/1099554.1099739. 84, 528, 529, 535
- Lester, Nicholas, Justin Zobel, and Hugh E. Williams. 2006. Efficient online index maintenance for contiguous inverted lists. *IP&M* 42(4):916–933. DOI: dx.doi.org/10.1016/j.ipm.2005.09.005. 84, 528, 535
- Levenshtein, Vladimir I. 1965. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission* 1:8–17. 65, 528
- Lew, Michael S. 2001. *Principles of Visual Information Retrieval*. Springer. xxiv, 528
- Lewis, David D. 1995. Evaluating and optimizing autonomous text classification systems. In *Proc. SIGIR*. ACM Press. 286, 528
- Lewis, David D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proc. ECML*, pp. 4–15. Springer. 286, 528

- Lewis, David D., and Karen Spärck Jones. 1996. Natural language processing for information retrieval. *CACM* 39(1):92–101. DOI: doi.acm.org/10.1145/234173.234210. [xxxiv](#), [527](#), [528](#)
- Lewis, David D., and Marc Ringuette. 1994. A comparison of two learning algorithms for text categorization. In *Proc. SDAIR*, pp. 81–93. [286](#), [528](#), [531](#)
- Lewis, David D., Robert E. Schapire, James P. Callan, and Ron Papka. 1996. Training algorithms for linear text classifiers. In *Proc. SIGIR*, pp. 298–306. ACM Press. DOI: doi.acm.org/10.1145/243199.243277. [315](#), [523](#), [528](#), [530](#), [532](#)
- Lewis, David D., Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *JMLR* 5:361–397. [84](#), [287](#), [528](#), [532](#), [535](#)
- Li, Fan, and Yiming Yang. 2003. A loss function analysis for classification methods in text categorization. In *Proc. ICML*, pp. 472–479. [xvii](#), [282](#), [347](#), [528](#), [535](#)
- Liddy, Elizabeth D. 2005. Automatic document retrieval. In *Encyclopedia of Language and Linguistics*, 2nd edition. Elsevier. [17](#), [528](#)
- List, Johan, Vojkan Mihajlovic, Georgina Ramírez, Arjen P. Vries, Djoerd Hiemstra, and Henk Ernst Blok. 2005. TIJAH: Embracing IR methods in XML databases. *IR* 8(4):547–570. DOI: dx.doi.org/10.1007/s10791-005-0747-2. [216](#), [522](#), [526](#), [528](#), [529](#), [531](#), [534](#)
- Lita, Lucian Vlad, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. tRuE-casIng. In *Proc. ACL*, pp. 152–159. [46](#), [526](#), [527](#), [528](#), [532](#)
- Littman, Michael L., Susan T. Dumais, and Thomas K. Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In Gregory Grefenstette (ed.), *Proc. Cross-Language Information Retrieval*. Kluwer. URL: cite-seer.ist.psu.edu/littman98automatic.html. [417](#), [524](#), [528](#)
- Liu, Tie-Yan, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma. 2005. Support vector machines classification with very large scale taxonomy. *ACM SIGKDD Explorations* 7(1):36–43. [347](#), [523](#), [528](#), [529](#), [534](#), [535](#)
- Liu, Xiaoyong, and W. Bruce Croft. 2004. Cluster-based retrieval using language models. In *Proc. SIGIR*, pp. 186–193. ACM Press. DOI: doi.acm.org/10.1145/1008992.1009026. [252](#), [351](#), [372](#), [524](#), [528](#)
- Lloyd, Stuart P. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2):129–136. [373](#), [529](#)
- Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *JMLR* 2:419–444. [347](#), [523](#), [529](#), [532](#), [534](#)
- Lombard, Matthew, Cheryl C. Bracken, and Jennifer Snyder-Duch. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research* 28:587–604. [174](#), [522](#), [529](#), [533](#)
- Long, Xiaohui, and Torsten Suel. 2003. Optimized query execution in large search engines with global page ordering. In *Proc. VLDB*. URL: cite-seer.ist.psu.edu/long03optimized.html. [149](#), [529](#), [533](#)

- Lovins, Julie Beth. 1968. Development of a stemming algorithm. *Translation and Computational Linguistics* 11(1):22–31. 33, 529
- Lu, Wei, Stephen E. Robertson, and Andrew MacFarlane. 2007. CISR at INEX 2006. In Fuhr et al. (2007), pp. 57–63. 216, 529, 531
- Luhn, Hans Peter. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 1(4):309–317. 133, 529
- Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2):159–165, 317. 133, 529
- Luk, Robert W. P., and Kui-Lam Kwok. 2002. A comparison of Chinese document indexing strategies and retrieval models. *ACM Transactions on Asian Language Information Processing* 1(3):225–268. 45, 528, 529
- Lunde, Ken. 1998. *CJKV Information Processing*. O'Reilly. 45, 529
- MacFarlane, A., J.A. McCann, and S.E. Robertson. 2000. Parallel search using partitioned inverted files. In *Proc. SPIRE*, pp. 209–220. 458, 529, 531
- MacQueen, James B. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. Berkeley Symposium on Mathematics, Statistics and Probability*, pp. 281–297. University of California Press. 373, 529
- Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press. xxxiv, 40, 105, 251, 252, 286, 372, 529, 532
- Maron, M. E., and J. L. Kuhns. 1960. On relevance, probabilistic indexing, and information retrieval. *JACM* 7(3):216–244. 235, 286, 528, 529
- Mass, Yosi, Matan Mandelbrod, Einat Amitay, David Carmel, Yoëlle S. Maarek, and Aya Soffer. 2003. JuruXML – An XML retrieval system at INEX'02. In Fuhr et al. (2003b), pp. 73–80. URL: inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf. 216, 521, 523, 529, 533
- McBryan, Oliver A. 1994. GENVL and WWW: Tools for Taming the Web. In *Proc. WWW*. URL: citeseer.ist.psu.edu/mcbryan94genvl.html. 442, 480, 529
- McCallum, Andrew, and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. In *AAAI/ICML Workshop on Learning for Text Categorization*, pp. 41–48. 286, 529, 530
- McCallum, Andrew, Ronald Rosenfeld, Tom M. Mitchell, and Andrew Y. Ng. 1998. Improving text classification by shrinkage in a hierarchy of classes. In *Proc. ICML*, pp. 359–367. Morgan Kaufmann. 347, 529, 530, 532
- McCallum, Andrew Kachites. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. www.cs.cmu.edu/~mccallum/bow. 316, 529
- McKeown, Kathleen, and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *Proc. SIGIR*, pp. 74–82. ACM Press. DOI: [doi.acm.org/10.1145/215206.215334](https://doi.org/10.1145/215206.215334). 400, 529, 531
- McKeown, Kathleen R., Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman.

2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proc. Human Language Technology Conference*. 351, 373, 522, 524, 526, 527, 529, 530, 532
- McLachlan, Geoffrey J., and Thiriyambakam Krishnan. 1996. *The EM Algorithm and Extensions*. John Wiley & Sons. 373, 528, 529
- Meadow, Charles T., Donald H. Kraft, and Bert R. Boyce. 1999. *Text Information Retrieval Systems*. Academic Press. xxxiv, 522, 528, 529
- Meilă, Marina. 2005. Comparing clusterings – An axiomatic view. In *Proc. ICML*. 373, 529
- Melnik, Sergey, Sriram Raghavan, Beverly Yang, and Hector Garcia-Molina. 2001. Building a distributed full-text index for the web. In *Proc. WWW*, pp. 396–406. ACM Press. DOI: doi.acm.org/10.1145/371920.372095. 83, 525, 529, 531, 535
- Mihajlović, Vojkan, Henk Ernst Blok, Djoerd Hiemstra, and Peter M. G. Apers. 2005. Score region algebra: Building a transparent XML-R database. In *Proc. CIKM*, pp. 12–19. DOI: doi.acm.org/10.1145/1099554.1099560. 216, 521, 522, 526, 529
- Miller, David R. H., Tim Leek, and Richard M. Schwartz. 1999. A hidden Markov model information retrieval system. In *Proc. SIGIR*, pp. 214–221. ACM Press. 246, 252, 528, 529, 532
- Minsky, Marvin Lee, and Seymour Papert (eds.). 1988. *Perceptrons: An introduction to computational geometry*. MIT Press. Expanded edition. 315, 529, 530
- Mitchell, Tom M. 1997. *Machine Learning*. McGraw Hill. 286, 529
- Moffat, Alistair, and Timothy A. H. Bell. 1995. In situ generation of compressed inverted files. *JASIS* 46(7):537–550. 83, 522, 529
- Moffat, Alistair, and Lang Stuiver. 1996. Exploiting clustering in inverted file compression. In *Proc. Conference on Data Compression*, pp. 82–91. IEEE Computer Society. 106, 529, 533
- Moffat, Alistair, and Justin Zobel. 1992. Parameterised compression for sparse bitmaps. In *Proc. SIGIR*, pp. 274–285. ACM Press. DOI: doi.acm.org/10.1145/133160.133210. 106, 530, 535
- Moffat, Alistair, and Justin Zobel. 1996. Self-indexing inverted files for fast text retrieval. *TOIS* 14(4):349–379. 46, 47, 530, 535
- Moffat, Alistair, and Justin Zobel. 1998. Exploring the similarity space. *SIGIR Forum* 32(1). 133, 530, 535
- Mooers, Calvin. 1961. From a point of view of mathematical etc. techniques. In R. A. Fairthorne (ed.), *Towards information retrieval*, pp. xvii–xxiii. Butterworths. 17, 530
- Mooers, Calvin E. 1950. Coding, information retrieval, and the rapid selector. *American Documentation* 1(4):225–229. 17, 530
- Moschitti, Alessandro. 2003. A study on optimal parameter tuning for Rocchio text classifier. In *Proc. ECIR*, pp. 420–435. 315, 530
- Moschitti, Alessandro, and Roberto Basili. 2004. Complex linguistic features for text classification: A comprehensive study. In *Proc. ECIR*, pp. 181–196. 347, 522, 530

- Murata, Masaki, Qing Ma, Kiyotaka Uchimoto, Hiromi Ozaku, Masao Utiyama, and Hitoshi Isahara. 2000. Japanese probabilistic information retrieval using location and category information. In *International Workshop on Information Retrieval With Asian Languages*, pp. 81–88. URL: portal.acm.org/citation.cfm?doid=355214.355226. 340, 526, 529, 530, 534
- Muresan, Gheorghe, and David J. Harper. 2004. Topic modeling for mediated access to very large document collections. *JASIST* 55(10):892–910. DOI: [dx.doi.org/10.1002/asi.20034](https://doi.org/10.1002/asi.20034). 372, 526, 530
- Murtagh, Fionn. 1983. A survey of recent advances in hierarchical clustering algorithms. *Computer Journal* 26(4):354–359. 399, 530
- Najork, Marc, and Allan Heydon. 2001. High-performance web crawling. Technical Report 173, Compaq Systems Research Center. 458, 526, 530
- Najork, Marc, and Allan Heydon. 2002. High-performance web crawling. In James Abello, Panos Pardalos, and Mauricio Resende (eds.), *Handbook of Massive Data Sets*, chapter 2. Kluwer. 458, 526, 530
- Navarro, Gonzalo, and Ricardo Baeza-Yates. 1997. Proximal nodes: A model to query document databases by content and structure. *TOIS* 15(4):400–435. DOI: [doi.acm.org/10.1145/263479.263482](https://doi.org/10.1145/263479.263482). 217, 521, 530
- Newsam, Shawn, Sitaram Bhagavathy, and B. S. Manjunath. 2001. Category-based image retrieval. In *Proc. IEEE International Conference on Image Processing, Special Session on Multimedia Indexing, Browsing and Retrieval*, pp. 596–599. 179, 522, 529, 530
- Ng, Andrew Y., and Michael I. Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proc. NIPS*, pp. 841–848. URL: www-2.cs.cmu.edu/Groups/NIPS/NIPS2001/papers/psgz/AA28.ps.gz. 286, 336, 527, 530
- Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. 2001a. On spectral clustering: Analysis and an algorithm. In *Proc. NIPS*, pp. 849–856. 400, 527, 530, 534
- Ng, Andrew Y., Alice X. Zheng, and Michael I. Jordan. 2001b. Link analysis, eigenvectors and stability. In *Proc. IJCAI*, pp. 903–910. URL: citeseer.ist.psu.edu/ng01link.html. 481, 527, 530, 535
- Nigam, Kamal, Andrew McCallum, and Tom Mitchell. 2006. Semi-supervised text classification using EM. In *Chapelle et al. (2006)*, pp. 33–56. 347, 529, 530
- Ntoulas, Alexandros, and Junghoo Cho. 2007. Pruning policies for two-tiered inverted index with correctness guarantee. In *Proc. SIGIR*, pp. 191–198. ACM Press. 105, 523, 530
- Oard, Douglas W., and Bonnie J. Dorr. 1996. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA. xxxiv, 524, 530
- Ogilvie, Paul, and Jamie Callan. 2005. Parameter estimation for a simple hierarchical generative model for XML retrieval. In *Proc. INEX*, pp. 211–224. DOI: [dx.doi.org/10.1007/11766278_16](https://doi.org/10.1007/11766278_16). 216, 523, 530

- O’Keefe, Richard A., and Andrew Trotman. 2004. The simplest query language that could possibly work. In Fuhr et al. (2005), pp. 167–174. 217, 530, 534
- Osiński, Stanisław, and Dawid Weiss. 2005. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems* 20(3):48–54. 400, 530, 534
- Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The Page-Rank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project. URL: citeseer.ist.psu.edu/page98pagerank.html. 480, 522, 530, 535
- Paice, Chris D. 1990. Another stemmer. *SIGIR Forum* 24(3):56–61. 33, 530
- Papineni, Kishore. 2001. Why inverse document frequency? In *Proc. North American Chapter of the Association for Computational Linguistics*, pp. 1–8. 133, 530
- Pavlov, Dmitry, Ramnath Balasubramanyan, Byron Dom, Shyam Kapur, and Jignashu Parikh. 2004. Document preprocessing for naive Bayes classification and clustering with mixture of multinomials. In *Proc. KDD*, pp. 829–834. 286, 521, 524, 527, 530
- Pelleg, Dan, and Andrew Moore. 1999. Accelerating exact k-means algorithms with geometric reasoning. In *Proc. KDD*, pp. 277–281. ACM Press. DOI: [doi.acm.org/10.1145/312129.312248](https://doi.org/10.1145/312129.312248). 373, 530, 531
- Pelleg, Dan, and Andrew Moore. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proc. ICML*, pp. 727–734. Morgan Kaufmann. 373, 530, 531
- Perkins, Simon, Kevin Lacker, and James Theiler. 2003. Grafting: Fast, incremental feature selection by gradient descent in function space. *JMLR* 3:1333–1356. 286, 528, 531, 533
- Persin, Michael. 1994. Document filtering for fast ranking. In *Proc. SIGIR*, pp. 339–348. ACM Press. 149, 531
- Persin, Michael, Justin Zobel, and Ron Sacks-Davis. 1996. Filtered document retrieval with frequency-sorted indexes. *JASIS* 47(10):749–764. 149, 531, 532, 535
- Peterson, James L. 1980. Computer programs for detecting and correcting spelling errors. *CACM* 23(12):676–687. DOI: [doi.acm.org/10.1145/359038.359041](https://doi.org/10.1145/359038.359041). 65, 531
- Picca, Davide, Benoît Curdy, and François Bavaud. 2006. Non-linear correspondence analysis in text retrieval: A kernel view. In *Proc. JADT*. 308, 522, 524, 531
- Pinski, Gabriel, and Francis Narin. 1976. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of Physics. *IP&M* 12:297–326. 480, 530, 531
- Pirolli, Peter L. T. 2007. *Information Foraging Theory: Adaptive Interaction With Information*. Oxford University Press. 373, 531
- Platt, John. 2000. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press. 325, 531

- Ponte, Jay M., and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. SIGIR*, pp. 275–281. ACM Press. [xxii](#), [246](#), [247](#), [249](#), [252](#), [524](#), [531](#)
- Popescul, Alexandrin, and Lyle H. Ungar. 2000. Automatic labeling of document clusters. Unpublished MS, U. Pennsylvania. URL: <http://www.cis.upenn.edu/~popescul/Publications/popescul00labeling.pdf>. [400](#), [531](#), [534](#)
- Porter, Martin F. 1980. An algorithm for suffix stripping. *Program* 14(3):130–137. [33](#), [531](#)
- Pugh, William. 1990. Skip lists: A probabilistic alternative to balanced trees. *CACM* 33(6):668–676. [46](#), [531](#)
- Qin, Tao, Tie-Yan Liu, Wei Lai, Xu-Dong Zhang, De-Sheng Wang, and Hang Li. 2007. Ranking with multiple hyperplanes. In *Proc. SIGIR*. ACM Press. [348](#), [528](#), [531](#), [534](#), [535](#)
- Qiu, Yonggang, and H.P. Frei. 1993. Concept based query expansion. In *Proc. SIGIR*, pp. 160–169. ACM Press. [194](#), [525](#), [531](#)
- R Development Core Team. 2005. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. URL: www.R-project.org. ISBN 3-900051-07-0. [374](#), [400](#), [531](#)
- Radev, Dragomir R., Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. 2001. Interactive, domain-independent identification and summarization of topically related news articles. In *Proc. European Conference on Research and Advanced Technology for Digital Libraries*, pp. 225–238. [373](#), [522](#), [531](#), [535](#)
- Rahm, Erhard, and Philip A. Bernstein. 2001. A survey of approaches to automatic schema matching. *VLDB Journal* 10(4):334–350. URL: citeseer.ist.psu.edu/rahm01survey.html. [216](#), [522](#), [531](#)
- Rand, William M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336):846–850. [373](#), [531](#)
- Rasmussen, Edie. 1992. Clustering algorithms. In *Frakes and Baeza-Yates (1992)*, pp. 419–442. [372](#), [531](#)
- Rennie, Jason D., Lawrence Shih, Jaime Teevan, and David R. Karger. 2003. Tackling the poor assumptions of naive Bayes text classifiers. In *Proc. ICML*, pp. 616–623. [286](#), [527](#), [531](#), [532](#), [533](#)
- Ribeiro-Neto, Berthier, Edleno S. Moura, Marden S. Neubert, and Nivio Ziviani. 1999. Efficient distributed algorithms to build inverted files. In *Proc. SIGIR*, pp. 105–112. ACM Press. DOI: [doi.acm.org/10.1145/312624.312663](https://doi.org/10.1145/312624.312663). [83](#), [530](#), [531](#), [535](#)
- Ribeiro-Neto, Berthier A., and Ramurti A. Barbosa. 1998. Query performance for tightly coupled distributed digital libraries. In *Proc. ACM Conference on Digital Libraries*, pp. 182–190. [459](#), [521](#), [531](#)
- Rice, John A. 2006. *Mathematical Statistics and Data Analysis*. Duxbury Press. [99](#), [235](#), [276](#), [531](#)
- Richardson, M., A. Prakash, and E. Brill. 2006. Beyond PageRank: machine learning for static ranking. In *Proc. WWW*, pp. 707–715. [348](#), [522](#), [531](#)

- Riezler, Stefan, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proc. ACL*, pp. 464–471. Association for Computational Linguistics. URL: www.aclweb.org/anthology/P/P07/P07-1059. 194, 529, 531, 534
- Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press. 222, 235, 531
- Robertson, Stephen. 2005. How Okapi came to TREC. In *Voorhees and Harman (2005)*, pp. 287–299. 235, 531
- Robertson, Stephen, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proc. CIKM*, pp. 42–49. DOI: doi.acm.org/10.1145/1031171.1031181. 235, 531, 533, 535
- Robertson, Stephen E., and Karen Spärck Jones. 1976. Relevance weighting of search terms. *JASIS* 27:129–146. 133, 235, 527, 531
- Rocchio, J. J. 1971. Relevance feedback in information retrieval. In *Salton (1971b)*, pp. 313–323. 181, 193, 314, 532
- Roget, P. M. 1946. *Roget's International Thesaurus*. Thomas Y. Crowell. 194, 532
- Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proc. UAI*, pp. 487–494. 418, 525, 532, 533
- Ross, Sheldon. 2006. *A First Course in Probability*. Pearson Prentice Hall. 99, 235, 532
- Rusmevichientong, Paat, David M. Pennock, Steve Lawrence, and C. Lee Giles. 2001. Methods for sampling pages uniformly from the world wide web. In *Proc. AAAI Fall Symposium on Using Uncertainty Within Computation*, pp. 121–128. URL: citeseer.ist.psu.edu/rusmevichientong01methods.html. 442, 525, 528, 531, 532
- Ruthven, Ian, and Mounia Lalmas. 2003. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review* 18(1). 194, 528, 532
- Sahoo, Nachiketa, Jamie Callan, Ramayya Krishnan, George Duncan, and Rema Padman. 2006. Incremental hierarchical clustering of text documents. In *Proc. CIKM*, pp. 357–366. DOI: doi.acm.org/10.1145/1183614.1183667. 400, 523, 524, 528, 530, 532
- Sakai, Tetsuya. 2007. On the reliability of information retrieval metrics based on graded relevance. *IP&M* 43(2):531–548. 174, 532
- Salton, Gerard. 1971a. Cluster search strategies and the optimization of retrieval effectiveness. In *The SMART Retrieval System – Experiments in Automatic Document Processing* Salton (1971b), pp. 223–242. 351, 372, 532
- Salton, Gerard (ed.). 1971b. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall. 133, 173, 193, 499, 511, 532
- Salton, Gerard. 1975. *Dynamic information and library processing*. Prentice Hall. 372, 532
- Salton, Gerard. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison Wesley. 46, 194, 532

- Salton, Gerard. 1991. The Smart project in automatic document retrieval. In *Proc. SIGIR*, pp. 356–358. ACM Press. 173, 532
- Salton, Gerard, James Allan, and Chris Buckley. 1993. Approaches to passage retrieval in full text information systems. In *Proc. SIGIR*, pp. 49–58. ACM Press. DOI: doi.acm.org/10.1145/160688.160693. 217, 521, 522, 532
- Salton, Gerard, and Chris Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report, Cornell University, Ithaca, NY, USA. 133, 522, 532
- Salton, Gerard, and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *IP&M* 24(5):513–523. 133, 522, 532
- Salton, Gerard, and Chris Buckley. 1990. Improving retrieval performance by relevance feedback. *JASIS* 41(4):288–297. 194, 522, 532
- Saracevic, Tefko, and Paul Kantor. 1988. A study of information seeking and retrieving. II: Users, questions and effectiveness. *JASIS* 39:177–196. 173, 527, 532
- Saracevic, Tefko, and Paul Kantor. 1996. A study of information seeking and retrieving. III: Searchers, searches, overlap. *JASIS* 39(3):197–216. 173, 527, 532
- Savaresi, Sergio M., and Daniel Boley. 2004. A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis* 8(4):345–362. 400, 522, 532
- Schamber, Linda, Michael Eisenberg, and Michael S. Nilan. 1990. A re-examination of relevance: toward a dynamic, situational definition. *IP&M* 26(6):755–776. 174, 524, 530, 532
- Schapire, Robert E. 2003. The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, and B. Yu (eds.), *Nonlinear Estimation and Classification*. Springer. 347, 532
- Schapire, Robert E., and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning* 39(2/3):135–168. 347, 532, 533
- Schapire, Robert E., Yoram Singer, and Amit Singhal. 1998. Boosting and Rocchio applied to text filtering. In *Proc. SIGIR*, pp. 215–223. ACM Press. 314, 315, 532, 533
- Schlieder, Torsten, and Holger Meuss. 2002. Querying and ranking XML documents. *JASIST* 53(6):489–503. DOI: dx.doi.org/10.1002/asi.10060. 216, 529, 532
- Scholer, Falk, Hugh E. Williams, John Yiannis, and Justin Zobel. 2002. Compression of inverted indexes for fast query evaluation. In *Proc. SIGIR*, pp. 222–229. ACM Press. DOI: doi.acm.org/10.1145/564376.564416. 106, 532, 535
- Schölkopf, Bernhard, and Alexander J. Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press. 346, 532, 533
- Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1):97–124. 192, 194, 532
- Schütze, Hinrich, David A. Hull, and Jan O. Pedersen. 1995. A comparison of classifiers and document representations for the routing problem. In *Proc. SIGIR*, pp. 229–237. ACM Press. 193, 286, 315, 526, 531, 532

- Schütze, Hinrich, and Jan O. Pedersen. 1995. Information retrieval based on word senses. In *Proc. SDAIR*, pp. 161–175. 374, 531, 532
- Schütze, Hinrich, and Craig Silverstein. 1997. Projections for efficient document clustering. In *Proc. SIGIR*, pp. 74–81. ACM Press. 373, 417, 532
- Schwarz, Gideon. 1978. Estimating the dimension of a model. *Annals of Statistics* 6 (2):461–464. 373, 532
- Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47. 286, 532
- Shawe-Taylor, John, and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press. 346, 523, 532
- Shkapenyuk, Vladislav, and Torsten Suel. 2002. Design and implementation of a high-performance distributed web crawler. In *Proc. International Conference on Data Engineering*. URL: citeseer.ist.psu.edu/shkapenyuk02design.html. 458, 532, 533
- Siegel, Sidney, and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition. McGraw Hill. 174, 523, 532
- Sifry, Dave. 2007. The state of the Live Web, April 2007. URL: technorati.com/weblog/2007/04/328.html. 30, 532
- Sigurbjörnsson, Börkur, Jaap Kamps, and Maarten de Rijke. 2004. Mixture models, overlap, and structural hints in XML element retrieval. In *Proc. INEX*, pp. 196–210. 216, 527, 531, 532
- Silverstein, Craig, Monika Rauch Henzinger, Hannes Marais, and Michael Moricz. 1999. Analysis of a very large web search engine query log. *SIGIR Forum* 33(1): 6–12. 47, 526, 529, 530, 532
- Silvestri, Fabrizio. 2007. Sorting out the document identifier assignment problem. In *Proc. ECIR*, pp. 101–112. 106, 533
- Silvestri, Fabrizio, Raffaele Perego, and Salvatore Orlando. 2004. Assigning document identifiers to enhance compressibility of web search engines indexes. In *Proc. ACM Symposium on Applied Computing*, pp. 600–605. 106, 530, 531, 533
- Sindhwani, V., and S. S. Keerthi. 2006. Large scale semi-supervised linear SVMs. In *Proc. SIGIR*, pp. 477–484. 348, 527, 533
- Singhal, Amit, Chris Buckley, and Mandar Mitra. 1996a. Pivoted document length normalization. In *Proc. SIGIR*, pp. 21–29. ACM Press. URL: citeseer.ist.psu.edu/singhal96pivoted.html. 133, 522, 529, 533
- Singhal, Amit, Mandar Mitra, and Chris Buckley. 1997. Learning routing queries in a query zone. In *Proc. SIGIR*, pp. 25–32. ACM Press. 193, 522, 529, 533
- Singhal, Amit, Gerard Salton, and Chris Buckley. 1995. Length normalization in degraded text collections. Technical report, Cornell University, Ithaca, NY. 133, 522, 532, 533
- Singhal, Amit, Gerard Salton, and Chris Buckley. 1996b. Length normalization in degraded text collections. In *Proc. SDAIR*, pp. 149–162. 133, 522, 532, 533

- Singitham, Pavan Kumar C., Mahathi S. Mahabhashyam, and Prabhakar Raghavan. 2004. Efficiency-quality tradeoffs for vector score aggregation. In *Proc. VLDB*, pp. 624–635. URL: www.vldb.org/conf/2004/RS17P1.PDF. 149, 372, 529, 531, 533
- Smeulders, Arnold W. M., Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(12):1349–1380. DOI: [dx.doi.org/10.1109/34.895972](https://doi.org/10.1109/34.895972). xxxiv, 525, 526, 532, 533, 535
- Sneath, Peter H.A., and Robert R. Sokal. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W.H. Freeman. 399, 533
- Snedecor, George Waddel, and William G. Cochran. 1989. *Statistical methods*. Iowa State University Press. 286, 523, 533
- Somogyi, Zoltan. 1990. The Melbourne University bibliography system. Technical Report 90/3, Melbourne University, Parkville, Victoria, Australia. 83, 533
- Song, Ruihua, Ji-Rong Wen, and Wei-Ying Ma. 2005. Viewing term proximity from a different perspective. Technical Report MSR-TR-2005-69, Microsoft Research. 149, 529, 533, 534
- Sornil, Ohm. 2001. *Parallel Inverted Index for Large-Scale, Dynamic Digital Libraries*. PhD thesis, Virginia Tech. URL: scholar.lib.vt.edu/theses/available/etd-02062001-114915/. 459, 533
- Spärck Jones, Karen. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1):11–21. 133, 527
- Spärck Jones, Karen. 2004. Language modelling's generative model: Is it rational? MS, Computer Laboratory, University of Cambridge. URL: www.cl.cam.ac.uk/~ksj21/langmodnote4.pdf. 252, 527
- Spärck Jones, Karen, S. Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments. *IP&M* 36(6): 779–808, 809–840. 232, 234, 235, 527, 531, 534
- Spink, Amanda, and Charles Cole (eds.). 2005. *New Directions in Cognitive Information Retrieval*. Springer. 175, 523, 533
- Spink, Amanda, Bernard J. Jansen, and H. Cenk Ozmultu. 2000. Use of query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policy* 10(4):317–328. URL: ist.psu.edu/faculty_pages/jjansen/academic/pubs/internetresearch2000.pdf. 185, 526, 530, 533
- Sproat, Richard, and Thomas Emerson. 2003. The first international Chinese word segmentation bakeoff. In *SIGHAN Workshop on Chinese Language Processing*. 46, 524, 533
- Sproat, Richard, William Gale, Chilin Shih, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics* 22(3):377–404. 46, 523, 525, 532, 533
- Sproat, Richard William. 1992. *Morphology and computation*. MIT Press. 46, 533

- Stein, Benno, and Sven Meyer zu Eissen. 2004. Topic identification: Framework and application. In *Proc. International Conference on Knowledge Management*. 400, 524, 533
- Stein, Benno, Sven Meyer zu Eissen, and Frank Wißbrock. 2003. On cluster validity and the information need of users. In *Proc. Artificial Intelligence and Applications*. 373, 524, 533, 535
- Steinbach, Michael, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*. 400, 527, 528, 533
- Strang, Gilbert (ed.). 1986. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press. 417, 533
- Strehl, Alexander. 2002. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. PhD thesis, The University of Texas at Austin. 373, 533
- Strohman, Trevor, and W. Bruce Croft. 2007. Efficient document retrieval in main memory. In *Proc. SIGIR*, pp. 175–182. ACM Press. 47, 524, 533
- Swanson, Don R. 1988. Historical note: Information retrieval and the future of an illusion. *JASIS* 39(2):92–98. 173, 193, 533
- Tague-Sutcliffe, Jean, and James Blustein. 1995. A statistical analysis of the TREC-3 data. In *Proc. TREC*, pp. 385–398. 174, 522, 533
- Tan, Songbo, and Xueqi Cheng. 2007. Using hypothesis margin to boost centroid text classifier. In *Proc. ACM Symposium on Applied Computing*, pp. 398–403. ACM Press. DOI: doi.acm.org/10.1145/1244002.1244096. 314, 523, 533
- Tannier, Xavier, and Shlomo Geva. 2005. XML retrieval with a natural language interface. In *Proc. SPIRE*, pp. 29–40. 217, 525, 533
- Tao, Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. 2006. Language model information retrieval with document expansion. In *Proc. Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics*, pp. 407–414. 252, 529, 533, 534, 535
- Taube, Mortimer, and Harold Wooster (eds.). 1958. *Information storage and retrieval: Theory, systems, and devices*. Columbia University Press. 17, 533, 535
- Taylor, Michael, Hugo Zaragoza, Nick Craswell, Stephen Robertson, and Chris Burges. 2006. Optimisation methods for ranking functions with multiple parameters. In *Proc. CIKM*. ACM Press. 348, 522, 523, 531, 533, 535
- Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101(476): 1566–1581. 418, 522, 527, 533
- Theobald, Martin, Holger Bast, Debapriyo Majumdar, Ralf Schenkel, and Gerhard Weikum. 2008. TopX: Efficient and versatile top-*k* query processing for semistructured data. *VLDB Journal* 17(1):81–115. 216, 522, 529, 532, 533, 534
- Theobald, Martin, Ralf Schenkel, and Gerhard Weikum. 2005. An efficient and versatile query engine for TopX search. In *Proc. VLDB*, pp. 625–636. VLDB Endowment. 216, 532, 533, 534

- Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B* 63:411–423. 374, 526, 533, 534
- Tishby, Naftali, and Noam Slonim. 2000. Data clustering by Markovian relaxation and the information bottleneck method. In *Proc. NIPS*, pp. 640–646. 374, 533
- Toda, Hiroyuki, and Ryoji Kataoka. 2005. A search result clustering method using informatively named entities. In *International Workshop on Web Information and Data Management*, pp. 81–86. ACM Press. DOI: doi.acm.org/10.1145/1097047.1097063. 372, 527, 533
- Tomasic, Anthony, and Hector Garcia-Molina. 1993. Query processing and inverted indices in shared-nothing document information retrieval systems. *VLDB Journal* 2(3):243–275. 458, 525, 533
- Tombros, Anastasios, and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *Proc. SIGIR*, pp. 2–10. ACM Press. DOI: doi.acm.org/10.1145/290941.290947. 174, 532, 533
- Tombros, Anastasios, Robert Villa, and Cornelis Joost van Rijsbergen. 2002. The effectiveness of query-specific hierarchic clustering in information retrieval. *IP&M* 38(4):559–582. DOI: dx.doi.org/10.1016/S0306-4573(01)00048-6. 372, 531, 533, 534
- Tomlinson, Stephen. 2003. Lexical and algorithmic stemming compared for 9 European languages with Hummingbird Searchserver at CLEF 2003. In *Proc. Cross-Language Evaluation Forum*, pp. 286–300. 46, 533
- Tong, Simon, and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *JMLR* 2:45–66. 348, 527, 533
- Toutanova, Kristina, and Robert C. Moore. 2002. Pronunciation modeling for improved spelling correction. In *Proc. ACL*, pp. 144–151. 65, 530, 533
- Treeratpituk, Pucktada, and Jamie Callan. 2006. An experimental study on automatically labeling hierarchical clusters using statistical features. In *Proc. SIGIR*, pp. 707–708. ACM Press. DOI: doi.acm.org/10.1145/1148170.1148328. 400, 523, 534
- Trotman, Andrew. 2003. Compressing inverted files. *IR* 6(1):5–19. DOI: dx.doi.org/10.1023/A:1022949613039. 106, 534
- Trotman, Andrew, and Shlomo Geva. 2006. Passage retrieval and other XML-retrieval tasks. In *SIGIR 2006 Workshop on XML Element Retrieval Methodology*, pp. 43–50. 217, 525, 534
- Trotman, Andrew, Shlomo Geva, and Jaap Kamps (eds.). 2007. *SIGIR Workshop on Focused Retrieval*. University of Otago. 217, 525, 527, 534
- Trotman, Andrew, Nils Pharo, and Miro Lehtonen. 2006. XML-IR users and use cases. In *Proc. INEX*, pp. 400–412. 216, 528, 531, 534
- Trotman, Andrew, and Börkur Sigurbjörnsson. 2004. Narrowed Extended XPath I (NEXI). In Fuhr et al. (2005), pp. 16–40. DOI: dx.doi.org/10.1007/11424550_2. 217, 532, 534
- Tseng, Huihsin, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *SIGHAN Workshop on Chinese Language Processing*. 46, 521, 523, 527, 529, 534

- Tsochantaridis, Ioannis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. *JMLR* 6:1453–1484. 347, 521, 526, 527, 534
- Turpin, Andrew, and William R. Hersh. 2001. Why batch and user evaluations do not give the same results. In *Proc. SIGIR*, pp. 225–231. 175, 526, 534
- Turpin, Andrew, and William R. Hersh. 2002. User interface effects in past batch versus user experiments. In *Proc. SIGIR*, pp. 431–432. 175, 526, 534
- Turpin, Andrew, Yohannes Tsegay, David Hawking, and Hugh E. Williams. 2007. Fast generation of result snippets in web search. In *Proc. SIGIR*, pp. 127–134. ACM Press. 174, 526, 534, 535
- Turtle, Howard. 1994. Natural language vs. Boolean query evaluation: A comparison of retrieval performance. In *Proc. SIGIR*, pp. 212–220. ACM Press. 15, 534
- Turtle, Howard, and W. Bruce Croft. 1989. Inference networks for document retrieval. In *Proc. SIGIR*, pp. 1–24. ACM Press. 234, 524, 534
- Turtle, Howard, and W. Bruce Croft. 1991. Evaluation of an inference network-based retrieval model. *TOIS* 9(3):187–222. 234, 524, 534
- Turtle, Howard, and James Flood. 1995. Query evaluation: strategies and optimizations. *IP&M* 31(6):831–850. DOI: dx.doi.org/10.1016/0306-4573(95)00020-H. 133, 524, 534
- Vaithyanathan, Shivakumar, and Byron Dom. 2000. Model-based hierarchical clustering. In *Proc. UAI*, pp. 599–608. Morgan Kaufmann. 400, 524, 534
- van Rijsbergen, Cornelis Joost. 1979. *Information Retrieval*, 2nd edition. Butterworths. 173, 216, 221, 231, 235, 531
- van Rijsbergen, Cornelis Joost. 1989. Towards an information logic. In *Proc. SIGIR*, pp. 77–86. ACM Press. DOI: doi.acm.org/10.1145/75334.75344. xxiv, 531
- van Zwol, Roelof, Jeroen Baas, Herre van Oostendorp, and Frans Wiering. 2006. Bricks: The building blocks to tackle query formulation in structured document retrieval. In *Proc. ECIR*, pp. 314–325. 217, 521, 530, 534, 535
- Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley-Interscience. 346, 534
- Vittaut, Jean-Noël, and Patrick Gallinari. 2006. Machine learning ranking for structured information retrieval. In *Proc. ECIR*, pp. 338–349. 216, 525, 534
- Voorhees, Ellen M. 1985a. The cluster hypothesis revisited. In *Proc. SIGIR*, pp. 188–196. ACM Press. 372, 534
- Voorhees, Ellen M. 1985b. The effectiveness and efficiency of agglomerative hierarchical clustering in document retrieval. Technical Report TR 85-705, Cornell. 399, 534
- Voorhees, Ellen M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *IP&M* 36:697–716. 174, 534
- Voorhees, Ellen M., and Donna Harman (eds.). 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press. 173, 314, 499, 511, 526, 534

- Wagner, Robert A., and Michael J. Fischer. 1974. The string-to-string correction problem. *JACM* 21(1):168–173. DOI: doi.acm.org/10.1145/321796.321811. 65, 524, 534
- Ward Jr., J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58:236–244. 399, 534
- Wei, Xing, and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proc. SIGIR*, pp. 178–185. ACM Press. DOI: doi.acm.org/10.1145/1148170.1148204. 418, 524, 534
- Weigend, Andreas S., Erik D. Wiener, and Jan O. Pedersen. 1999. Exploiting hierarchy in text categorization. *IR* 1(3):193–216. 347, 531, 534
- Weston, Jason, and Chris Watkins. 1999. Support vector machines for multi-class pattern recognition. In *Proc. European Symposium on Artificial Neural Networks*, pp. 219–224. 347, 534
- Williams, Hugh E., and Justin Zobel. 2005. Searchable words on the web. *International Journal on Digital Libraries* 5(2):99–105. DOI: dx.doi.org/10.1007/s00799-003-0050-z. 105, 535
- Williams, Hugh E., Justin Zobel, and Dirk Bahle. 2004. Fast phrase querying with combined indexes. *TOIS* 22(4):573–594. 43, 521, 535
- Witten, Ian H., and Timothy C. Bell. 1990. Source models for natural language text. *International Journal Man-Machine Studies* 32(5):545–579. 105, 522, 535
- Witten, Ian H., and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. Morgan Kaufmann. 374, 525, 535
- Witten, Ian H., Alistair Moffat, and Timothy C. Bell. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd edition. Morgan Kaufmann. 18, 83, 105, 106, 522, 530, 535
- Wong, S. K. Michael, Yiyu Yao, and Peter Bollmann. 1988. Linear structure in information retrieval. In *Proc. SIGIR*, pp. 219–232. ACM Press. 348, 522, 535
- Woodley, Alan, and Shlomo Geva. 2006. NLPX at INEX 2006. In *Proc. INEX*, pp. 302–311. 217, 525, 535
- Xu, Jinxi, and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proc. SIGIR*, pp. 4–11. ACM Press. 194, 524, 535
- Xu, Jinxi, and W. Bruce Croft. 1999. Cluster-based language models for distributed retrieval. In *Proc. SIGIR*, pp. 254–261. ACM Press. DOI: doi.acm.org/10.1145/312624.312687. 372, 524, 535
- Yang, Hui, and Jamie Callan. 2006. Near-duplicate detection by instance-level constrained clustering. In *Proc. SIGIR*, pp. 421–428. ACM Press. DOI: doi.acm.org/10.1145/1148170.1148243. 373, 523, 535
- Yang, Yiming. 1994. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proc. SIGIR*, pp. 13–22. ACM Press. 314, 535
- Yang, Yiming. 1999. An evaluation of statistical approaches to text categorization. *IR* 1:69–90. 347, 535

- Yang, Yiming. 2001. A study of thresholding strategies for text categorization. In *Proc. SIGIR*, pp. 137–145. ACM Press. DOI: doi.acm.org/10.1145/383952.383975. 315, 535
- Yang, Yiming, and Bryan Kisiel. 2003. Margin-based local regression for adaptive filtering. In *Proc. CIKM*, pp. 191–198. DOI: doi.acm.org/10.1145/956863.956902. 315, 527, 535
- Yang, Yiming, and Xin Liu. 1999. A re-examination of text categorization methods. In *Proc. SIGIR*, pp. 42–49. ACM Press. 287, 347, 529, 535
- Yang, Yiming, and Jan Pedersen. 1997. Feature selection in statistical learning of text categorization. In *Proc. ICML*. 286, 531, 535
- Yue, Yisong, Thomas Finley, Filip Radlinski, and Thorsten Joachims. 2007. A support vector method for optimizing average precision. In *Proc. SIGIR*. ACM Press. 348, 524, 527, 531, 535
- Zamir, Oren, and Oren Etzioni. 1999. Grouper: A dynamic clustering interface to web search results. In *Proc. WWW*, pp. 1361–1374. Elsevier North-Holland. DOI: dx.doi.org/10.1016/S1389-1286(99)00054-7. 372, 400, 524, 535
- Zaragoza, Hugo, Djoerd Hiemstra, Michael Tipping, and Stephen Robertson. 2003. Bayesian extension to the language model for ad hoc information retrieval. In *Proc. SIGIR*, pp. 4–9. ACM Press. 252, 526, 531, 533, 535
- Zavrel, Jakub, Peter Berck, and Willem Lavrijsen. 2000. Information extraction by text classification: Corpus mining for features. In *Workshop Information Extraction Meets Corpus Linguistics*. URL: www.cnts.ua.ac.be/Publications/2000/ZBL00. Held in conjunction with LREC-2000. 315, 522, 528, 535
- Zha, Hongyuan, Xiaofeng He, Chris H. Q. Ding, Ming Gu, and Horst D. Simon. 2001. Bipartite graph partitioning and data clustering. In *Proc. CIKM*, pp. 25–32. 374, 400, 524, 525, 526, 533, 535
- Zhai, Chengxiang, and John Lafferty. 2001a. Model-based feedback in the language modeling approach to information retrieval. In *Proc. CIKM*. ACM Press. 250, 528, 535
- Zhai, Chengxiang, and John Lafferty. 2001b. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. SIGIR*, pp. 334–342. ACM Press. 252, 528, 535
- Zhai, ChengXiang, and John Lafferty. 2002. Two-stage language models for information retrieval. In *Proc. SIGIR*, pp. 49–56. ACM Press. DOI: doi.acm.org/10.1145/564376.564387. 252, 528, 535
- Zhang, Jiangong, Xiaohui Long, and Torsten Suel. 2007. Performance of compressed inverted list caching in search engines. In *Proc. CIKM*. 106, 529, 533, 535
- Zhang, Tong, and Frank J. Oles. 2001. Text categorization based on regularized linear classification methods. *IR* 4(1):5–31. URL: citeseer.ist.psu.edu/zhang00text.html. 347, 530, 535
- Zhao, Ying, and George Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *Proc. CIKM*, pp. 515–524. ACM Press. DOI: doi.acm.org/10.1145/584792.584877. 399, 527, 535

- Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Addison Wesley. 105, 535
- Zobel, Justin. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proc. SIGIR*, pp. 307–314. 174, 535
- Zobel, Justin, and Philip Dart. 1995. Finding approximate matches in large lexicons. *Software Practice and Experience* 25(3):331–345. URL: cite-seer.ifi.unizh.ch/zobel95finding.html. 65, 524, 535
- Zobel, Justin, and Philip Dart. 1996. Phonetic string matching: Lessons from information retrieval. In *Proc. SIGIR*, pp. 166–173. ACM Press. 65, 524, 535
- Zobel, Justin, and Alistair Moffat. 2006. Inverted files for text search engines. *ACM Computing Surveys* 38(2). 18, 83, 106, 133, 530, 535
- Zobel, Justin, Alistair Moffat, Ross Wilkinson, and Ron Sacks-Davis. 1995. Efficient retrieval of partial documents. *IP&M* 31(3):361–377. DOI: dx.doi.org/10.1016/0306-4573(94)00052-5. 217, 530, 532, 534, 535
- Zukowski, Marcin, Sandor Heman, Niels Nes, and Peter Boncz. 2006. Super-scalar RAM-CPU cache compression. In *Proc. International Conference on Data Engineering*, p. 59. IEEE Computer Society. DOI: dx.doi.org/10.1109/ICDE.2006.150. 106, 522, 526, 530, 535

Author Index

- Aberer: Aberer (2001)
Ahn: Ittner et al. (1995)
Aizerman: Aizerman et al. (1964)
Akaike: Akaike (1974)
Allan: Allan (2005), Allan et al. (1998),
Buckley et al. (1994a), Buckley
et al. (1994b), Salton et al. (1993)
Allwein: Allwein et al. (2000)
Alonso: Alonso et al. (2006)
Altingövd: Can et al. (2004)
Altingövd: Altingövd et al. (2007)
Altun: Tsochantaridis et al. (2005)
Amer-Yahia: Amer-Yahia et al. (2006),
Amer-Yahia et al. (2005),
Amer-Yahia and Lalmas (2006)
Amitay: Mass et al. (2003)
Anagnostopoulos: Anagnostopoulos
et al. (2006)
Anderberg: Anderberg (1973)
Anderson: Burnham and Anderson
(2002)
Andoni: Andoni et al. (2006)
Andrew: Tseng et al. (2005)
Anh: Anh et al. (2001), Anh and
Moffat (2005), Anh and Moffat
(2006a), Anh and Moffat (2006b),
Anh and Moffat (2006c)
Aone: Larsen and Aone (1999)
Apers: Mihajlović et al. (2005)
Apté: Apté et al. (1994)
Arabie: Hubert and Arabie (1985)
Arthur: Arthur and Vassilvitskii
(2006)
Arvola: Arvola et al. (2005)
Aslam: Aslam and Yilmaz (2005)
Ault: Ault and Yang (2002)
Baas: van Zwol et al. (2006)
Badue: Badue et al. (2001)
Baeza-Yates: Badue et al. (2001),
Baeza-Yates et al. (2005),
Baeza-Yates and Ribeiro-Neto
(1999), de Moura et al. (2000),
Frakes and Baeza-Yates (1992),
Harman et al. (1992), Navarro
and Baeza-Yates (1997)
Bahle: Bahle et al. (2002), Williams
et al. (2004)
Bai: Cao et al. (2005)
Bakiri: Dietterich and Bakiri (1995)
Balasubramanyan: Pavlov et al.
(2004)
Baldrige: Baldrige and Osborne
(2004)
Baldwin: Hughes et al. (2006)
Ball: Ball (1965)
Banerjee: Alonso et al. (2006), Basu
et al. (2004)
Banko: Banko and Brill (2001)
Bar-Ilan: Bar-Ilan and Gutman (2005)
Bar-Yossef: Bar-Yossef and Gurevich
(2006)
Barbosa: Ribeiro-Neto and Barbosa
(1998)
Barreiro: Blanco and Barreiro (2006),
Blanco and Barreiro (2007)
Barroso: Barroso et al. (2003)
Bartell: Bartell (1994), Bartell et al.
(1998)

- Barzilay: Barzilay and Elhadad (1997), McKeown et al. (2002)
- Basili: Moschitti and Basili (2004)
- Bast: Bast and Majumdar (2005), Theobald et al. (2008)
- Basu: Basu et al. (2004)
- Bavaud: Picca et al. (2006)
- Beal: Teh et al. (2006)
- Beesley: Beesley (1998), Beesley and Karttunen (2003)
- Belew: Bartell et al. (1998)
- Belkin: Koenemann and Belkin (1996)
- Bell: Moffat and Bell (1995), Witten and Bell (1990), Witten et al. (1999)
- Bennett: Bennett (2000)
- Berck: Zavrel et al. (2000)
- Berger: Berger and Lafferty (1999)
- Berkhin: Berkhin (2005), Berkhin (2006a), Berkhin (2006b)
- Berners-Lee: Berners-Lee et al. (1992)
- Bernstein: Rahm and Bernstein (2001)
- Berry: Berry and Young (1995), Berry et al. (1995), Kent et al. (1955)
- Betsi: Betsi et al. (2006)
- Bhagavathy: Newsam et al. (2001)
- Bharat: Bharat and Broder (1998), Bharat et al. (1998), Bharat et al. (2000), Bharat and Henzinger (1998)
- Bienenstock: Geman et al. (1992)
- Bird: Hughes et al. (2006)
- Bishop: Bishop (2006)
- Blair: Blair and Maron (1985)
- Blair-Goldensohn: Radev et al. (2001)
- Blanco: Blanco and Barreiro (2006), Blanco and Barreiro (2007)
- Blandford: Blandford and Blelloch (2002)
- Blei: Blei et al. (2003), Teh et al. (2006)
- Blelloch: Blandford and Blelloch (2002)
- Blok: List et al. (2005), Mihajlović et al. (2005)
- Blustein: Tague-Sutcliffe and Blustein (1995)
- Boldi: Baeza-Yates et al. (2005), Boldi et al. (2002), Boldi et al. (2005), Boldi and Vigna (2004a), Boldi and Vigna (2004b), Boldi and Vigna (2005)
- Boley: Boley (1998), Savaresi and Boley (2004)
- Bollmann: Wong et al. (1988)
- Boncz: Zukowski et al. (2006)
- Borodin: Borodin et al. (2001)
- Botev: Amer-Yahia et al. (2006)
- Bourne: Bourne and Ford (1961)
- Boyce: Meadow et al. (1999)
- Bracken: Lombard et al. (2002)
- Bradley: Bradley and Fayyad (1998), Bradley et al. (1998), Fayyad et al. (1998)
- Braverman: Aizerman et al. (1964)
- Brill: Banko and Brill (2001), Brill and Moore (2000), Cucerzan and Brill (2004), Richardson et al. (2006)
- Brin: Brin and Page (1998), Page et al. (1998)
- Brisaboa: Brisaboa et al. (2007)
- Broder: Anagnostopoulos et al. (2006), Bharat and Broder (1998), Bharat et al. (1998), Bharat et al. (2000), Broder (2002), Broder et al. (2000), Broder et al. (1997)
- Brown: Brown (1995), Coden et al. (2002)
- Buckley: Buckley et al. (1994a), Buckley and Salton (1995), Buckley et al. (1994b), Buckley et al. (1995), Buckley and Voorhees (2000), Hersh et al. (1994), Salton et al. (1993), Salton and Buckley (1987), Salton and Buckley (1988), Salton and Buckley (1990), Singhal et al. (1996a), Singhal et al. (1997), Singhal et al. (1995), Singhal et al. (1996b)
- Burges: Burges et al. (2005), Burges (1998), Taylor et al. (2006)
- Burner: Burner (1997)

- Burnham: Burnham and Anderson (2002)
- Bush: Bush (1945)
- Büttcher: Büttcher and Clarke (2005a), Büttcher and Clarke (2005b), Büttcher and Clarke (2006), Büttcher et al. (2006)
- Cacheda: Cacheda et al. (2003)
- Cailliau: Berners-Lee et al. (1992)
- Callan: Callan (2000), Lewis et al. (1996), Ogilvie and Callan (2005), Sahoo et al. (2006), Treeratpituk and Callan (2006), Yang and Callan (2006)
- Campbell: Crestani et al. (1998)
- Can: Altingövde et al. (2007), Can et al. (2004), Can and Ozkarahan (1990)
- Candela: Harman and Candela (1990)
- Cannane: Garcia et al. (2004)
- Cao: Cao et al. (2005), Cao et al. (2006), Gao et al. (2004)
- Carbonell: Carbonell and Goldstein (1998)
- Carletta: Carletta (1996)
- Carmel: Carmel et al. (2001), Carmel et al. (2003), Mass et al. (2003)
- Carneiro: Cacheda et al. (2003)
- Caruana: Caruana and Niculescu-Mizil (2006)
- Case: Amer-Yahia et al. (2005)
- Castellan: Siegel and Castellan (1988)
- Castillo: Baeza-Yates et al. (2005)
- Castro: Castro et al. (2004)
- Cavnar: Cavnar and Trenkle (1994)
- Chakrabarti: Chakrabarti (2002), Chakrabarti et al. (1998)
- Chan: Hersh et al. (2000a), Hersh et al. (2001), Hersh et al. (2000b)
- Chang: Sproat et al. (1996), Tseng et al. (2005)
- Chapelle: Chapelle et al. (2006)
- Chaudhuri: Chaudhuri et al. (2006)
- Cheeseman: Cheeseman and Stutz (1996)
- Chen: Chen and Lin (2000), Chen et al. (2005), Cooper et al. (1994), Dumais and Chen (2000), Kishida et al. (2005), Kishida et al. (2005), Kupiec et al. (1995), Liu et al. (2005)
- Cheng: Tan and Cheng (2007)
- Chiaramella: Chiaramella et al. (1996)
- Chierichetti: Chierichetti et al. (2007)
- Cho: Cho and Garcia-Molina (2002), Cho et al. (1998), Ntoulas and Cho (2007)
- Chu-Carroll: Chu-Carroll et al. (2006)
- Church: Kernighan et al. (1990)
- Clarke: Büttcher and Clarke (2005a), Büttcher and Clarke (2005b), Büttcher and Clarke (2006), Büttcher et al. (2006), Clarke et al. (2000)
- Cleverdon: Cleverdon (1991)
- Coates: Castro et al. (2004)
- Cochran: Snedecor and Cochran (1989)
- Coden: Coden et al. (2002)
- Codenotti: Boldi et al. (2002)
- Cohen: Carmel et al. (2001), Cohen (1995), Cohen (1998), Cohen et al. (1998), Cohen and Singer (1999), Forman and Cohen (2004)
- Cole: Spink and Cole (2005)
- Comtet: Comtet (1974)
- Cooper: Cooper et al. (1994)
- Cormack: Clarke et al. (2000)
- Cormen: Cormen et al. (1990)
- Cottrell: Bartell et al. (1998)
- Cover: Cover and Hart (1967), Cover and Thomas (1991)
- Crammer: Crammer and Singer (2001)
- Craswell: Taylor et al. (2006)
- Creedy: Creedy et al. (1992)
- Crestani: Crestani et al. (1998)
- Cristianini: Cristianini and Shawe-Taylor (2000), Lodhi et al. (2002), Shawe-Taylor and Cristianini (2004)
- Croft: Croft (1978), Croft and Harper

- (1979), Croft and Lafferty (2003), Lavrenko and Croft (2001), Liu and Croft (2004), Ponte and Croft (1998), Strohman and Croft (2007), Turtle and Croft (1989), Turtle and Croft (1991), Wei and Croft (2006), Xu and Croft (1996), Xu and Croft (1999)
- Crouch: Crouch (1988)
- Cucerzan: Cucerzan and Brill (2004)
- Curdy: Picca et al. (2006)
- Cutting: Cutting et al. (1993), Cutting et al. (1992)
- Czuba: Chu-Carroll et al. (2006)
- Damerau: Apté et al. (1994), Damerau (1964)
- Dart: Zobel and Dart (1995), Zobel and Dart (1996)
- Das: Chaudhuri et al. (2006)
- Datar: Andoni et al. (2006)
- Davidson: Davidson and Satyanarayana (2003)
- Day: Day and Edelsbrunner (1984)
- Dean: Barroso et al. (2003), Bharat et al. (2000), Dean and Ghemawat (2004)
- Deeds: Burges et al. (2005)
- Deerwester: Deerwester et al. (1990)
- Demir: Can et al. (2004)
- Dempster: Dempster et al. (1977)
- Dhillon: Dhillon (2001), Dhillon and Modha (2001)
- Di Eugenio: Di Eugenio and Glass (2004)
- Dietterich: Dietterich (2002), Dietterich and Bakiri (1995)
- Ding: Zha et al. (2001)
- Dom: Chakrabarti et al. (1998), Dom (2002), Pavlov et al. (2004), Vaithyanathan and Dom (2000)
- Domingos: Domingos (2000), Domingos and Pazzani (1997)
- Dorr: Oard and Dorr (1996)
- Doursat: Geman et al. (1992)
- Downie: Downie (2006)
- Drake: Alonso et al. (2006)
- Dubes: Jain and Dubes (1988)
- Duboue: Chu-Carroll et al. (2006)
- Duda: Duda et al. (2000)
- Dumais: Berry et al. (1995), Deerwester et al. (1990), Dumais et al. (1998), Dumais (1993), Dumais (1995), Dumais and Chen (2000), Littman et al. (1998)
- Duncan: Sahoo et al. (2006)
- Dunning: Dunning (1993), Dunning (1994)
- Dörre: Amer-Yahia et al. (2006)
- Eckart: Eckart and Young (1936)
- Edelsbrunner: Day and Edelsbrunner (1984)
- Eisenberg: Schamber et al. (1990)
- Eissen: Stein and zu Eissen (2004), Stein et al. (2003)
- El-Hamdouchi: El-Hamdouchi and Willett (1986)
- Elhadad: Barzilay and Elhadad (1997)
- Elias: Elias (1975)
- Elkan: Hamerly and Elkan (2003)
- Emerson: Sproat and Emerson (2003)
- Etzioni: Zamir and Etzioni (1999)
- Evans: McKeown et al. (2002)
- Eyheramendy: Eyheramendy et al. (2003)
- Fagin: Carmel et al. (2001)
- Fallows: Fallows (2004)
- Farchi: Carmel et al. (2001)
- Fariña: Brisaboa et al. (2007)
- Fayyad: Bradley and Fayyad (1998), Bradley et al. (1998), Fayyad et al. (1998)
- Feldmann: Kammenhuber et al. (2006)
- Fellbaum: Fellbaum (1998)
- Ferragina: Ferragina and Venturini (2007)
- Ferrucci: Chu-Carroll et al. (2006)
- Finley: Yue et al. (2007)
- Fischer: Wagner and Fischer (1974)
- Flach: Gaertner et al. (2002)
- Flake: Glover et al. (2002b)
- Flood: Turtle and Flood (1995)

- Flynn: Jain et al. (1999)
 Ford: Bourne and Ford (1961)
 Forman: Forman (2004), Forman (2006), Forman and Cohen (2004)
 Fourel: Chiaramella et al. (1996)
 Fowlkes: Fowlkes and Mallows (1983)
 Fox: Fox and Lee (1991), Harman et al. (1992), Lee and Fox (1988)
 Fraenkel: Fraenkel and Klein (1985)
 Frakes: Frakes and Baeza-Yates (1992)
 Fraley: Fraley and Raftery (1998)
 Frank: Witten and Frank (2005)
 Frei: Qiu and Frei (1993)
 Frieder: Grossman and Frieder (2004)
 Friedl: Friedl (2006)
 Friedman: Friedman (1997), Friedman and Goldszmidt (1996), Hastie et al. (2001)
 Fuhr: Fuhr (1989), Fuhr (1992), Fuhr et al. (2003a), Fuhr and Großjohann (2004), Fuhr and Lalmas (2007), Fuhr et al. (2006), Fuhr et al. (2005), Fuhr et al. (2007), Fuhr et al. (2003b), Fuhr and Pfeifer (1994), Fuhr and Rölleke (1997)
 Furnas: Deerwester et al. (1990)
 Gaertner: Gaertner et al. (2002)
 Gale: Kernighan et al. (1990), Sproat et al. (1996)
 Gallinari: Vittaut and Gallinari (2006)
 Gao: Gao et al. (2005), Gao et al. (2004)
 Garcia: Garcia et al. (2004)
 Garcia-Molina: Cho and Garcia-Molina (2002), Cho et al. (1998), Garcia-Molina et al. (1999), Hirai et al. (2000), Melnik et al. (2001), Tomasic and Garcia-Molina (1993)
 Garfield: Garfield (1955), Garfield (1976)
 Gay: Joachims et al. (2005)
 Geman: Geman et al. (1992)
 Geng: Geng et al. (2007)
 Gerrand: Gerrand (2007)
 Geva: Tannier and Geva (2005), Trotman and Geva (2006), Trotman et al. (2007), Woodley and Geva (2006)
 Gey: Cooper et al. (1994), Gey (1994)
 Ghamrawi: Ghamrawi and McCallum (2005)
 Ghemawat: Dean and Ghemawat (2004)
 Gibson: Chakrabarti et al. (1998)
 Giles: Lawrence and Giles (1998), Lawrence and Giles (1999), Rusmevichientong et al. (2001)
 Glass: Di Eugenio and Glass (2004)
 Glassman: Broder et al. (1997)
 Glover: Glover et al. (2002a), Glover et al. (2002b)
 Goldstein: Carbonell and Goldstein (1998)
 Goldszmidt: Friedman and Goldszmidt (1996)
 Grabs: Grabs and Schek (2002)
 Graepel: Herbrich et al. (2000)
 Granka: Joachims et al. (2005)
 Gravano: Hatzivassiloglou et al. (2000)
 Greiff: Greiff (1998)
 Griffiths: Rosen-Zvi et al. (2004)
 Grinstead: Grinstead and Snell (1997)
 Groff: Berners-Lee et al. (1992)
 Grossman: Grossman and Frieder (2004)
 Großjohann: Fuhr and Großjohann (2004)
 Gu: Zha et al. (2001)
 Guerrero: CACHED et al. (2003)
 Gupta: Smeulders et al. (2000)
 Gurevich: Bar-Yossef and Gurevich (2006)
 Gusfield: Gusfield (1997)
 Gutman: Bar-Ilan and Gutman (2005)
 Gövert: Fuhr et al. (2003a), Gövert and Kazai (2003)
 Hamerly: Hamerly and Elkan (2003)
 Hamilton: Burges et al. (2005)

- Han: Han and Karypis (2000)
 Hand: Hand (2006), Hand and Yu (2001)
 Harman: Harman (1991), Harman (1992), Harman et al. (1992), Harman and Candela (1990), Voorhees and Harman (2005)
 Harold: Harold and Means (2004)
 Harper: Croft and Harper (1979), Muresan and Harper (2004)
 Harshman: Deerwester et al. (1990)
 Hart: Cover and Hart (1967), Duda et al. (2000)
 Harter: Harter (1998)
 Hartigan: Hartigan and Wong (1979)
 Hastie: Hastie et al. (2001), Tibshirani et al. (2001)
 Hatzivassiloglou: Hatzivassiloglou et al. (2000), McKeown et al. (2002)
 Haveliwala: Haveliwala (2003), Haveliwala (2002)
 Hawking: Turpin et al. (2007)
 Hayes: Hayes and Weinstein (1990)
 He: Zha et al. (2001)
 Heaps: Heaps (1978)
 Hearst: Hearst (1997), Hearst (2006), Hearst and Pedersen (1996), Hearst and Plaunt (1993)
 Heckerman: Dumaïs et al. (1998)
 Heinz: Heinz and Zobel (2003), Heinz et al. (2002)
 Heman: Zukowski et al. (2006)
 Hembrooke: Joachims et al. (2005)
 Henzinger: Bharat et al. (1998), Bharat et al. (2000), Bharat and Henzinger (1998), Henzinger et al. (2000), Silverstein et al. (1999)
 Herbrich: Herbrich et al. (2000)
 Herscovici: Carmel et al. (2001)
 Hersh: Hersh et al. (1994), Hersh et al. (2000a), Hersh et al. (2001), Hersh et al. (2000b), Turpin and Hersh (2001), Turpin and Hersh (2002)
 Heydon: Henzinger et al. (2000), Najork and Heydon (2001), Najork and Heydon (2002)
 Hickam: Hersh et al. (1994)
 Hiemstra: Hiemstra (1998), Hiemstra (2000), Hiemstra and Kraaij (2005), Kraaij et al. (2002), List et al. (2005), Mihajlović et al. (2005), Zaragoza et al. (2003)
 Hirai: Hirai et al. (2000)
 Hofmann: Hofmann (1999a), Hofmann (1999b), Tsochantaridis et al. (2005)
 Hollink: Hollink et al. (2004)
 Hon: Cao et al. (2006)
 Hopcroft: Hopcroft et al. (2000)
 Hristidis: Chaudhuri et al. (2006)
 Huang: Cao et al. (2006), Gao et al. (2005), Huang and Mitchell (2006)
 Hubert: Hubert and Arabie (1985)
 Hughes: Hughes et al. (2006)
 Hull: Hull (1993), Hull (1996), Schütze et al. (1995)
 Hullender: Burges et al. (2005)
 Hölzle: Barroso et al. (2003)
 Ide: Ide (1971)
 Immorlica: Andoni et al. (2006)
 Indyk: Andoni et al. (2006), Indyk (2004)
 Ingwersen: Ingwersen and Järvelin (2005)
 Isahara: Murata et al. (2000)
 Ittner: Ittner et al. (1995)
 Ittycheriah: Lita et al. (2003)
 Iwayama: Iwayama and Tokunaga (1995)
 Järvelin: Ingwersen and Järvelin (2005)
 Jackson: Jackson and Moulinier (2002)
 Jacobs: Jacobs and Rau (1990)
 Jain: Jain et al. (1999), Jain and Dubes (1988), Smeulders et al. (2000)
 Jansen: Spink et al. (2000)

- Jardine: Jardine and van Rijsbergen (1971)
- Jeh: Jeh and Widom (2003)
- Jensen: Jensen and Jensen (2001), Jensen and Jensen (2001)
- Jeong: Jeong and Omiecinski (1995)
- Ji: Ji and Xu (2006)
- Jing: Jing (2000)
- Joachims: Joachims (1997), Joachims (1998), Joachims (1999), Joachims (2002a), Joachims (2002b), Joachims (2006a), Joachims (2006b), Joachims et al. (2005), Tsochantaridis et al. (2005), Yue et al. (2007)
- Johnson: Johnson et al. (2006)
- Jones: Lewis and Jones (1996), Robertson and Jones (1976), Spärck Jones (1972), Spärck Jones (2004), Spärck Jones et al. (2000)
- Jordan: Blei et al. (2003), Ng and Jordan (2001), Ng et al. (2001a), Ng et al. (2001b), Teh et al. (2006)
- Jr: Kent et al. (1955)
- Junkkari: Arvola et al. (2005)
- Jurafsky: Jurafsky and Martin (2008), Tseng et al. (2005)
- Järvelin: Järvelin and Kekäläinen (2002), Kekäläinen and Järvelin (2002)
- Kalita: Kołcz et al. (2000)
- Kambhatla: Lita et al. (2003)
- Kammenhuber: Kammenhuber et al. (2006)
- Kamps: Hollink et al. (2004), Kamps et al. (2004), Kamps et al. (2006), Lalmas et al. (2007), Sigurbjörnsson et al. (2004), Trotman et al. (2007)
- Kamvar: Kamvar et al. (2002)
- Kando: Kishida et al. (2005)
- Kannan: Kannan et al. (2000)
- Kantor: Saracevic and Kantor (1988), Saracevic and Kantor (1996)
- Kapur: Pavlov et al. (2004)
- Karger: Cutting et al. (1993), Cutting et al. (1992), Rennie et al. (2003)
- Karttunen: Beesley and Karttunen (2003)
- Karypis: Han and Karypis (2000), Steinbach et al. (2000), Zhao and Karypis (2002)
- Kaszkiel: Kaszkiel and Zobel (1997)
- Kataoka: Toda and Kataoka (2005)
- Kaufman: Kaufman and Rousseeuw (1990)
- Kazai: Fuhr et al. (2003a), Fuhr et al. (2006), Gövert and Kazai (2003), Kazai and Lalmas (2006), Lalmas et al. (2007)
- Keerthi: Sindhvani and Keerthi (2006)
- Kekäläinen: Arvola et al. (2005), Järvelin and Kekäläinen (2002), Kekäläinen (2005), Kekäläinen and Järvelin (2002)
- Kemeny: Kemeny and Snell (1976)
- Kent: Kent et al. (1955)
- Kernighan: Kernighan et al. (1990)
- Khachiyani: Kozlov et al. (1979)
- King: King (1967)
- Kishida: Kishida et al. (2005)
- Kisiel: Yang and Kisiel (2003)
- Klavans: McKeown et al. (2002)
- Klein: Fraenkel and Klein (1985), Kamvar et al. (2002), Klein and Manning (2002)
- Kleinberg: Chakrabarti et al. (1998), Kleinberg (1997), Kleinberg (1999), Kleinberg (2002)
- Knuth: Knuth (1997)
- Ko: Ko et al. (2004)
- Koenemann: Koenemann and Belkin (1996)
- Koller: Koller and Sahami (1997), Tong and Koller (2001)
- Konheim: Konheim (1981)
- Korfhage: Korfhage (1997)
- Kozlov: Kozlov et al. (1979)
- Kołcz: Kołcz et al. (2000), Kołcz and Yih (2007)

- Kraaij: Hiemstra and Kraaij (2005),
Kraaij and Spitters (2003), Kraaij
et al. (2002)
- Kraemer: Hersh et al. (2000a), Hersh
et al. (2001), Hersh et al. (2000b)
- Kraft: Meadow et al. (1999)
- Kretser: Anh et al. (2001)
- Krippendorff: Krippendorff (2003)
- Krishnan: McLachlan and Krishnan
(1996), Sahoo et al. (2006)
- Krovetz: Glover et al. (2002a),
Krovetz (1995)
- Kuhns: Maron and Kuhns (1960)
- Kukich: Kukich (1992)
- Kumar: Bharat et al. (1998), Broder
et al. (2000), Kumar et al. (1999),
Kumar et al. (2000), Steinbach
et al. (2000)
- Kupiec: Kupiec et al. (1995)
- Kuriyama: Kishida et al. (2005)
- Kurland: Kurland and Lee (2004)
- Kwok: Luk and Kwok (2002)
- Käki: Käki (2005)
- Lacker: Perkins et al. (2003)
- Lafferty: Berger and Lafferty (1999),
Croft and Lafferty (2003),
Lafferty and Zhai (2001), Lafferty
and Zhai (2003), Zhai and
Lafferty (2001a), Zhai and
Lafferty (2001b), Zhai and
Lafferty (2002)
- Lai: Qin et al. (2007)
- Laird: Dempster et al. (1977)
- Lalmas: Amer-Yahia and Lalmas
(2006), Betsi et al. (2006), Crestani
et al. (1998), Fuhr et al. (2003a),
Fuhr and Lalmas (2007), Fuhr
et al. (2006), Fuhr et al. (2005),
Fuhr et al. (2007), Fuhr et al.
(2003b), Kazai and Lalmas
(2006), Lalmas et al. (2007),
Lalmas and Tombros (2007),
Ruthven and Lalmas (2003)
- Lance: Lance and Williams (1967)
- Landauer: Deerwester et al. (1990),
Littman et al. (1998)
- Langville: Langville and Meyer
(2006)
- Larsen: Larsen and Aone (1999)
- Larson: Larson (2005)
- Lavrenko: Allan et al. (1998),
Lavrenko and Croft (2001)
- Lavrijssen: Zavrel et al. (2000)
- Lawrence: Glover et al. (2002a),
Glover et al. (2002b), Lawrence
and Giles (1998), Lawrence and
Giles (1999), Rusmevichientong
et al. (2001)
- Lazier: Burges et al. (2005)
- Lee: Fox and Lee (1991), Harman
et al. (1992), Kishida et al. (2005),
Kurland and Lee (2004), Lee and
Fox (1988)
- Leek: Miller et al. (1999)
- Lehtonen: Trotman et al. (2006)
- Leiserson: Cormen et al. (1990)
- Lempel: Lempel and Moran (2000)
- Leone: Hersh et al. (1994)
- Lesk: Lesk (1988), Lesk (2004)
- Lester: Lester et al. (2005), Lester
et al. (2006)
- Levenshtein: Levenshtein (1965)
- Lew: Lew (2001)
- Lewis: Eyheramendy et al. (2003),
Ittner et al. (1995), Lewis (1995),
Lewis (1998), Lewis and Jones
(1996), Lewis and Ringuette
(1994), Lewis et al. (1996), Lewis
et al. (2004)
- Li: Cao et al. (2006), Gao et al. (2005),
Geng et al. (2007), Lewis et al.
(2004), Li and Yang (2003), Qin
et al. (2007)
- Liddy: Liddy (2005)
- Lin: Chen and Lin (2000), Chen et al.
(2005)
- List: List et al. (2005)
- Lita: Lita et al. (2003)
- Littman: Littman et al. (1998)
- Liu: Cao et al. (2006), Geng et al.
(2007), Liu et al. (2005), Liu and
Croft (2004), Qin et al. (2007),

- Riezler et al. (2007), Yang and Liu (1999)
- Lloyd: Gaertner et al. (2002), Lloyd (1982)
- Lodhi: Lodhi et al. (2002)
- Lombard: Lombard et al. (2002)
- Long: Long and Suel (2003), Zhang et al. (2007)
- Lovins: Lovins (1968)
- Lu: Lu et al. (2007)
- Luehrs: Kent et al. (1955)
- Luhn: Luhn (1957), Luhn (1958)
- Luk: Luk and Kwok (2002)
- Lunde: Lunde (1998)
- Lushman: Büttcher et al. (2006)
- Luxenburger: Kammenhuber et al. (2006)
- Ma: Liu et al. (2005), Murata et al. (2000), Song et al. (2005)
- Maarek: Carmel et al. (2001), Carmel et al. (2003), Mass et al. (2003)
- MacFarlane: Lu et al. (2007), MacFarlane et al. (2000)
- MacKinlay: Hughes et al. (2006)
- MacQueen: MacQueen (1967)
- Madigan: Eyheramendy et al. (2003)
- Maganti: Hatzivassiloglou et al. (2000)
- Maghoul: Broder et al. (2000)
- Mahabhashyam: Singitham et al. (2004)
- Majumdar: Bast and Majumdar (2005), Theobald et al. (2008)
- Malhotra: Johnson et al. (2006)
- Malik: Fuhr et al. (2006), Fuhr et al. (2005), Fuhr et al. (2003b)
- Mallows: Fowlkes and Mallows (1983)
- Manasse: Broder et al. (1997)
- Mandelbrod: Carmel et al. (2003), Mass et al. (2003)
- Manjunath: Newsam et al. (2001)
- Manning: Kamvar et al. (2002), Klein and Manning (2002), Manning and Schütze (1999), Tseng et al. (2005)
- Marais: Silverstein et al. (1999)
- Maron: Blair and Maron (1985), Maron and Kuhns (1960)
- Martin: Jurafsky and Martin (2008)
- Marx: Kamps et al. (2006)
- Masand: Creecy et al. (1992)
- Mass: Carmel et al. (2003), Mass et al. (2003)
- McBryan: McBryan (1994)
- McCallum: Ghamrawi and McCallum (2005), McCallum and Nigam (1998), McCallum et al. (1998), McCallum (1996), Nigam et al. (2006)
- McCann: MacFarlane et al. (2000)
- McKeown: McKeown and Radev (1995), McKeown et al. (2002)
- McLachlan: McLachlan and Krishnan (1996)
- Meadow: Meadow et al. (1999)
- Means: Harold and Means (2004)
- Mei: Tao et al. (2006)
- Meilă: Meilă (2005)
- Melnik: Melnik et al. (2001)
- Meuss: Schlieder and Meuss (2002)
- Meyer: Langville and Meyer (2006)
- Mihajlović: Mihajlović et al. (2005)
- Mihajlovic: List et al. (2005)
- Miller: Miller et al. (1999)
- Minsky: Minsky and Papert (1988)
- Mirroknî: Andoni et al. (2006)
- Mitchell: Huang and Mitchell (2006), McCallum et al. (1998), Mitchell (1997), Nigam et al. (2006)
- Mitra: Buckley et al. (1995), Singhal et al. (1996a), Singhal et al. (1997)
- Mittal: Riezler et al. (2007)
- Mitzenmacher: Henzinger et al. (2000)
- Modha: Dhillon and Modha (2001)
- Moffat: Anh et al. (2001), Anh and Moffat (2005), Anh and Moffat (2006a), Anh and Moffat (2006b), Anh and Moffat (2006c), Lester et al. (2005), Moffat and Bell (1995), Moffat and Stuiver (1996),

- Moffat and Zobel (1992), Moffat and Zobel (1996), Moffat and Zobel (1998), Witten et al. (1999), Zobel and Moffat (2006), Zobel et al. (1995)
- Monz: Hollink et al. (2004)
- Mooers: Mooers (1961), Mooers (1950)
- Mooney: Basu et al. (2004)
- Moore: Brill and Moore (2000), Pelleg and Moore (1999), Pelleg and Moore (2000), Toutanova and Moore (2002)
- Moran: Lempel and Moran (2000)
- Moricz: Silverstein et al. (1999)
- Moschitti: Moschitti (2003), Moschitti and Basili (2004)
- Motwani: Hopcroft et al. (2000), Page et al. (1998)
- Moulinier: Jackson and Moulinier (2002)
- Moura: de Moura et al. (2000), Ribeiro-Neto et al. (1999)
- Mulhem: Chiaramella et al. (1996)
- Murata: Murata et al. (2000)
- Muresan: Muresan and Harper (2004)
- Murtagh: Murtagh (1983)
- Murty: Jain et al. (1999)
- Myaeng: Kishida et al. (2005)
- Najork: Henzinger et al. (2000), Najork and Heydon (2001), Najork and Heydon (2002)
- Narin: Pinski and Narin (1976)
- Navarro: Brisaboa et al. (2007), de Moura et al. (2000), Navarro and Baeza-Yates (1997)
- Nenkova: McKeown et al. (2002)
- Nes: Zukowski et al. (2006)
- Neubert: Ribeiro-Neto et al. (1999)
- Newsam: Newsam et al. (2001)
- Ng: Blei et al. (2003), McCallum et al. (1998), Ng and Jordan (2001), Ng et al. (2001a), Ng et al. (2001b)
- Nicholson: Hughes et al. (2006)
- Niculescu-Mizil: Caruana and Niculescu-Mizil (2006)
- Nie: Cao et al. (2005), Gao et al. (2004)
- Nigam: McCallum and Nigam (1998), Nigam et al. (2006)
- Nilan: Schamber et al. (1990)
- Nowak: Castro et al. (2004)
- Ntoulas: Ntoulas and Cho (2007)
- O'Brien: Berry et al. (1995)
- O'Keefe: O'Keefe and Trotman (2004)
- Oard: Oard and Dorr (1996)
- Obermayer: Herbrich et al. (2000)
- Ocalan: Altingövde et al. (2007)
- Ogilvie: Ogilvie and Callan (2005)
- Oles: Zhang and Oles (2001)
- Olson: Hersh et al. (2000a), Hersh et al. (2001), Hersh et al. (2000b)
- Omiecinski: Jeong and Omiecinski (1995)
- Oostendorp: van Zwol et al. (2006)
- Orlando: Silvestri et al. (2004)
- Osborne: Baldridge and Osborne (2004)
- Osiński: Osiński and Weiss (2005)
- Ozaku: Murata et al. (2000)
- Ozcan: Altingövde et al. (2007)
- Ozkarahan: Can and Ozkarahan (1990)
- Ozmultu: Spink et al. (2000)
- Padman: Sahoo et al. (2006)
- Paepcke: Hirai et al. (2000)
- Page: Brin and Page (1998), Cho et al. (1998), Page et al. (1998)
- Paice: Paice (1990)
- Pan: Joachims et al. (2005)
- Panconesi: Chierichetti et al. (2007)
- Papert: Minsky and Papert (1988)
- Papineni: Papineni (2001)
- Papka: Allan et al. (1998), Lewis et al. (1996)
- Paramá: Brisaboa et al. (2007)
- Parikh: Pavlov et al. (2004)
- Park: Ko et al. (2004)
- Pavlov: Pavlov et al. (2004)
- Pazzani: Domingos and Pazzani (1997)
- Pedersen: Cutting et al. (1993), Cutting et al. (1992), Hearst and

- Pedersen (1996), Kupiec et al. (1995), Schütze et al. (1995), Schütze and Pedersen (1995), Weigend et al. (1999), Yang and Pedersen (1997)
- Pehcevski: Lalmas et al. (2007)
- Pelleg: Pelleg and Moore (1999), Pelleg and Moore (2000)
- Pennock: Glover et al. (2002a), Glover et al. (2002b), Rusmevichientong et al. (2001)
- Perego: Silvestri et al. (2004)
- Perkins: Perkins et al. (2003)
- Perry: Kent et al. (1955)
- Persin: Persin (1994), Persin et al. (1996)
- Peterson: Peterson (1980)
- Pfeifer: Fuhr and Pfeifer (1994)
- Pharo: Trotman et al. (2006)
- Picca: Picca et al. (2006)
- Pinski: Pinski and Narin (1976)
- Pirolli: Pirolli (2007)
- Piwowarski: Lalmas et al. (2007)
- Platt: Dumais et al. (1998), Platt (2000)
- Plaunt: Hearst and Plaunt (1993)
- Pollermann: Berners-Lee et al. (1992)
- Ponte: Ponte and Croft (1998)
- Popescul: Popescul and Ungar (2000)
- Porter: Porter (1980)
- Prabakarmurthi: Kołcz et al. (2000)
- Prager: Chu-Carroll et al. (2006)
- Prakash: Richardson et al. (2006)
- Price: Hersh et al. (2000a), Hersh et al. (2001), Hersh et al. (2000b)
- Pugh: Pugh (1990)
- Punera: Anagnostopoulos et al. (2006)
- Qin: Geng et al. (2007), Qin et al. (2007)
- Qiu: Qiu and Frei (1993)
- R Development Core Team: R Development Core Team (2005)
- Radev: McKeown and Radev (1995), Radev et al. (2001)
- Radlinski: Yue et al. (2007)
- Raftery: Fraley and Raftery (1998)
- Raghavan: Broder et al. (2000), Chakrabarti et al. (1998), Chierichetti et al. (2007), Hirai et al. (2000), Kumar et al. (1999), Kumar et al. (2000), Melnik et al. (2001), Radev et al. (2001), Singitham et al. (2004)
- Rahm: Rahm and Bernstein (2001)
- Rajagopalan: Broder et al. (2000), Chakrabarti et al. (1998), Kumar et al. (1999), Kumar et al. (2000)
- Ramírez: List et al. (2005)
- Rand: Rand (1971)
- Rasmussen: Rasmussen (1992)
- Rau: Jacobs and Rau (1990)
- Reina: Bradley et al. (1998), Fayyad et al. (1998)
- Rennie: Rennie et al. (2003)
- Renshaw: Burges et al. (2005)
- Ribeiro-Neto: Badue et al. (2001), Baeza-Yates and Ribeiro-Neto (1999), Ribeiro-Neto et al. (1999), Ribeiro-Neto and Barbosa (1998)
- Rice: Rice (2006)
- Richardson: Richardson et al. (2006)
- Riezler: Riezler et al. (2007)
- Rijke: Hollink et al. (2004), Kamps et al. (2004), Kamps et al. (2006), Sigurbjörnsson et al. (2004)
- Rijsbergen: Crestani et al. (1998), Jardine and van Rijsbergen (1971), Tombros et al. (2002), van Rijsbergen (1979), van Rijsbergen (1989)
- Ringuette: Lewis and Ringuette (1994)
- Ripley: Ripley (1996)
- Rivest: Cormen et al. (1990)
- Roberts: Borodin et al. (2001)
- Robertson: Lalmas et al. (2007), Lu et al. (2007), MacFarlane et al. (2000), Robertson (2005), Robertson et al. (2004), Robertson and Jones (1976), Spärck Jones et al. (2000), Taylor et al. (2006), Zaragoza et al. (2003)

- Rocchio: Rocchio (1971)
 Roget: Roget (1946)
 Rose: Lewis et al. (2004)
 Rosen-Zvi: Rosen-Zvi et al. (2004)
 Rosenfeld: McCallum et al. (1998)
 Rosenthal: Borodin et al. (2001)
 Ross: Ross (2006)
 Roukos: Lita et al. (2003)
 Rousseeuw: Kaufman and Rousseeuw (1990)
 Rozonoér: Aizerman et al. (1964)
 Rubin: Dempster et al. (1977)
 Rusmevichientong:
 Rusmevichientong et al. (2001)
 Ruthven: Ruthven and Lalmas (2003)
 Rölleke: Amer-Yahia et al. (2005), Fuhr and Rölleke (1997)
 Sable: McKeown et al. (2002)
 Sacherek: Hersh et al. (2000a), Hersh et al. (2001), Hersh et al. (2000b)
 Sacks-Davis: Persin et al. (1996), Zobel et al. (1995)
 Sahami: Dumais et al. (1998), Koller and Sahami (1997)
 Sahoo: Sahoo et al. (2006)
 Sakai: Sakai (2007)
 Salton: Buckley et al. (1994a), Buckley and Salton (1995), Buckley et al. (1994b), Salton (1971a), Salton (1971b), Salton (1975), Salton (1989), Salton (1991), Salton et al. (1993), Salton and Buckley (1987), Salton and Buckley (1988), Salton and Buckley (1990), Singhal et al. (1995), Singhal et al. (1996b)
 Sanderson: Tombros and Sanderson (1998)
 Santini: Boldi et al. (2002), Boldi et al. (2005), Smeulders et al. (2000)
 Saracevic: Saracevic and Kantor (1988), Saracevic and Kantor (1996)
 Satyanarayana: Davidson and Satyanarayana (2003)
 Saunders: Lodhi et al. (2002)
 Savaresi: Savaresi and Boley (2004)
 Schamber: Schamber et al. (1990)
 Schapire: Allwein et al. (2000), Cohen et al. (1998), Lewis et al. (1996), Schapire (2003), Schapire and Singer (2000), Schapire et al. (1998)
 Schek: Grabs and Schek (2002)
 Schenkel: Theobald et al. (2008), Theobald et al. (2005)
 Schiffman: McKeown et al. (2002)
 Schlieder: Schlieder and Meuss (2002)
 Scholer: Scholer et al. (2002)
 Schwartz: Miller et al. (1999)
 Schwarz: Schwarz (1978)
 Schölkopf: Chen et al. (2005), Schölkopf and Smola (2001)
 Schütze: Manning and Schütze (1999), Schütze (1998), Schütze et al. (1995), Schütze and Pedersen (1995), Schütze and Silverstein (1997)
 Sebastiani: Sebastiani (2002)
 Seo: Ko et al. (2004)
 Shaked: Burges et al. (2005)
 Shanmugasundaram: Amer-Yahia et al. (2006), Amer-Yahia et al. (2005)
 Shawe-Taylor: Cristianini and Shawe-Taylor (2000), Lodhi et al. (2002), Shawe-Taylor and Cristianini (2004)
 Shih: Rennie et al. (2003), Sproat et al. (1996)
 Shkapenyuk: Shkapenyuk and Suel (2002)
 Siegel: Siegel and Castellan (1988)
 Sifry: Sifry (2007)
 Sigelman: McKeown et al. (2002)
 Sigurbjörnsson: Kamps et al. (2004), Kamps et al. (2006), Sigurbjörnsson et al. (2004), Trotman and Sigurbjörnsson (2004)
 Silverstein: Schütze and Silverstein (1997), Silverstein et al. (1999)

- Silvestri: Silvestri (2007), Silvestri et al. (2004)
- Simon: Zha et al. (2001)
- Sindhwani: Sindhwani and Keerthi (2006)
- Singer: Allwein et al. (2000), Cohen et al. (1998), Cohen and Singer (1999), Crammer and Singer (2001), Schapire and Singer (2000), Schapire et al. (1998)
- Singhal: Buckley et al. (1995), Schapire et al. (1998), Singhal et al. (1996a), Singhal et al. (1997), Singhal et al. (1995), Singhal et al. (1996b)
- Singitham: Singitham et al. (2004)
- Sivakumar: Kumar et al. (2000)
- Slonim: Tishby and Slonim (2000)
- Smeulders: Smeulders et al. (2000)
- Smith: Creecy et al. (1992)
- Smola: Schölkopf and Smola (2001)
- Smyth: Rosen-Zvi et al. (2004)
- Sneath: Sneath and Sokal (1973)
- Snedecor: Snedecor and Cochran (1989)
- Snell: Grinstead and Snell (1997), Kemeny and Snell (1976)
- Snyder-Duch: Lombard et al. (2002)
- Soffer: Carmel et al. (2001), Carmel et al. (2003), Mass et al. (2003)
- Sokal: Sneath and Sokal (1973)
- Somogyi: Somogyi (1990)
- Song: Song et al. (2005)
- Sornil: Sornil (2001)
- Sozio: Chierichetti et al. (2007)
- Spink: Spink and Cole (2005), Spink et al. (2000)
- Spitters: Kraaij and Spitters (2003)
- Sproat: Sproat and Emerson (2003), Sproat et al. (1996), Sproat (1992)
- Srinivasan: Coden et al. (2002)
- Stata: Broder et al. (2000)
- Stein: Stein and zu Eissen (2004), Stein et al. (2003)
- Steinbach: Steinbach et al. (2000)
- Steyvers: Rosen-Zvi et al. (2004)
- Stork: Duda et al. (2000)
- Strang: Strang (1986)
- Strehl: Strehl (2002)
- Strohman: Strohman and Croft (2007)
- Stuiver: Moffat and Stuiver (1996)
- Stutz: Cheeseman and Stutz (1996)
- Suel: Long and Suel (2003), Shkapenyuk and Suel (2002), Zhang et al. (2007)
- Swanson: Swanson (1988)
- Szlávik: Fuhr et al. (2005)
- Tague-Sutcliffe: Tague-Sutcliffe and Blustein (1995)
- Tan: Tan and Cheng (2007)
- Tannier: Tannier and Geva (2005)
- Tao: Tao et al. (2006)
- Tarasov: Kozlov et al. (1979)
- Taube: Taube and Wooster (1958)
- Taylor: Robertson et al. (2004), Taylor et al. (2006)
- Teevan: Rennie et al. (2003)
- Teh: Teh et al. (2006)
- Theiler: Perkins et al. (2003)
- Theobald: Theobald et al. (2008), Theobald et al. (2005)
- Thomas: Cover and Thomas (1991)
- Tiberi: Chierichetti et al. (2007)
- Tibshirani: Hastie et al. (2001), Tibshirani et al. (2001)
- Tipping: Zaragoza et al. (2003)
- Tishby: Tishby and Slonim (2000)
- Toda: Toda and Kataoka (2005)
- Tokunaga: Iwayama and Tokunaga (1995)
- Tomasic: Tomasic and Garcia-Molina (1993)
- Tombros: Betsi et al. (2006), Lalmas and Tombros (2007), Tombros and Sanderson (1998), Tombros et al. (2002)
- Tomkins: Broder et al. (2000), Kumar et al. (1999), Kumar et al. (2000)
- Tomlinson: Tomlinson (2003)
- Tong: Tong and Koller (2001)
- Toutanova: Toutanova and Moore (2002)

- Treeratpituk: Treeratpituk and Callan (2006)
- Trenkle: Cavnar and Trenkle (1994)
- Trotman: Fuhr et al. (2007), O'Keefe and Trotman (2004), Trotman (2003), Trotman and Geva (2006), Trotman et al. (2007), Trotman et al. (2006), Trotman and Sigurbjörnsson (2004)
- Tsaparas: Borodin et al. (2001)
- Tsegay: Turpin et al. (2007)
- Tseng: Tseng et al. (2005)
- Tsikrika: Betsi et al. (2006)
- Tsioutsoulis: Glover et al. (2002b)
- Tsochantaridis: Riezler et al. (2007), Tsochantaridis et al. (2005)
- Tudhope: Clarke et al. (2000)
- Tukey: Cutting et al. (1992)
- Turpin: Hersh et al. (2000a), Hersh et al. (2001), Hersh et al. (2000b), Turpin and Hersh (2001), Turpin and Hersh (2002), Turpin et al. (2007)
- Turtle: Turtle (1994), Turtle and Croft (1989), Turtle and Croft (1991), Turtle and Flood (1995)
- Uchimoto: Murata et al. (2000)
- Ullman: Garcia-Molina et al. (1999), Hopcroft et al. (2000)
- Ulusoy: Altingövd et al. (2007)
- Ungar: Popescu and Ungar (2000)
- Upfal: Chierichetti et al. (2007), Kumar et al. (2000)
- Utiyama: Murata et al. (2000)
- Vaithyanathan: Vaithyanathan and Dom (2000)
- Vamplew: Johnson et al. (2006)
- Vapnik: Vapnik (1998)
- Vasserman: Riezler et al. (2007)
- Vassilvitskii: Arthur and Vassilvitskii (2006)
- Vempala: Kannan et al. (2000)
- Venkatasubramanian: Bharat et al. (1998)
- Venturini: Ferragina and Venturini (2007)
- Veta: Kannan et al. (2000)
- Vigna: Boldi et al. (2002), Boldi et al. (2005), Boldi and Vigna (2004a), Boldi and Vigna (2004b), Boldi and Vigna (2005)
- Villa: Tombros et al. (2002)
- Vittaut: Vittaut and Gallinari (2006)
- Viña: Cacheda et al. (2003)
- Voorhees: Buckley and Voorhees (2000), Voorhees (1985a), Voorhees (1985b), Voorhees (2000), Voorhees and Harman (2005)
- Vries: List et al. (2005)
- Wagner: Wagner and Fischer (1974)
- Walker: Spärck Jones et al. (2000)
- Walther: Tibshirani et al. (2001)
- Waltz: Creecy et al. (1992)
- Wan: Liu et al. (2005)
- Wang: Qin et al. (2007), Tao et al. (2006)
- Ward Jr.: Ward Jr. (1963)
- Watkins: Lodhi et al. (2002), Weston and Watkins (1999)
- Wei: Wei and Croft (2006)
- Weigend: Weigend et al. (1999)
- Weikum: Amer-Yahia et al. (2005), Chaudhuri et al. (2006), Kammenhuber et al. (2006), Theobald et al. (2008), Theobald et al. (2005)
- Weinstein: Hayes and Weinstein (1990)
- Weiss: Apté et al. (1994), Ng et al. (2001a), Osifski and Weiss (2005)
- Wen: Song et al. (2005)
- Westerveld: Kraaij et al. (2002)
- Weston: Weston and Watkins (1999)
- Widom: Garcia-Molina et al. (1999), Jeh and Widom (2003)
- Wiener: Broder et al. (2000), Weigend et al. (1999)
- Wiering: van Zwol et al. (2006)
- Wilkinson: Zobel et al. (1995)
- Willett: El-Hamdouchi and Willett (1986)

- Williams: Bahle et al. (2002), Garcia et al. (2004), Heinz et al. (2002), Lance and Williams (1967), Lester et al. (2006), Scholer et al. (2002), Turpin et al. (2007), Williams and Zobel (2005), Williams et al. (2004)
- Winograd: Page et al. (1998)
- Witten: Witten and Bell (1990), Witten and Frank (2005), Witten et al. (1999)
- Wißbrock: Stein et al. (2003)
- Wong: Hartigan and Wong (1979), Wong et al. (1988)
- Woodley: Woodley and Geva (2006)
- Wooster: Taube and Wooster (1958)
- Worring: Smeulders et al. (2000)
- Wu: Gao et al. (2005), Gao et al. (2004)
- Xu: Cao et al. (2006), Ji and Xu (2006), Xu and Croft (1996), Xu and Croft (1999)
- Yang: Ault and Yang (2002), Lewis et al. (2004), Li and Yang (2003), Liu et al. (2005), Melnik et al. (2001), Yang and Callan (2006), Yang (1994), Yang (1999), Yang (2001), Yang and Kisiel (2003), Yang and Liu (1999), Yang and Pedersen (1997)
- Yao: Wong et al. (1988)
- Yiannis: Scholer et al. (2002)
- Yih: Kolcz and Yih (2007)
- Yilmaz: Aslam and Yilmaz (2005)
- Young: Berry and Young (1995), Eckart and Young (1936)
- Yu: Hand and Yu (2001)
- Yue: Yue et al. (2007)
- Zamir: Zamir and Etzioni (1999)
- Zaragoza: Robertson et al. (2004), Taylor et al. (2006), Zaragoza et al. (2003)
- Zavrel: Zavrel et al. (2000)
- Zeng: Liu et al. (2005)
- Zha: Zha et al. (2001)
- Zhai: Lafferty and Zhai (2001), Lafferty and Zhai (2003), Tao et al. (2006), Zhai and Lafferty (2001a), Zhai and Lafferty (2001b), Zhai and Lafferty (2002)
- Zhang: Qin et al. (2007), Radev et al. (2001), Zhang et al. (2007), Zhang and Oles (2001)
- Zhao: Zhao and Karypis (2002)
- Zheng: Ng et al. (2001b)
- Zien: Chapelle et al. (2006)
- Zipf: Zipf (1949)
- Ziviani: Badue et al. (2001), de Moura et al. (2000), Ribeiro-Neto et al. (1999)
- Zobel: Bahle et al. (2002), Heinz and Zobel (2003), Heinz et al. (2002), Kaszkiel and Zobel (1997), Lester et al. (2005), Lester et al. (2006), Moffat and Zobel (1992), Moffat and Zobel (1996), Moffat and Zobel (1998), Persin et al. (1996), Scholer et al. (2002), Williams and Zobel (2005), Williams et al. (2004), Zobel (1998), Zobel and Dart (1995), Zobel and Dart (1996), Zobel and Moffat (2006), Zobel et al. (1995)
- Zukowski: Zukowski et al. (2006)
- Zweig: Broder et al. (1997)
- Zwol: van Zwol et al. (2006)
- del Bimbo: del Bimbo (1999)

Index

- L_2 distance, 131
- χ^2 feature selection, 275
- δ codes, 104
- γ encoding, 99
- k nearest neighbor classification, 297
- k -gram index, 54, 60
- 1/0 loss, 221
- 11-point interpolated average
 - precision, 159
- 20 Newsgroups, 154
- A/B test, 170
- access control lists, 81
- accumulator, 113, 125
- accuracy, 155
- active learning, 336
- ad hoc retrieval, 5, 253
- add-one smoothing, 260
- adjacency table, 455
- adversarial information retrieval, 429
- Akaike Information Criterion, 367
- algorithmic search, 430
- anchor text, 425
- any-of classification, 257, 306
- authority score, 474
- auxiliary index, 78
- average-link clustering, 389
- B-tree, 50
- bag of words, 117, 267
- bag-of-words, 269
- balanced F measure, 156
- Bayes error rate, 300
- Bayes Optimal Decision Rule, 222
- Bayes risk, 222
- Bayes' Rule, 220
- Bayesian networks, 234
- Bayesian prior, 226
- Bernoulli model, 263
- best-merge persistence, 388
- bias, 311
- bias-variance tradeoff, 241, 312, 321
- biclustering, 374
- bigram language model, 240
- Binary Independence Model, 222
- binary tree, 50, 377
- biword index, 39, 43
- blind relevance feedback, *see* pseudo relevance feedback
- blocked sort-based indexing
 - algorithm, 71
- blocked storage, 92
- blog, 195
- BM25 weights, 232
- boosting, 286
- bottom-up clustering, *see* hierarchical agglomerative clustering
- bowtie, 426
- break-even, 334
- break-even point, 161
- BSBI, 71
- Buckshot algorithm, 399
- buffer, 69
- caching, 9, 68, 146, 447, 450
- capture-recapture method, 435
- cardinality
 - in clustering, 355
- CAS topics, 211
- case-folding, 30

- category, 256
- centroid, 292, 360
 - in relevance feedback, 181
- centroid-based classification, 314
- chain rule, 220
- chaining
 - in clustering, 385
- champion lists, 143
- class boundary, 303
- classification, 253, 344
- classification function, 256
- classifier, 183
- CLEF, 154
- click spam, 431
- clickstream mining, 170, 188
- clickthrough log analysis, 170
- clique, 384
- cluster, 74, 349
 - in relevance feedback, 184
- cluster hypothesis, 350
- cluster-based classification, 314
- cluster-internal labeling, 396
- CO topics, 211
- co-clustering, 374
- collection, 4
- collection frequency, 27
- combination similarity, 378, 384, 393
- complete-link clustering, 382
- complete-linkage clustering, *see* complete-link clustering
- component coverage, 212
- compound-splitter, 25
- compounds, 25
- concept drift, 269, 283, 286, 336
- conditional independence
 - assumption, 224, 266
- confusion matrix, 307
- connected component, 384
- connectivity queries, 455
- connectivity server, 455
- content management system, 84
- context
 - XML, 199
- context resemblance, 208
- contiguity hypothesis, 289
- continuation bit, 96
- corpus, 4
- cosine similarity, 121, 372
- CPC, 430
- CPM, 430
- Cranfield, 153
- cross-entropy, 251
- cross-language information retrieval, 154, 417
- cumulative gain, 162
- data-centric XML, 196, 214
- database
 - relational, 1, 195, 214
- decision boundary, 292, 303
- decision hyperplane, 290, 302
- decision trees, 282, 286
- dendrogram, 378
- development set, 283
- development test collection, 153
- Dice coefficient, 163
- dictionary, 6, 7
- differential cluster labeling, 396
- digital libraries, 195
- distortion, 366
- distributed index, 74, 458
- distributed indexing, 74
- distributed information retrieval, *see* distributed crawling, 458
- divisive clustering, 395
- DNS resolution, 450
- DNS server, 450
- docID, 7
- document, 4, 20
- document collection, *see* collection
- document frequency, 7, 118
- document likelihood model, 250
- document partitioning, 454
- document space, 256
- document vector, 119, 120
- document-at-a-time, 126, 140
- document-partitioned index, 75
- dot product, 121
- East Asian languages, 45
- edit distance, 58
- effectiveness, 5, 280
- eigen decomposition, 406

- eigenvalue, **404**
- EM algorithm, **369**
- email sorting, **254**
- enterprise resource planning, **84**
- enterprise search, **67**
- entropy, **99, 106, 358**
- equivalence classes, **28**
- Ergodic Markov Chain, **467**
- Euclidean distance, **131, 372**
- Euclidean length, **121**
- evidence accumulation, **146**
- exclusive clustering, **355**
- exhaustive clustering, **355**
- expectation step, **370**
- Expectation-Maximization algorithm, **336, 369**
- expected edge density, **373**
- extended query, **205**
- Extensible Markup Language, **196**
- external criterion of quality, **356**
- external sorting algorithm, **70**

- F measure, **156, 173**
 - as an evaluation measure in clustering, **359**
- false negative, **359**
- false positive, **359**
- feature engineering, **338**
- feature selection, **271**
- field, **110**
- filtering, **253, 314**
- first story detection, **395, 399**
- flat clustering, **350**
- focused retrieval, **217**
- free text, **109, 148**
- free text query, *see* query, free text, **124, 145, 196**
- frequency-based feature selection, **277**
- Frobenius norm, **410**
- front coding, **93**
- functional margin, **322**

- GAAC, **388**
- generative model, **237, 309, 311**
- geometric margin, **323**
- gold standard, **152**
- Golomb codes, **106**

- GOV2, **154**
- greedy feature selection, **279**
- grep, **3**
- ground truth, **152**
- group-average agglomerative clustering, **388**
- group-average clustering, **389**

- HAC, **378**
- hard assignment, **350**
- hard clustering, **350, 355**
- harmonic number, **101**
- Heaps' law, **88**
- held-out, **298**
- held-out data, **283**
- hierarchic clustering, **377**
- hierarchical agglomerative clustering, **378**
- hierarchical classification, **337, 347**
- hierarchical clustering, **350, 377**
- Hierarchical Dirichlet Processes, **418**
- hierarchy
 - in clustering, **377**
- highlighting, **203**
- HITS, **477**
- HTML, **421**
- http, **421**
- hub score, **474**
- hyphens, **24**

- i.i.d., **283**, *see* independent and identically distributed
- Ide dec-hi, **183**
- idf, **83, 204, 227, 232**
- iid, *see* independent and identically distributed
- impact, **81**
- implicit relevance feedback, **187**
- in-links, **425, 461**
- incidence matrix, **3, 408**
- independence, **275**
- independent and identically distributed, **283**
 - in clustering, **367**
- index, **3**, *see* permuterm index, *see also* parametric index, zone index
- index construction, **67**

- indexer, **67**
- indexing, **67**
 - sort-based, **7**
- indexing granularity, **21**
- indexing unit, **201**
- INEX, **210**
- information gain, **285**
- information need, **5, 152**
- information retrieval, **1**
- informational queries, **432**
- inner product, **121**
- instance-based learning, **300**
- inter-similarity, **381**
- internal criterion of quality, **356**
- interpolated precision, **158**
- intersection
 - postings list, **10**
- inverse document frequency, **118, 125**
- inversion, **71, 378, 391**
- inverted file, *see* inverted index
- inverted index, **6**
- inverted list, *see* postings list
- inverter, **76**
- IP address, **449**

- Jaccard coefficient, **61, 438**

- K-medoids, **365**
- kappa statistic, **165, 174, 373**
- kernel, **332**
- kernel function, **332**
- kernel trick, **331**
- key-value pairs, **75**
- keyword-in-context, **171**
- kNN classification, **297**
- Kruskal's algorithm, **399**
- Kullback-Leibler divergence, **251, 317, 372**
- KWIC, *see* keyword-in-context

- label, **256**
- labeling, **255**
- language, **237**
- language identification, **24, 46**
- language model, **238**
- Laplace smoothing, **260**
- Latent Dirichlet Allocation, **418**

- latent semantic indexing, **192, 413**
- LDA, **418**
- learning algorithm, **256**
- learning error, **310**
- learning method, **256**
- lemma, **32**
- lemmatization, **32**
- lemmatizer, **33**
- length-normalization, **121**
- Levenshtein distance, **58**
- lexicalized subtree, **206**
- lexicon, **6**
- likelihood, **221**
- likelihood ratio, **239**
- linear classifier, **301, 343**
- linear problem, **303**
- linear separability, **304**
- link farms, **481**
- link spam, **429, 461**
- LM, **243**
- logarithmic merging, **79**
- lossless, **87**
- lossy compression, **87**
- low-rank approximation, **410**
- LSA, **413**
- LSI as soft clustering, **417**

- machine translation, **240, 243, 251**
- machine-learned relevance, **113, 342**
- macroaveraging, **280**
- MAP, **159, 227, 258**
- map phase, **75**
- MapReduce, **75**
- margin, **320**
- marginal relevance, **167**
- marginal statistic, **165**
- master node, **75**
- matrix decomposition, **406**
- maximization step, **370**
- maximum a posteriori, **227, 265**
- maximum a posteriori class, **258**
- maximum likelihood estimate, **226, 259**
- maximum likelihood estimation, **244**
- Mean Average Precision, *see* MAP
- medoid, **365**
- memory capacity, **312**

- memory-based learning, 300
- Mercator, 445
- Mercer kernel, 332
- merge
 - postings, 10
- merge algorithm, 10
- metadata, 24, 110, 171, 197, 373, 428
- microaveraging, 280
- minimum spanning tree, 399, 401
- minimum variance clustering, 399
- MLE, *see* maximum likelihood estimate
- ModApte split, 279, 286
- model complexity, 312, 366
- model-based clustering, 368
- monotonicity, 378
- multiclass classification, 306
- multiclass SVM, 347
- multilabel classification, 306
- multimodal class, 296
- multinomial classification, 306
- multinomial distribution, 241
- multinomial model, 263, 270
- multinomial Naive Bayes, 258
- multinomial NB, *see* multinomial Naive Bayes
- multivalue classification, 306
- multivariate Bernoulli model, 263
- mutual information, 272, 358

- Naive Bayes assumption, 224
- named entity tagging, 195, 339
- National Institute of Standards and Technology, 153
- natural language processing, xxxiv, 33, 171, 217, 249, 372
- navigational queries, 432
- NDCG, 163
- nested elements, 203
- NEXI, 200
- next word index, 44
- nibble, 98
- NLP, *see* natural language processing
- NMI, 358
- noise document, 303
- noise feature, 271
- nonlinear classifier, 305
- nonlinear problem, 305
- normal vector, 293
- normalized discounted cumulative gain, 163
- normalized mutual information, 358
- novelty detection, 395
- NTCIR, 154, 174

- objective function, 354, 360
- odds, 221
- odds ratio, 225
- Okapi weighting, 232
- one-of classification, 257, 284, 306
- optimal classifier, 270, 310
- optimal clustering, 393
- optimal learning method, 310
- ordinal regression, 344
- out-links, 425
- outlier, 363
- overfitting, 271, 312

- PageRank, 464
- paid inclusion, 428
- parameter tuning, 153, 314, 315, 348
- parameter tying, 340
- parameter-free compression, 100
- parameterized compression, 106
- parametric index, 110
- parametric search, 197
- parser, 75
- partition rule, 220
- partitional clustering, 355
- passage retrieval, 217
- patent databases, 195
- perceptron algorithm, 286, 315
- performance, 280
- permuterm index, 53
- personalized PageRank, 471
- phrase index, 40
- phrase queries, 39, 47
- phrase search, 15
- pivoted document length
 - normalization, 129
- pointwise mutual information, 286
- polychotomous, 306
- polytomous classification, 306
- polytope, 298

- pooling, 164, 174
- pornography filtering, 338
- Porter stemmer, 33
- positional independence, 267
- positional index, 41
- posterior probability, 220
- posting, 6, 7, 71, 86
- postings list, 6
- power law, 89, 426
- precision, 5, 155
- precision at k , 161
- precision-recall curve, 158
- prefix-free code, 100
- principal direction divisive partitioning, 400
- principal left eigenvector, 465
- prior probability, 220
- Probability Ranking Principle, 221
- probability vector, 466
- prototype, 290
- proximity operator, 14
- proximity weighting, 145
- pseudo relevance feedback, 187
- pseudocounts, 226
- pull model, 314
- purity, 356
- push model, 314
- Quadratic Programming, 324
- query, 5
 - free text, 14, 16, 117
 - simple conjunctive, 10
- query expansion, 189
- query likelihood model, 242
- query optimization, 11
- query-by-example, 201, 249
- R-precision, 161, 174
- Rand index, 359
 - adjusted, 373
- random variable, 220
- random variable C , 268
- random variable U , 266
- random variable X , 266
- rank, 403
- Ranked Boolean retrieval, 112
- ranked retrieval, 81, 107
 - model, 14
- ranking SVM, 345
- recall, 5, 155
- reduce phase, 75
- reduced SVD, 409, 412
- regression, 344
- regular expressions, 3, 18
- regularization, 328
- relational database, 195, 214
- relative frequency, 226
- relevance, 5, 152
- relevance feedback, 178
- residual sum of squares, 360
- results snippets, 146
- retrieval model
 - Boolean, 4
- Retrieval Status Value, 225
- retrieval systems, 81
- Reuters-21578, 154
- Reuters-RCV1, 69, 154
- RF, 178
- Robots Exclusion Protocol, 447
- ROC curve, 162
- Rocchio algorithm, 181
- Rocchio classification, 292
- routing, 253, 314
- RSS, 360
- rule of 30, 86
- rules in text classification, 255
- Scatter-Gather, 351
- schema, 199
- schema diversity, 204
- schema heterogeneity, 204
- search advertising, 430
- search engine marketing, 431
- Search Engine Optimizers, 429
- search result clustering, 351
- search results, 351
- security, 81
- seed, 361
- seek time, 68
- segment file, 75
- semi-supervised learning, 336
- semistructured query, 197
- semistructured retrieval, 2, 197
- sensitivity, 162

- sentiment detection, **254**
- sequence model, **267**
- shingling, **438**
- single-label classification, **306**
- single-link clustering, **382**
- single-linkage clustering, *see*
single-link clustering
- single-pass in-memory indexing, **73**
- singleton, **378**
- singleton cluster, **363**
- singular value decomposition, **407**
- skip list, **36, 46**
- slack variables, **327**
- SMART, **182**
- smoothing, **127, 226**
 - add α , **226**
 - add $\frac{1}{2}$, **232**
 - add $\frac{1}{2}$, **226–229, 262**
 - Bayesian prior, **226, 228, 245**
 - linear interpolation, **245**
- snippet, **170**
- soft assignment, **350**
- soft clustering, **350, 355, 377**
- sorting
 - in index construction, **7**
- soundex, **63**
- spam, **338, 427**
 - email, **254**
 - web, **254**
- sparseness, **241, 244, 260**
- specificity, **162**
- spectral clustering, **400**
- speech recognition, **240**
- spelling correction, **147, 240, 242**
- spider, **443**
- spider traps, **433**
- SPIMI, **73**
- splits, **75**
- sponsored search, **430**
- standing query, **253**
- static quality scores, **138**
- static web pages, **424**
- statistical significance, **276**
- statistical text classification, **255**
- steady-state, **467, 468**
- stemming, **32, 46**
- stochastic matrix, **465**
- stop words, **117**
- stop list, **27**
 - stop words, **117**
- stop words, **23, 27, 45, 127**
- structural SVM, **345**
- structural SVMs, **330**
- structural term, **207**
- structured document retrieval
 - principle, **201**
- structured query, **197**
- structured retrieval, **195, 197**
- summarization, **400**
- summary
 - dynamic, **171**
 - static, **171**
- supervised learning, **256**
- support vector, **320**
- support vector machine, **319, 346**
 - multiclass, **330**
- SVD, **373, 400, 408**
- SVM, *see* support vector machine
- symmetric diagonal decomposition, **407, 408**
- synonymy, **177**
- teleport, **464**
- term, **3, 19, 22**
- term frequency, **16, 117**
- term normalization, **28**
- term partitioning, **454**
- term-at-a-time, **125, 140**
- term-document matrix, **123**
- term-partitioned index, **74**
- termID, **69**
- test data, **256**
- test set, **256, 283**
- text categorization, **253**
- text classification, **253**
- text summarization, **171**
- text-centric XML, **214**
- tf, *see* term frequency
- tf-idf, **119**
- tiered indexes, **143**
- token, **19, 22**
- token normalization, **28**
- top docs, **149**

- top-down clustering, 395
- topic, 153, 253
 - in XML retrieval, 211
- topic classification, 253
- topic spotting, 253
- topic-specific PageRank, 471
- topical relevance, 212
- training set, 256, 283
- transactional query, 433
- transductive SVMs, 336
- translation model, 251
- TREC, 153, 314
- trec_eval, 174
- truecasing, 30, 46
- truncated SVD, 409, 412, 415
- two-class classifier, 279
- type, 22

- unary code, 99
- unigram language model, 240
- union-find algorithm, 395, 440
- universal code, 100
- unsupervised learning, 349
- URL, 422
- URL normalization, 447
- utility measure, 286

- variable byte encoding, 96
- variance, 311
- vector space model, 120
- vertical search engine, 254
- vocabulary, 6
- Voronoi tessellation, 297

- Ward's method, 399
- web crawler, 443
- weight vector, 322
- weighted zone scoring, 110
- Wikipedia, 211
- wildcard query, 3, 49, 52
- within-point scatter, 375
- word segmentation, 25

- XML, 20, 196
- XML attribute, 197
- XML DOM, 197
- XML DTD, 199
- XML element, 197
- XML fragment, 216
- XML Schema, 199
- XML tag, 197
- XPath, 199

- Zipf's law, 89
- zone, 110, 337, 339, 340
- zone index, 110
- zone search, 197