

## 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 Conference on Information and Knowledge Management.  
*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.

- 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. [319]
- Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE Transactions on automatic control* 19(6):716–723. [345]
- Allan, James. 2005. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proc. TREC*. [160]
- Allan, James, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proc. SIGIR*, pp. 37–45. ACM Press. doi: <http://doi.acm.org/10.1145/290941.290954>. [367]
- Allwein, Erin L., Robert E. Schapire, and Yoram Singer. 2000. Reducing multi-class to binary: A unifying approach for margin classifiers. *JMLR* 1:113–141. URL: [www.jmlr.org/papers/volume1/allwein00a/allwein00a.pdf](http://www.jmlr.org/papers/volume1/allwein00a/allwein00a.pdf). [292]
- 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: <http://doi.acm.org/10.1145/1135777.1135913>. [344]
- Altıngövd, İsmail Sengör, Engin Demir, Fazlı Can, and Özgür Ulusoy. 2008. Incremental cluster-based retrieval using compressed cluster-skipping inverted files. *TOIS*. To appear. 372
- Amer-Yahia, Sihem, Chavdar Botev, Jochen Dörre, and Jayavel Shanmugasundaram. 2006. XQuery full-text extensions explained. *IBM Systems Journal* 45(2):335–352. [200]
- 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: <http://doi.acm.org/10.1145/1107499.1107514>. [200]
- Amer-Yahia, Sihem, and Mounia Lalmas. 2006. XML search: Languages, INEX and scoring. *SIGMOD Record* 35(4):16–23. doi: <http://doi.acm.org/10.1145/1228268.1228271>. [200]
- 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: <http://doi.acm.org/10.1145/1183614.1183648>. [292]
- Anderberg, Michael R. 1973. *Cluster analysis for applications*. Academic Press. [344]
- 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. [291]
- 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. [137]
- Anh, Vo Ngoc, and Alistair Moffat. 2005. Inverted index compression using word-aligned binary codes. *IR* 8(1):151–166. doi: <http://dx.doi.org/10.1023/B:INRT.0000048490.99518.5c>. [98]
- 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. [98]
- Anh, Vo Ngoc, and Alistair Moffat. 2006b. Pruned query evaluation using pre-computed impacts. In *Proc. SIGIR*, pp. 372–379. ACM Press. doi: <http://doi.acm.org/10.1145/1148170.1148235>. [137]
- Anh, Vo Ngoc, and Alistair Moffat. 2006c. Structured index organizations for high-throughput text querying. In *Proc. SPIRE*, pp. 304–315. Springer. [138]
- Apté, Chidanand, Fred Damerau, and Sholom M. Weiss. 1994. Automated learning of decision rules for text categorization. *TOIS* 12(1):233–251. [265]
- Arthur, David, and Sergei Vassilvitskii. 2006. How slow is the  $k$ -means method? In *Proc. Symposium on Computational Geometry*, pp. 144–153. [345]
- Arvola, Paavo, Marko Junkkari, and Jaana Kekäläinen. 2005. Generalized contextualization method for XML information retrieval. In *Proc. CIKM*, pp. 20–27. ACM Press. [199]

## Bibliography

443

- Aslam, Javed A., and Emine Yilmaz. 2005. A geometric interpretation and analysis of R-precision. In *Proc. CIKM*, pp. 664–671. ACM Press. [160]
- Ault, Thomas Galen, and Yiming Yang. 2002. Information filtering in TREC-9 and TDT-3: A comparative analysis. *IR* 5(2-3):159–187. [292]
- 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. [420]
- 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. [439]
- Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley. [xviii, 76, 97, 161, 368]
- Bahle, Dirk, Hugh E. Williams, and Justin Zobel. 2002. Efficient phrase querying with an auxiliary index. In *Proc. SIGIR*, pp. 215–221. ACM Press. [44]
- Baldrige, Jason, and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proc. EMNLP*, pp. 9–16. [320]
- 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. [345]
- Banko, Michele, and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proc. ACL*. [309]
- 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. [43]
- Bar-Yossef, Ziv, and Maxim Gurevich. 2006. Random sampling from a search engine's index. In *Proc. WWW*, pp. 367–376. ACM Press. doi: <http://doi.acm.org/10.1145/1135777.1135833>. [404]
- Barroso, Luiz André, Jeffrey Dean, and Urs Hölzle. 2003. Web search for a planet: The Google cluster architecture. *IEEE Micro* 23(2):22–28. <http://dx.doi.org/10.1109/MM.2003.1196112>. [420]
- 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. [138]
- Bartell, Brian T., Garrison W. Cottrell, and Richard K. Belew. 1998. Optimizing similarity using multi-query relevance feedback. *JASIS* 49(8):742–761. [138]
- Barzilay, Regina, and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Workshop on Intelligent Scalable Text Summarization*, pp. 10–17. [161]
- Bast, Holger, and Debapriyo Majumdar. 2005. Why spectral retrieval works. In *Proc. SIGIR*, pp. 11–18. ACM Press. doi: <http://doi.acm.org/10.1145/1076034.1076040>. [384]
- 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. [344]
- Beesley, Kenneth R. 1998. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Languages at Crossroads: Proceedings of the Annual Conference of the American Translators Association*, pp. 47–54. [43]
- Beesley, Kenneth R., and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications. [43]
- 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. [265]
- Berger, Adam, and John Lafferty. 1999. Information retrieval as statistical translation. In *Proc. SIGIR*, pp. 222–229. ACM Press. [232]
- Berkhin, Pavel. 2005. A survey on pagerank computing. *Internet Mathematics* 2(1):73–120. [439]

- Berkhin, Pavel. 2006a. Bookmark-coloring algorithm for personalized pagerank computing. *Internet Mathematics* 3(1):41–62. [439]
- 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. [343]
- 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](http://citeseer.ist.psu.edu/article/berners-lee92worldwide.html). [404]
- Berry, Michael, and Paul Young. 1995. Using latent semantic indexing for multilanguage information retrieval. *Computers and the Humanities* 29(6):413–429. [384]
- 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. [383]
- Betsi, Stamatina, Mounia Lalmas, Anastasios Tombros, and Theodora Tsikrika. 2006. User expectations from XML element retrieval. In *Proc. SIGIR*, pp. 611–612. ACM Press. [199]
- 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: [http://dx.doi.org/10.1016/S0169-7552\(98\)00127-5](http://dx.doi.org/10.1016/S0169-7552(98)00127-5). [404]
- 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. [420]
- 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](http://citeseer.ist.psu.edu/bharat99comparison.html). [404]
- 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](http://citeseer.ist.psu.edu/bharat98improved.html). [439]
- Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer. [292]
- 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. [177]
- Blanco, Roi, and Alvaro Barreiro. 2006. TSP and cluster-based solutions to the reassignment of document identifiers. *IR* 9(4):499–517. [98]
- Blanco, Roi, and Alvaro Barreiro. 2007. Boosting static pruning of inverted files. In *Proc. SIGIR*. ACM Press. [97]
- Blandford, Dan, and Guy Blelloch. 2002. Index compression through document reordering. In *Proc. Data Compression Conference*, p. 342. IEEE Computer Society. [98]
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR* 3:993–1022. [384]
- 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](http://citeseer.ist.psu.edu/article/boldi03ubicrawler.html). [419]
- 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](http://citeseer.ist.psu.edu/boldi05pagerank.html). [439]
- Boldi, Paolo, and Sebastiano Vigna. 2004a. Codes for the World-Wide Web. *Internet Mathematics* 2(4):405–427. [420]
- Boldi, Paolo, and Sebastiano Vigna. 2004b. The WebGraph framework I: Compression techniques. In *Proc. WWW*, pp. 595–601. ACM Press. [420]
- Boldi, Paolo, and Sebastiano Vigna. 2005. Compressed perfect embedded skip lists for quick inverted-index lookups. In *Proc. SPIRE*. Springer. [44]
- Boley, Daniel. 1998. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery* 2(4):325–344. DOI: <http://dx.doi.org/10.1023/A:1009740529316>. [368]

## Bibliography

445

- 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. [439]
- 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: <http://doi.acm.org/10.1145/321088.321094>. [60]
- Bradley, Paul S., and Usama M. Fayyad. 1998. Refining initial points for k-means clustering. In *Proc. ICML*, pp. 91–99. [345]
- Bradley, Paul S., Usama M. Fayyad, and Cory Reina. 1998. Scaling clustering algorithms to large databases. In *Proc. KDD*, pp. 9–15. [345]
- Brill, Eric, and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proc. ACL*, pp. 286–293. [60]
- Brin, Sergey, and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proc. WWW*, pp. 107–117. [137, 419, 439]
- Brisaboa, Nieves R., Antonio Fariña, Gonzalo Navarro, and José R. Paramá. 2007. Lightweight natural language text compression. *IR* 10(1):1–33. [99]
- Broder, Andrei. 2002. A taxonomy of web search. *SIGIR Forum* 36(2):3–10. doi: <http://doi.acm.org/10.1145/792550.792552>. [404]
- 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. [404]
- Broder, Andrei Z., Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. In *Proc. WWW*, pp. 391–404. [404]
- Brown, Eric W. 1995. *Execution Performance Issues in Full-Text Information Retrieval*. PhD thesis, University of Massachusetts, Amherst. [137]
- Buckley, Chris, James Allan, and Gerard Salton. 1994a. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In *Proc. TREC*, pp. 45–55.
- Buckley, Chris, and Gerard Salton. 1995. Optimization of relevance feedback weights. In *Proc. SIGIR*, pp. 351–357. ACM Press. doi: <http://doi.acm.org/10.1145/215206.215383>. [292]
- 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. [170, 177]
- Buckley, Chris, Amit Singhal, and Mandar Mitra. 1995. New retrieval approaches using SMART: TREC 4. In *Proc. TREC*. [172]
- Buckley, Chris, and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proc. SIGIR*, pp. 33–40. [160]
- 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*. [320]
- Burges, Christopher J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2):121–167. [318]
- Burner, Mike. 1997. Crawling towards eternity: Building an archive of the World Wide Web. *Web Techniques Magazine* 2(5). [419]
- Burnham, Kenneth P., and David Anderson. 2002. *Model Selection and Multi-Model Inference*. Springer. [345]
- Bush, Vannevar. 1945. As we may think. *The Atlantic Monthly*. URL: [www.theatlantic.com/doc/194507/bush](http://www.theatlantic.com/doc/194507/bush). [16, 404]
- 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: <http://doi.acm.org/10.1145/1099554.1099645>. [76]
- Büttcher, Stefan, and Charles L. A. Clarke. 2005b. A security model for full-text file system search in multi-user environments. In *FAST*. URL: [www.usenix.org/events/fast05/tech/buettcher.html](http://www.usenix.org/events/fast05/tech/buettcher.html). [77]

- 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. ACM Press. doi: <http://doi.acm.org/10.1145/1183614.1183644>. [97]
- 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: <http://doi.acm.org/10.1145/1148170.1148233>. [76]
- 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. [344]
- Callan, Jamie. 2000. Distributed information retrieval. In W. Bruce Croft (ed.), *Advances in information retrieval*, pp. 127–150. Kluwer. [76]
- 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: [http://dx.doi.org/10.1016/S0306-4379\(03\)00062-0](http://dx.doi.org/10.1016/S0306-4379(03)00062-0). [344]
- Can, Fazli, and Esen A. Ozkaran. 1990. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. *ACM Trans. Database Syst.* 15(4):483–517. [344]
- Cao, Guihong, Jian-Yun Nie, and Jing Bai. 2005. Integrating word relationships into language models. In *Proc. SIGIR*, pp. 298–305. ACM Press.
- 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. [320]
- 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: <http://doi.acm.org/10.1145/290941.291025>. [154]
- Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22:249–254. [160]
- 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: <http://doi.acm.org/10.1145/383952.383958>. [97, 138]
- 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: <http://doi.acm.org/10.1145/860435.860464>. [199]
- Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proc. ICML*. [319]
- Castro, R. M., M. J. Coates, and R. D. Nowak. 2004. Likelihood based hierarchical clustering. *IEEE Transactions in Signal Processing* 52(8):2308–2321. [368]
- Cavnar, William B., and John M. Trenkle. 1994. N-gram-based text categorization. In *Proc. SDAIR*, pp. 161–175. [43]
- Chakrabarti, Soumen. 2002. *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufman. [404]
- 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](http://cite-seer.ist.psu.edu/chakrabarti98automatic.html). [439]
- Chapelle, Olivier, Bernhard Schölkopf, and Alexander Zien (eds.). 2006. *Semi-Supervised Learning*. MIT Press. [319, 459]
- Chaudhuri, Surajit, Gautam Das, Vagelis Hristidis, and Gerhard Weikum. 2006. Probabilistic information retrieval approach for ranking of database query results. *ACM Trans. Database Syst.* 31(3):1134–1168. doi: <http://doi.acm.org/10.1145/1166074.1166085>. [200]
- 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. [345]

## Bibliography

447

- Chen, Hsin-Hsi, and Chuan-Jie Lin. 2000. A multilingual news summarizer. In *Proc. COLING*, pp. 159–165. [344]
- Chen, Pai-Hsuen, Chih-Jen Lin, and Bernhard Schölkopf. 2005. A tutorial on  $\nu$ -support vector machines. *Applied Stochastic Models in Business and Industry* 21:111–136. [318]
- Chiararamella, Yves, Philippe Mulhem, and Franck Fourel. 1996. A model for multimedia information retrieval. Technical Report 4-96, University of Glasgow. [198]
- Chierichetti, Flavio, Alessandro Panconesi, Prabhakar Raghavan, Mauro Sozio, Alessandro Tiberi, and Eli Upfal. 2007. Finding near neighbors through cluster pruning. In *Proc. PODS*. [137]
- Cho, Junghoo, and Hector Garcia-Molina. 2002. Parallel crawlers. In *Proc. WWW*, pp. 124–135. ACM Press. doi: <http://doi.acm.org/10.1145/511446.511464>. [419]
- Cho, Junghoo, Hector Garcia-Molina, and Lawrence Page. 1998. Efficient crawling through URL ordering. In *Proc. WWW*, pp. 161–172. [419]
- 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: <http://doi.acm.org/10.1145/1148170.1148247>. [198]
- 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. [138]
- Cleverdon, Cyril W. 1991. The significance of the Cranfield tests on index languages. In *Proc. SIGIR*, pp. 3–12. ACM Press. [159]
- Coden, Anni R., Eric W. Brown, and Savitha Srinivasan (eds.). 2002. *Information Retrieval Techniques for Speech Applications*. Springer. [xviii]
- Cohen, Paul R. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press. [265]
- Cohen, William W. 1998. Integration of heterogeneous databases without common domains using queries based on textual similarity. In *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 201–212. ACM Press. [200]
- Cohen, William W., Robert E. Schapire, and Yoram Singer. 1998. Learning to order things. In *Proc. NIPS*. The MIT Press. URL: [citeseer.ist.psu.edu/article/cohen98learning.html](http://citeseer.ist.psu.edu/article/cohen98learning.html). [138]
- Cohen, William W., and Yoram Singer. 1999. Context-sensitive learning methods for text categorization. *TOIS* 17(2):141–173. [312]
- Comtet, Louis. 1974. *Advanced Combinatorics*. Reidel. [327]
- 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. [138]
- Cormen, Thomas H., Charles Eric Leiserson, and Ronald L. Rivest. 1990. *Introduction to Algorithms*. MIT Press. [10, 72, 367]
- Cover, Thomas M., and Peter E. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1):21–27. [292]
- Cover, Thomas M., and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley. [98]
- Crammer, Koby, and Yoram Singer. 2001. On the algorithmic implementation of multiclass kernel-based machines. *JMLR* 2:265–292. [319]
- Creedy, 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: <http://doi.acm.org/10.1145/135226.135228>. [291]
- 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: <http://doi.acm.org/10.1145/299917.299920>. [216]
- Cristianini, Nello, and John Shawe-Taylor. 2000. *Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge. [319]

- Croft, W. Bruce. 1978. A file organization for cluster-based retrieval. In *Proc. SIGIR*, pp. 65–82. ACM Press. [344]
- Croft, W. Bruce, and David J. Harper. 1979. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation* 35(4):285–295. [122, 209]
- Croft, W. Bruce, and John Lafferty (eds.). 2003. *Language Modeling for Information Retrieval*. Springer. [232]
- Crouch, Carolyn J. 1988. A cluster-based approach to thesaurus construction. In *Proc. SIGIR*, pp. 309–320. ACM Press. doi: <http://doi.acm.org/10.1145/62437.62467>. [345]
- 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*. [60]
- 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. [367]
- 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. [344, 367]
- Damerau, Fred J. 1964. A technique for computer detection and correction of spelling errors. *CACM* 7(3):171–176. doi: <http://doi.acm.org/10.1145/363958.363994>. [59]
- Davidson, Ian, and Ashwin Satyanarayana. 2003. Speeding up k-means clustering by bootstrap averaging. In *ICDM 2003 Workshop on Clustering Large Data Sets*. [345]
- Day, William H., and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification* 1:1–24. [367]
- 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: <http://doi.acm.org/10.1145/348751.348754>. [99]
- Dean, Jeffrey, and Sanjay Ghemawat. 2004. MapReduce: Simplified data processing on large clusters. In *Symposium on Operating System Design and Implementation*. [69, 76]
- 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. [383]
- del Bimbo, Alberto. 1999. *Visual Information Retrieval*. Morgan Kaufmann. [xviii]
- 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. [345]
- Dhillon, Inderjit S. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. KDD*, pp. 269–274. [345, 368]
- 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: <http://dx.doi.org/10.1023/A:1007612920971>. [345]
- Di Eugenio, Barbara, and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics* 30(1):95–101. doi: <http://dx.doi.org/10.1162/089120104773633402>. [160]
- Dietterich, Thomas G. 2002. Ensemble learning. In Michael A. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks*. 2nd edition. MIT Press. [319]
- Dietterich, Thomas G., and Ghulum Bakiri. 1995. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research* 2: 263–286. [292]
- Dom, Byron E. 2002. An information-theoretic external cluster-validity measure. In *Proc. UAI*. [344]
- Domingos, Pedro. 2000. A unified bias-variance decomposition for zero-one and squared loss. In *Proc. National Conference on Artificial Intelligence and Proc. Confer-*



## Bibliography

449

- ence *Innovative Applications of Artificial Intelligence*, pp. 564–569. AAAI Press/The MIT Press. [292]
- 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](http://citeseer.ist.psu.edu/domingos97optimality.html). [265]
- Downie, J. Stephen. 2006. The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine* 12(12). [xviii]
- Duda, Richard O., Peter E. Hart, and David G. Stork. 2000. *Pattern Classification*, 2nd edition. Wiley-Interscience. [264, 343]
- 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: <http://doi.acm.org/10.1145/288627.288651>. [261, 306, 319]
- Dumais, Susan T. 1993. Latent semantic indexing (LSI) and TREC-2. In *Proc. TREC*, pp. 105–115. [382, 383]
- Dumais, Susan T. 1995. Latent semantic indexing (LSI): TREC-3 report. In *Proc. TREC*, pp. 219–230. [382, 383]
- Dumais, Susan T., and Hao Chen. 2000. Hierarchical classification of Web content. In *Proc. SIGIR*, pp. 256–263. ACM Press. [319]
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1):61–74. [265]
- Dunning, Ted. 1994. *Statistical identification of language*. Technical Report 94-273, Computing Research Laboratory, New Mexico State University. [43]
- Eckart, Carl, and Gale Young. 1936. The approximation of a matrix by another of lower rank. *Psychometrika* 1:211–218. [383]
- El-Hamdouchi, Abdelmoula, and Peter Willett. 1986. Hierarchic document classification using Ward's clustering method. In *Proc. SIGIR*, pp. 149–156. ACM Press. DOI: <http://doi.acm.org/10.1145/253168.253200>. [367]
- Elias, Peter. 1975. Universal code word sets and representations of the integers. *IEEE Transactions on Information Theory* 21(2):194–203. [98]
- Eyheramendy, Susana, David Lewis, and David Madigan. 2003. On the Naive Bayes model for text categorization. In *Proc. International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics. [265]
- Fallows, Deborah. 2004. The internet and daily life. URL: [www.pewinternet.org/pdfs/PIP\\_Internet\\_and\\_Daily\\_Life.pdf](http://www.pewinternet.org/pdfs/PIP_Internet_and_Daily_Life.pdf). Pew / Internet and American Life Project. [xv]
- Fayyad, Usama M., Cory Reina, and Paul S. Bradley. 1998. Initialization of iterative refinement clustering algorithms. In *Proc. KDD*, pp. 194–198. [345]
- Fellbaum, Christiane D. 1998. *WordNet – An Electronic Lexical Database*. MIT Press. [177]
- Ferragina, Paolo, and Rossano Venturini. 2007. Compressed permuterm indexes. In *Proc. SIGIR*. ACM Press. [59]
- Forman, George. 2004. A pitfall and solution in multi-class feature selection for text classification. In *Proc. ICML*. [265]
- Forman, George. 2006. Tackling concept drift by temporal inductive transfer. In *Proc. SIGIR*, pp. 252–259. ACM Press. DOI: <http://doi.acm.org/10.1145/1148170.1148216>. [265]
- Forman, George, and Ira Cohen. 2004. Learning from little: Comparison of classifiers given little training. In *PKDD*, pp. 161–172. [308]
- 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/di985957/98p0926/0](http://www.jstor.org/view/01621459/di985957/98p0926/0). [368]
- 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. [76]
- 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. [98]
- Frakes, William B., and Ricardo Baeza-Yates (eds.). 1992. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall. [451, 461]
- 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. [345]
- Friedl, Jeffrey E. F. 2006. *Mastering Regular Expressions*, 3rd edition. O'Reilly. [17]
- Friedman, Jerome H. 1997. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* 1(1):55–77. [265, 292]
- Friedman, Nir, and Moises Goldszmidt. 1996. Building classifiers using bayesian networks. In *Proc. National Conference on Artificial Intelligence*, pp. 1277–1284. [213]
- Fuhr, Norbert. 1989. Optimum polynomial retrieval functions based on the probability ranking principle. *TOIS* 7(3):183–204. [138]
- Fuhr, Norbert. 1992. Probabilistic models in information retrieval. *Computer Journal* 35(3):243–255. [216, 320]
- Fuhr, Norbert, Norbert Gövert, Gabriella Kazai, and Mounia Lalmas (eds.). 2003a. *INitiative for the Evaluation of XML Retrieval (INEX)*. *Proc. First INEX Workshop*. ERCIM. [198]
- Fuhr, Norbert, and Kai Großjohann. 2004. XIRQL: An XML query language based on information retrieval concepts. *TOIS* 22(2):313–356. URL: <http://doi.acm.org/10.1145/984321.984326>. [198]
- Fuhr, Norbert, and Mounia Lalmas. 2007. Advances in XML retrieval: The INEX initiative. In *Proc. International Workshop on Research Issues in Digital Libraries*. [198]
- 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. [198]
- 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. [198, 460, 465]
- 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. [198, 456, 458]
- Fuhr, Norbert, Saadia Malik, and Mounia Lalmas (eds.). 2003b. *INEX 2003 Workshop Proceedings*. URL: <http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf>. [198, 451, 458]
- 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: <http://doi.acm.org/10.1145/174608.174612>. [138]
- 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: <http://doi.acm.org/10.1145/239041.239045>. [200]
- Gaertner, Thomas, John W. Lloyd, and Peter A. Flach. 2002. Kernels for structured data. In *International Conference on Inductive Logic Programming*, pp. 66–83. [319]
- 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. [43]
- 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.
- Garcia, Steven, Hugh E. Williams, and Adam Cannane. 2004. Access-ordered indexes. In *Proc. Australasian conference on Computer science*, pp. 7–14. [137]
- Garcia-Molina, Hector, Jennifer Widom, and Jeffrey D. Ullman. 1999. *Database System Implementation*. Prentice-Hall. [77]
- Garfield, Eugene. 1955. Citation indexes to science: A new dimension in documentation through association of ideas. *Science* 122:108–111. [439]

## Bibliography

451

- Garfield, Eugene. 1976. The permuted subject index: An autobiographic review. *JASIS* 27(5-6):288–291. [59]
- Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Computation* 4(1):1–58. [292]
- Geng, Xiubo, Tie-Yan Liu, Tao Qin, and Hang Li. 2007. Feature selection for ranking. In *Proc. SIGIR*, pp. 407–414. ACM Press. [320]
- Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication* 12(4). URL: <http://jcmc.indiana.edu/vol12/issue4/gerrand.html>. article 8. [29]
- Gey, Fredric C. 1994. Inferring probability of relevance using the method of logistic regression. In *Proc. SIGIR*, pp. 222–231. ACM Press. [320]
- Ghamrawi, Nadia, and Andrew McCallum. 2005. Collective multi-label classification. In *Proc. CIKM*, pp. 195–200. ACM Press. doi: <http://doi.acm.org/10.1145/1099554.1099591>. [292]
- Glover, Eric, David M. Pennock, Steve Lawrence, and Robert Krovetz. 2002a. Inferring hierarchical descriptions. In *Proc. CIKM*, pp. 507–514. ACM Press. doi: <http://doi.acm.org/10.1145/584792.584876>. [368]
- 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: <http://doi.acm.org/10.1145/511446.511520>. [367]
- 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: <http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf>. [198]
- 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*. [199]
- Greiff, Warren R. 1998. A theory of term weighting based on exploratory data analysis. In *Proc. SIGIR*, pp. 11–19. ACM Press. [209]
- 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](http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf). [216]
- Grossman, David A., and Ophir Frieder. 2004. *Information Retrieval: Algorithms and Heuristics*, 2nd edition. Springer. [xviii, 76, 200]
- Gusfield, Dan. 1997. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press. [60]
- Hamerly, Greg, and Charles Elkan. 2003. Learning the  $k$  in  $k$ -means. In *NIPS*. URL: [http://books.nips.cc/papers/files/nips16/NIPS2003\\_AA36.pdf](http://books.nips.cc/papers/files/nips16/NIPS2003_AA36.pdf). [345]
- Han, Eui-Hong, and George Karypis. 2000. Centroid-based document classification: Analysis and experimental results. In *PKDD*, pp. 424–431. [291]
- Hand, David J. 2006. Classifier technology and the illusion of progress. *Statistical Science* 21:1–14. [265]
- Hand, David J., and Keming Yu. 2001. Idiot's Bayes: Not so stupid after all. *International Statistical Review* 69(3):385–398. [265]
- Harman, Donna. 1991. How effective is suffixing? *JASIS* 42:7–15. [43]
- Harman, Donna. 1992. Relevance feedback revisited. In *Proc. SIGIR*, pp. 1–10. ACM Press. [170, 177]
- Harman, Donna, Ricardo Baeza-Yates, Edward Fox, and W. Lee. 1992. Inverted files. In Frakes and Baeza-Yates (1992), pp. 28–43. [76]
- Harman, Donna, and Gerald Candela. 1990. Retrieving records from a gigabyte of text on a minicomputer using statistical ranking. *JASIS* 41(8):581–589. [76]
- Harold, Elliotte Rusty, and Scott W. Means. 2004. *XML in a Nutshell*, 3rd edition. O'Reilly. [198]
- Harter, Stephen P. 1998. Variations in relevance assessments and the measurement of retrieval effectiveness. *JASIS* 47:37–49. [160]

- Hartigan, J. A., and M. A. Wong. 1979. A K-means clustering algorithm. *Applied Statistics* 28:100–108. [345]
- Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. [264, 265, 291, 292, 319]
- Hatzivassiloglou, Vasileios, Luis Gravano, and Anke Needu Maganti. 2000. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proc. SIGIR*, pp. 224–231. ACM Press. doi: <http://doi.acm.org/10.1145/345508.345582>. [344]
- Haveliwalla, 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/haveliwalla03topicsensitive.html](http://citeseer.ist.psu.edu/article/haveliwalla03topicsensitive.html). [439]
- Haveliwalla, Taher H. 2002. Topic-sensitive PageRank. In *Proc. WWW*. URL: [citeseer.ist.psu.edu/haveliwalla02topicsensitive.html](http://citeseer.ist.psu.edu/haveliwalla02topicsensitive.html). [439]
- 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. [308]
- Heaps, Harold S. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press. [97]
- Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64. [199]
- Hearst, Marti A. 2006. Clustering versus faceted categories for information exploration. *CACM* 49(4):59–61. doi: <http://doi.acm.org/10.1145/1121949.1121983>. [344]
- Hearst, Marti A., and Jan O. Pedersen. 1996. Reexamining the cluster hypothesis. In *Proc. SIGIR*, pp. 76–84. ACM Press. [344]
- Hearst, Marti A., and Christian Plaunt. 1993. Subtopic structuring for full-length document access. In *Proc. SIGIR*, pp. 59–68. ACM Press. doi: <http://doi.acm.org/10.1145/160688.160695>. [199]
- Heinz, Steffen, and Justin Zobel. 2003. Efficient single-pass index construction for text databases. *JASIST* 54(8):713–729. doi: <http://dx.doi.org/10.1002/asi.10268>. [76]
- 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: <http://doi.acm.org/10.1145/506309.506312>. [77]
- 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: [http://dx.doi.org/10.1016/S1389-1286\(00\)00055-4](http://dx.doi.org/10.1016/S1389-1286(00)00055-4). [404]
- 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. [320]
- 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. [160]
- 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.
- 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.
- 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*.
- Hiemstra, Djoerd. 1998. A linguistically motivated probabilistic model of information retrieval. In *Proc. ECDL*, pp. 569–584.

## Bibliography

453

- Hiemstra, Djoerd. 2000. A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries* 3(2):131–139.
- Hiemstra, Djoerd, and Wessel Kraaij. 2005. A language-modeling approach to TREC. In Voorhees and Harman (2005), pp. 373–395. [232, 233]
- Hirai, Jun, Sriram Raghavan, Hector Garcia-Molina, and Andreas Paepcke. 2000. WebBase: A repository of web pages. In *Proc. WWW*, pp. 277–293. [419]
- Hofmann, Thomas. 1999a. Probabilistic Latent Semantic Indexing. In *UAI*. URL: [citeseer.ist.psu.edu/hofmann99probabilistic.html](http://citeseer.ist.psu.edu/hofmann99probabilistic.html).
- Hofmann, Thomas. 1999b. Probabilistic Latent Semantic Indexing. In *Proc. SIGIR*, pp. 50–57. ACM Press. URL: [citeseer.ist.psu.edu/article/hofmann99probabilistic.html](http://citeseer.ist.psu.edu/article/hofmann99probabilistic.html).
- Hollink, Vera, Jaap Kamps, Christof Monz, and Maarten de Rijke. 2004. Monolingual document retrieval for European languages. *IR* 7(1):33–52. [43]
- Hopcroft, John E., Rajeev Motwani, and Jeffrey D. Ullman. 2000. *Introduction to Automata Theory, Languages, and Computation*, 2nd edition. Addison Wesley. [17]
- Huang, Yifen, and Tom M. Mitchell. 2006. Text clustering with extended user feedback. In *Proc. SIGIR*, pp. 413–420. ACM Press. doi: <http://doi.acm.org/10.1145/1148170.1148242>. [345]
- Hubert, Lawrence, and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2:193–218. [344]
- Hughes, Baden, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. 2006. Reconsidering language identification for written language resources. In *International Conference on Language Resources and Evaluation*, pp. 485–488. [43]
- Hull, David. 1993. Using statistical testing in the evaluation of retrieval performance. In *Proc. SIGIR*, pp. 329–338. ACM Press. [159]
- Hull, David. 1996. Stemming algorithms – A case study for detailed evaluation. *JASIS* 47(1):70–84. [43]
- Ide, E. 1971. New experiments in relevance feedback. In Salton (1971b), pp. 337–354. [177]
- 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. [291]
- Ingwersen, Peter, and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer. [xviii]
- Ittner, David J., David D. Lewis, and David D. Ahn. 1995. Text categorization of low quality images. In *Proc. Annual Symposium on Document Analysis and Information Retrieval*, pp. 301–315. [291]
- Iwayama, Makoto, and Takenobu Tokunaga. 1995. Cluster-based text categorization: A comparison of category search strategies. In *Proc. SIGIR*, pp. 273–280. ACM Press. [291]
- Jackson, Peter, and Isabelle Moulinier. 2002. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. John Benjamins. [307]
- Jacobs, Paul S., and Lisa F. Rau. 1990. SCISOR: Extracting information from on-line news. *CACM* 33:88–97. [308]
- Jain, Anil, M. Narasimha Murty, and Patrick Flynn. 1999. Data clustering: A review. *ACM Computing Surveys* 31(3):264–323. [367]
- Jain, Anil K., and Richard C. Dubes. 1988. *Algorithms for Clustering Data*. Prentice-Hall. [367]
- Jardine, N., and C. J. van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval* 7:217–240. [344]
- Järvelin, Kalervo, and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *TOIS* 20(4):422–446. [160]
- Jeh, Glen, and Jennifer Widom. 2003. Scaling personalized web search. In *Proc. WWW*, pp. 271–279. ACM Press. [439]

- Jensen, Finn V., and Finn B. Jensen. 2001. *Bayesian Networks and Decision Graphs*. Springer. [215]
- 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. [419]
- Ji, Xiang, and Wei Xu. 2006. Document clustering with prior knowledge. In *Proc. SIGIR*, pp. 405–412. ACM Press. doi: <http://doi.acm.org/10.1145/1148170.1148241>. [345]
- Jing, Hongyan. 2000. Sentence reduction for automatic text summarization. In *Proc. Conference on applied natural language processing*, pp. 310–315. [161]
- Joachims, Thorsten. 1997. A probabilistic analysis of the Rocchio algorithm with tfidf for text categorization. In *Proc. ICML*, pp. 143–151. Morgan Kaufmann. [291]
- Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proc. ECML*, pp. 137–142. Springer. [261, 306, 307]
- 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. [319]
- Joachims, Thorsten. 2002a. *Learning to Classify Text Using Support Vector Machines*. Kluwer. [306, 307, 319]
- Joachims, Thorsten. 2002b. Optimizing search engines using clickthrough data. In *Proc. KDD*, pp. 133–142. [161, 170, 320]
- Joachims, Thorsten. 2006a. Training linear SVMs in linear time. In *Proc. KDD*, pp. 217–226. ACM Press. doi: <http://doi.acm.org/10.1145/1150402.1150429>. [265, 302, 319]
- Joachims, Thorsten. 2006b. Transductive support vector machines. In Chapelle et al. (2006), pp. 105–118. [320]
- 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. [161, 170]
- 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](http://www.webology.ir/2006/v3n4/a35.html). [44]
- 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. [xviii]
- Käki, Mika. 2005. Findex: Search result categories help users when document ranking fails. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pp. 131–140. ACM Press. doi: <http://doi.acm.org/10.1145/1054972.1054991>. [344, 368]
- Kammenhuber, Nils, Julia Luxenburger, Anja Feldmann, and Gerhard Weikum. 2006. Web search clickstreams. In *ACM SIGCOMM on Internet Measurement*, pp. 245–250. ACM Press. [44]
- Kamps, Jaap, Maarten de Rijke, and Börkur Sigurbjörnsson. 2004. Length normalization in XML retrieval. In *Proc. SIGIR*, pp. 80–87. ACM Press. doi: <http://doi.acm.org/10.1145/1008992.1009009>. [199]
- Kamps, Jaap, Maarten Marx, Maarten de Rijke, and Börkur Sigurbjörnsson. 2006. Articulating information needs in XML query languages. *TOIS* 24(4):407–436. doi: <http://doi.acm.org/10.1145/1185877.1185879>. [198]
- 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. [368]
- Kannan, Ravi, Santosh Vempala, and Adrian Vetta. 2000. On clusterings – Good, bad and spectral. In *Proc. Annual Symposium on Foundations of Computer Science*, p. 367. IEEE Computer Society. [368]
- Kaszkiel, Marcin, and Justin Zobel. 1997. Passage retrieval revisited. In *Proc. SIGIR*, pp. 178–185. ACM Press. doi: <http://doi.acm.org/10.1145/258525.258561>. [199]

## Bibliography

455

- Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding groups in data*. Wiley. [345]
- Kazai, Gabriella, and Mounia Lalmas. 2006. eXtended cumulated gain measures for the evaluation of content-oriented XML retrieval. *TOIS* 24(4):503–542. doi: <http://doi.acm.org/10.1145/1185883>. [199]
- Kekäläinen, Jaana. 2005. Binary and graded relevance in IR evaluations – Comparison of the effects on ranking of IR systems. *IP&M* 41:1019–1033. [160]
- Kekäläinen, Jaana, and Kalervo Järvelin. 2002. Using graded relevance assessments in IR evaluation. *JASIST* 53(13):1120–1129. [160]
- Kemeny, John G., and J. Laurie Snell. 1976. *Finite Markov Chains*. Springer. [439]
- 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. [159]
- 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. [60]
- King, Benjamin. 1967. Step-wise clustering procedures. *Journal of the American Statistical Association* 69:86–101. [367]
- 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 *Proc. NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*. National Institute of Informatics. [43]
- 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. [308]
- Kleinberg, Jon M. 1997. Two algorithms for nearest-neighbor search in high dimensions. In *Proc. Annual ACM Symposium on Theory of Computing*, pp. 599–608. ACM Press. doi: <http://doi.acm.org/10.1145/258533.258653>. [291]
- Kleinberg, Jon M. 1999. Authoritative sources in a hyperlinked environment. *JACM* 46(5):604–632. URL: [citeseer.ist.psu.edu/article/kleinberg98authoritative.html](http://citeseer.ist.psu.edu/article/kleinberg98authoritative.html). [439]
- Kleinberg, Jon M. 2002. An impossibility theorem for clustering. In *Proc. NIPS*. [345]
- Knuth, Donald E. 1997. *The Art of Computer Programming, Volume 3: Sorting and Searching*, 3rd edition. Addison-Wesley. [59]
- Ko, Youngjoong, Jinwoo Park, and Jungyun Seo. 2004. Improving text categorization using the importance of sentences. *IP&M* 40(1):65–79. [313]
- Koenemann, Jürgen, and Nicholas J. Belkin. 1996. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pp. 205–212. ACM Press. doi: <http://doi.acm.org/10.1145/238386.238487>. [177]
- Kołcz, Aleksander, Vidya Prabakarmurthi, and Jugal Kalita. 2000. Summarization as feature selection for text categorization. In *Proc. CIKM*, pp. 365–370. ACM Press. [313]
- Kołcz, Aleksander, and Wen-Tau Yih. 2007. Raising the baseline for high-precision text classifiers. In *Proc. KDD*. [265]
- Koller, Daphne, and Mehran Sahami. 1997. Hierarchically classifying documents using very few words. In *Proc. ICML*, pp. 170–178. [319]
- Konheim, Alan G. 1981. *Cryptography: A Primer*. John Wiley & Sons. [43]
- Korfhage, Robert R. 1997. *Information Storage and Retrieval*. Wiley. [161]
- 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). [302]
- 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. [231]
- 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. [233]

- Krippendorff, Klaus. 2003. *Content Analysis: An Introduction to its Methodology*. Sage. [160]
- Krovetz, Bob. 1995. *Word sense disambiguation for large text databases*. PhD thesis, University of Massachusetts Amherst. [43]
- Kukich, Karen. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys* 24(4):377–439. doi: <http://doi.acm.org/10.1145/146370.146380>. [59]
- 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](http://citeseer.ist.psu.edu/kumar99trawling.html). [404]
- 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](http://citeseer.ist.psu.edu/article/kumar00web.html). [404]
- Kupiec, Julian, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proc. SIGIR*, pp. 68–73. ACM Press. [160]
- Kurland, Oren, and Lillian Lee. 2004. Corpus structure, language models, and ad hoc information retrieval. In *Proc. SIGIR*, pp. 194–201. ACM Press. doi: <http://doi.acm.org/10.1145/1008992.1009027>. [344]
- 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. [231]
- 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 and Information Retrieval*. Kluwer. [233]
- 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. [199]
- Lalmas, Mounia, and Anastasios Tombros. 2007. Evaluating XML retrieval effectiveness at INEX. *SIGIR Forum* 41(1):40–57. doi: <http://doi.acm.org/10.1145/1273221.1273225>. [198]
- Lance, G. N., and W. T. Williams. 1967. A general theory of classificatory sorting strategies 1. Hierarchical systems. *Computer Journal* 9(4):373–380. [367]
- Langville, Amy, and Carl Meyer. 2006. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press. [439]
- 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: <http://doi.acm.org/10.1145/312129.312186>. [367, 368]
- Larson, Ray R. 2005. A fusion approach to XML structured document retrieval. *IR* 8(4):601–629. doi: <http://dx.doi.org/10.1007/s10791-005-0749-0>. [199]
- Lavrenko, Victor, and W. Bruce Croft. 2001. Relevance-based language models. In *Proc. SIGIR*, pp. 120–127. ACM Press. [231]
- Lawrence, Steve, and C. Lee Giles. 1998. Searching the World Wide Web. *Science* 280(5360):98–100. URL: [citeseer.ist.psu.edu/lawrence98searching.html](http://citeseer.ist.psu.edu/lawrence98searching.html). [404]
- Lawrence, Steve, and C. Lee Giles. 1999. Accessibility of information on the web. *Nature* 500:107–109. [404]
- 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]
- 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](http://citeseer.ist.psu.edu/lempel00stochastic.html). [440]
- Lesk, Michael. 1988. Grab – Inverted indexes with low storage overhead. *Computing Systems* 1:207–220. [76]
- Lesk, Michael. 2004. *Understanding Digital Libraries*, 2nd edition. Morgan Kaufmann. [xviii]



## Bibliography

457

- 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: <http://doi.acm.org/10.1145/1099554.1099739>. [76]
- Lester, Nicholas, Justin Zobel, and Hugh E. Williams. 2006. Efficient online index maintenance for contiguous inverted lists. *IP&M* 42(4):916–933. doi: <http://dx.doi.org/10.1016/j.ipm.2005.09.005>. [76]
- Levenshtein, Vladimir I. 1965. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission* 1:8–17. [59]
- Lew, Michael S. 2001. *Principles of Visual Information Retrieval*. Springer. [xviii]
- Lewis, David D. 1995. Evaluating and optimizing autonomous text classification systems. In *Proc. SIGIR*. ACM Press. [265]
- Lewis, David D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *ECML*, pp. 4–15. Springer. [265]
- Lewis, David D., and Karen Spärck Jones. 1996. Natural language processing for information retrieval. *CACM* 39(1):92–101. doi: <http://doi.acm.org/10.1145/234173.234210>. [xviii]
- Lewis, David D., and Marc Ringuette. 1994. A comparison of two learning algorithms for text categorization. In *SDAIR*, pp. 81–93. [265]
- 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: <http://doi.acm.org/10.1145/243199.243277>. [292]
- 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. [77, 265]
- Li, Fan, and Yiming Yang. 2003. A loss function analysis for classification methods in text categorization. In *Proc. ICML*, pp. 472–479. [261, 319]
- Liddy, Elizabeth D. 2005. Automatic document retrieval. In *Encyclopedia of Language and Linguistics*, 2nd edition. Elsevier.
- 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: <http://dx.doi.org/10.1007/s10791-005-0747-2>. [199]
- Lita, Lucian Vlad, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. tRuE-casIng. In *Proc. ACL*, pp. 152–159. [43]
- Littman, Michael L., Susan T. Dumais, and Thomas K. Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In Gregory Grefenstette (ed.), *Cross Language Information Retrieval*. Kluwer. URL: [citeseer.ist.psu.edu/littman98automatic.html](http://citeseer.ist.psu.edu/littman98automatic.html). [384]
- 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. [319]
- Liu, Xiaoyong, and W. Bruce Croft. 2004. Cluster-based retrieval using language models. In *Proc. SIGIR*, pp. 186–193. ACM Press. doi: <http://doi.acm.org/10.1145/1008992.1009026>. [233, 323, 344]
- Lloyd, Stuart P. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2):129–136. [345]
- Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *JMLR* 2:419–444. [319]
- 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. [160]
- Long, Xiaohui, and Torsten Suel. 2003. Optimized query execution in large search engines with global page ordering. In *Proc. VLDB*. URL: [citeseer.ist.psu.edu/long03optimized.html](http://citeseer.ist.psu.edu/long03optimized.html). [137]
- Lovins, Julie Beth. 1968. Development of a stemming algorithm. *Translation and Computational Linguistics* 11(1):22–31. [31]

- Lu, Wei, Stephen E. Robertson, and Andrew MacFarlane. 2007. CISR at INEX 2006. In Fuhr et al. (2007), pp. 57–63. [199]
- 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. [122]
- Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2):159–165, 317. [122]
- 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. [43]
- Lunde, Ken. 1998. *CJKV Information Processing*. O'Reilly. [43]
- MacFarlane, A., J.A. McCann, and S.E. Robertson. 2000. Parallel search using partitioned inverted files. In *Proc. SPIRE*, pp. 209–220. [419]
- 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. [345]
- Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press. [xviii, 37, 97, 264, 342]
- Maron, M. E., and J. L. Kuhns. 1960. On relevance, probabilistic indexing, and information retrieval. *JACM* 7(3):216–244. [216, 265]
- 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: <http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf>. [199]
- McBryan, Oliver A. 1994. GENVL and WWW: Tools for Taming the Web. In *Proc. WWW*. URL: [citeseer.ist.psu.edu/mcbryan94genvl.html](http://citeseer.ist.psu.edu/mcbryan94genvl.html). [404, 439]
- McCallum, Andrew, and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. In *Working Notes of the 1998 AAAI/ICML Workshop on Learning for Text Categorization*, pp. 41–48. [265]
- 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. [319]
- McCallum, Andrew Kachites. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. URL: [www.cs.cmu.edu/~mccallum/bow](http://www.cs.cmu.edu/~mccallum/bow). [289]
- McKeown, Kathleen, and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *Proc. SIGIR*, pp. 74–82. ACM Press. DOI: <http://doi.acm.org/10.1145/215206.215334>. [368]
- 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*. [323, 344]
- McLachlan, Geoffrey J., and Thiriyambakam Krishnan. 1996. *The EM Algorithm and Extensions*. John Wiley & Sons. [345]
- Meadow, Charles T., Donald H. Kraft, and Bert R. Boyce. 1999. *Text Information Retrieval Systems*. Academic Press.
- Meilă, Marina. 2005. Comparing clusterings – An axiomatic view. In *Proc. ICML*. [345]
- 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: <http://doi.acm.org/10.1145/371920.372095>. [76]
- 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. ACM Press. DOI: <http://doi.acm.org/10.1145/1099554.1099560>. [199]
- 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.

## Bibliography

459

- Minsky, Marvin Lee, and Seymour Papert (eds.). 1988. *Perceptrons: An Introduction to Computational Geometry*. MIT Press. Expanded edition. [292]
- Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill. [264]
- Moffat, Alistair, and Timothy A. H. Bell. 1995. In situ generation of compressed inverted files. *JASIS* 46(7):537–550. [76]
- Moffat, Alistair, and Lang Stuiver. 1996. Exploiting clustering in inverted file compression. In *Proc. Conference on Data Compression*, pp. 82–91. IEEE Computer Society. [98]
- Moffat, Alistair, and Justin Zobel. 1992. Parameterised compression for sparse bitmaps. In *Proc. SIGIR*, pp. 274–285. ACM Press. doi: <http://doi.acm.org/10.1145/133160.133210>. [98]
- Moffat, Alistair, and Justin Zobel. 1996. Self-indexing inverted files for fast text retrieval. *TOIS* 14(4):349–379. [44]
- Moffat, Alistair, and Justin Zobel. 1998. Exploring the similarity space. *SIGIR Forum* 32(1). [123]
- 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]
- Mooers, Calvin E. 1950. Coding, information retrieval, and the rapid selector. *American Documentation* 1(4):225–229. [16]
- Moschitti, Alessandro. 2003. A study on optimal parameter tuning for Rocchio text classifier. In *Proc. ECIR*, pp. 420–435. [292]
- Moschitti, Alessandro, and Roberto Basili. 2004. Complex linguistic features for text classification: A comprehensive study. In *Proc. ECIR*, pp. 181–196. [319]
- Murata, Masaki, Qing Ma, Kiyotaka Uchimoto, Hiromi Ozaku, Masao Utiyama, and Hitoshi Isahara. 2000. Japanese probabilistic information retrieval using location and category information. In *Proc. International Workshop on Information Retrieval With Asian Languages*, pp. 81–88. URL: <http://portal.acm.org/citation.cfm?doid=355214.355226>. [312]
- Muresan, Gheorghe, and David J. Harper. 2004. Topic modeling for mediated access to very large document collections. *JASIST* 55(10):892–910. doi: <http://dx.doi.org/10.1002/asi.20034>. [344]
- Murtagh, Fionn. 1983. A survey of recent advances in hierarchical clustering algorithms. *Computer Journal* 26(4):354–359. [367]
- Najork, Marc, and Allan Heydon. 2001. *High-performance web crawling*. Technical Report 173, Compaq Systems Research Center. [419]
- Najork, Marc, and Allan Heydon. 2002. High-performance web crawling. In Panos Pardalos, James Abello and Mauricio Resende (eds.), *Handbook of Massive Data Sets*, chapter 2. Kluwer. [419]
- 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: <http://doi.acm.org/10.1145/263479.263482>. [200]
- Newsam, Shawn, Sitaram Bhagavathy, and B. S. Manjunath. 2001. Category-based image retrieval. In *IEEE International Conference on Image Processing, Special Session on Multimedia Indexing, Browsing and Retrieval*, pp. 596–599. [164]
- Ng, Andrew Y., and Michael I. Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. In *NIPS*, pp. 841–848. URL: [www-2.cs.cmu.edu/Groups/NIPS/NIPS2001/papers/psgz/AA28.ps.gz](http://www-2.cs.cmu.edu/Groups/NIPS/NIPS2001/papers/psgz/AA28.ps.gz). [265, 308]
- Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. 2001a. On spectral clustering: Analysis and an algorithm. In *Proc. NIPS*, pp. 849–856. [368]
- 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](http://citeseer.ist.psu.edu/ng01link.html). [439, 440]
- Nigam, Kamal, Andrew McCallum, and Tom Mitchell. 2006. Semi-supervised text classification using EM. In Chapelle et al. (2006), pp. 33–56. [320]

- Ntoulas, Alexandros, and Junghoo Cho. 2007. Pruning policies for two-tiered inverted index with correctness guarantee. In *Proc. SIGIR*, pp. 191–198. ACM Press. [97]
- 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. [xviii]
- Ogilvie, Paul, and Jamie Callan. 2005. Parameter estimation for a simple hierarchical generative model for XML retrieval. In *Proc. INEX*, pp. 211–224. doi: <http://dx.doi.org/10.1007/11766278.16>. [199]
- O’Keefe, Richard A., and Andrew Trotman. 2004. The simplest query language that could possibly work. In Fuhr et al. (2005), pp. 167–174. [199]
- Osiński, Stanisław, and Dawid Weiss. 2005. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems* 20(3):48–54. [368]
- 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](http://citeseer.ist.psu.edu/page98pagerank.html). [439]
- Paice, Chris D. 1990. Another stemmer. *SIGIR Forum* 24(3):56–61. [31]
- Papineni, Kishore. 2001. Why inverse document frequency? In *North American Chapter of the Association for Computational Linguistics*, pp. 1–8. [122]
- 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. [265]
- Pelleg, Dan, and Andrew Moore. 1999. Accelerating exact k-means algorithms with geometric reasoning. In *Proc. KDD*, pp. 277–281. ACM Press. doi: <http://doi.acm.org/10.1145/312129.312248>. [345]
- 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. [345]
- Perkins, Simon, Kevin Lacker, and James Theiler. 2003. Grafting: Fast, incremental feature selection by gradient descent in function space. *JMLR* 3:1333–1356. [265]
- Persin, Michael. 1994. Document filtering for fast ranking. In *Proc. SIGIR*, pp. 339–348. ACM Press. [137]
- Persin, Michael, Justin Zobel, and Ron Sacks-Davis. 1996. Filtered document retrieval with frequency-sorted indexes. *JASIS* 47(10):749–764. [137]
- Peterson, James L. 1980. Computer programs for detecting and correcting spelling errors. *CACM* 23(12):676–687. doi: <http://doi.acm.org/10.1145/359038.359041>. [59]
- Picca, Davide, Benoît Curdy, and François Bavaud. 2006. Non-linear correspondence analysis in text retrieval: A kernel view. In *Proc. JADT*. [283]
- 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. [439]
- Pirolli, Peter L. T. 2007. *Information Foraging Theory: Adaptive Interaction With Information*. Oxford University Press. [344]
- 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. [298]
- Ponte, Jay M., and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. SIGIR*, pp. 275–281. ACM Press. [227, 228, 229]
- Popescul, Alexandrin, and Lyle H. Ungar. 2000. Automatic labeling of document clusters. Unpublished. [367]
- Porter, Martin F. 1980. An algorithm for suffix stripping. *Program* 14(3):130–137. [31]
- Pugh, William. 1990. Skip lists: A probabilistic alternative to balanced trees. *CACM* 33(6):668–676. [44]

## Bibliography

461

- 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. [320]
- Qiu, Yonggang, and H.P. Frei. 1993. Concept based query expansion. In *Proc. SIGIR*, pp. 160–169. ACM Press. [177]
- R Development Core Team. 2005. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: [www.R-project.org](http://www.R-project.org). ISBN 3-900051-07-0. [342, 368]
- 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. [344]
- 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](http://citeseer.ist.psu.edu/rahm01survey.html). [198]
- Rand, William M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336):846–850. [344]
- Rasmussen, Edie. 1992. Clustering algorithms. In Frakes and Baeza-Yates (1992), pp. 419–442. [344]
- 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. [265]
- 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: <http://doi.acm.org/10.1145/312624.312663>. [76]
- Ribeiro-Neto, Berthier A., and Ramurti A. Barbosa. 1998. Query performance for tightly coupled distributed digital libraries. In *ACM Conference on Digital Libraries*, pp. 182–190. [420]
- Rice, John A. 2006. *Mathematical Statistics and Data Analysis*. Duxbury Press. [91, 216, 256]
- Richardson, M., A. Prakash, and E. Brill. 2006. Beyond PageRank: machine learning for static ranking. In *Proc. WWW*, pp. 707–715. [320]
- 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](http://www.aclweb.org/anthology/P/P07/P07-1059). [177]
- Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press. [204, 216]
- Robertson, Stephen. 2005. How Okapi came to TREC. In Voorhees and Harman (2005), pp. 287–299. [216]
- Robertson, Stephen, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proc. CIKM*, pp. 42–49. ACM Press. DOI: <http://doi.acm.org/10.1145/1031171.1031181>. [217]
- Robertson, Stephen E., and Karen Spärck Jones. 1976. Relevance weighting of search terms. *JASIS* 27:129–146. [122, 216]
- Rocchio, J. J. 1971. Relevance feedback in information retrieval. In Salton (1971b), pp. 313–323. [166, 177, 291]
- Roget, P. M. 1946. *Roget's International Thesaurus*. Thomas Y. Crowell. [177]
- 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. AUAI Press. [384]
- Ross, Sheldon. 2006. *A First Course in Probability*. Pearson Prentice-Hall. [91, 216]
- 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](http://citeseer.ist.psu.edu/rusmevichientong01methods.html). [404]

- Ruthven, Ian, and Mounia Lalmas. 2003. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review* 18(1). [177]
- Sahoo, Nachiketa, Jamie Callan, Ramayya Krishnan, George Duncan, and Rema Padman. 2006. Incremental hierarchical clustering of text documents. In *Proc. CIKM*, pp. 357–366. ACM Press. doi: <http://doi.acm.org/10.1145/1183614.1183667>. [368]
- Sakai, Tetsuya. 2007. On the reliability of information retrieval metrics based on graded relevance. *IP&M* 43(2):531–548. [160]
- 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. [323, 344]
- Salton, Gerard (ed.). 1971b. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice-Hall. [122, 159, 177, 453, 461, 462]
- Salton, Gerard. 1975. *Dynamic information and library processing*. Prentice-Hall. [344]
- Salton, Gerard. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison Wesley. [43, 177]
- Salton, Gerard. 1991. The Smart project in automatic document retrieval. In *Proc. SIGIR*, pp. 356–358. ACM Press. [159]
- 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: <http://doi.acm.org/10.1145/160688.160693>. [199]
- Salton, Gerard, and Chris Buckley. 1987. *Term weighting approaches in automatic text retrieval*. Technical report, Cornell University, Ithaca, NY. [122]
- Salton, Gerard, and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *IP&M* 24(5):513–523. [123]
- Salton, Gerard, and Chris Buckley. 1990. Improving retrieval performance by relevance feedback. *JASIS* 41(4):288–297. [177]
- Saracevic, Tefko, and Paul Kantor. 1988. A study of information seeking and retrieving. II: Users, questions and effectiveness. *JASIS* 39:177–196. [159]
- Saracevic, Tefko, and Paul Kantor. 1996. A study of information seeking and retrieving. III: Searchers, searches, overlap. *JASIS* 39(3):197–216. [159]
- 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. [368]
- 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. [160]
- 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. [319]
- Schapire, Robert E., and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning* 39(2/3):135–168. [319]
- Schapire, Robert E., Yoram Singer, and Amit Singhal. 1998. Boosting and Rocchio applied to text filtering. In *Proc. SIGIR*, pp. 215–223. ACM Press. [291, 292]
- Schlieder, Torsten, and Holger Meuss. 2002. Querying and ranking XML documents. *JASIST* 53(6):489–503. doi: <http://dx.doi.org/10.1002/asi.10060>. [199]
- 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: <http://doi.acm.org/10.1145/564376.564416>. [98]
- Schölkopf, Bernhard, and Alexander J. Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press. [319]
- Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1):97–124. [176, 177]
- 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. [177, 265, 292]

## Bibliography

463

- Schütze, Hinrich, and Jan O. Pedersen. 1995. Information retrieval based on word senses. In *Proc. SDAIR*, pp. 161–175. [345]
- Schütze, Hinrich, and Craig Silverstein. 1997. Projections for efficient document clustering. In *Proc. SIGIR*, pp. 74–81. ACM Press. [344, 383]
- Schwarz, Gideon. 1978. Estimating the dimension of a model. *Annals of Statistics* 6(2): 461–464. [345]
- Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47. [264]
- Shawe-Taylor, John, and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press. [319]
- 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](http://citeseer.ist.psu.edu/shkapenyuk02design.html). [419]
- Siegel, Sidney, and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition. McGraw-Hill. [160]
- Sifry, Dave. 2007. The state of the Live Web, April 2007. URL: <http://technorati.com/weblog/2007/04/328.html>. [29]
- 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. [199]
- 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. [44]
- Silvestri, Fabrizio. 2007. Sorting out the document identifier assignment problem. In *Proc. ECIR*, pp. 101–112. [98]
- 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. [98]
- Sindhwani, V., and S. S. Keerthi. 2006. Large scale semi-supervised linear SVMs. In *Proc. SIGIR*, pp. 477–484. [320]
- 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](http://citeseer.ist.psu.edu/singhal96pivoted.html). [122]
- Singhal, Amit, Mandar Mitra, and Chris Buckley. 1997. Learning routing queries in a query zone. In *Proc. SIGIR*, pp. 25–32. ACM Press. [177]
- Singhal, Amit, Gerard Salton, and Chris Buckley. 1995. *Length normalization in degraded text collections*. Technical report, Cornell University, Ithaca, NY. [123]
- Singhal, Amit, Gerard Salton, and Chris Buckley. 1996b. Length normalization in degraded text collections. In *Proc. SDAIR*, pp. 149–162. [123]
- 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: <http://www.vldb.org/conf/2004/RS17P1.PDF>. [137, 344]
- 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: <http://dx.doi.org/10.1109/34.895972>. [xviii]
- Sneath, Peter H.A., and Robert R. Sokal. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W.H. Freeman. [367]
- Snedecor, George Waddel, and William G. Cochran. 1989. *Statistical Methods*. Iowa State University Press. [265]
- Somogyi, Zoltan. 1990. *The Melbourne University bibliography system*. Technical Report 90/3, Melbourne University, Parkville, Victoria, Australia. [76]
- 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. [138]

- Sornil, Ohm. 2001. *Parallel Inverted Index for Large-Scale, Dynamic Digital Libraries*. PhD thesis, Virginia Tech. URL: <http://scholar.lib.vt.edu/theses/available/etd-02062001-114915/>. [420]
- Spärck Jones, Karen. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1):11–21. [122]
- Spärck Jones, Karen. 2004. Language modelling's generative model: Is it rational? MS, Computer Laboratory, University of Cambridge. URL: <http://www.cl.cam.ac.uk/~ksj21/langmodnote4.pdf>. [233]
- 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. [214, 215, 216]
- Spink, Amanda, and Charles Cole (eds.). 2005. *New Directions in Cognitive Information Retrieval*. Springer. [161]
- 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: <http://ist.psu.edu/faculty-pages/jjansen/academic/pubs/internetresearch2000.pdf>. [170]
- Sproat, Richard, and Thomas Emerson. 2003. The first international Chinese word segmentation bakeoff. In *SIGHAN Workshop on Chinese Language Processing*. [43]
- 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. [43]
- Sproat, Richard William. 1992. *Morphology and Computation*. MIT Press. [43]
- Stein, Benno, and Sven Meyer zu Eissen. 2004. Topic identification: Framework and application. In *Proc. International Conference on Knowledge Management*. [367]
- 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*. [344]
- Steinbach, Michael, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*. [368]
- Strang, Gilbert (ed.). 1986. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press. [383]
- Strehl, Alexander. 2002. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. PhD thesis, The University of Texas at Austin. [344]
- Strohman, Trevor, and W. Bruce Croft. 2007. Efficient document retrieval in main memory. In *Proc. SIGIR*, pp. 175–182. ACM Press. [44]
- Swanson, Don R. 1988. Historical note: Information retrieval and the future of an illusion. *JASIS* 39(2):92–98. [159, 177]
- Tague-Sutcliffe, Jean, and James Blustein. 1995. A statistical analysis of the TREC-3 data. In *Proc. TREC*, pp. 385–398. [160]
- 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: <http://doi.acm.org/10.1145/1244002.1244096>. [291]
- Tannier, Xavier, and Shlomo Geva. 2005. XML retrieval with a natural language interface. In *Proc. SPIRE*, pp. 29–40. [200]
- 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. [233]
- Taube, Mortimer, and Harold Wooster (eds.). 1958. *Information Storage and Retrieval: Theory, Systems, and Devices*. Columbia University Press. [16]
- 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. [320]



## Bibliography

465

- 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. [384]
- 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. [199]
- 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. [199]
- 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. [345]
- Tishby, Naftali, and Noam Slonim. 2000. Data clustering by Markovian relaxation and the information bottleneck method. In *Proc. NIPS*, pp. 640–646. [345]
- Toda, Hiroyuki, and Ryoji Kataoka. 2005. A search result clustering method using informatively named entities. In *Proc. Annual ACM International Workshop on Web Information and Data Management*, pp. 81–86. ACM Press. doi: <http://doi.acm.org/10.1145/1097047.1097063>. [344]
- 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. [419]
- Tombros, Anastasios, and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *Proc. SIGIR*, pp. 2–10. ACM Press. doi: <http://doi.acm.org/10.1145/290941.290947>. [161]
- Tombros, Anastasios, Robert Villa, and C. J. van Rijsbergen. 2002. The effectiveness of query-specific hierarchic clustering in information retrieval. *IP&M* 38(4):559–582. doi: [http://dx.doi.org/10.1016/S0306-4573\(01\)00048-6](http://dx.doi.org/10.1016/S0306-4573(01)00048-6). [344]
- 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. [43]
- Tong, Simon, and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *JMLR* 2:45–66. [320]
- Toutanova, Kristina, and Robert C. Moore. 2002. Pronunciation modeling for improved spelling correction. In *Proc. ACL*, pp. 144–151. [60]
- 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: <http://doi.acm.org/10.1145/1148170.1148328>. [368]
- Trotman, Andrew. 2003. Compressing inverted files. *IR* 6(1):5–19. doi: <http://dx.doi.org/10.1023/A:1022949613039>. [98]
- 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. [199]
- Trotman, Andrew, Shlomo Geva, and Jaap Kamps (eds.). 2007. *Proc. SIGIR 2007 Workshop on Focused Retrieval*. University of Otago, Dunedin, New Zealand. [199]
- Trotman, Andrew, Nils Pharo, and Miro Lehtonen. 2006. XML-IR users and use cases. In *Proc. INEX*, pp. 400–412. [198]
- Trotman, Andrew, and Börkur Sigurbjörnsson. 2004. Narrowed Extended XPath I (NEXI). In Fuhr et al. (2005), pp. 16–40. doi: [http://dx.doi.org/10.1007/11424550\\_2](http://dx.doi.org/10.1007/11424550_2). [199]
- 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*. [43]
- Tsochantaridis, Ioannis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. *JMLR* 6:1453–1484. [319]

- Turpin, Andrew, and William R. Hersh. 2001. Why batch and user evaluations do not give the same results. In *Proc. SIGIR*, pp. 225–231.
- Turpin, Andrew, and William R. Hersh. 2002. User interface effects in past batch versus user experiments. In *Proc. SIGIR*, pp. 431–432.
- 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. [161]
- Turtle, Howard. 1994. Natural language vs. Boolean query evaluation: A comparison of retrieval performance. In *Proc. SIGIR*, pp. 212–220. ACM Press. [15]
- Turtle, Howard, and W. Bruce Croft. 1989. Inference networks for document retrieval. In *Proc. SIGIR*, pp. 1–24. ACM Press. [215]
- Turtle, Howard, and W. Bruce Croft. 1991. Evaluation of an inference network-based retrieval model. *TOIS* 9(3):187–222. [215]
- Turtle, Howard, and James Flood. 1995. Query evaluation: strategies and optimizations. *IP&M* 31(6):831–850. doi: [http://dx.doi.org/10.1016/0306-4573\(95\)00020-H](http://dx.doi.org/10.1016/0306-4573(95)00020-H). [123]
- Vaithyanathan, Shivakumar, and Byron Dom. 2000. Model-based hierarchical clustering. In *Proc. UAI*, pp. 599–608. Morgan Kaufmann. [368]
- van Rijsbergen, C. J. 1979. *Information Retrieval*, 2nd edition. Butterworths. [159, 198, 203, 213, 216]
- van Rijsbergen, C. J. 1989. Towards an information logic. In *SIGIR*, pp. 77–86. ACM Press. doi: <http://doi.acm.org/10.1145/75334.75344>. [xviii]
- 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. [200]
- Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley-Interscience. [319]
- Vittaut, Jean-Noël, and Patrick Gallinari. 2006. Machine learning ranking for structured information retrieval. In *Proc. ECIR*, pp. 338–349. [199]
- Voorhees, Ellen M. 1985a. The cluster hypothesis revisited. In *Proc. SIGIR*, pp. 188–196. ACM Press. [344]
- Voorhees, Ellen M. 1985b. *The effectiveness and efficiency of agglomerative hierarchical clustering in document retrieval*. Technical Report TR 85-705, Cornell. [367]
- Voorhees, Ellen M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *IP&M* 36:697–716. [160]
- Voorhees, Ellen M., and Donna Harman (eds.). 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press. [159, 453, 461]
- Wagner, Robert A., and Michael J. Fischer. 1974. The string-to-string correction problem. *JACM* 21(1):168–173. doi: <http://doi.acm.org/10.1145/321796.321811>. [59]
- Ward Jr., J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58:236–244. [367]
- Wei, Xing, and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proc. SIGIR*, pp. 178–185. ACM Press. doi: <http://doi.acm.org/10.1145/1148170.1148204>. [384]
- Weigend, Andreas S., Erik D. Wiener, and Jan O. Pedersen. 1999. Exploiting hierarchy in text categorization. *IR* 1(3):193–216. [319]
- 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. [319]
- Williams, Hugh E., and Justin Zobel. 2005. Searchable words on the web. *International Journal on Digital Libraries* 5(2):99–105. doi: <http://dx.doi.org/10.1007/s00799-003-0050-z>. [97]
- Williams, Hugh E., Justin Zobel, and Dirk Bahle. 2004. Fast phrase querying with combined indexes. *TOIS* 22(4):573–594. [41]
- Witten, Ian H., and Timothy C. Bell. 1990. Source models for natural language text. *International Journal Man-Machine Studies* 32(5):545–579. [97]

## Bibliography

467

- Witten, Ian H., and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. Morgan Kaufmann. [342]
- Witten, Ian H., Alistair Moffat, and Timothy C. Bell. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd edition. Morgan Kaufmann. [76, 97, 98]
- Wong, S. K. Michael, Yiyu Yao, and Peter Bollmann. 1988. Linear structure in information retrieval. In *Proc. SIGIR*, pp. 219–232. ACM Press. [320]
- Woodley, Alan, and Shlomo Geva. 2006. NLPX at INEX 2006. In *Proc. INEX*, pp. 302–311. [200]
- Xu, Jinxi, and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proc. SIGIR*, pp. 4–11. ACM Press. [177]
- Xu, Jinxi, and W. Bruce Croft. 1999. Cluster-based language models for distributed retrieval. In *Proc. SIGIR*, pp. 254–261. ACM Press. doi: <http://doi.acm.org/10.1145/312624.312687>. [344]
- Yang, Hui, and Jamie Callan. 2006. Near-duplicate detection by instance-level constrained clustering. In *Proc. SIGIR*, pp. 421–428. ACM Press. doi: <http://doi.acm.org/10.1145/1148170.1148243>. [344]
- 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. [291]
- Yang, Yiming. 1999. An evaluation of statistical approaches to text categorization. *IR* 1:69–90. [319]
- Yang, Yiming. 2001. A study of thresholding strategies for text categorization. In *Proc. SIGIR*, pp. 137–145. ACM Press. doi: <http://doi.acm.org/10.1145/383952.383975>. [292]
- Yang, Yiming, and Bryan Kisiel. 2003. Margin-based local regression for adaptive filtering. In *Proc. CIKM*, pp. 191–198. ACM Press. doi: <http://doi.acm.org/10.1145/956863.956902>. [292]
- Yang, Yiming, and Xin Liu. 1999. A re-examination of text categorization methods. In *Proc. SIGIR*, pp. 42–49. ACM Press. [265, 319]
- Yang, Yiming, and Jan Pedersen. 1997. Feature selection in statistical learning of text categorization. In *Proc. ICML*. [265]
- Yue, Yisong, Thomas Finley, Filip Radlinski, and Thorsten Joachims. 2007. A support vector method for optimizing average precision. In *Proc. SIGIR*. ACM Press. [320]
- 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: [http://dx.doi.org/10.1016/S1389-1286\(99\)00054-7](http://dx.doi.org/10.1016/S1389-1286(99)00054-7). [344, 368]
- 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. [232]
- Zavrel, Jakub, Peter Berck, and Willem Lavrijssen. 2000. Information extraction by text classification: Corpus mining for features. In *Proc. Workshop Information Extraction meets Corpus Linguistics*. URL: <http://www.cnts.ua.ac.be/Publications/2000/ZBL00>. Held in conjunction with LREC-2000. [292]
- 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. ACM Press. [345, 368]
- Zhai, Chengxiang, and John Lafferty. 2001a. Model-based feedback in the language modeling approach to information retrieval. In *Proc. CIKM*. ACM Press. [231]
- 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. [232]
- Zhai, Chengxiang, and John Lafferty. 2002. Two-stage language models for information retrieval. In *Proc. SIGIR*, pp. 49–56. ACM Press. doi: <http://doi.acm.org/10.1145/564376.564387>. [233]

- Zhang, Jiangong, Xiaohui Long, and Torsten Suel. 2007. Performance of compressed inverted list caching in search engines. In *Proc. CIKM*. ACM Press. [98]
- 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](http://citeseer.ist.psu.edu/zhang00text.html). [319]
- Zhao, Ying, and George Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *Proc. CIKM*, pp. 515–524. ACM Press. doi: <http://doi.acm.org/10.1145/584792.584877>. [367]
- Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley. [97]
- Zobel, Justin. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proc. SIGIR*, pp. 307–314. [160]
- Zobel, Justin, and Philip Dart. 1995. Finding approximate matches in large lexicons. *Software Practice and Experience* 25(3):331–345. URL: [citeseer.ifi.unizh.ch/zobel95finding.html](http://citeseer.ifi.unizh.ch/zobel95finding.html). [60]
- Zobel, Justin, and Philip Dart. 1996. Phonetic string matching: Lessons from information retrieval. In *Proc. SIGIR*, pp. 166–173. ACM Press. [60]
- Zobel, Justin, and Alistair Moffat. 2006. Inverted files for text search engines. *ACM Computing Surveys* 38(2). [17, 76, 98, 122]
- Zobel, Justin, Alistair Moffat, Ross Wilkinson, and Ron Sacks-Davis. 1995. Efficient retrieval of partial documents. *IP&M* 31(3):361–377. doi: [http://dx.doi.org/10.1016/0306-4573\(94\)00052-5](http://dx.doi.org/10.1016/0306-4573(94)00052-5). [199]
- 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: <http://dx.doi.org/10.1109/ICDE.2006.150>. [98]

## Index

- A/B test, 156
- Accents, 27–28
- Access control lists, 74
- Accumulator, 103, 115
- Accuracy, 143
- Active learning, 309
- Add-one smoothing, 240
- Ad hoc retrieval
  - defined, 4–5
  - evaluation of, 139–141
  - machine learning methods, 314–318, 320
- Adjacency tables, 417
- Adjusted Rand index, 330
- Adversarial information retrieval, 392
- Akaike information criterion (AIC), 337
- Algebra, linear, review, 369–373
- Algorithmic search, 393
- Anchor text, 389, 422–423
- Any-of classification, 238, 281
- Auxiliary index, 71
- Average-link clustering, 350, 356–358
- Back queues, 412–415
- Bag of words model. *See also* Unigram language model
  - defined, 107, 113–114
- Balanced F measure, 144. *See also* F measure
- Bayes error rate, 277
- Bayesian networks, 215–216
- Bayesian prior, 208, 210
- Bayesian smoothing, 226
- Bayes Optimal Decision Rule, 203
- Bayes risk, 203
- Bayes' Rule, 202
- Bernoulli model, 243–251
- Best-merge persistence, 355
- Bias, 286
- Bias-variance tradeoff, 284–289, 292
- Biclustering, 345
- Bigram language model, 221–222
- Binary Independence Model (BIM), 204–212, 229–230
- Binary search tree, 46, 47
- Biword indexes, 36–38
- Blind relevance feedback, 171–172
- Blocked sort-based indexing algorithm (BSBI), 63–66, 75
- Blocked storage described, 85–87
- Blogs, 178
- BM25 weights, 213–215
- Boolean retrieval
  - defined, xvi
  - model, 4
  - principles, 3–6
  - query processing, 9–13
  - ranked retrieval *vs.*, 13–16
  - tokenization, 26
  - vector space model interactions, 136
- Boosting, 264
- Bottom-up clustering. *See* hierarchical agglomerative clustering (HAC)
- Bowtie structure, WWW, 389
- Break-even point, 148, 261, 306
- BSBI (blocked sort-based indexing algorithm), 66, 75
- B-trees, 47–48
- Buckshot algorithm, 366
- Buffer, 62
- Caching
  - compression and, 78
  - defined, 62
  - in search systems, 135, 409, 411
  - variable length arrays and, 9
- Capitalization, 26
- Capture-recapture method, 396–400

- Cardinality in clustering, 327, 336–338
- Case-folding, 26
- CAS topics, 193
- Category, 237
- Centroid-based classification, 291
- Centroids
  - defined, 331
  - HAC, 350, 358–359
  - Rocchio classification, 269, 271
- Chaining in clustering, 352
- Chain rule, 202
- Champion lists, 127–128
- Character sequence decoding, 18–21
- $\chi^2$  feature selection, 256, 258
- Chinese, 23–24, 47
- Class boundary, 279
- Classes
  - defined, 238
  - maximum a posteriori, 239
- Classification. *See also* Text classification
  - any-of, 281
  - centroid-based, 291
  - defined, 234, 235
  - kNN (*See* k nearest neighbor classification (kNN))
  - multivalued, 281
  - one-of, 282
  - one-versus-all, 303
  - Rocchio (*See* Rocchio classification)
- Classification function, 237
- Classifiers
  - choosing, 308–309
  - defined, 237
  - performance, improving, 309–313
  - two-class, 259, 267, 292
- CLEF collection, 142
- Click spam, 394
- Clickstream mining, 172
- Clickthrough log analysis, 156, 172
- Cliques, 351
- Cloaking, in spamming, 391
- Cluster-based classification, 291
- Cluster hypothesis, 322–323, 325, 344
- Clustering
  - average-link, 350, 358
  - cardinality in, 327, 338
  - centroid-based, 362, 367
  - chaining in, 352
  - complete-link HAC, 360
  - divisive, 362–363
  - exclusive *vs.* exhaustive, 327
  - flat (*See* Flat clustering)
  - function notations, xi
  - group-average agglomerative, 350, 358, 360, 362, 367
  - hard, 322
  - hierarchical, 346 (*See also* Hierarchical clustering)
  - minimum variance, 367
  - model-based, 338–342
  - optimal, 362
  - overview, 322–326
  - single-link HAC, 359, 360, 362
  - spectral, 368
  - top-down, 363
- Clusters
  - defined, 68, 321
  - labeling, 363–365, 367–368
  - pruning, 129–131
- Co-clustering, 345
- Collections
  - clustering, 325
  - defined, 4
  - frequency, 25, 108–109
  - residual defined, 171
  - statistics, large, 75
- Combination schemes, 40–42
- Combination similarity, 347, 351, 360
- Complete-linkage clustering. *See* Complete-link clustering
- Component coverage, 193–194
- Compound nouns, 24
- Compound-splitter, 24
- Compression
  - of dictionaries, 82–87, 102
  - of docIDs, 88
  - lossless/lossy, 80
  - parameter-free, 92
  - parameterized, 98
  - of postings list, 87–95
- Compression/indexes
  - Heaps' law, 80–82, 276–277
  - overview, 78
  - Zipf's law, 82–83
- Concept drift, 249
- Conditional independence assumption, 246
- Confusion matrix, 283
- Connected components, 351
- Connectivity queries, 416
- Connectivity servers, 419
- Content management systems, 77
- Content seen module, 410–411
- Context, XML, 181
- Context resemblance, 190
- Contiguity hypothesis, 266
- Continuation bit, 89
- Corpus, 4
- Cosine similarity, 111, 112, 121, 344

*Index*

471

- CO topics, 193
- CPC (cost per click), 393
- CPM (cost per mil), 393
- Cranfield collection, 141–142
- Cross-entropy, 232
- Cross-language information retrieval, 142, 384
- Cumulative gain, 149
- Databases
  - communication with, 77
  - relational, 178–179, 197
- $\Delta$ -codes, 96, 98
- Decision boundaries
  - defined, 269
  - kNN, 274
- Decision hyperplanes, 267, 278
- Decision trees, 261
- Dendrograms
  - complete-link clustering, 352
  - described, 347, 348
- Development sets, 262
- Development test collection, 141
- Diacritics, 28
- Dice coefficient, 150
- Dictionaries
  - compression of, 87, 102
  - in inverted indexes, 5–7
  - search structures for, 45–47
- Differential cluster labeling, 365
- Digital libraries, 178
- Discrete-time stochastic processes, 425
- Disk seek, 62
- Distortion, 336
- Distributed crawling, 419
- Distributed index, 68
- Distributed indexing, 67–70, 415–416
- Distributed information retrieval, 70, 416
- Divisive clustering, 363
- DNS resolution, 411–412
- DNS resolution module, 408
- DNS server, 411
- DocIDs
  - compression of, 88
  - in inverted indexes, 7
  - in postings list intersection operations, 10
- Document-at-a-time scoring, 129
- Document collection. *See* Collections
- Document likelihood model, 231
- Document-partitioned index, 68
- Documents
  - character sequence decoding, 21
  - classification of (*See* Text classification)
  - defined, 4
  - delineation of, 21
  - frequency defined, 7
  - function notations, xi
  - partitioning, 416
  - relevant, retrieving, xvii
  - unit, choosing, 20–21
  - vector, defined, 109–110
- Document space, 237
- Document zones, 312–313
- Doorway pages, 392
- Dot products
  - described, 110–113
  - in SVMs, 298
- Duplicate elimination modules, 408
- Dynamic indexing, 71
- Dynamic summary, 157
- East Asian languages, 43. *See also* Chinese; Japanese
- Edit distance, 53–55
- Effectiveness
  - assessment of, 5
  - text classification, 259, 261
- Efficiency, 259
- Eigen decomposition, 372
- Eigenvalues, 370, 425
- 11-point interpolated average precision, 146
- Email
  - document units, 20
  - sorting, 2, 235
- EM algorithm, 339–341
- Enterprise resource planning, 77
- Enterprise search, 61
- Entropy, 91, 330
- Equivalence classes, 26
- Ergodic Markov Chain, 427
- Euclidean distance, 121, 296–297, 344
- Euclidean length, 111
- Evaluation of retrieval systems
  - A/B test, 156
  - ad hoc, 141
  - clustering, 327–331
  - F measure, 144, 331
  - interpolated precision, 145
  - kappa statistic, 151, 152, 160
  - keyword-in-context snippets, 158
  - MAP, 147
  - marginal relevance, 154
  - normalized discounted cumulative gain, 149
  - overview, 141
  - pooling, 151
  - precision at k, 148
  - precision-recall curve, 145, 146

- Evaluation of retrieval systems (*cont.*)
  - probabilistic information retrieval, 212–213
  - ranked sets, 145–151
  - relevance assessment, 154
  - relevance feedback, 170–171
  - results snippets, 159
  - ROC curve, 149
  - R-precision, 148, 160
  - sensitivity, 149
  - specificity, 149
  - summarization, static *vs.* dynamic, 157
  - system quality/user utility, 156
  - test collections, standard, 142
  - text classification, 258–263
  - text summarization, 157
  - unranked sets, 142–145
  - XML retrieval, 192–196
- Evidence accumulation, 134
- Exclusive clustering, 327
- Exhaustive clustering, 327
- Expectation-Maximization (EM)
  - algorithm, 340, 341
- Expectation step, 340
- Expected edge density, 344–345
- Extended query, 187–188
- Extensible Markup Language. *See* XML
- External criterion of quality, 328, 329
- External sorting algorithm, 63
- False negative, 330
- False positive, 330
- Feature engineering, 311
- Feature selection/text classification
  - $\chi^2$ , 255–256
  - frequency-based, 257
  - greedy, 258
  - method comparison, 257–258
  - multiple classifiers, 257
  - mutual information, 252–255
  - noise feature, 251
  - overfitting, 251
  - overview, 251–252
  - in performance improvement, 310–312
  - statistical significance, 256
- Fetch modules, 408
- Field, 101
- Filtering, 234, 291
- First story detection, 362
- Flat clustering
  - Akaike information criterion, 337
  - cardinality in, 327, 338
  - classification *vs.*, 321
  - collections, 325
  - defined, 321
  - distortion, 336
  - evaluation of, 331
  - exhaustive, 327
  - Expectation-Maximization algorithm, 340, 341
  - expectation step, 340
  - external criterion of quality, 328, 329
  - HAC *vs.*, 367
  - internal criterion of quality, 327
  - K means, 331–338
  - K-medoids, 336
  - in language models, 325
  - maximization step, 340
  - model complexity, 336
  - normalized mutual information, 329
  - objective functions, 326
  - outliers, 334
  - partitional, 326–327
  - purity, 328, 329
  - Rand index, adjusted, 330
  - residual sum of squares, 337
  - scatter-gather, 323, 324, 344
  - search result, 323
  - seeds, 332
  - singleton, 334
  - soft, 322, 382
  - unsupervised learning, 321
- F measure, 144, 160, 331
- Focused retrieval, 199
- Free text, 100, 136–137
- Free text query
  - parsing functions, designing, 133–134
  - tokenization, 26
  - in vector retrieval models, 13
- Frequency-based feature selection, 257
- Frobenius norm, 376
- Front coding, 86, 87
- Front queues, 415
- Functional margins, 296
- GAAC. *See* Group-average
  - agglomerative clustering
- $\gamma$  encoding, 90–95
- Gaps, encoding, 88
- Generative model, 218–220
- Geometric margin, 297
- Global champion list, 128
- Gold standard, 140
- Golomb codes, 98
- GOV2 collection, 142
- Greedy feature selection, 258
- Grepping, 3



*Index*

473

- Ground truth, 140
- Group-average agglomerative clustering, 350, 358, 360, 362, 367
- Group-average clustering. *See* Group-average agglomerative clustering
- HAC. *See* hierarchical agglomerative clustering (HAC)
- Hard assignment, 322
- Hard clustering, 326
- Harmonic numbers, 93
- Hashing, 46, 86–87
- Heaps' law, 82, 277
- Held-out data, 262
- Hierarchical agglomerative clustering (HAC)
  - algorithm comparison, 362
  - best-merge persistence, 355
  - Buckshot algorithm, 366
  - centroids, 350, 359, 362, 367
  - chaining in, 352
  - cliques, 351
  - cluster-internal labeling, 365
  - combination similarity, 347, 360
  - complete-link clustering, 359, 360, 362, 367
  - connected components, 351
  - dendrograms, 347, 348, 352
  - differential cluster labeling, 365
  - divisive, 363
  - first story detection, 362
  - flat *vs.*, 367
  - group-average, 350, 358, 360, 362, 367
  - inversions, 347, 359
  - monotonicity, 347
  - next-best merge (NBM) arrays, 355
  - novelty detection, 362
  - optimality, 362
  - outliers, 353
  - overview, 347–349
  - priority queue algorithm, 353, 354
  - single-link clustering, 359–360, 362
  - time complexity, 353–356
  - top-down, 363
- Hierarchical classification, 319
- Hierarchical clustering, 346
  - agglomerative (*See* hierarchical agglomerative clustering (HAC))
  - applications, 346–347
  - defined, 321
  - probabilistic interpretation of, 368
  - R environment support for, 368
- Hierarchical Dirichlet processes, 384
- Hierarchy, 346
- Highlighting, 186
- HITS (hyperlink-induced topic search), 435, 437, 439
- Host splitters, 410
- HTML, 385
- http, 385
- Hub score, 433–439
- Hyperlink-induced topic search (HITS), 435, 437, 439
- Hyperlinks, 389. *See also* Link analysis
- Hyphenation and tokenization, 24
- Ide dec-hi, 167
- IDF. *See* Inverse document frequency (IDF)
- IID. *See* Independent and identically distributed (IID)
- Images, searching for. *See* Relevance feedback
- Impact ordering, 129
- Implicit relevance feedback, 172
- Incidence matrix, 374
- Independence, 255
- Independent and identically distributed (IID), 262
- Index construction
  - BSBI, 66, 75
  - distributed indexes, 68, 419
  - resources, 76
- Indexer, 61
- Indexes
  - biword, 38
  - defined, 3
  - document-partitioned, 70, 416
  - k-gram, 50–51, 55–57, 311
  - next word, 41
  - parametric, 101–107
  - permuterm, 49–50
  - positional, 38–40
  - size/estimation, 400
  - term-partitioned, 70
  - zone, 107
- Indexing
  - defined, 61
  - distributed, 70, 416
  - granularity, 20
  - latent semantic, 378–382
  - unit defined, 184
- INEX, 196
- Informational queries, 395
- Information gain, 264
- Information need, 5, 140

- Information retrieval
  - hardware issues, 63
  - history of, 17
  - overview, 3, xvi
  - search system components, 135, 135
  - terms, statistical properties of, 82
- In-links, 389
- Inner product. *See* Dot products
- Instance-based learning, 276
- Internal criterion of quality, 327
- Interpolated precision, 145
- Intersection, postings list, 10, 36
- Inter-similarity, 350
- Inverse document frequency (IDF), 109, 190, 209
- Inversions
  - defined, 64
  - in HAC, 347, 358
- Inverted file. *See* Inverted index;
- Postings list
- Inverted index
  - Boolean query processing, 13
  - building principles, 9
  - described, 6
  - $\gamma$  encoding, 90, 95
  - kNN classification in, 277
- Inverter, 69–70
- IP address, 411
- Jaccard coefficient, 56, 401
- Japanese, 29–30
- Journal influence weight, 439
- Kappa statistic, 151, 152, 160
- Kernel function, 305
- Kernels
  - Mercer, 305
  - polynomial, 305
  - quadratic, 305
  - radial basis functions, 305
- Kernel trick, 304
- Keys, 46
- Key-value pairs, 68
- Keyword-in-context (KWIC) snippets, 158
- k-gram index
  - described, 51
  - spelling correction in, 57
  - word matching in, 311
- K means, 338
- K-medoids, 336
- k nearest neighbor classification (kNN)
  - algorithm, 273–275
  - Bayes error rate, 277
  - bias in, 286–287
  - decision boundaries, 274
  - described, 267, 291–292
  - effectiveness, 261, 292
  - instance-based learning, 276
  - memory-based learning, 276
  - memory capacity, 287
  - multinomial Naive Bayes *vs.*, 249
  - as nonlinear classification, 280–281
  - testing/training capacity, 302
  - time complexity/optimality, 275–277
  - variance, 287
  - Voronoi tessellation, 273, 274
- KNN classification. *See* K nearest neighbor classification (kNN)
- Kruskal's algorithm, 367
- Kullback-Leibler divergence, 231, 344
- KWIC (keyword-in-context), 158
- Labeling
  - of clusters, 368
  - defined, 236
- Language, of an automaton, 219
- Language identification, 22
- Language issues, relevance feedback, 169–170
- Language models
  - Bayesian smoothing, 226
  - BIM/XML *vs.*, 230
  - clustering in, 325
  - defined, 219, 224
  - distributions, multinomial, 222–223
  - document likelihood, 231
  - extended approaches, 230–232
  - finite automata and, 220
  - Kullback-Leibler divergence, 231
  - likelihood ratio, 220
  - linear interpolation, 226
  - overview, 218
  - query likelihood, 223–229
  - tf-idf weighting *vs.*, 228
  - translation, 232
  - types of, 222
- Laplace smoothing, 240
- Latent Dirichlet Allocation (LDA), 384
- Latent semantic analysis (LSA), 379
- Latent semantic indexing (LSI), 382
- LDA (Latent Dirichlet Allocation), 384
- L2 distance, 121, 297, 344
- Learning algorithm described, 103–106.  
*See also* Weighted zone scoring
- Learning error, 285
- Learning method, 237
- Lemma, 31
- Lemmatization described, 30–33
- Length-normalization, 111

*Index*

475

- Lemmatizer, 32
- Levenshtein distance, 55
- Lexicalized subtrees, 188–189
- Lexicons in inverted indexes, 6
- Likelihood, 202
- Likelihood ratio, 220
- Linear algebra review, 373
- Linear classifiers, 267, 277–281
- Linear interpolation, 226
- Linear problem, 279
- Linear separability, 280, 294–300
- Link analysis
  - anchor text, 389, 423
  - authority score, 439
  - ergodic Markov chain, 427
  - HITS, 435, 437
  - hub score, 439
  - Markov chains, 427
  - overview, 421
  - PageRank (*See* PageRank)
  - steady-state theorem, 427
- Link farms, 439
- Link spam, 421
- LLRUN, 98
- LM, 224. *See* Language models
- Logarithmic merging, 72
- Lossless compression, 80
- Lossy compression, 80
- Lovins stemmer, 32
- Low-rank approximation, 376–378
- LSA (latent semantic analysis), 379
- LSI (latent semantic indexing), 382
- Machine-learned relevance described, 106
- Machine learning methods, 318, 320
- Machine translation, 224
- Macroaveraging, 259–261
- MAP (mean average precision), 239
- Map phase, 69
- MapReduce, 69, 70, 76
- Marginal relevance, 154
- Marginal statistic, 152
- Margins, 295, 298
- Markov chains, 427
- Master node, 68
- Matrix decomposition
  - eigen, 372
  - eigenvalues, 370
  - Frobenius norm, 376
  - latent semantic indexing, 382
  - linear algebra review, 373
  - low-rank approximation, 378
  - reduced SVD, 374
  - singular value, 373–376
  - symmetric diagonal, 373, 374
  - theorems, 372–373
  - truncated SVD, 374
- Maximization step, 340
- Maximum a posteriori, 208
- Maximum likelihood estimate (MLE), 208, 224–227, 240, 252
- Mean average precision, 147
- Medoids, 336
- Memory-based learning, 276
- Memory capacity, 287
- Mercator crawler, 407, 419
- Mercer kernels, 305
- Merge algorithm, 10
- Merge postings list, 10, 65
- Metadata, 101
- Microaveraging, 261
- Minimum spanning tree, 367
- Minimum variance clustering, 367
- ModApte split, 259, 265
- Model-based clustering, 342
- Model complexity, 336
- Monotonicity, 347
- Multiclass classification, 282
- Multiclass SVMs, 303
- Multilabel classification, 281
- Multimodal class, 272
- Multinomial classification, 282
- Multinomial model, 242–243
- Multinomial Naive Bayes
  - Bernoulli model, 245, 251
  - bias in, 286
  - concept drift, 249
  - conditional independence
    - assumption, 246
  - as linear classifier, 278
  - optimal classifier, 250
  - positional independence assumption, 240, 247
  - properties, 251
  - in query likelihood models, 224
  - random variables  $X$  and  $U$ , 246
  - semi-supervised learning, 308
  - sparseness, 240
  - testing/training capacity, 302
  - in text classification, 243
  - variance, 287
- Multinomial NB. *See* Multinomial Naive Bayes
- Multivalued classification, 281
- Multivariate Bernoulli model, 245
- Mutual information, 255, 258

- Naive Bayes assumption, 168, 206
- Naive Bayes learning method, 237.
  - See also* Multinomial Naive Bayes;
  - Multivariate Bernoulli model
- Named entity tagging, 178
- National Institute of Standards and Technology, 141
- Natural language processing
  - issues in, 342
  - lemmatizers in, 32
  - text summarization, 313
  - XML retrieval, 230
- Navigational queries, 395
- NDCG (normalized discounted cumulative gain), 149
- Near-duplicate search results, 400–403
- Nested elements, 185–186
- NEXI, 182
- Next-best merge (NBM) arrays, 355
- Next word index, 41
- N-gram language model, 43. *See also*
  - Bigram language model; Unigram language model
- Nibble, 90
- NLP. *See* Natural language processing
- NMI. *See* Normalized mutual information (NMI)
- Noise documents, 279–280
- Noise feature, 251
- Nonlinear classifiers, 280, 303–306
- Nonlinear problem, 281
- Normalization
  - in probability theory, 225
  - term, 26–30
  - tf weighting, 117
  - URL, 409
- Normalized discounted cumulative gain (NDCG), 149
- Normalized mutual information (NMI), 329
- Normalized tokens in inverted indexes, 7
- Normal vectors, 270
- Notation, table of, xi
- Novelty detection, 362
- NTCIR collection, 142
- Objective function, 326, 332
- Odds, 203
- Odds ratio, 207
- Okapi BM25 weighting, 213
- 1/0 loss, 203
- One-of classification, 238, 263
- One-versus-all (OVA) classification, 303
- Optimal classifier, 285
- Optimal clustering, 362
- Optimal learning method, 285
- Optimal weight, 106
- Ordering, 127–129
- Ordinal regression, 317
- Outliers, 334, 353
- Out-links, 389
- Overfitting, 287
- Overlap score measure, 109
- Oxford English Dictionary, 80
- PageRank
  - computation, 427–430, 439
  - described, 424–425
  - ergodic Markov chain, 427
  - Markov chains, 427
  - personalized, 431
  - principal left eigen vector, 425
  - probability vectors, 426
  - steady-state theorem, 427
  - stochastic matrix, 425
  - teleport operation, 424
  - topic-specific, 430–432
- Paice stemmer, 32
- Paid inclusion, 391
- Parameter-free compression, 92
- Parameterized compression, 98
- Parameter tuning, 141, 291
- Parameter tying, 312
- Parametric indexes, 107
- Parametric search, 180
- Parser, 69
- Parsing functions, designing, 134
- Parsing modules, 408
- Partitional clustering, 327
- Partition rule, 202
- Passage retrieval, 199
- Patent databases, 178
- Performance, 259
- Permuterm index, 50
- Personalized PageRank, 431
- Pew Internet Survey 2004, xv
- Phonetic correction, 58–59
- Phrase index, 37
- Phrase queries, 36–42, 44, 137
- Phrase search, 14
- Pivoted document length normalization, 121
- Pivot length, 120
- Pointwise mutual information, 265
- Polytymous classification, 282
- Polytopes, 274
- Pooling, 160
- Pornography filtering, 311

*Index*

477

- Porter stemmer, 31, 32
- Positional independence assumption, 240, 247
- Positional indexes, 40
- Posterior probability, 202
- Postfiltering, in k-gram indexes, 51
- Postings
  - in block sort-based indexing, 64
  - compression and, 79
  - defined, 6, 79
  - in inverted indexes, 7
  - positional, 42
- Postings list
  - compression of, 95
  - described, 6
  - intersection/merging, 10
  - skip pointers, 36
  - storage of, 9
- Power law, 389
- Precision, 5, 142
- Precision at k, 148
- Precision-recall curve, 145, 146
- Prefix-free code, 92
- Preprocessing, effects of, 80
- Principal direction divisive partitioning, 368
- Priority queue algorithm, HAC, 353, 354
- Prior probability, 202
- Probabilistic information retrieval
  - Bayesian networks, 215
  - Bayesian prior, 208
  - Bayes Optimal Decision Rule, 203
  - Binary Independence Model, 212
  - evaluation, 213
  - maximum a posteriori, 208
  - maximum likelihood estimate, 208, 227, 240, 252
  - Naive Bayes assumption, 168, 206
  - odds ratio, 207
  - overview, 201
  - probability theory principles, 202–203
  - pseudocounts, 208
  - query generation, estimating, 227
  - relative frequency, 208
  - relevance feedback, 209–211
  - Retrieval Status Value, 207
  - tree-structured dependencies, 213
- Probability Ranking Principle, 204
- Probability vectors, 426
- Prototypes, 267
- Proximity operator, 14
- Proximity weighting, 132–133
- Pseudocounts, 208
- Pseudo-relevance feedback described, 172
- Pull model, 291
- Purity, 328, 329
- Push model, 291
- Quadratic optimization, 298
- Queries. *See also* Terms
  - BIM ranking function, deriving, 205–207
  - Boolean, 4, 13
  - defined, 5
  - expansion, 173–175
  - extended, 188
  - free text (*See* Free text query)
  - generation probability, estimating, 227
  - informational, 395
  - navigational, 395
  - optimization of, 10
  - phrase (*See* Phrase queries)
  - semistructured, 180
  - simple conjunctive, 9
  - structured, 180
  - term highlighting, 159, 186
  - transactional, 396
  - user/web search, 395–396
  - as vectors, 114
- Query-by-example, 183, 230
- Query likelihood model, 229
- Query parser, 134
- Query reformulation
  - expansion, 175
  - local *vs.* global, 162
  - vocabulary tools for, 173
- Radial basis functions, 305
- Rand index, adjusted, 330, 344
- Random variables
  - C, 248
  - defined, 202
  - U, 246
  - X, 246
- Rank, of matrices, 369
- Ranked Boolean retrieval, 103. *See also* Weighted zone scoring
- Ranked retrieval models
  - Boolean retrieval *vs.*, 16
  - described, 74
  - evaluation of, 151
- Ranking/results
  - BIM function, deriving, 207
  - efficiency in, 124–125
  - machine learning, 316–318
- Ranking SVM, 317
- Recall, 5, 143

478

Index

- Reduced SVD, 378
- Reduce phase, 69
- Regression, 317
- Regular expressions, 3, 17
- Regularization, 301
- Relational databases, 179, 197
- Relative frequency, 208
- Relevance
  - assessment of, 154, 160
  - defined, 5
- Relevance feedback
  - applications, 170
  - evaluation of, 171
  - images, 163–164
  - implicit/indirect, 172
  - overview, 172–173
  - probabilistic models, 168, 211
  - pseudo-relevance, 172
  - Rocchio algorithm (*See* Rocchio algorithm)
  - text, 165
  - Web applications, 170
- R environment, 368
- Residual collection, 171
- Residual sum of squares (RSS), 332, 337
- Results snippets, 135
- Retrieval model, Boolean. *See* Boolean retrieval
- Retrieval Status Value, 207
- Reuters-21578 collection
  - confusion matrix, 283
  - described, 142
  - as linear, 279
  - text classification in, 259, 260, 261
- Reuters-RCV1 collection
  - blocked storage, 87
  - collection *vs.* document frequency, 109
  - construction of, 66, 75
  - described, 63–64, 77
  - dictionary-as-a-string storage, 83–85
  - dictionary compression, 95, 95
  - $\gamma$  - encoding, 92, 94
  - index compression, 95
  - preprocessing, effects of, 80
  - residual sum of squares, 337
  - Zipf's law, 82, 83
- RF. *See* Relevance feedback
- Robots Exclusion Protocol, 408
- Rocchio algorithm
  - applications, 170
  - overview, 163–168
- Rocchio classification
  - bias in, 286
  - centroids, 269–273
  - decision boundaries, 269
  - described, 273
  - effectiveness, 261, 292
  - as linear, 278
  - memory capacity, 287
  - multimodal class, 272
  - normal vectors, 270, 271
  - prototypes, 267
  - testing/training capacity, 302
  - variance, 287
- ROC curve, 149
- Routing, 234, 291
- R-precision, 148, 160
- RSS. *See* Residual sum of squares (RSS)
- Rule of 30, 79
- Rules in text classification, 236
- Scatter-Gather, 323, 324, 344
- Schema, 182
- Schema diversity/heterogeneity, 186–187
- Scoring
  - champion lists, 127, 128
  - cluster pruning, 131
  - document-at-a-time, 129
  - efficiency in, 125
  - functions, designing, 134
  - index elimination, 126–157
  - machine learning methods, 314–316
  - overview, 100
  - SimNoMerge, computing, 190, 191, 191
  - static quality scores, 129
  - top K document retrieval, 125–126
  - vector scores, computing, 114–116
  - vector space model (*See* Vector space model)
- Search advertising, 393, 394
- Search engines. *See also* Web index
  - components, 396
  - marketing, 394
  - optimizers, 392
- Search result clustering, 323
- Search results, 323
- Search system, complete, 135, 135. *See also* Web index
- Security, 74
- Seeds, 332
- Seed sets, 406
- Seek time, 62
- Segment file, 69
- Semistructured query, 180
- Semistructured retrieval, 179
- Semi-supervised learning, 308
- Sensitivity, 149

## Index

479

- Sentiment detection, 235
- Sequence model, 21–25, 28, 247
- Shingling, 403
- SimNoMerge, computing, 190, 191, 191
- Simple conjunctive queries, 9
- Single-label classification, 282
- Single-linkage clustering. *See* Single-link clustering
- Single-link clustering, 359, 360, 362
- Single-pass in-memory indexing (SPIMI), 67, 76
- Singleton cluster, 334, 347
- Singly-linked lists, 7
- Singular value decomposition (SVD), 376, 380
- Skip list, 36
- Skip pointers, 36
- Slack variables, 301
- SMART notation, 118
- Smoothing
  - add  $\alpha$ , 208
  - add-one, 240
  - add  $\frac{1}{2}$ , 208, 210, 211, 243
  - Bayesian, 226
  - Bayesian prior, 208, 210, 226
  - Laplace, 240
  - linear interpolation, 226
  - query generation estimation, 225–227
  - tf weighting, 117
- Snippet, 157
- Soft assignment, 322
- Soft clustering, 322, 382
- Soft margin classification, 300–303
- Sort-based multiway merge, 76
- Sorting, 7, 76
- Soundex algorithms, 59
- Spam
  - click, 394
  - filters, email, 2
  - link, 421
  - overview, 390–392
- Sparseness, 240
- Specificity, 149
- Spectral clustering, 368
- Speech recognition, 222
- Spelling correction, 51–58
- Spiders. *See* Web crawlers
- Spider traps, 405. *See also* Web crawlers
- SPIMI (single-pass in-memory indexing), 67, 76
- Splits, 68
- Sponsored search, 393
- Standing query, 234
- Static quality scores, 129
- Static summary, 157
- Static web pages, 388
- Statistical significance, 256
- Statistical text classification, 236
- Steady-state theorem, 427
- Stemming described, 33
- Stochastic matrix, 425
- Stop list, 25
- Stop words, 25–26
- Storage
  - blocked, 87
  - dictionary-as-a-string, 85
- Structural SVMs, 303
- Structural term, 189
- Structured document retrieval principle, 184
- Structured query, 180
- Structured retrieval, 178, 183–188
- Sublinear tf scaling, 117
- Summarization
  - in cluster labeling, 368
  - static *vs.* dynamic, 157
  - text, 157
- Supervised learning, 237
- Support vector, 294
- Support vector machines (SVMs)
  - active learning, 309
  - dot products in, 298
  - effectiveness, 262
  - Euclidean distance, 297
  - experimental results, 306–307
  - functional margins, 296
  - geometric margin, 297
  - kernel function, 305
  - kernels, polynomial, 305
  - kernel trick, 304
  - linear separability, 280, 300
  - margins, 294, 295
  - Mercer kernels, 305
  - multiclass, 303
  - nonlinear, 306
  - overview, 293
  - quadratic optimization, 298
  - radial basis functions, 305
  - ranking, 317
  - regularization, 301
  - slack variables, 301
  - soft margin classification, 303
  - structural, 303
  - testing/training capacity, 302
  - transductive, 309
  - weight vectors, 295
- SVD (singular value decomposition), 376, 380

480

Index

- SVMs. *See* Support vector machines (SVMs)
- Symmetric diagonal decomposition, 373, 374
- Synonymy, 162
- Table of notation, xi
- Taxonomies, performance improvement, 310
- Teleport operation, 424
- Term-at-a-time, 115
- Term-document matrix
  - defined, 4–5, 369
  - singular value decomposition, 376, 380
- Term frequency
  - benefits of, 15
  - defined, 107
  - weighting and, 107–110, 112
- TermID, 62
- Term normalization, 30
- Term-partitioned index, 70
- Terms. *See also* Queries
  - BIM ranking function, deriving, 207
  - defined, 3, 21
  - function notations, xi
  - partitioning, 416
  - statistical properties of, 82
  - tree-structured dependencies, 213
  - vectors, weighting and, 113
- Term weighting. *See* Weighting
- Test data, 237
- Test set, 262
- Text, grepping, 3
- Text categorization. *See* Text classification
- Text classification
  - Bernoulli model, 245, 251
  - classes, 237, 238
  - classifiers (*See* Classifiers; specific classifiers)
  - decision trees, 261
  - defined, 234
  - development sets, 262
  - document space, 237
  - document zones, 313
  - effectiveness, 259, 261
  - email sorting (*See* Email)
  - evaluation of, 263
  - feature selection, 251–258
  - held-out data, 262
  - issues in, 307–313
  - labeling, 236
  - learning method, 237
  - linear, 267, 281
  - macroaveraging, 261
  - microaveraging, 261
  - ModApte split, 259
  - multinomial Naive Bayes (*See* Multinomial Naive Bayes)
  - nonlinear, 281
  - overview, 237–238
  - parameter tying, 312
  - performance/efficiency, 259
  - rules in, 236
  - semi-supervised learning, 308
  - sentiment detection, 235
  - statistical, 236
  - supervised learning, 237
  - test sets, 237, 238
  - training sets, 237, 238
  - two-class classifier, 259, 267, 292
  - vertical search engines, 235
- Text summarization, 157, 313
- TF. *See* Term frequency
- Tf-idf weighting, 116–121
- Thesauri
  - automatic generation of, 175
  - query expansion in, 175
- Tiered indexes described, 132–133
- Time complexity in HAC, 356
- Tokenization
  - defined, 18
  - hyphenation and, 24
  - vocabulary/terms, determining, 25
- Tokens
  - defined, 21
  - in inverted indexes, 7
  - normalization of, 30
- Top docs, 137
- Top-down clustering, 363
  - classification of (*See* Text classification)
  - standing queries *vs.*, 234
  - in test collections, 142
  - in XML retrieval, 193
- Topic-specific PageRank, 432
- Topic spotting. *See* Text classification
- Trailing wildcard query, 48
- Training set, 237, 238
- Transactional query, 396
- Transductive SVMs, 309
- Translation model, 232
- TREC collection, 142, 147
- Trec\_eval, 160
- Truecasing, 28
- Truncated SVD, 378, 381
- 20 Newsgroups, 142
- Two-class classifier, 259, 267, 292
- Type, 21



## Index

481

- Unary code, 90, 95
- Unigram language model. *See also* Bag of words model
  - described, 222
  - distributions, multinomial, 223
  - multinomial Naive Bayes *vs.*, 243
- Union-find algorithm, 362, 403
- Universal code, 92
- Unsupervised learning, 321
- URLs
  - defined, 386
  - frontiers, 406, 407,
  - normalization of, 409
- User document matrix, access control lists, 74
- Utility measure, 265
- Variable byte encoding, 88–90
- Variable length arrays, 9
- Variance, 287
- Vector space model. *See also* k nearest neighbor classification (kNN); Rocchio classification
  - any-of classification, 281
  - bias defined, 286
  - bias-variance tradeoff, 289, 292
  - class boundaries, 279
  - confusion matrix, 283
  - contiguity hypothesis, 266
  - decision hyperplanes, 267, 278
  - described, 110–116, 125
  - document representation, 267–269
  - learning error, 285
  - linear classifiers, 267, 281
  - linear separability, 280
  - memory capacity, 287
  - noise documents, 280
  - nonlinear classifiers, 281
  - one-of classification, 282
  - optimal classifier, 250, 285
  - optimal learning method, 285
  - overfitting, 251, 287
  - query operator interactions, 137
  - relatedness measures, 269
  - 3+ classes, 281–283
  - variance, 287
  - XML retrieval, 188–192
- Vertical search engines, 235
- Vocabulary
  - controlled, query expansion and, 175
  - function notations, xi
  - in inverted indexes, 6
  - issues, relevance feedback, 170
  - permuterm, 50
  - Vocabulary/terms, determining
    - common terms, dropping, 26
    - lemmatization/stemming, 33
    - normalization, 30
    - tokenization, 25
- Voronoi tessellation, 273–274
- Ward's method, 367
- Web crawlers
  - adjacency tables, 417
  - back queues, 415
  - connectivity servers, 419
  - content seen module, 411
  - distributed indexing, 70, 416
  - distributing, 411
  - DNS resolution, 412
  - DNS resolution module, 408
  - duplicate elimination modules, 408
  - fetch modules, 408
  - front queues, 415
  - host splitters, 410
  - Mercator, 407, 419
  - operation/architecture, 406–410
  - overview, 405–406
  - parsing modules, 408
  - Robots Exclusion Protocol, 408
  - seed sets, 406
  - URL frontiers, 406, 407, 415
- Web graphs, 389–390
- Web index. *See also* Search engines
  - adversarial information retrieval, 392
  - advertising/economimc model, 392–394
  - algorithmic search results, 393
  - caching in, 135, 409, 411
  - capture-recapture method, 400
  - click spam, 394
  - distributed indexing, 70, 416
  - engine components, 396
  - index size/estimation, 400
  - informational queries, 395
  - issues in, 2
  - navigational queries, 395
  - near-duplicate results, 403
  - paid inclusion, 391
  - query expansion, 175
  - relevance feedback, 170
  - search engine marketing, 394
  - search engine optimizers, 392
  - shingling, 403
  - spam, 392
  - sponsored, 393, 394
  - transactional queries, 396
  - user queries, 396

482

Index

- Web pages
  - anchor text, 389
  - doorway, 392
  - dynamically generated, 388
  - hyperlinks, 389
  - power law, 82, 389
  - static, 388
- Weighted zone scoring
  - described, 102–104
  - learning algorithm, 106
  - optimal weight, 106
- Weighting
  - inverse document frequency, 109, 190, 209
  - Okapi BM25, 215
  - proximity, 133
  - SMART notation, 118
  - tf-idf (*See* Tf-idf weighting)
- Weight vectors, 295
- Westlaw, 14
- Wikipedia, 411
- Wildcard queries
  - defined, 45, 48
  - general, 48–50
  - k-gram index, 51
  - vector space model interactions, 136–137
- Within-point scatter, 343
- Word segmentation, 24
- World Wide Web. *See also under* Web
  - advertising/economimc model, 394
  - background/history, 385–387
  - bowtie structure, 389, 390
  - characteristics, 387–392
  - HTML, 385
  - http, 385
  - paid inclusion, 391
  - spam, 392
  - URL, 386
  - web graphs, 390, 423
- XML, 19
  - attributes, 180
  - concepts, basic, 180–183
  - contexts, 181
  - data-centric, 179, 196–197
  - documents, decoding, 19
  - DOM, 181
  - DTD, 182
  - elements, 180
  - extended queries, 188
  - fragments, 199
  - nested elements, 186
  - NEXI, 182
  - overview, 179–180
  - schema, 182
  - schema diversity/heterogeneity, 187
  - structured document retrieval
    - principle, 184
  - tag, 180
  - text-centric, 197
- XML retrieval
  - challenges in, 188
  - context resemblance, 190
  - data-centric, 179, 197
  - evaluation of, 196
  - focused, 199
  - language models *vs.*, 230
  - lexicalized subtrees, 189
  - natural language processing, 230
  - SimNoMerge, computing, 190, 191
  - structural terms, 189
  - text-centric, 197
  - topics in, 193
  - vector space model, 192
- XPath, 181
- Zipf's law, 82–83, 92
- Zone indexes, 107
- Zones, 102–103
- Zone search, 180