References

1. S. Adali, T. Liu, and M. Magdon-Ismail. Optimal Link Bombs are Uncoordinated. In *Proc. of 1st Intl. Workshop on Adversarial Information Retrieval on the Web*, 2005.
2. R. Agarwal, C. Aggarwal, and V. Prasad. A Tree Projection Algorithm for Generation of Frequent Itemsets. In *Proc. of the High Performance Data Mining Workshop*, 1999.
3. C. Aggarwal, F. Al-Garawi, and P. Yu. Intelligent Crawling on the World Wide Web with Arbitrary Predicates. In *Proc. of 10th Intl. World Wide Web Conf.* (*WWW'01*), pp. 96–105, 2001.
4. C. Aggarwal, C. Propiuc, J. L. Wolf, P. S. Yu, and J. S. Park. A Framework for Finding Projected Clusters in High Dimensional Spaces. In *Proc. of Intl. Conf. on Management of Data* (*SIGMOD'99*), pp. 407–418, 1999.
5. C. Aggarwal, and P. S. Yu. Finding Generalized Projected Clusters in High Dimensional Spaces. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'00*), pp. 70–81, 2000.
6. E. Agichtein. Confidence Estimation Methods for Partially Supervised Relation Extraction. In *Proc. of SIAM Intl. Conf. on Data Mining* (*SDM06*), 2006
7. R. Agrawal, J. R. Bayardo, and R. Srikant. Athena: Mining-Based Interactive Management of Text Databases. In *Proc. of Extending Database Technology* (*EDBT'00*), pp. 365–379, 2000
8. R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan. Automatic Subspace Clustering for High Dimensional Data for Data Mining Applications. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'98*), pp. 94–105, 1998.
9. R. Agrawal, T. Imielinski, and A. N. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'93*), pp. 207–216, 1993.
10. R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining Newsgroups Using Networks Arising from Social Behavior. In *Proc. of the 12th Intl. World Wide Web Conf.* (*WWW'03*), pp. 529–535, 2003.
11. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. of the 20th Intl. Conf. on Very Large Data Bases* (*VLDB'94*), pp. 487–499, 1994.
12. R. Agrawal and R. Srikant. Mining Sequential Patterns. In *Proc. of the Intl. Conf. on Data Engineering* (*ICDE'95*), pp. 3–14, 1995.
13. R. Agrawal and R. Srikant. On Integrating Catalogs. In *Proc. of the Tenth Intl. World Wide Web Conf.* (*WWW'01*), pp. 603–612, 2001.
14. L. von Ahn, M. Blum, N. Hopper, and J. Langford. CAPTCHA: Using Hard AI Problems for Security. In *Proc. of Eurocrypt*, pp. 294–311, 2003.
15. R. Akavipat, L.-S. Wu, and F. Menczer. Small World Peer Networks in Distributed Web Search. In *Alt. Track Papers and Posters Proc. 13th Intl. World Wide Web Conf.*, pp. 396–397, 2004.
16. B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, L. Ding, P. Kolari, A. Sheth, I. Arpinar, A. Joshi, and T. Finin. Semantic Analytics on Social Networks: Experiences

in Addressing the Problem of Conflict of Interest Detection. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
17. C. Alpert, A. Kahng and S. Yao. Spectral Partitioning: The More Eigenvectors, the Better. *Discrete Applied Mathematics*, 90, pp. 3–5, 1999.
18. B. Amento, L. Terveen, and W. Hill. Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents. In *Proc. of the 23rd ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 296–303, 2000.
19. E. Amitay, D. Carmel, A. Darlow, R. Lempel and A. Soffer. The Connectivity Sonar: Detecting Site Functionality by Structural Patterns. In *Proc. of the 14th ACM Conf. on Hypertext and Hypermedia*, pp. 38–47, 2003.
20. S. S. Anand, and B. Mobasher. Intelligent Techniques for Web Personalization. In *Intelligent Techniques for Web Personalization*, B. Mobasher and S. S. Anand (eds.), *Lecture Notes in AI* (*LNAI 3169*), Springer, 2005.
21. R. Andersen, and K. J. Lang. Communities from Seed Sets. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
22. A. Andreevskaia and S. Bergler. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In *Proc. of 11th Conf. of the European Chapter of the Association for Computational Linguistics* (*EACL'06*), pp. 209–216, 2006.
23. M. Antonie, and O. Zaiane. Text Document Categorization by Term Association. In *Proc. of IEEE Intl. Conf. on Data Mining*, 2002.
24. P. Arabie and L. Hubert. An Overview of Combinatorial Data Analysis. In P. Arabie, L. Hubert and G. D. Soets (eds.). *Clustering and Classification*, pp. 5–63, 1996.
25. A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the Web. *ACM Trans. Internet Technology*, 1 (1), pp. 2–43, 2001.
26. A. Arasu and H. Garcia-Molina. Extracting Structured Data from Web Pages. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'03*), pp. 337–348, 2003.
27. L. Arllota, V. Crescenzi, G. Mecca, and P. Merialdo. Automatic Annotation of Data Extraction from Large Web Sites. In *Intl. Workshop on Web and Databases*, 2003.
28. J. A. Aslam and M. Montague. Models for Metasearch. In *Proc. of the 24th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (*SIGIR'01*), pp. 276–284, 2001.
29. J. Ayres, J. Gehrke, T. Yiu, and J. Flannick. Sequential Pattern Mining Using Bitmaps. In *Proc. of the Eighth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'02*), pp. 429–435, 2002.
30. R. Baeza-Yates, C. Castillo and V. Lopez. PageRank Increase under Different Collusion Topologies. In *Proc. of the 1st Intl. Workshop on Adversarial Information Retrieval on the Web*, 2005.
31. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison–Wesley, 1999.
32. R. Baeza-Yates, F. Saint-Jean, and C. Castillo. Web Dynamics, Age and Page Quality. In *Proc. of String Processing and Information Retrieval*. pp. 117–130, 2002.
33. P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. Wiley, 2003.
34. L. Barabasi and R. Albert. Emergence of Scaling in Random Walk. *Science*, 286, pp. 509–512, 1999.
35. D. Barbará, C. Domeniconi, and N. Kang. Classifying Documents without Labels. In *Proc. of the SIAM Intl. Conf. on Data Mining* (*SDM'04*), 2004.
36. D. Barbará, Y. Li and J. Couto. COOLCAT: an Entropy-Based Algorithm for Categorical Clustering. In *Proc. of the 11th Intl. Conf. on Information and knowledge management* (*CIKM'02*), pp. 582–589, 2002.

37. Z. Bar-Yossef, and M. Gurevich. Random Sampling from a Search Engine's Index. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
38. Z. Bar-Yossef, S. Rajagopalan. Template Detection via Data Mining and its Applications. In *Proc. of the 11th Intl World Wide Web Conf.* (*WWW'02*), pp. 580–591, 2002.
39. S. Basu, A. Banerjee, and R. J. Mooney: Semi-supervised Clustering by Seeding. In *Proc. of the Nineteenth Intl. Conf. on Machine Learning* (*ICML'02*), pp. 27–34, 2002.
40. C. Batini, M. Lenzerini, and S. Navathe. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Survey* 18(4), pp. 323–364, 1986.
41. R. Baumgartner, S. Flesca, and G. Gottlob. Visual Web Information Extraction with Lixto. In *Proc. of the Intl. Conf. on Very Large Data Bases* (*VLDB'01*), pp. 119–128, 2001.
42. R. J. Bayardo. Efficiently Mining Long Patterns from Databases. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'98*), pp. 85–93, 1998.
43. R. J. Bayardo, and R. Agrawal. Mining the Most Interesting Rules. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'99*), pp. 145-154, 1999.
44. P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan. An Exploration of Sentiment Summarization. In: *Proc. of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2003.
45. T. C. Bell, A. Moffat, C. G. Nevill-Manning, I. H. Witten, and J. Zobel. Data Compression in Full-Text Retrieval Systems. *Journal of the American Society for Information Science*, 44(9), pp. 508–531, 1993.
46. B. Berendt, B. Mobasher, M. Nakagawa, and M. Spiliopoulou. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. In *Proc. of the KDD'02 WebKDD Workshop*, 2002.
47. B. Berendt and M. Spiliopoulou. Analyzing Navigation Behavior in Web Sites Integrating Multiple Information Systems. *VLDB Journal*, 9(1), pp. 56–75, 2000.
48. M. Berry, S. T. Dumais, and G. W. O'Brien. Using Linear Algebra for Intelligent Information Retrieval. *SIAM: Review*, 37(4), pp. 573–595, 1995.
49. M. J. A. Berry and G. Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley Computer Publishing, 2004.
50. J. C. Bezdek. Cluster Validity with Fuzzy Sets. *J. of Cybernetics*, 3, pp. 58–72. 1974.
51. K. Bharat, and A. Z. Broder: A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. *Computer Networks*, 30(1–7), pp. 379–388, 1998.
52. K. Bharat and M. Henzinger. Improved Algorithms for Topic Distillation in Hyperlinked Environments. In *Proc. of the 21st ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 104–111, 1998
53. A. Bilke, and F. Naumann. Schema Matching Using Duplicates. In *Proc. of Intl. Conf. on Data Engineering* (*ICDE'05*), pp. 69–80, 2005.
54. A. Blum, and S. Chawla. Learning from Labeled and Unlabeled Data Using Graph Mincuts. In *Proc. of Intl. Conf. on Machine Learning* (*ICML'01*), pp.19–26, 2001.
55. A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proc. of Computational Learning Theory*, pp. 92–100, 1998.
56. J. C. De Borda. Mémoire sur les élections au scrutin. *Mémoires de l'Académie Royale des Sciences année*, 1781.
57. J. Borges and M. Levene. Data Mining of User Navigation Patterns. In *Web Usage Analysis and User Profiling*, LNAI 1836, Springer, pp. 92–111, 1999.
58. C. L. Borgman, (ed.) *Scholarly Communication and Bibliometrics*. Sage Publications, Inc., 1990.

59. B. E. Boser, I. Guyon, and V. N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proc. of the Fifth Annual Workshop on Computational Learning Theory*, 5: pp. 144–152, 1992.
60. P. De Bra and R. Post. Information Retrieval in the World Wide Web: Making Client-Based Searching Feasible, In *Proc. of the 1st Intl. World Wide Web Conf.*, pp. 183–192, 1994.
61. P. S. Bradley, U. Fayyad and C. Reina. Scaling Clustering Algorithms to Large Databases. In *Proc. of Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'98*), pp. 9–15, 1998.
62. L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, New York, 1984.
63. L. Breiman. Bagging Predictors. *Machine Learning*, 24(2), 123–140, 1996.
64. L. Breiman. Prediction Games and Arcing Classifiers. *Technical Report 504*, Statistics Department, University of California at Berkeley, 1997.
65. L. Breiman: Random Forests. *Machine Learning*, 45(1), pp. 5–32, 2001.
66. B. E. Brewington, and G. Cybenko. How Dynamic is the Web? In *Proc. of the 9th Intl. World Wide Web Conf.,* 2000.
67. BrightPlanet.com. *The Deep Web: Surfacing Hidden Value*. Accessible at http://brightplanet.com, July 2000.
68. S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Sngine. *Computer Networks*, 30(1–7), pp. 107–117, 1998.
69. A. Broder, S. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structure in the Web. *Computer Networks*, 33(1–6), pp. 309–320, 2000.
70. C. A. Brunk and M. J. Pazzani. An Investigation of Noise-Tolerant Relational Concept Learning Algorithms. In *Proc. of the 8th Intl. Workshop on Machine Learning*, pp. 389–393, 1991.
71. A. Buchner and M. D. Mulvenna. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'99*), pp. 54–61, 1999.
72. G. Buehrer, S. Parthasarathy, and A. Ghoting. Out-of-core frequent pattern mining on a commodity PC. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'06*), pp. 86 – 95, 2006.
73. D. Burdick, M. Calimlim, and J. Gehrke. MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases. In *Proc. of the Intl. Conf. on Data Engineering* (*ICDE'01*), pp. 443, 2001.
74. C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), pp. 955–974, 1998.
75. D. Buttler, L. Liu, and C. Pu. A Fully Automated Object Extraction System for the World Wide Web. In *Proc. of Intl. Conf. on Distributed Computing Systems* (*ICDCS'01*), pp. 361–370, 2001.
76. I. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Model-Based Clustering and Visualization of Navigation Patterns on a Web Site. *Data Mining Knowledge Discovery* 7(4), pp. 399–424, 2003.
77. D. Cai, S. Yu, J.-R. Wen and W.-Y Ma. Extracting Content Structure for Web Pages based on Visual Representation. In *Proc. of the APWeb'03 Conf.*, Number 2642 in Lecture notes in Computer Science (LNCS), pp. 406–417, 2003
78. D. Cai, S. Yu, J.-R. Wen and W.-Y. Ma. Block-Based Web Search. In *Proc. of the ACM SIGIR Research and Development in Information Retrieval* (*SIGIR'04*), pp. 456–463, 2004.

79. Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon. Adapting Ranking SVM to Document Retrieval. In *Proc. of the 29th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (*SIGIR'06*), pp. 186-193, 2006.
80. G. Carenini, R. Ng, and E. Zwart. Extracting Knowledge from Evaluative Text. In *Proc. of the Third Intl. Conf. on Knowledge Capture* (*K-CAP'05*), pp. 11–18. 2005.
81. G. Carenini, R. Ng, and A. Pauls. Interactive Multimedia Summaries of Evaluative Text. In *Proc. of the 10th Intl. Conf. on Intelligent User Interfaces* (*IUI'06*), pp. 305–312, 2006.
82. H. Carrillo and D. Lipman. The Multiple Sequence Alignment Problem in Biology. *SIAM Journal Applied Mathematics*, 48(5), pp. 1073–1082, 1988.
83. V. Castelli and T. M. Cover. Classification Rules in the Unknown Mixture Parameter Case: Relative Value of Labeled and Unlabeled Samples. In *Proc. of 1994 IEEE Intern. Symp. Inform. Theory*, 111, 1994.
84. S. Chakrabarti. Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation and Information Extraction. In *Proc. of the 13th Intl. World Wide Web Conf.* (*WWW'01*), pp. 211–220, 2001.
85. S. Chakrabarti. *Mining the Web. Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, 2003.
86. S. Chakrabarti, B. Dom, S. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's Link Structure. *IEEE Computer*, 32(8), pp. 60–67, 1999.
87. S. Chakrabarti, M. van den Berg, and B. Dom. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. *Computer Networks*, 31(11–16), pp. 1623–1640, 1999.
88. S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. *Computer Networks* 30(1–7), pp. 65–74, 1998.
89. S. Chakrabarti, K. Puniyani, and S. Das. Optimizing Scoring Functions and Indexes for Proximity Search in Type-annotated Corpora. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
90. C-H. Chang, M. Kayed, M. R. Girgis, and K. Shaalan. A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10), pp. 1411–1428, 2006.
91. C-H. Chang and S. Lui. IEPAD. Information Extraction Based on Pattern Discovery. In *Proc. of the Tenth Intl. World Wide Web Conf.* (*WWW'01*), pp. 681–688, 2001.
92. K. C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang. Structured Databases on the Web: Observations and Implications. *SIGMOD Record*, 33(3), pp. 61-70, 2004.
93. O. Chapelle, B. Schölkopf and A. Zien. (eds.) *Semi-Supervised Learning*. MIT Press, 2006.
94. P. Cheeseman, and J. Stutz. Bayesian Classification (AutoClass): Theory and Results. In *Advances in Knowledge Discovery and Data Mining*, 1996.
95. W. Chen. New Algorithm for Ordered Tree-to-Tree Correction Problem. *J. Algorithms*, 40(2), pp. 135–158, 2001.
96. H. Chen, Y.-M. Chung, M. Ramsey, and C. Yang. A Smart Itsy Bitsy Spider for the Web. *Journal of the American Society for Information Science* 49 (7), 604–618, 1998.
97. S. F. Chen, and J. Goodman. *An Empirical Study of Smoothing Techniques for Language Modeling.* Tech. Rep. TR-10-98, Harvard University, 1998.
98. Y.-Y. Chen, Q. Gan and T. Suel. Local Methods for Estimating PageRank Values. In *Proc. of the Intl. Conf. on Information and Knowledge Management* (*CIKM'04*), pp. 381–389, 2004.

99. C. H. Cheng, A. W. Fu and Y Zhang. Entropy-Based Subspace Clustering for Mining Numerical Data. In *Proc. of Knowledge Discovery and Data Mining* (*KDD'99*), pp. 84–93, 1999.
100. Y. Cheng and G. Church. Biclustering of Expression Data. In *Proc. ISMB*, pp. 93–103. AAAI Press, 2000.
101. J. Cho and H. Garcia-Molina. The Evolution of the Web and Implications for an Incremental Crawler. In *Proc. of the 26th Intl. Conf. on Very Large Data Bases* (*VLDB'00*), 2000.
102. J. Cho, H. Garcia-Molina, and L. Page. Efficient Crawling through URL Ordering. *Computer Networks* 30 (1–7), pp. 161–172, 1998.
103. J. Cho, and S. Roy. Impact of Web Search Engines on Page Popularity. In *Proc. of the 13th Intl. World Wide Web Conf.* (*WWW'04*), pp. 20–29, 2004.
104. P. Clark, and T. Niblett. The CN2 Induction Algorithm. *Machine Learning*, 3, pp. 261–283, 1989.
105. C. Clifton, E. Housman, and A. Rosenthal. Experience with a Combined Approach to Attribute-Matching Across Heterogenenous Databases. In: *Proc, IFIP 2.6 Working Conf. Database Semantics*, 1997.
106. W. W. Cohen. Fast Effective Rule Induction. In *Proc. of 12th Intl. Conf. on Machine Learning* (*ICML'95*), pp. 115–123, 1995.
107. W. W. Cohen. Integration of Heterogeneous Databases without Common Domains Using Queries Based on Textual Similarity. In *Proc ACM SIGMOD Conf. on Management of Data* (*SIGMOD'08*), pp. 201–212, 1998.
108. W. W. Cohen, M. Hurst, and L. S. Jensen. A Flexible Learning System for Wrapping Tables and Lists in Html Documents. In *Proc. of the 11th Intl. World Wide Web Conf.* (*WWW'02*), pp. 232–241, 2002.
109. M. Collins and Y. Singer. Unsupervised Models for Named Entity Classification. In *Proc. of Intl. Conf. on Empirical Methods in Natural Language Processing* (*EMNLP'99*), pp. 100–110, 1999.
110. M. de Condorcet. *Essai sur l'application de l'analyse a la probabilitie des decisions rendues a la pluralite des voix*, Paris, 1785.
111. G. Cong, W. S. Lee, H. Wu, and B. Liu. Semi-Supervised Text Classification Using Partitioned EM. In *Proc. of Database Systems for Advanced Applications* (*DASFAA 2004*): 482–493, 2004.
112. G. Cong, K.-L. Tan, A. K. H. Tung, and X. Xu. Mining Top-k Covering Rule Groups for Gene Expression Data. In *Proc. of ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'05*), pp. 670–681, 2005.
113. G. Cong, A. K. H. Tung, X. Xu, F. Pan, and J. Yang. Farmer: Finding Interesting Rule Groups in Microarray Datasets. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'04*), pp. 143–154. 2004.
114. R. Cooley, B. Mobasher, and J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web. In *Proc. of the 9th IEEE Intl. Conf. on Tools With Artificial Intelligence* (*ICTAI'97*), pp. 558–567, 1997.
115. R. Cooley, B. Mobasher, and J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems*, 1(1), pp. 5–32, 1999.
116. T. Corman, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2001.
117. V. Crescenzi, G. Mecca, and P. Merialdo. Roadrunner: Towards Automatic Data Extraction from Large Web Sites. In *Proc. of Very Large Data Bases* (*VLDB'01*), pp. 109–118, 2001.
118. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

119. W. B. Croft. Combining Approaches to Information Retrieval. In W. B. Croft (eds.), *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, Kluwer Academic Publishers, 2000.
120. S. Das and M. Chen. Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards. *APFA'01*, 2001.
121. S. Dasgupta, M.L. Littman, and D. McAllester. PAC Generalization Bounds for Co-Training. *Advances in Neural Information Processing Systems* (*NIPS*), 2001.
122. K. Dave, S. Lawrence, and D. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proc. of the 12th Intl. World Wide Web Conference* (*WWW'03*), pp. 519–528, 2003.
123. B. Davison. Topical Locality in the Web. In *Proc. 23rd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 272–279, 2000.
124. S. Debnath, P. Mitra, and C. L. Giles. Automatic Extraction of Informative Blocks from Webpages. In *Proc. of the 2005 ACM Symposium on Applied Computing*, pp. 1722–1726, 2005.
125. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, pp. 391–407, 1990.
126. M. Degeratu, G. Pant, and F. Menczer. Latency-Dependent Fitness in Evolutionary Multithreaded Web Agents. In *Proc. of GECCO Workshop on Evolutionary Computation and Multi-Agent Systems*, pp. 313–316, 2001.
127. A. P. Dempster, N. M. Laird, and D. B. Rubin Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1), pp. 1–38, 1977.
128. L. Deng, X. Chai, Q. Tan, W. Ng, and D. L. Lee. Spying Out Real User Preferences for Metasearch Engine Personalization. In *Proc. of the Workshop on WebKDD*, 2004.
129. F. Denis. PAC Learning from Positive Statistical Queries. In *Proc. of Intl. Conf. on Algorithmic Learning Theory* (*ALT'98*), pp. 112–126, 1998.
130. F. Denis, R. Gilleron and M. Tommasi. Text Classification from Positive and Unlabeled Examples. *IPMU*, 2002.
131. M. Deshpande and G. Karypis. Using conjunction of attribute values for classification. In *Proc. of the ACM Intl. Conf. on Information and Knowledge Management* (*CIKM'02*), pp. 356-364, 2002.
132. M. Deshpande and G. Karypis. Selective Markov Models for Predicting Web Page Accesses. *ACM Trans. on Internet Technology*, 4(2), 163–184, 2004.
133. R. Dhamankar, Y. Lee, A. Doan, A. Halevy, and P. Domingos. IMap: Discovering Complex Semantic Matches between Database Schemas. In *Proc. of ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'04*), pp. 383–394, 2004.
134. S. Dhillon. Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning. In *Proc. of the 7th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'01*), pp. 269–274, 2001.
135. S. Dhillon, S. Mallela, and D. S. Modha. Information-Theoretic Co-Clustering. In *Proc. of The Ninth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'03*), pp. 89–98, 2003.
136. L. Dice. Measures of the Amount of Ecologic Association between Species. *Ecology*, 26(3), 1945.
137. J. Diesner and K. M. Carley. Exploration of Communication Networks from the Enron Email Corpus. In *Workshop on Link Analysis, Counterterrorism and Security at SDM'05*, 2005.
138. T. G. Dietterich and G. Bakiri. Solving Multiclass Learning Problems via Error-Correcting Output Codes. *J. of Artificial Intelligence Research*, 2. pp. 263–286, 1995.

139. M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori. Focused Crawling Using Context Graphs. In *Proc. of Intl. Conf. on Very Large Databases* (*VLDB'00*), pp. 527–534, 2000.
140. M. Diligenti, M. Gori, and M. Maggini, Web Page Scoring Systems for Horizontal and Vertical Search. In *Proc. of the 11th Intl. World Wide Web Conference* (*WWW'02*), pp. 508–516. 2002.
141. C. Ding, and X. He. Linearized Cluster Assignment via Spectral Ordering. In *Proc. of Int'l Conf. Machine Learning* (*ICML'04*), 2004.
142. C. Ding, X. He, H. Zha, and H. Simon. PageRank, HITS and a Unified Framework for Link Analysis. In *Proc. of SIAM Data Mining Conf*., 2003.
143. C. Djeraba, O. R. Zaiane, and S. Simoff. (eds.). *Mining Multimedia and Complex Data*. Springer, 2003.
144. H. Do and E. Rahm. Coma: A System for Flexible Combination of Schema Matching Approaches. In *Proc. of the Intl. Conf. on Very Large Data Bases* (*VLDB'02*), pp. 610–621, 2002.
145. A. Doan, P. Domingos, and A. Y. Halevy. Reconciling Schemas of Disparate Data Sources: a Machine-Learning Approach. In *Proc. of the 2001 ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'01*), pp. 509–520, 2001.
146. A. Doan and A. Halevy, Semantic Integration Research in the Database Community: A Brief Survey. *AI Magzine*, 26(1), pp. 83–94, 2005.
147. A. Doan, J. Madhavan, P. Domingos, and A. Y. Halevy: Learning to Map between Ontologies on the Semantic Web. In P*roc. of the 11th Intl. World Wide Web Conference* (*WWW'02*), pp. 662–673, 2002.
148. P. Domingos, and M. J. Pazzani. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* 29(2–3), pp. 103–130, 1997.
149. G. Dong, X. Zhang, L. Wong, and J. Li. CAEP: Classification by Aggregating Emerging Patterns. In *Proc. of Intl. Conf. on Discovery Science*, pp. 30–42, 1999.
150. C. Doran, D. Egedi, B. A. Hockey, B. Srinivas, and M. Zaidel. XTAG System-A Wide Coverage Grammar for English. In *Proc. of Intl. Conf. on Computational Linguistics* (*COLING'94*), pp. 922–928, 1994.
151. J. Dougherty, R. Kohavi, and M. Sahami. Supervised and Unsupervised Discretization of Continuous Features. In *Proc. of the 12th Intl. Conf. on Machine Learning* (*ICML'95*), 1995.
152. A. Douglis. B. Feldmann, Krishnamurthy, and J. C. Mogul. Rate of Change and Other Metrics: a Live Study of the World Wide Web. In *Proc. of USENIX Symp. on Internet Technologies and Systems*, pp. 147–158, 1997.
153. E. Dragut, W. Wu, P. Sistla, C. Yu, and W. Meng. Merging Source Query Interfaces on Web Databases. In *Proc. of the International Conference on Data Engineering* (*ICDE'06*), 2006.
154. E. Dragut, C. Yu, and W. Meng. Meaningful Labeling of Integrated Query Interfaces. In *Proceedings of Very Large Data Bases* (*VLDB'06*), 2006.
155. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons Inc., 2nd edition, 2001
156. M. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall, 2002.
157. J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3, pp. 32–57, 1974.
158. J. Eckmann, and E. Moses. Curvature of Co-Links Uncovers Hidden Thematic Layers in the World Wide Web. In *Proc. of the National Academy of Sciences*, pp. 5825–5829, 2002.
159. K. Eguchi, and V. Lavrenko. Sentiment Retrieval Using Generative Models. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing* (*EMNLP'06*),

pp. 345–354, 2006.
160. D. Eichmann. Ethical Web Agents. *Computer Networks*, 28(1–2), pp. 127–136, 1995.
161. P. Elias. Universal Codeword Sets and Representations of the Integers. *IEEE Transactions on Information Theory*, IT–21(2), pp.194–203, 1975.
162. D. W. Embley, D. Jackman, and L. Xu. Multifaceted Exploitation of Metadata for Attribute Match Discovery in Information Integration. In: *Proc Intl. Workshop on Information Integration on the Web*, pp. 110–117, 2001.
163. D. W. Embley, Y. Jiang, and Y. K. Ng. Record-Boundary Discovery in Web Documents. In *Proc. of ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'99*), pp. 467–478, 1999.
164. M. Ester, H.-P. Kriegal, J. Sander and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. of Knowledge Discovery and Data Mining* (*KDD'96*), pp. 226–231, 1996.
165. A. Esuli, and F. Sebastiani. Determining Term Subjectivity and Term Orientation for Opinion Mining. In: *Proc. of Conf. of the European Chapter of the Association for Computational Linguistics* (*EACL'06*), 2006.
166. O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Web-Scale Information Extraction in Knowitall. In *Proc. of the 13th Intl. World Wide Web Conference* (*WWW'04*), pp. 100–110, 2004.
167. B. S. Everitt. *Cluster Analysis*. Heinemann, London, 1974.
168. R. Fagin, R. Kumar, K. S. McCurley, J. Novak, D. Sivakumar, J. A. Tomlin, and D. P. Williamson. Searching the Workplace Web. In *Proc. of the 12th Intl. World Wide Web Conference* (*WWW'03*), pp. 366–375, 2003.
169. W. Fan. On the Optimality of Probability Estimation by Random Decision Trees. In *Proc. of National Conf. on Artificial Intelligence* (*AAAI'04*), pp. 336-341, 2004.
170. W. Fan, S. J. Stolfo, J. Zhang, P. K. Chan: AdaCost: Misclassification Cost-Sensitive Boosting. In *Proc. of the 16th Intl. Conf. on Machine Learning* (*ICML'99*), pp. 97-105, 1999.
171. A. Farahat, T. LoFaro, J. C. Miller, G Rae and L Ward. Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization. *SIAM Journal on Scientific Computing*, pp. 1181–1201, 2005.
172. U. M. Fayyad, and K. B. Irani. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *Proc. of the Intl. Joint Conf. on Artificial Intelligence*, pp. 102–1027, 1993.
173. U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, pp. 1–34, 1996.
174. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurasamy (eds.). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
175. C. Fellbaum. *WordNet*: *An On-Line Lexical Database*. MIT Press, 1998.
176. D. Fetterly, M. Manasse and M. Najork. Detecting Phrase-Level Duplication on the World Wide Web. In *Proc. of the 28th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 170–177, 2005.
177. D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A Large-Scale Study of the Evolution of Web Pages. In *Proc. of the 12th Intl. World Wide Web Conf.* (*WWW'03*), pp. 669–678, 2003.
178. D. Fisher. Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning,* 2, pp. 139–172, 1987.
179. G. W. Flake, S. Lawrence, and C. L. Giles, Efficient Identification of Web Communities. In *Proc. of the sixth ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pp.150–160, 2000.

180. G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-Organization of the Web and Identification of Communities. *IEEE Computer* 35(3), pp. 66–71, 2002.
181. L. R. Ford Jr. and D. R. Fulkerson. Maximal Flow through a Network. *Canadian Journal Mathematics*, 8: pp. 399–404, 1956.
182. S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Topical Interests and the Mitigation of Search Engine Bias. In *Proc. Natl. Acad. Sci*. USA, 103(34), pp. 12684-12689, 2006.
183. S. Fortunato, A. Flammini, and F. Menczer. Scale-Free Network Growth by Ranking. *Phys. Rev. Lett*. 96(21), 2006.
184. E. Fox and J. Shaw. Combination of Multiple Searches. In *Proc. of the Second Text REtrieval Conf.*, pp. 243-252, 1993.
185. D. Freitag and A. McCallum. Information Extraction with HMM Structures Learned by Stochastic Optimization. In *Proc. of National Conf. on Artificial Intelligence* (*AAAI'00*), 2000.
186. Y. Freund, and R. E. Schapire. Experiments with a New Boosting Algorithm. In *Proc. of the 13th Intl. Conf. on Machine Learning* (*ICML'96*), pp. 148–156, 1996.
187. X. Fu, J. Budzik, and K. J. Hammond. Mining Navigation History for Recommendation. In *Proc. of the Intl. Conf. on Intelligent User Interfaces*, pp. 106–112, 2000.
188. G. P. C. Fung, J. X. Yu, H. Lu, and P. S. Yu. Text Classification without Labeled Negative Documents. In *Proc. 21st Intl. Conf. on Data Engineering* (*ICDE'05*), pp. 594–605, 2005.
189. J. Furnkranz and G. Widmer. Incremental Reduced Error Pruning. In *Proc. of the Eleventh Intl. Conf. Machine Learning*, pp. 70–77, 1994.
190. A. Gal, G. Modica, H. Jamil, and A. Eyal. Automatic Ontology Matching Using Application Semantics. *AI Magazine*, 26(1), pp. 21–32, Spring 2005.
191. M. Gamon. Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. In *Proc. of the 20th Intl. Conf. on Computational Linguistics*, pp. 841–847, 2004.
192. M. Gamon, A. Aue, S. Corston-Oliver, and E. K. Ringger. Pulse: Mining Customer Opinions from Free Text. In *Proc. of Intelligent Data Analysis* (*IDA' 05*), pp. 121–132, 2005.
193. V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS – Clustering Categorical Data Using Summaries. In *Proc. of Knowledge Discovery and Data Mining* (*KDD'99*), pp. 73–83, 1999.
194. F. Gasparetti and A. Micarelli. Swarm Intelligence: Agents for Adaptive Web Search. In *Proc. of the 16th European Conf. on Artificial Intelligence* (*ECAI'04*), 2004.
195. J. Gehrke, R. Ramakrishnan, and V. Ganti. RainForest - A Framework for Fast Decision Tree Construction of Large Datasets. In *Proc. of Intl. Conf. on Very Large Data Bases* (*VLDB'98*), pp. 416-427, 1998.
196. R. Ghani, Combining Labeled and Unlabeled Data for MultiClass Text Categorization. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'02*), pp. 187–194, 2002.
197. D. Gibson, J. M. Kleinberg, and P. Raghavan, Clustering Categorical Data: An Approach Based on Dynamical Systems. In *Proc. of the Intl. Conf. on Very Large Data Bases* (*VLDB'98*), pp.311–322, 1998.
198. D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web Communities from Link Topology. In *Proc. of the 9th ACM Conf. on Hypertext and Hypermedia*, 1998.
199. D. Gibson, K. Punera, and A. Tomkins. The Volume and Evolution of Web Page Templates. In *Special Interest Tracks and Posters of the 14th Intl. Conf. on World Wide Web* (*WWW '05*). pp. 830–839, 2005.
200. M. Girvan and M. Newman. Community Structure in Social and Biological Network. In *Proc. of the National Academy of Sciences*, 2001.

201. S. Goldman and Y. Zhou. Enhanced Supervised Learning with Unlabeled Data. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'00*), pp. 327–334, 2000.
202. S. W. Golomb. Run-Length Encodings. *IEEE Transactions on Information Theory*, 12(3), pp. 399–401, July 1966.
203. G. H. Golub, and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1983.
204. I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, 1965.
205. J. C. Gower. A General Coefficient of Similarity and Some of its Properties. *Biometrics*, 27, pp. 857–871, 1971.
206. G. Grefenstette, Y. Qu, D. A. Evans, and J. G. Shanahan. Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. In *Proc. of AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.
207. G. Grimmett and D. Stirzaker. *Probability and Random Process*. Oxford University Press, 1989.
208. R. L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, (eds.). *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers, 2001.
209. D. A. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics*, Springer, 2004.
210. S. Grumbach and G. Mecca. In Search of the Lost Schema. In *Proc. of the Intl. Conf. on Database Theory*, pp. 314–331, 1999.
211. S. Guha, R. Rastogi, and K. Shim. CURE: An Efficient Clustering Algorithm for Large Databases. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'98*), pp. 73–84, 1998.
212. S. Guha, R. Rastogi, and K. Shim. ROCK: a Robust Clustering Algorithm for Categorical Attributes. In *Proc. of the 15th Intl. Conf. on Data Engineering*, pp. 345–366. 2000.
213. D. Gusfield. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, 1997.
214. Z. Gyongyi and H. Garcia-Molina. *Web Spam Taxonomy*. Technical Report, Stanford University, 2004.
215. Z. Gyöngyi, and H. Garcia-Molina. Link Spam Alliances. In *Proc. of the 31st Intl Conf. on Very Large Data Bases* (*VLDB'05*), pp. 517–528, 2005.
216. Z. Gyongyi, H. Garcia-Molina and J. Pedersen. Combating Web Spam with TrustRank. In *Proc. of 30th Intl. Conf. on Very Large Data Bases* (*VLDB'04*), pp. 576–587, 2004.
217. J. Han and Y. Fu. Discovery of Multi-Level Association Rules from Large Databases. In *Proc. of the 21st Intl. Conf. on Very Large Data Bases* (*VLDB'05*), pp. 420–431, 1995.
218. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, 2001.
219. J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu. Freespan: Frequent Pattern-Projected Sequential Pattern Mining. In *Proc. of the 2000 Int. Conf. Knowledge Discovery and Data Mining* (*KDD'00*), pp. 355–359, 2000.
220. J. Han, J. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data* (*SIGMOD'00*), pp. 1–12, 2000.
221. D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.
222. J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons Inc. 1975.

223. J. A. Hartigan. Direct Clustering of a Data Matrix. *Journal of the American Statistical Association*, 67(337): pp. 123–129, 1972.
224. V. Hatzivassiloglou and K. McKeown. Predicting the Semantic Orientation of Adjectives. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics* (*ACL-EACL'97*), pp. 174−181, 1997.
225. V. Hatzivassiloglou, and J. Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In *Proc. of the Intl. Conf. on Computational Linguistics* (*COLING'00*), pp. 299–305. 2000.
226. T. Haveliwala. Extrapolation Methods for Accelerating PageRank Computations. In *Proc. of the 12th Intl. World Wide Web Conf.* (*WWW'03*), pp. 261–270, 2003.
227. B. He, and K. C.-C. Chang. Statistical Schema Matching across Web Query Interfaces. In *Proc. of the 2003 ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'03*), pp. 217–228, 2003.
228. B. He and K. C.-C. Chang. Making Holistic Schema Matching Robust: An Ensemble Approach. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'05*), pp. 429-438, 2005.
229. B. He, K. C.-C. Chang, and J. Han. Discovering Complex Matchings across Web Query Interfaces: A Correlation Mining Approach. In *Proc. of the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining* (*KDD'04*), pp. 148–157, 2004.
230. H. He, W. Meng, C. T. Yu, and Z. Wu. WISE-Integrator: An Automatic Integrator of Web Search Interfaces for E-commerce. In *Proc. of Very Large Data Bases* (*VLDB'03*), 2003.
231. H. He, W. Meng, C. T. Yu, and Z. Wu. Automatic extraction of web search interfaces for interface schema integration. In *Proc. of WWW Alternate Track Papers and Posters,* pp. 414-415, 2004.
232. M. A. Hearst. Direction-based Text Interpretation as an Information Access Refinement. In P. Jacobs (eds.), *Text-Based Intelligent Systems*. Lawrence Erlbaum Associates, 1992.
233. M. A. Hearst, and J. O. Pedersen. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In *Proc. of the 19th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (*SIGIR'96*), 1996.
234. M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. Measuring Search Engine Quality Using Random Walks on the Web. In *Proc. of the 8th Intl. World Wide Web Conf.*, pp. 213–225, 1999.
235. M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On Near-Uniform URL Sampling. In *Proc. of the 9th Intl. World Wide Web Conf.* (*WWW'00*), pp. 295–308, 2000.
236. J. L. Herlocker, J. Konstan, L. Terveen, and J. Riedl. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems*, 22(1), pp. 5–53, 2004
237. M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur. The Shark-Search Algorithm˙ An Application: Tailored Web Site Mapping. In *Proc. of the 7th Intl. World Wide Web Conf.* (*WWW7*), pp. 317–326, 1998.
238. A. Heydon and M. Najork. Mercator: A Scalable, Extensible Web Crawler. *World Wide Web* 2(4), pp. 219–229, 1999.
239. A. Hinneburg and D. A. Keim. An Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. In *Proc. of Very Large Data Bases* (*VLDB'99*), pp. 506–517, 1999.
240. T. Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1): pp. 177–196, 2001.
241. A. Hogue and D. Karger. Thresher: Automating the Unwrapping of Semantic Con-

tent from the World Wide Web. In *Proc. of the 14th Intl. World Wide Web Conference* (*WWW'05*), pp. 86–95, 2005.
242. S. C. H. Hoi, R. Jin, M. R. Lyu. Large-Scale Text Categorization by Batch Mode Active Learning. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
243. V. Honavar and G. Slutzki. (eds.). Grammatical Inference. In *Proc. of the Fourth Intl Colloquium on Grammatical Inference*. LNCS 1433. Springer-Verlag, 1998.
244. C.-N. Hsu and M.-T. Dung. Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web. *Inf. System*, 23(9), pp. 521–538, 1998.
245. M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *Proc. of ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'04*), pp. 168–177, 2004.
246. M. Hu and B. Liu. Mining Opinion Features in Customer Reviews. In *Proc. of the 19th National Conf. on Artificial Intelligence* (*AAAI'04*), pp. 755–760, 2004.
247. M. Hu and B. Liu. Opinion Feature Extraction Using Class Sequential Rules. In *Proc. of the Spring Symposia on Computational Approaches to Analyzing Weblogs*, 2006.
248. L. Hyafil, and R. L. Rivest. Constructing Optimal Binary Decision Trees is NP-Complete. *Information Processing Letters* 5, pp. 15–17, 1976.
249. H. Ino, M. Kudo, and A. Nakamura. Partitioning of Web Graphs by Community Topology. In *Proc. of the 14th Intl. Conf. on World Wide Web* (*WWW'05*), pp. 661–66, 2005.
250. U. Irmak, and T. Suel. Interactive Wrapper Generation with Minimal User Effort. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
251. T. Jagatic, N. Johnson, M. Jakobsson, and F. Menczer. Social Phishing. *Communications of the ACM*. In press, 2006.
252. A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
253. T. Jiang, L. Wang, and K. Zhang. Alignment of Trees - an Alternative to Tree edit. In *Proc. of Combinatorial Pattern Matching*, pp. 75–86, 1994.
254. X. Jin, Y. Zhou, and B. Mobasher. Web Usage Mining Based on Probabilistic Latent Semantic Analysis. In *Proc. of the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining* (*KDD'04*), pp. 197–205, 2004.
255. N. Jindal, and B. Liu. Identifying Comparative Sentences in Text Documents. In *Proc. of ACM SIGIR Intl. Conf. on Research and Development in Information Retrieval* (*SIGIR'06*), pp. 244–251, 2006.
256. N. Jindal, and B. Liu. Mining Comparative Sentences and Relations. In *Proc. of National Conference on Artificial Intelligence* (*AAAI'06*), 2006.
257. T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Machine Learning: In *Proc. of Tenth European Conf. on Machine Learning* (*ECML'98*), pp. 137–142, 1998.
258. T. Joachims. Making Large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (eds.), MIT Press, 1999.
259. T. Joachims. Transductive Inference for Text Classification Using Support Vector Machines. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'99*), pp. 200–209, 1999.
260. T. Joachims. Optimizing Search Engines Using Click-through Data. In *Proc. of the ACM Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'02*), pp. 133-142, 2002.
261. T. Joachims, Transductive Learning via Spectral Graph Partitioning. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'03*), pp. 290–297, 2003.
262. R. Jones, B. Rey, O. Madani, and W. Greiner. Generating Query Substitutions. In

Proc. of the 15th Intl. Conf. on World Wide Web (*WWW'06*), 2006.
263. L. P. Kaelbling, M. Littman, and A. Moore. Reinforcement Learning: A survey. *Journal of Artificial Intelligence Research* 4, pp. 237–285, 1996.
264. N. Kaji and M. Kitsuregawa. Automatic Construction of Polarity-Tagged Corpus from HTML Documents. In *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 452–459, 2006.
265. Y. Kalfoglou and M. Schorlemmer. Ontology Mapping: the State of the Art. *The Knowledge Engineering Review Journal,* 18(1), pp. 1–31, 2003.
266. S. D. Kamar, T. Haveliwala, C. D. Manning, and G. H. Golub, Extrapolation Methods for Accelerating PageRank Computations. In *Proc. of the 12th Intl. World Wide Web Conference* (*WWW'03*), pp. 261–270, 2003.
267. H. Kanayama and T. Nasukawa. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. In *Proc. of the 2006 Conf. on Empirical Methods in Natural Language Processing* (*EMNLP'06*), pp. 355–363, 2006.
268. J. Kang, and J. F. Naughton, On Schema Matching with Opaque Column Names and Data Values. In *Proc. of the 2003 ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'03*), 2003.
269. V. Kashyap and A. Sheth. Semantic and Schematic Similarities between Database Objects: a Context-Based Approach. In *Proc. of the Intl. Journal on Very Large Data Bases* (*VLDB'96*), 5(4): pp. 276–304, 1996.
270. G. V. Kass. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29, pp. 119–127, 1980.
271. L. Kaufman, and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, 1990.
272. M. Kearns. Efficient Noise-Tolerant Learning from Statistical Queries. *Journal of the ACM*, 45, pp. 983–1006, 1998.
273. J. S. Kelly. *Social Choice Theory: An Introduction*. Springer-Verlag, 1988.
274. C. Kennedy. Comparatives, Semantics of. In *Encyclopedia of Language and Linguistics*, Second Edition, Elsevier, 2005.
275. M. M. Kessler. Bibliographic Coupling between Scientific Papers. *American Documentation*, 14, 1963.
276. S. Kim and E. Hovy. Determining the Sentiment of Opinions. In *Proc. of the Intl. Conf. on Computational Linguistics* (*COLING'04*), 2004.
277. S.-M. Kim and E. Hovy. Automatic Identification of Pro and Con Reasons in Online Reviews. In *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions,* pp. 483–490, 2006.
278. S.-M. Kim and E. Hovy. Identifying and Analyzing Judgment Opinions. In *Proc. of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 200–207, 2006.
279. R. Kimball and R. Merz. *The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse.* John Wiley & Sons, 2000.
280. J. L. Klavans, and S. Muresan. DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and Their Associated Definitions from On-line Text. In *Proc. of American Medical Informatics Assoc*., 2000.
281. J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proc. of the 9th ACM SIAM Symposium on Discrete Algorithms* (*SODA'98*), pp. 668–677, 1998.
282. J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM* 46 (5), pp. 604–632, 1999.
283. M. Klemetinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding Interesting Rules from Large Sets of Discovered Association Rules. In *Proc. of the ACM Intl. Conf. on Information and Knowledge Management* (*CIKM'94*), pp. 401-

407, 1994.
284. N. Kobayashi, R. Iida, K. Inui and Y. Matsumoto. Opinion Mining on the Web by Extracting Subject-Attribute-Value Relations. In *Proc. of AAAI-CAAW'06*, 2006.
285. R. Kohavi, B. Becker, and D. Sommerfield, Improving Simple Bayes. In *Proc. of European Conference on Machine Learning* (*ECML'97*), 1997.
286. R. Kohavi, L. Mason, R. Parekh, and Z. Zheng. Lessons and Challenges from Mining Retail E-Commerce Data. *Machine Learning*, 57(1–2), pp. 83–113, 2004.
287. T. Kohonen. Self-Organizing Maps. *Series in Information Sciences*, 30, Springer, Heidelberg, Second Edition. 1995.
288. F. Korn, H. V. Jagadish and C. Faloutsos. Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences. In *Proc. ACM SIGMOD International Conference on Management of Data* (*SIGMOD'97*), pp 289–300, 1997.
289. R. Kraft, C. C. Chang, F. Maghoul, and Ravi Kumar. Searching with Context. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
290. L.-W. Ku, H.-W. Ho, and H.-H. Chen. Novel Relationship Discovery Using Opinions Mined from the Web. In *Proc. of the Twenty-First National Conf. on Artificial Intelligence* (*AAAI'06*), 2006.
291. L.-W. Ku, Y.-T. Liang and H.-H. Chen. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *Proc. of the AAAI-CAAW'06*, 2006.
292. V. Kumar, A. Grama, A. Gupta, and G. Karypis. *Introduction to Parallel Computing: Design and Analysis of Algorithms*. Benjamin/Cummings, 1994.
293. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for Emerging Cyber-Communities. In *Proc. of the 8th Intl. World Wide Web Conference* (*WWW8*), pp. 1481–1493, 1999.
294. K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram. 2004. A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. In *Proc. of the 13th Intl. Conf. on World Wide Web* (*WWW'04*), pp. 658–665, 2004.
295. N. Kushmerick. Wrapper Induction: Efficiency and Expressiveness. *Artificial Intelligence*, 118: pp. 15–68, 2000.
296. N. Kushmerick. *Wrapper Induction for Information Extraction*. Ph.D Thesis. Dept. of Computer Science, University of Washington, TR UW-CSE-97-11-04, 1997.
297. S. H. Kwok and C. C. Yang. Searching the Peer–to–Peer Networks: The Community and their Queries. *Journal of the American Society for Information Science and Technology*, *Special Topic Issue on Research on Information Seeking*, 55(9), pp.783–793, 2004.
298. J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling or Sequence Data. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'01*), pp. 282–289, 2001.
299. S. Lam, and J. Reidl. Shilling Recommender Systems for Fun and Profit. In *Proc. of the 13th Int.l World Wide Web Conf.* (*WWW'04*), pp. 393-402, 2004.
300. K. Lang. Newsweeder: Learning to Filter Netnews. In *Proc. of the International Conference on Machine Learning* (*ICML'95*), pp. 331–339, 1995.
301. P. Langley, W. Iba, and K. Thompson. An Analysis of Bayesian Classifiers. In *Proc. of the 10th National Conf. on Artificial Intelligence* (*AAAI'92*), pp. 223–228, 1992.
302. P. Langley. *Elements of Machine Learning*. Morgan Kauffmann, 1996.
303. A. N. Langville and Carl D. Meyer. Deeper Inside PageRank. *Internet Mathematics*, 1(3), pp. 335–380, 2005.
304. A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.
305. D. T. Larose. *Discovering Knowledge in Data: An Introduction to Data Mining*.

John Wiley, 2004.
306. J. A Larson, S. B Navathe, and R. ElMasri. A Theory of Attribute Equivalence in Databases with Application to Schema Integration. In *Proc. of IEEE Trans Software Engineering* 16(4): pp. 449–463, 1989.
307. S. Lawrence and C. Giles. Accessibility of Information on the Web. *Nature* 400, pp. 107–109, 1999.
308. S. Lawrence, C. L. Giles, and K. Bollaker. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer* 32(6), pp. 67–71, 1999.
309. W. S. Lee, and B. Liu. Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression. In *Proc. of the Twentieth Intl. Conf. on Machine Learning* (*ICML'03*), pp. 448–455, 2003.
310. R. Lempel and S. Moran. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. In *Proc. of the Ninth Intl. World Wide Web Conf.* (*WWW'9*), pp. 387–401, 2000.
311. A. V. Leouski and W. B. Croft. *An Evaluation of Techniques for Clustering Search Results*. Technical Report IR–76, Department of Computer Science, University of Massachusetts, Amherst, 1996.
312. K. Lerman, L. Getoor, S. Minton, and C. Knoblock. Using the Structure of Web Sites for Automatic Segmentation of Tables. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'04*), pp. 119–130, 2004.
313. J. Lerner and M. Pinkal. *Comparatives and Nested Quantification*. CLAUS-Report 21, 1992.
314. N. Lesh, M. J. Zaki, and M. Ogihara. Mining Features for Sequence Classification. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'99*), 1999.
315. F. Letouzey, F. Denis, and R. Gilleron. Learning from Positive and Unlabeled Examples. In *Proc. of the 11th Intl. Conf. on Algorithmic Learning Theory* (*ALT'00*), pp. 71–85, 2000.
316. D. Lewis. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proc. of the ACM SIGIR Intl. Conf. on Research and Development in Information Retrieval* (*SIGIR'92*), pp. 37–50, 1992.
317. D. Lewis and W. Gale. A Sequential Algorithm for Training Text Classifiers. In *Proc. of the ACM SIGIR Intl. Conf. on Research and Development in Information Retrieval* (*SIGIR'94*), pp. 3–12, 1994.
318. H. Li and K. Yamanishi. Document Classification Using a Finite Mixture Model. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 39–47, 1997.
319. J. Li, G. Dong, K. Ramamohanarao. Making Use of the Most Expressive Jumping Emerging Patterns for Classification. In *Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining* (*PAKDD'00*), pp. 220–232, 2000.
320. J. Li, G. Dong, K. Ramamohanarao, and L. Wong. DeEPs: A New Instance-Based Lazy Discovery and Classification System. *Machine Learning*, 54(2), pp. 99–124 2004.
321. X. L. Li, and B. Liu. Learning to Classify Text Using Positive and Unlabeled Data. In *Proc. of the Eighteenth Intl. Joint Conf. on Artificial Intelligence* (*IJCAI'03*), pp. 587–594, 2003.
322. X. L. Li, and B. Liu. Learning from Positive and Unlabeled Examples with Different Data Distributions. In *Proc. of the European Conf. on Machine Learning* (*ECML'05*), 2005.
323. X. L. Li, B. Liu and S-K. Ng. Learning to Identify Unexpected Instances in the Test Set. To appear in *Proc. of Intl. Joint Conf. on Artificial Intelligence* (*IJCAI'06*), 2006.

324. X. L. Li, T.-H. Phang, M. Hu, and B. Liu. Using Micro Information Units for Internet Search. In *Proc. of the ACM Intl. Conf. on Information and Knowledge Management* (*CIKM'02*), pp. 566–573, 2002.
325. X. Li, B. Liu and P. S. Yu. Discovering Overlapping Communities of Named Entities. In *Proc. of Conf. on Practical Knowledge Discovery and Data Mining* (*PKDD'06*), 2006.
326. X. Li, B. Liu and P. S. Yu. Time Sensitive Ranking with Application to Publication Search. *Forthcoming paper*, 2006.
327. W. Li, and C. Clifton. SemInt: a Tool for Identifying Attribute Correspondences in Heterogeneous Databases Using Neural Network. *Data Knowledge Engineering* 33(1), pp. 49–84, 2000.
328. W. Li, J. Han, and J. Pei. CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In *Proc. of the 2001 IEEE Intl. Conf. on Data Mining* (*ICDM'01*), pp. 369–376, 2001.
329. C. Li, J.-R. Wen, and H. Li. Text Classification Using Stochastic Keyword Generation. In *Proc. of Intl. Conf. on Machine Learning* (*ICML'03*), pp. 464-471, 2003.
330. G. Lidstone. Note on the General Case of the Bayes-Laplace formula for Inductive or a Posteriori Probabilities. *Transactions of the Faculty of Actuaries*, 8, pp. 182–192, 1920.
331. W. Lin, S. A. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6, pp. 83-105, 2002.
332. D. Lin. PRINCIPAR-An Efficient, Broad-Coverage, Principle-Based Parser. In *Proc. of the 15th Conf. on Computational Linguistics*, pp. 482–488., 1994.
333. C.-R Lin, and M.-S. Chen: A Robust and Efficient Clustering Algorithm Based on Cohesion Self-merging. In *Proc. of the SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'02*), pp. 582–587, 2002.
334. D.-I. Lin and Z. M. Kedem. Pincer-Search: A New Algorithm for Discovering the Max Mum Frequent Set. In *Proc. of the 6th Intl. Conf. Extending Database Technology* (*EDBT'98*), 1998.
335. L.-J. Lin. Self-Improving Reactive Agents Based on Reinforcement Learning, Planning, and Teaching. *Machine Learning* 8, pp. 293–321, 1992.
336. S.-H. Lin, and J.-M. Ho. Discovering Informative Content Blocks from Web Documents. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'02*), pp. 588–593, 2002
337. W. Lin, S. A. Alvarez, and C. Ruiz. Efficient Adaptive-Support Association Rule Mining for Recommender Systems. *Data Mining and Knowledge Discovery*, 6, pp. 83–105, 2002.
338. G. S. Linoff, and M. J. Berry. *Mining the Web: Transforming Customer Data into Customer Value*. John Wiley & Sons. 2002.
339. B. Liu, C. W. Chin, and H. T. Ng. Mining Topic-Specific Concepts and Definitions on the Web. In *Proc. of the 12th Intl. World Wide Web Conf.* (*WWW'03*), pp. 251–260, 2003.
340. B. Liu, Yang Dai, Xiaoli Li, Wee Sun Lee and Philip Yu. Building Text Classifiers Using Positive and Unlabeled Examples. In *Proc. of the 3rd IEEE Intl. Conf. on Data Mining* (*ICDM'03*), pp. 179–188, 2003.
341. B. Liu, R. Grossman, and Y. Zhai. Mining Data Records in Web Pages. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'03*), pp. 601–606. 2003.
342. B. Liu, W. Hsu, and S. Chen. Using General Impressions to Analyze Discovered Classification Rules. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Dis-*

covery and Data Mining (*KDD'97*), pp. 31-36, 1997.
343. B. Liu, W. Hsu, and Y. Ma. Integrating Classification and Association Rule Mining. In *Proc. of Knowledge Discovery and Data Mining* (*KDD'98*), pp. 80–86, 1998.
344. B. Liu, W. Hsu, and Y. Ma. Mining Association Rules with Multiple Minimum Supports. In *Proc. of Intl. Conf. Knowledge Discovery and Data Mining* (*KDD'99*), pp. 337–341, 1999.
345. B. Liu, W. Hsu, and Y. Ma. Pruning and Summarizing the Discovered Associations. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'99*), pp. 125-134, 1999.
346. B. Liu, W. Hsu, L. Mun, and H. Lee. Finding Interesting Patterns Using User Expectations. *IEEE Transactions on Knowledge and Data Engineering*, 11(6), pp.817–832, 1999.
347. B. Liu, M. Hu, and J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proc. of the 14th Intl. World Wide Web Conf.* (*WWW'05*), pp. 342–351, 2005.
348. B. Liu, W. S. Lee, Philip S. Yu and Xiaoli Li. Partially Supervised Classification of Text Documents. In *Proc. of the Nineteenth Intl. Conf. on Machine Learning* (*ICML'02*), pp. 8–12, 2002.
349. B. Liu, Y. Ma, and C-K Wong, Classification Using Association Rules: Weaknesses and Enhancements. In Vipin Kumar, et al, (eds), *Data Mining for Scientific Applications*, 2001.
350. B. Liu, Y. Xia, and P. S. Yu. Clustering through Decision Tree Construction. In *Proc. of the ACM Intl. Conf. on Information and Knowledge Management* (*CIKM'00*), pp. 20–29, 2000.
351. B. Liu and Y. Zhai. NET – A System for Extracting Web Data from Flat and Nested Data Records. In *Proc. of 6th Intl. Conf. on Web Information Systems Engineering* (*WISE'05*), pp. 487–495, 2005.
352. B. Liu, K. Zhao, J. Benkler and W. Xiao. Rule Interestingness Analysis Using OLAP Operations. In *Proc. of the Twelfth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'06*), pp. 297–306, 2006.
353. H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, 1998.
354. J. Lu and J. Callan. Content-Based Retrieval in Hybrid Peer-to-Peer Networks. In *Proc. 12th ACM Intl. Conf. on Information and Knowledge Management* (*CIKM'03*), pp. 199–206, 2003
355. L. Ma, N. Goharian, and A. Chowdhury. Extracting Unstructured Data from Template Generated Web Document. In *Proc. of the ACM Intl. Conf. on Information and Knowledge Management* (*CIKM'03*), pp. 512–515, 2003
356. J. Madhavan, P. A. Bernstein, A. Doan, and A. Y. Halevy. Corpus-Based Schema Matching. In *Proc. of International Conference on Data Engineering* (*ICDE'05*), pp. 57–68, 2005.
357. J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.
358. J. Madhavan, P. A. Bernstein, and E. Rahm. Generic Schema Matching with Cupid. In *Proc 27th Int. Conf. on Very Large Data Bases* (*VLDB'01*), pp. 49–58, 2001.
359. A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani. Algorithmic Detection of Semantic Similarity. In *Proc. 14th Intl. World Wide Web Conf.* (*WWW'05*), pp. 107–116, 2005.
360. L. Manevitz and M. Yousef. One-Class SVMs for Document Classification. *Journal of Machine Learning Research*, 2, pp. 139–154, 2001.

361. H. Mannila, H. Toivonen, and I. Verkamo. Efficient Algorithms for Discovering Association Rules. In *Proc. of Knowledge Discovery in Databases* (*KDD'94*), pp. 181–19, AAAI Press 1994
362. C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts, 1999.
363. B. Markines, L. Stoilova, and F. Menczer. Social Bookmarks for Collaborative Search and Recommendation. In *Proc. of the 21st National Conf. on Artificial Intelligence* (*AAAI'06*), 2006.
364. O. McBryan. Genvl and WWWW: Tools for Taming the Web. In O. Nierstrasz (Ed.). In *Proc. of the First Intl. World Wide Web Conf.*, Geneva. CERN, 1994.
365. A. McCallum, and K. Nigam. A Comparison of Event Models for Naïve Bayes Text Classification. In *Proc. of the AAAI–98 Workshop on Learning for Text Categorization*. 1998.
366. A. McCallum, K. Nigam, J. Rennie, and K. Seymore. A Machine Learning Approach to Building Domain-Specific Search Engines. In *Proc. 16th Intl. Joint Conf. on Artificial Intelligence* (*IJCAI'99*), pp. 662–667, 1999.
367. R. McCann, B. K. AlShebli, Q. Le, H. Nguyen, L. Vu, and A. Doan: Mapping Maintenance for Data Integration Systems. In *Proc. of Intl. Conf. on Very Large Data Bases* (*VLDB'05*): pp. 1018–1030, 2005.
368. F. McSherry. A Uniform Approach to Accelerated PageRank Computation. In *Proc. of the 14th Intl. World Wide Web Conference* (*WWW'05*), pp. 575–582, 2005.
369. F. Menczer. ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery. In *Proc. of the 14th Intl. Conf. on Machine Learning*, pp. 227–235, 1997.
370. F. Menczer. Growing and Navigating the Small World Web by Local Content. In *Proc. Natl. Acad. Sci*. USA, 99(22), pp. 14014-14019, 2002
371. F. Menczer. The Evolution of Document Networks. In *Proc. Natl. Acad. Sci*. USA 101, pp. 5261-5265, 2004
372. F. Menczer. Lexical and Semantic Clustering by Web Links. *Journal of the American Society for Information Science and Technology*, 55(14), pp. 1261–1269, 2004.
373. F. Menczer. Mapping the Semantics of Web Text and Links. *IEEE Internet Computing* 9 (3), pp. 27–36, 2005.
374. F. Menczer and R. Belew. Adaptive Information Agents in Distributed Textual Environments. In *Proc. of the 2nd Intl. Conf. on Autonomous Agents*, pp. 157–164, 1998.
375. F. Menczer and R. Belew. Adaptive Retrieval Agents: Internalizing Local Context and Scaling up to the Web. *Machine Learning*, 39(2–3), 203–242, 2000.
376. F. Menczer, G. Pant, M. Ruiz, and P. Srinivasan. Evaluating Topic Driven Web Crawlers. In *Proc. 24th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 241–249, 2001.
377. F. Menczer, G. Pant, and P. Srinivasan. Topical Web Crawlers: Evaluating Adaptive Algorithms. *ACM Transactions on Internet Technology* 4(4), pp. 378–419, 2004.
378. W. Meng, C. Yu, and K.-L. Liu. Building Efficient and Effective Metasearch Engines. *ACM Computing Surveys*, 34(1), pp. 48–84, 2002.
379. D. Meretakis, and B. Wüthrich. Extending Naïve Bayes Classifiers Using Long Itemsets. In *Proc. of the Fifth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'99*), pp. 165–174, 1999.
380. A. Micarelli and F. Gasparetti (in press). Adaptive Focused Crawling. In P. Brusilovsky, W. Nejdl, and A. Kobsa (eds.), *Adaptive Web*. Springer.
381. R. S. Michalski, I. Mozetic, J. Hong and N. Lavrac. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In *Proc. of the National Conf. on Artificial Intelligence* (*AAAI'86*), pp. 1041–1047,

1986.
382. T. Milo, S. Zohar. Using schema matching to simplify heterogeneous data translation. In: *Proc. of Intl Conf on Very Large Data Bases* (*VLDB'98*), pp. 122–133, 1998.
383. B. Mirkin. *Clustering for Data Mining: A Data Recovery Approach*. Chapman & Hall/CRC, April 29, 2005.
384. N. Misha, D. Ron, and R. Swaminathan. A New Conceptual Clustering Framework. *Machine Learning*, 56(1–3): pp. 115–151, 2004.
385. T. Mitchell. *Machine Learning*. McGraw Hill, 1997.
386. B. Mobasher. Web Usage Mining and Personalization. In Munindar P. Singh (ed.), *Practical Handbook of Internet Computing*. CRC Press, 2005.
387. B. Mobasher. Web Usage Mining. In John Wang (eds.), *Encyclopedia of Data Warehousing and Mining*, Idea Group, 2006.
388. B. Mobasher, R. Cooley and J. Srivastava. Automatic Personalization based on Web Usage Mining. *Communications of the ACM*, 43(8), pp. 142–151, 2000.
389. B. Mobasher, H. Dai, T. Luo, and N. Nakagawa. Effective Personalization Based on Association Rule Discovery from Web Usage Data. In *Proc. of the 3rd ACM Workshop on Web Information and Data Management* (*WIDM01*), pp. 9–15, 2001.
390. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery*, 6, pp. 61–82, 2002.
391. B. Mobasher and R. Burke and J. J Sandvig. Model-Based Collaborative Filtering as a Defense against Profile Injection Attacks. In *Proc. of the 21st National Conf. on Artificial Intelligence (AAAI'06)*, 2006.
392. A. Moffat, R. Neal, and I. Witten. Arithmetic Coding Revisited. *ACM Transactions on Information Systems*, pp. 256–294, 1998.
393. F. Moltmann, *Coordination and Comparatives*. Ph.D. dissertation. MIT, Cambridge Ma., 1987.
394. M. Montague, and J. Aslam. Condorcet Fusion for Improved Retrieval. *In Proc. of the Intl. Conf. on Information and Knowledge Management* (*CIKM'02*), pp. 538–548, 2002.
395. R. J. Mooney and R. Bunescu. Mining Knowledge from Text Using Information Extraction. *SIGKDD Explorations*, pp. 3–10. 2005.
396. A. Moore. Very Fast EM-based Mixture Model Clustering Using Multiresolution Kd-Trees. In *Proc. of the Neural Info. Processing Systems* (*NIPS'98*), pp. 543–549, 1998.
397. S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining Product Reputations on the Web. In *Proc. of the SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'02*), pp. 341–349, 2002.
398. S. Muggleton. Learning from the Positive Data. *Inductive Logic Programming Workshop*, pp. 358–376, 1996.
399. I. Muslea, S. Minton, and C. A. Knoblock. A Hierarchical Approach to Wrapper Induction. In *Proc. of the Intl. Conf. on Autonomous Agents* (*AGENTS'99*), pp. 190–197, 1999.
400. I. Muslea, S. Minton, and C. A. Knoblock. Active Learning with Multiple Views. *Journal of Artificial Intelligence Research*, 1, pp. 1–31, 2006.
401. M. Najork and J. L. Wiener. Breadth-First Search Crawling Yields High Quality Pages. In *Proc. of the 10th Intl. World Wide Web Conf.* (*WWW'01*), pp. 114–118, 2001.
402. T. Nasukawa and J. Yi. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In *Proc. of the K-CAP-03, 2nd Intl. Conf. on Knowledge Capture*, pp. 70–77, 2003.
403. T. Nelson. A File Structure for the Complex, the Changing and the Indeterminate. In

Proc. of the ACM National Conf., pp. 84–100, 1965.
404. A. Ng, M. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Proc. of the 14th Advances in Neural Information Processing Systems*, pp. 849–856, 2001.
405. A. Ng, A. X. Zheng, and M. I. Jordan. Stable Algorithms for Link Analysis. In *Proc. of the 24th Annual ACM SIGIR Intl. Conf on Research and Development on Information Retrieval* (*SIGIR'01*), 2001.
406. R. T. Ng and J. Han. Efficient and Effective Clustering Methods for Spatial Data Mining. In *Proc. of Conf. on Very Large Data Bases* (*VLDB'94*), pp. 144–155, 1994.
407. R. T. Ng, J. Han. CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEE Transactions Knowledge Data Engineering* 14(5), pp. 1003-1016, 2002.
408. R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory Mining and Pruning Optimizations of Constrained Association Rules. In *Proc. of ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'98*), pp. 13–24, 1998.
409. V. Ng, S. Dasgupta and S. M. Niaz Arifin. Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. In. *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 611–618, 2006.
410. Z. Nie, Y. Zhang, J-R. Wen, and W-Y Ma. Object Level Ranking: Bringing Order to Web Objects. In *Proc. of the 14th Intl. World Wide Web Conference* (*WWW'05*), pp. 567–574, 2005
411. K. Nigam and R. Ghani. Analyzing the Effectiveness and Applicability of Co-training. In *Proc. of the ACM Intl. Conf. on Information and Knowledge Management* (*CIKM'00*), pp. 86–93, 2000.
412. K. Nigam and M. Hurst. Towards a Robust Metric of Opinion. *AAAI Spring Symp. on Exploring Attitude and Affect in Text*, 2004.
413. K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learning*, 39(2/3), pp. 103–134, 2000.
414. Z. Niu, D. Ji, and C. L. Tan. Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning. In *Proc. of the Meeting of the Association for Computational Linguistics* (*ACL'05*), 2005.
415. C. Notredame. Recent Progresses in Multiple Sequence Alignment: a Survey. Technical report, *Information Génétique et*, 2002.
416. A. Ntoulas, J. Cho, and C. Olston. What's New on the Web? The Evolution of the Web from a Search Engine Perspective. In *Proc. of the 13th Intl. World Wide Web Conference* (*WWW'04*), pp. 1–12, 2004.
417. A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting Spam Web Pages through Content Analysis. In *Proc. of the 15th Intl. World Wide Web Conference* (*WWW'06*), pp. 83–92, 2006.
418. R. Nuray, and F. Can. Automatic Ranking of Information Retrieval Systems Using Data Fusion. *Information Processing and Management*, 4(3), pp. 595–614, 2006.
419. M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre. Collaborative Recommendation: A Robustness Analysis. *ACM Transactions on Internet Technology* 4(4):344–377, 2004.
420. B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic Association Rules. In *Proc. 1998 Int. Conf. Data Engineering* (*ICDE'98*), pp. 412–421, 1998.
421. B. Padmanabhan, and A. Tuzhilin. Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'00*), pp. 54-63, 2000.
422. L. Page, S. Brin, R. Motwami, and T. Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999–0120, Computer Science De-

partment, Stanford University, 1999.
423. G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C. D. Spyropoulos. Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques. *Interacting with Computers Journal*, 14(6), pp. 761–791, 2002.
424. L. Palopoli, D. Sacca, and D. Ursino. An Automatic Technique for Detecting Type Conflicts in Database Schemas. In: *Proc of ACM Intl. Conf on Information and Knowledge Management* (*CIKM'98*), pp. 306–313, 1998.
425. S. Pandey, S. Roy, C. Olston, J. Cho and S. Chakrabarti, Shuffling a Stacked Deck: The Case for Partially Randomized Ranking of Search Engine Results. In *Proc. of Very Large Data Bases* (*VLDB'05*), pp. 781–792, 2005.
426. B. Pang and L. Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization based on Minimum Cuts. In *Proc. of the 42nd Meeting of the Association for Computational Linguistics* (*ACL'04*), pp. 271—278, 2004.
427. B. Pang and L. Lee, Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proc. of the Meeting of the Association for Computational Linguistics* (*ACL'05*), pp. 115–124, 2005.
428. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proc. of the EMNLP'02*, 2002.
429. G. Pant. Deriving Link-Context from Html Tag Tree. In *Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (*DMKD'03*), pp. 49–55, 2003.
430. G. Pant, S. Bradshaw, and F. Menczer. Search Engine – Crawler Symbiosis. In *Proc. of the 7th European Conf. on Research and Advanced Technology for Digital Libraries* (*ECDL'03*), 2003.
431. G. Pant and F. Menczer. MySpiders: Evolve your Own Intelligent Web Crawlers. *Autonomous Agents and Multi-Agent Systems*, 5(2), pp. 221–229, 2002.
432. G. Pant and F. Menczer. Topical Crawling for Business Intelligence. In *Proc. of the 7th European Conf. on Research and Advanced Technology for Digital Libraries* (*ECDL'03*), pp. 233–244, 2003.
433. G. Pant, and P. Srinivasan. Learning to Crawl: Comparing Classification Schemes. ACM Trans. *Information Systems*, 23(4), pp. 430–462, 2005.
434. G. Pant, P. Srinivasan, and F. Menczer. Exploration versus Exploitation in Topic Driven Crawlers. In *Proc. of the WWW-02 Workshop on Web Dynamics*, 2002.
435. J. S. Park, M.-S. Chen, and P. S. Yu: An Effective Hash Based Algorithm for Mining Association Rules. In *Proc. of the 1995 ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'95*), pp. 175–186, 1995.
436. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering Frequent Closed Itemsets for Association Rules. In *Proc. of the 7th Intl. Conf. on Database Theory*, pp. 398–416, 1999.
437. R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet*. Cambridge University Press, 2004.
438. M. J. Pazzani, C. Brunk, and G. Silverstein. A Knowledge-Intensive Approach to Learning Relational Concepts. In *Proc. of the Eighth Intl. Workshop on Machine Learning* (*ML'91*), pp. 432–436, 1991.
439. J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. Prefix-Span: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In *Proc. of the 2001 Int. Conf. Data Engineering* (*ICDE'01*), pp. 215–224, 2001.
440. M. Pennock, G. W. Flakes, S. Lawrence, C.L Giles, and E. J Gloves. Winners Don't Take All: Characterizing the Competition for Links on the Web. In *Proc. of National Academy of Science*, 99(8), pp. 5207–5211, 2002.
441. P. Perner. *Data Mining on Multimedia Data*. Springer, 2003.

442. T. P. Pham, H. T. Ng, and W. S. Lee. Word Sense Disambiguation with Semi-Supervised Learning. In *Proc. of the National Conference on Artificial Intelligence* (*AAAI'05*). pp. 1093–1098, 2005.
443. G. Piatetsky-Shapiro, and B. Masand. Estimating Campaign Benefits and Modeling Lift. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'99*), pp. 185-193, 1999.
444. G. Piatesky-Shapiro, and C. Matheus. The Interestingness of Deviations. In *Proc. of Knowledge Discovery and Data Mining* (*KDD'94*), 1994.
445. G. Pierrakos, G. Paliouras, C. Papatheodorou, and C. Spyropoulos. Web Usage Mining as a Tool for Personalization: a Survey. *User Modeling and User-Adapted Interaction*, 13, pp. 311–372, 2003.
446. J. Pitkow and P. Pirolli. Mining Longest Repeating Subsequences to Predict WWW Surfing. In *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems*, 1999.
447. A.-M. Popescu, and O. Etzioni. Extracting Product Features and Opinions from Reviews. In *Proc. of Conference on Empirical Methods in Natural Language Processing* (*EMNLP'05*), 2005.
448. J. Ponte, and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *Proc. of the Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (*SIGIR'98*), pp. 275-281, 1998.
449. M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3), pp 130−137, 1980.
450. D. Pyle. *Business Modeling and Data Mining*. Morgan Kaufmann, 2003.
451. F. Qiu and J. Cho. Automatic Identification of User Interest for Personalized Search. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
452. J. R. Quinlan. Learning Logical Definitions from Relations. *Machine Learning*, 5, pp. 239–266, 1990.
453. J. R. Quinlan. *C4.5: Program for Machine Learning*. Morgan Kaufmann, 1992.
454. J. R. Quinlan. Bagging, Boosting, and C4.5. In *Proc. of National Conf. on Artificial Intelligence* (*AAAI-96*), pp. 725-730, 1996.
455. E. Rahm, and P. A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *VLDB Journal*, 10, pp. 334–35, 2001.
456. L. Ramaswamy, A. Lyengar, L. Liu, and F. Douglis. Automatic Detection of Fragments in Dynamically Generated Web Pages. In *Proc. of the 13th Intl. World Wide Web Conference* (*WWW'04*), pp. 443–454, 2004
457. J. Raposo, A. Pan, M. Alvarez, J. Hidalgo, and A. Vina. The Wargo System: Semi-Automatic Wrapper Generation in Presence of Complex Data Access Modes. In *Proc. of the 13th Intl. Work-shop on Database and Expert Systems Applications*, pp. 313–320, 2002.
458. D. de Castro Reis, P. B. Golgher, A. S. da Silva, and A. H. F. Laender. Automatic Web News Extraction Using Tree Edit Distance. In *Proc. of the 13th Intl. World Wide Web Conference* (*WWW'04*), pp. 502–511, 2004
459. J. Rennie and A. McCallum. Using Reinforcement Learning to Spider the Web Efficiently. In *Proc. of the 16th Intl. Conf. on Machine Learning* (*ICML'99*), pp. 335–343, 1999.
460. M. Richardson, A. Prakash, and E. Brill. Beyond PageRank: Machine Learning for Static Ranking. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
461. C. van Rijsbergen. *Information Retrieval*, Chapter 3, London: Butterworths. Second edition, 1979.
462. E. Riloff and J. Wiebe. Learning Extraction Patterns for Subjective Expressions. In *Proc. of Conference on Empirical Methods in Natural Language Processing* (*EMNLP'03*), 2003.

463. R. L. Rivest. Learning Decision Lists. *Machine Learning*, 2(3), pp. 229–246, 1987.
464. S. E. Robertson and K. Sparck-Jones. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27, pp. 129–146, 1976.
465. S. E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC–7: Automatic Ad Hoc, Filtering, VLC and Filtering Tracks. In *Proc. of the Seventh Text REtrieval Conference* (*TREC-7*), pp. 253–264, 1999.
466. J. Rocchio. Relevant Feedback in Information Retrieval. In G. Salton (eds.). *The Smart Retrieval System – Experiments in Automatic Document Processing*, Englewood Cliffs, NJ, 1971
467. R. Roiger and M. Geatz. *Data Mining: A Tutorial Based Primer*. Addison-Wesley, 2002.
468. O. Parr Rud. *Data Mining Cookbook*. John Wiley & Sons, 2003.
469. D. Rumelhart, G. Hinton, and R. Williams. Learning Internal Representations by Error Propagation. In D. Rumelhart and J. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1, Chapter 8, pp. 318–362, 1996.
470. G. Salton and C. Buckley. Term-Weighting Approaches in Automatic Retrieval. *Information Processing and Management*, 24(5), pp. 513–525, 1988.
471. G. Salton and M. McGill. *An Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill, 1983.
472. B. Santorini. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990.
473. R. R. Sarukkai. Link Prediction and Path Analysis Using Markov Chains. In *Proc. of the 9th Intl. World Wide Web Conf.* (*WWW'00*), pp. 377–386. 2000.
474. B. Sarwar, G. Karypis, J. Konstan and J. Riedl. Application of Dimensionality Reduction in Recommender Systems – A Case Study. In *Proc. of the KDD Workshop on WebKDD'2000*, 2000.
475. B. Sarwar, G. Karypis, J. Konstan and J. Riedl. Item-Based Collaborative Filtering Recommendation Algorithms. In *Proc. of the 10th Intl. World Wide Web Conference* (*WWW'01*), pp. 285–295, 2001.
476. A. Savasere, E. Omiecinski, and S. B. Navathe. Mining for Strong Negative Associations in a Large Database of Customer Transactions. In *Proc. of the Fourteenth Intl. Conf. on Data Engineering* (*ICDE'98*), pp. 494–502, 1998.
477. R. E. Schapire. The Strength of Weak Learnability. *Machine Learning*, 5(2), pp. 197–227, 1990.
478. S. Scholkopf, J. Platt, J. Shawe, A. Smola, and R. Williamson. *Estimating the Support of a High-Dimensional Distribution*. Technical Report MSR-TR-99-87, Microsoft Research, pp. 1443–1471, 1999.
479. B. Scholkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.
480. A. Scime (eds.). *Web Mining: Applications and Techniques*. Idea Group Inc., 2005.
481. G. L. Scott and H. C. Longuet-Higgins. Feature Grouping by Relocalisation of Eigenvectors of the Proximity Matrix. In *Proc. British Machine Vision Conf*., pp. 103–108, 1990.
482. M. Seno, and G. Karypis. Finding Frequent Patterns Using Length-Decreasing Support Constraints. *Data Mining and Knowledge Discovery*, 10(3), pp 197–228, 2005.
483. J. G. Shanahan, Y. Qu, and J. Wiebe, (eds.). *Computing Attitude and Affect in Text*: *Theory and Applications*. Springer. 2005.
484. E. Shannon. A Mathematical Theory of Communication. In *Bell System Technical Journal*, 27: pp. 379–423, 1948.
485. G. Sheikholeslami, S. Chatterjee and A. Zhang. WaveCluster: a Multi-resolution

Clustering Approach for Very Large Spatial Databases. In *Proc. of Very Large Data Bases* (*VLDB'98*), pp. 428–439, 1998.
486. D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. A Comparison of Implicit and Explicit Links for Web Page Classification. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
487. X. Shen, B. Tan, and C. Zhai. Context-Sensitive Information Retrieval with Implicit Feedback. In *Proc. of the Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (*SIGIR'05*), pp. 43-50, 2005.
488. A. Sheth and J. Larson. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys*, 22(3):183–236, 1990.
489. J. Shi and J. Malik. Normalized Cuts and Image Segmentation. In *Proc. of the IEEE Conf. Computer Vision and Pattern Recognition*, pp. 731–737, 1997.
490. X. Shi and C. C. Yang, Mining Related Queries from Search Engines Query Logs. In *Proc. of the Intl. World Wide Web Conf.* (*WWW'06*), pp. 943-944, 2006.
491. P. Shvaiko, and J. Euzenat. A Survey of Schema-Based Matching Approaches. *Journal on Data Semantics*, IV, LNCS 3730, pp. 146–171, 2005.
492. A. Silberschatz, and A. Tuzhilin. What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), pp. 970–974, 1996.
493. A. Singhal. Modern Information Retrieval: A Brief Overview. *IEEE Data Engineering Bulletin* 24(4), pp. 35-43, 2001.
494. H. Small. Co-Citation in the Scientific Literature: a New Measure of the Relationship between Two Documents. *Journal of American Society for Information Science*, 24(4), pp. 265–269, 1973.
495. R. Song, H. Liu, J. R. Wen, and W. Y. Ma. Learning Block Importance Models for Web Pages. In *Proc. of the 13th Conf. on World Wide Web* (*WWW'04*), pp. 203–211, 2004.
496. M. Spiliopoulou. Web Usage Mining for Web Site Evaluation. *Communications of ACM*, 43(8), pp. 127–134, 2000.
497. M. Spiliopoulou, and L. Faulstich. WUM: A Tool for Web Utilization Analysis. In *Proc. of EDBT Workshop at WebDB'98*, pp. 184–203, 1999.
498. M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis. *INFORMS Journal of Computing*, 15(2), pp. 171–190, 2003.
499. R. Srikant and R. Agrawal. Mining Generalized Association Rules. In *Proc. of the 21st Int'l Conf. on Very Large Data Bases* (*VLDB'95*), pp. 407–419, 1995.
500. R. Srikant and R. Agrawal. Mining Sequential Patterns: Generalizations and Performance Improvements. In *Proc. of the 5th Intl. Conf. Extending Database Technology* (*EDBT'96*), pp. 3–17, 1996.
501. R. Srikant and R. Agrawal. Mining Quantitative Association Rules in Large Relational Tables. In *Proc. of the ACM SIGMOD Conf. on Management of Data* (*SIGMOD'96*), 1996.
502. R. Srikant, Q. Vu, and R. Agrawal. Mining Association Rules with Item Constraints. In *Proc. of the 3rd Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'97*), pp. 67–73, 1997.
503. P. Srinivasan, J. Mitchell, O. Bodenreider, G. Pant, and F. Menczer. Web Crawling Agents for Retrieving Biomedical Information. In *Proc. of the Intl. Workshop on Agents in Bioinformatics* (*NETTAB'02*), 2002.
504. P. Srinivasan, G. Pant, and F. Menczer. A General Evaluation Framework for Topical Crawlers. *Information Retrieval* 8 (3), pp. 417–447, 2005.

505. J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations*, 1(2), pp. 12–23, 2000.
506. M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. In *Proc. of the KDD Workshop on Text Mining*, 2000.
507. V. Stoyanov and C. Cardie. Toward Opinion Summarization: Linking the Sources. In *Proc. of the Workshop on Sentiment and Subjectivity in Text*, pp. 9–14, 2006.
508. J-T. Sun, X. Wang, D. Shen, H-J. Zeng, and Z. Chen. CWS: A Comparative Web Search System. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
509. K.-C. Tai. The Tree-to-Tree Correction Problem. *Journal of the ACM*, 26(3), pp. 422–433, 1979.
510. P. -N. Tan and V. Kumar. Discovery of Web Robot Sessions Based on their Navigational Patterns. *Data Mining and Knowledge Discovery*, 6(1), pp. 9–35, 2002.
511. P. -N. Tan, V. Kumar and J. Srivastava. Selecting the Right Interestingness Measure for Association Patterns. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'02*), pp. 32-41, 2002.
512. P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.
513. D. Tanasa and B. Trousse. Advanced Data Preprocessing for Intersite Web Usage Mining. *IEEE Intelligent Systems*, 19(2), pp. 59–65, 2004.
514. Z.-H. Tang, and J. MacLennan. *Data Mining with SQL Server 2005*. Wiley publishing, Inc. 2005.
515. B M. Thuraisingham. *Web Data Mining and Applications in Business Intelligence and Counter-Terrorism*, CRC Press, 2003.
516. J. Tomlin. A New Paradigm for Ranking Pages on the World Wide Web. In *Proc. of the 12th Intl. World Wide Web Conference* (*WWW'02*), pp. 350–355, 2003.
517. R. Tong. An Operational System for Detecting and Tracking Opinions in On-Line Discussion. In *Proc. of SIGIR Workshop on Operational Text Classification*, 2001.
518. M. Toyoda and M. Kitsuregawa. Creating a Web Community Chart for Navigating Related Communities. In *Proc. of the Twelfth ACM Conf. on Hypertext and Hypermedia*, pp. 103–112, 2001.
519. M. Toyoda and M. Kitsuregawa. Extracting Evolution of Web Communities from a Series of Web Archives. In *Proc. of the fourteenth ACM Conf. on Hypertext and Hypermedia*, pp. 28–37, 2003.
520. M. Toyoda, and M. Kitsuregawa. What's Really New on the Web?: Identifying New Pages from a Series of Unstable Web Snapshots. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
521. P. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proc. of the Meeting of the Association for Computational Linguistics* (*ACL'02*), pp. 417–424, 2002
522. A. Tuzhilin, and G. Adomavicius. Handling very Large Numbers of Association Rules in the Analysis of Microarray Data. In *Proc. of ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'02*), pp. 396–404, 2002.
523. J. R. Tyler, D. M. Wilkinson, and B. A. Huberman. Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. *Communities and Technologies*, pp. 81–96. 2003.
524. A. Valitutti, C. Strapparava, and O. Stock. Developing Affective Lexical Resources. *Psychnology Journal*, 2(1): pp. 61–83, 2004.
525. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
526. V. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.
527. H. Wache, T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and

S. Huebner. Ontology-Based Integration of Information – a Survey of Existing Approaches. In *Proc. of the IJCAI Workshop on Ontologies and Information Sharing*, pp. 108–117, 2001.
527. K. Wagstaff and C. Cardie. Clustering with Instance-Level Constraints. In *Proc. of the 17th Intl. Conf. on Machine Learning*, pp. 1103–1110, 2000.
528. J. Wang, J. Han, and J. Pei. Closet+: Searching for the Best Strategies for Mining Frequent Closed Itemsets. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'03*), pp. 236–245, 2003.
529. J. Wang and F. H. Lochovsky. Data Extraction and Label Assignment for Web Databases. In *Proc. of the 12th Intl. World Wide Web Conference* (*WWW'03*), pp. 187–196, 2003.
530. J. Wang, J-R. Wen, F. H. Lochovsky, and W-Y. Ma. Instance-Based Schema Matching for Web Databases by Domain-specific Query Probing. In *Proc. of the Intl. Conf. on Very Large Data Bases* (*VLDB'04*), pp. 408–419, 2004.
531. J. T.-L. Wang, B. A. Shapiro, D. Shasha, K. Zhang, and K. M. Currey. An algorithm for finding the largest approximately common substructures of two trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), pp. 889.895, 1998.
532. J. Wang, M. Zaki, H. Toivonen, and D. Shasha. (eds.). *Data Mining in Bioinformatics.* Springer, 2004.
533. K. Wang, Yu He, and J. Han. Mining Frequent Itemsets Using Support Constraints. In *Proc. of 26th Intl. Conf. on Very Large Data Bases* (*VLDB'00*), pp 43–52, 2000.
534. K. Wang, Y. Jiang, and L. V.S. Lakshmanan. Mining Unexpected Rules by Pushing User Dynamics. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'03*), pp. 246-255, 2003.
535. K. Wang, S. Zhou, and Y. He. Growing Decision Trees on Support-Less Association Rules. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining* (*KDD'00*), pp 265–269, 2000.
536. K. Wang, C. Xu, and B. Liu. 1999. Clustering Transactions Using Large Items. In *Proc. of the Eighth Intl. Conf. on information and Knowledge Management* (*CIKM'99*). 1999, pp. 483–490.
537. W. Wang, J. Yang and R. Muntz. STING: A Statistical Information Grid Approach to Spatial Data Mining. In *Proc. of Intl. Conf. on Very Large Data Bases* (*VLDB'97*), pp. 186–195, 1997.
538. W. Wang, J. Yang, and P. S. Yu. WAR: Weighted Association Rules for Item Intensities. *Knowledge and Information Systems*, 6(2), pp. 203–229, 2004.
539. S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.
540. I. G. Webb. Discovering Associations with Numeric Variables. In *Proc. of the SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'01*): pp. 383–388, 2001.
541. Y. Weiss. Segmentation Using Eigenvectors: a Unifying View. In *Proc. IEEE Intl. Conf. on Computer Vision*, pp. 975–982, 1999.
542. J. Wiebe. Learning Subjective Adjectives from Corpora. In *Proc. of 17th National Conf. on Artificial Intelligence*, pp. 735–740, Austin, USA, 2000.
543. J. Wiebe, and R. Mihalcea. Word Sense and Subjectivity. In *Proc. of the 21st Intl. Conf. on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 1065–1072, 2006.
544. J. Wiebe, and E. Riloff: Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Proc. of CICLing*, pp. 486–497, 2005.
545. T. Wilson, J. Wiebe and R. Hwa. Recognizing Strong and Weak Opinion Clauses.

Computational Intelligence, 22(2), pp. 73-99, 2006.
547. H. Williams and J. Zobel. Compressing Integers for Fast File Access. *Computer Journal,* 42(3), pp. 193—201, 1999.
548. T. Wilson, J. Wiebe, and J. Hwa. Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In *Proc. of the National Conference on Artificial Intelligence* (*AAAI'04*), 2004.
549. I. H. Witten and E. Frank. *Data Mining*: *Practical Machine Learning Tools and Techniques with Java Implementations*. Academic Press, 2000.
550. I. H. Witten, C. G. Nevill-Manning, and S. J. Cunningham. Building a Digital Library for Computer Science Research: Technical Issues. In *Proc. of the 19th Australasian Computer Science Conf*., pp. 534–542, 1996.
551. I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes*: *Compressing and Indexing Documents and Images.* Academic Press, 1999.
552. D. Wolpert. Stacked Generalization. *Neural Networks* 5, pp. 241–259, 1992.
553. L.-S. Wu, R. Akavipat, and F. Menczer. 6S: Distributing Crawling and Searching Across Web Peers. In *Proc. of the IASTED Int. Conf. on Web Technologies, Applications, and Services*, 2005.
554. L.-S. Wu, R. Akavipat, and F. Menczer. Adaptive Query Routing in Peer Web Search. In *Proc. of the 14th Intl. World Wide Web Conf.* (*WWW'05*), pp. 1074–1075, 2005.
555. B. Wu and B. Davison. Identifying Link Farm Spam Pages. In *Proc. of the 14th Intl. World Wide Web Conf.* (*WWW'05*), pp. 820–829, May 2005.
556. B. Wu and B. Davison. Cloaking and Redirection: a Preliminary Study. In *Proc. of the 1st Intl. Workshop on Adversarial Information Retrieval on the Web*, 2005.
557. B. Wu, V. Goel, and B. D. Davison. Topical TrustRank: Using Topicality to Combat Web Spam. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
558. W. Wu, A. Doan, and C. Yu. WebIQ: Learning from the Web to Match Query Interfaces on the Deep Web. In *Proc. of International Conference on Data Engineering* (*ICDE'06*), 2006.
559. W. Wu, C. Yu, A. Doan, and W. Meng. An Interactive Clustering-Based Approach to Integrating Source Query Interfaces on the Deep Web. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'04*), pp. 95–106, 2004.
560. X. Wu, C. Zhang and S. Zhang. Mining both Positive and Negative Association Rules. In *Proc. of 19th Intl. Conf. on Machine Learning*, pp. 658–665, 2002.
561. X. Wu, L. Zhang, and Y. Yu. Exploring Social Annotations for the Semantic Web. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
562. H. Xiong, P.-N. Tan, and V. Kumar. Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution. In *Proc. of the 3rd IEEE Intl. Conf. on Data Mining* (*ICDM'03*), pp. 387-394, 2003.
563. L. Xu and D. Embley. Discovering Direct and Indirect Matches for Schema Elements. In *Proc. of Intl. Conf. on Database Systems for Advanced Applications* (*DASFAA'03*), 2003.
564. X. Xu, M. Ester, H-P. Kriegel and J. Sander. A Non-Parametric Clustering Algorithm for Knowledge Discovery in Large Spatial Databases. In *Proc. of the Intl. Conf. on Data Engineering* (*ICDE'98*), 1998.
565. X. Yan, H. Cheng, J. Han, and D. Xin: Summarizing Itemset Patterns: a Profile-Based Approach. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'05*), pp. 314-323, 2005.
566. L. Yan, R. J. Miller, L. M. Haas, and R. Fagin. Data-Driven Understanding and Refinement of Schema Mappings. In *Proc ACM SIGMOD Intl. Conf. on Management of Data*, pp. 485–496, 2001.

567. C. C. Yang and K. Y. Chan. Retrieving Multimedia Web Objects Based on Page Rank Algorithm. *WWW'05 Poster*, 2005.
568. B. Yang and H. Garcia-Molina. Improving Search in Peer-to-Peer Networks. In *Proc. of the 22nd Intl. Conf. on Distributed Computing Systems* (*ICDCS'02*), pp. 5–14. IEEE Computer Society, 2002.
569. B. Yang, and G. Jeh. Retroactive Answering of Search Queries. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
570. Q. Yang, T. Y. Li, and K. Wang. Building Association-Rule Based Sequential Classifiers for Web-Document Prediction. *Data Mining Knowledge Discovery* 8(3), pp. 253-273, 2004.
571. J. Yang, W. Wang, and P. Yu. Mining Surprising Periodic Patterns. *Data Mining and Knowledge Discovery*, 9(2), pp. 189–216, 2004.
572. W. Yang. Identifying Syntactic Differences between Two Programs. *Software Practice Experiment*, 21(7), pp. 739–755, 1991.
573. Y. Yang. An Evaluation of Statistical Approaches to Text Categorization. *Journal of Information Retrieval*, 1, pp. 67–88, 1999.
574. Y. Yang, and X. Liu. A Re-Examination of Text Categorization Methods. In *Proc. of the ACM SIGIR Intl. Conf. Research and Development in Information Retrieval* (*SIGIR'99*), pp. 42–49, 1999.
575. Y. Yang and J. P. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'97*), pp. 412–420, 1997.
576. L. Yi, B. Liu, and X. L. Li. Eliminating Noisy Information in Web Pages for Data Mining. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*), pp. 296–305, 2003.
577. J. Yi, T. Nasukawa, R. C. Bunescu, and W. Niblack. Sentiment Analyzer: Extracting Sentiments about a Given Topic Using Natural Language Processing Techniques. In *Proc. of the IEEE Conf. on Data Mining* (*ICDM'03*), pp. 427–434, 2003.
578. X. Yin, and J. Han. CPAR: Classification based on Predictive Association Rules. In *Proc. of the SIAM Intl. Conf. on Data Mining* (*SDM'03*), 2003.
579. X. Yin and W. S. Lee. Using Link Analysis to Improve Layout on Mobile Devices. In *Proc. of the 13th Intl. Conf. on World Wide Web* (*WWW'04*), pp. 338–344, 2004.
580. A. Ypma and T. Heskes. Categorization of Web Pages and User Clustering with Mixtures of Hidden Markov Models. In *Proc. of the Workshop on WebKDD-2002*, pp. 35–49, 2002.
581. C. Yu and W. Meng. *Principles of Database Query Processing for Advanced Applications.* Morgan Kaufmann, 1998.
582. H. Yu. General MC: Estimating Boundary of Positive Class from Small Positive Data. In *Proc. of the Intl. Conf. on Data Mining* (*ICDM'03*), pp. 693–696, 2003.
583. H. Yu, J. Han and K. Chang. PEBL: Positive Example Based Learning for Web Page Classification Using SVM. In *Proc. of the Knowledge Discovery and Data Mining* (*KDD'02*), pp. 239–248., 2002.
584. H. Yu, and V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proc. of Intl. Conf. on Empirical Methods for Natural Language Processing* (*EMNLP'03*), 2003.
585. P. S. Yu, X. Li, and B. Liu. Adding the Temporal Dimension to Search – A Case Study in Publication Search. In *Proc. of Web Intelligence* (*WI'05*), pp. 543–549, 2005.
586. M. Zaki. SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning*, 40, pp. 31-60, 2001.

587. M. J. Zaki, and C. C. Aggarwal. XRules: an Effective Structural Classifier for XML Data. In *Proc. of the Ninth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'03*), pp. 316–325, 2003.
588. M. J. Zaki and C. Hsiao. Charm: An Efficient Algorithm for Closed Association Rule Mining. In *Proc. of SIAM Conf. on Data Mining*, 2002.
589. M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New Algorithms for Fast Discovery of Association Rules. In *Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining* (*KDD'97*), pp 283–286, 1997.
590. M. Zaki, M. Peters, I. Assent, and T. Seidl. CLICKS: an Effective Algorithm for Mining Subspace Clusters in Categorical Datasets. In *Proc. of the SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'05*), pp. 736–742, 2005.
591. O. Zamir, and O. Etzioni. Web Document Clustering: A Feasibility Demonstration. In *Proc. of the 19th Intl. ACM SIGIR Conf. on Research and Development of Information Retrieval* (*SIGIR'98*), pp. 46–54, 1998.
592. O. Zamir, and O. Etzioni. Grouper: A Dynamic Clustering Interface to Web Search Results. In *Proc. of the 8th Intl. World Wide Web Conf.* (*WWW8*), Toronto, Canada, pp. 1361–1374, 1999.
593. H. Zeng, Q. He, Z. Chen, W. Ma, and J. Ma. Learning to Cluster Web Search Results. In *Proc. of the 27th Intl. ACM SIGIR Conf. on Research and Development in information Retrieval* (*SIGIR'04*). pp. 210–217, 2004.
594. H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral Relaxation for K-means Clustering. In *Proc. of Neural Information Processing Systems* (*NIPS'01*), pp. 1057–1064, 2001.
595. C. Zhai. Statistical Language Model for Information Retrieval. *Tutorial Notes at the Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (*SIGIR'06*), 2006.
596. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. of the Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (*SIGIR'01*), pp. 334-342, 2001.
597. C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proc. of the ACM Intl. Conf. on Information and Knowledge Management* (*CIKM'01*), 2001.
598. C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In *Proc. of the Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (*SIGIR'02*), pp. 49-56, 2002.
599. Y. Zhai and B. Liu. Extracting Web Data Using Instance-Based Learning. In *Proc. of 6th Intl. Conf. on Web Information Systems Engineering* (*WISE'05*), pp. 318–331, 2005.
600. Y. Zhai and B. Liu. Web Data Extraction based on Partial Tree Alignment. In *Proc. of the 14th Intl. World Wide Web Conference* (*WWW'05*), pp. 76–85, 2005.
601. Y. Zhai and B. Liu. Structured Data Extraction from the Web Based on Partial Tree Alignment. To appear in *IEEE Transactions on Knowledge and Data Engineering*, 2006.
602. D. Zhang, and W. S. Lee: Web Taxonomy Integration Using Support Vector Machines. In *Proc. of the 13th Intl. World Wide Web Conference* (*WWW'04*), pp. 472–481, 2004.
603. D. Zhang, and W. S. Lee. A Simple Probabilistic Approach to Learning from Positive and Unlabeled Examples. In *Proc. of the 5th Annual UK Workshop on Computational Intelligence*, 2005.
604. H. Zhang, A. Goel, R. Govindan, K. Mason and B. Van Roy. Making Eigenvector-Based Systems Robust to Collusion. In *Proc. of the 3rd Intl. Workshop on Algorithms*

and Models for the Web Graph, pp. 92–104. 2004.
605. K. Zhang, R. Statman and D. Shasha. On the Editing Distance between Unordered Labeled Trees. *Information Processing Letters* 42(3), pp. 133–139, 1992.
606. T. Zhang. The Value of Unlabeled Data for Classification Problems. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'00*), 2000.
607. T. Zhang and F. Oles. A Probability Analysis on the Value of Unlabeled Data for Classification Problems. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'00*), 2000.
608. Z. Zhang, B. He, and K. C. -C. Chang. Understanding Web Query Interfaces: Best-Effort Parsing with Hidden Syntax. In *Proc. of International Conference on Management of Data* (*SIGMOD'04*), pp. 107–118, 2004.
609. Z. Zhang, B. He, and K. C.-C. Chang. Understanding Web Query Interfaces: Best-Effort Parsing with Hidden Syntax. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'04*), pp. 107-118, 2004.
610. T. Zhang, R. Ramakrishnan and M. Linvy. BIRCH: an Efficient Data Clustering Method for Very Large Data Bases. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data* (*SIGMOD'96*), pp. 103–114, 1996.
611. Q. Zhao, S. C. H. Hoi, T-Y. Liu, S. S Bhowmick, M. R. Lyu, and W-Y. Ma. Time-Dependent Semantic Similarity Measure of Queries Using Historical Click-through Data. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
612. H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu. Fully Automatic Wrapper Generation for Search Engines. In *Proc. of the 14th Intl. World Wide Web Conference* (*WWW'05*), pp. 66–75, 2005.
613. L. Zhao and N. K. Wee. WICCAP: From Semi-structured Data to Structured Data, pp. 86–93. In *Proc. of the 11th IEEE Intl. Conf. and Workshop on the Engineering of Computer-Based Systems* (*ECBS'04*), 1994.
614. Y. Zhao and G. Karypis. Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. *Machine Learning*, 55, pp. 311–331, 2003.
615. Y. Zhao and G. Karypis. Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, 10(2), pp.141–168, 2005.
616. Z. Zheng, R. Kohavi, and L. Mason. Real World Performance of Association Rule Algorithms. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'01*), pp. 401–406, 2001.
617. N. Zhong, Y. Yao, and J. Liu (eds.) *Web Intelligence*. Springer, 2003.
618. D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha. Probabilistic Models for Discovering E-Communities. In *Proc. of the 15th Intl. Conf. on World Wide Web* (*WWW'06*), 2006.
619. X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'03*), pp. 912–919, 2003.
620. J. Zhu, Z. Nie, J-R. Wen, B. Zhang, and W-Y Ma. 2D Conditional Random Fields for Web information extraction. In *Proc. of the Intl. Conf. on Machine Learning* (*ICML'05*), pp. 1044-1051, 2005.
621. J. Zhu, Z. Nie, J-R. Wen, B. Zhang, and W.-Y. Ma. Simultaneous record detection and attribute labeling in web data extraction. In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (*KDD'06*), pp. 494–503, 2006.
622. L. Zhuang, F. Jing, X.-Yan Zhu, and L. Zhang. Movie Review Mining and Summarization. To appear in *Proc. of the ACM 15th Conf, on Information and Knowledge Management* (*CIKM'06*), 2006.

Index

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

Y

Z