Machine Learning in Automated Text Categorization: a Bibliography

Fabrizio Sebastiani
Consiglio Nazionale delle Ricerche, Italy

This is the typeset version of a bibliography on automatic text categorization (ATC) that I have created and that I am mantaining. There is a fully searchable version of it at the address

`http://liinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html`

ATC is the activity of automatically building, by means of machine learning techniques, automatic text classifiers, i.e. systems capable of assigning to a text document one or more thematic categories from a predefined set. Everyone is welcome to let me know either additional references or corrections and additions (e.g. URLs and abstracts, where they are not already present) to the existing ones.

In general, only references specific to ATC are considered pertinent to this bibliography; in particular, references that **are** considered pertinent are:

—publications that discuss novel ATC methods, novel experimentation of previously known methods, or resources for ATC experimentation;
—publications that discuss applications of ATC (e.g. automated indexing for Boolean IR systems, filtering, etc.).

References that are *not* considered pertinent are:

—publications that discuss techniques in principle useful for ATC (e.g. machine learning techniques, information retrieval techniques) but do not explicitly discuss their application to ATC;
—publications that discuss related topics sometimes confused with ATC; these include, in particular, text clustering (i.e. text classification by unsupervised learning) and text indexing;
—technical reports and workshop papers. Only papers that have been the object of formal publication (i.e. conferences and journals) are to be included in the bibliography, so as to avoid its explosion and the inclusion of material bound to obsolescence.

The searchable version of this bibliography also contains URLs from which to download on-line copies of the papers. Where possible I have included URLs with unrestricted access (e.g. home pages of authors). When such URLs were not available, sometimes a URL with restricted access (e.g. the ACM Digital Library or the IEEE Computing Society Digital Library, which are accessible to subscribers only) is indicated. When this is the case, if you know of a URL with unrestricted access from which the paper is also available, please let me know and I will substitute the link.

REFERENCES

Adam, C. K., Ng, H. T., and Chieu, H. L. 2002. Bayesian online classifiers for text classification and filtering. In *Proceedings of SIGIR-02, 25th ACM International Conference on Research and Development in Information Retrieval* (Tampere, FI, 2002). ACM Press, New York, US.

Aggarwal, C. C., Gates, S. C., and Yu, P. S. 1999. On the merits of building catego-

---

Address: Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Via Giuseppe Moruzzi, 1 - 56124 Pisa (Italy). E-mail: `fabrizio@iei.pi.cnr.it`

rization systems by supervised clustering. In *Proceedings of EDBT-00, 7th International Conference on Extending Database Technology* (Konstanz, DE, 1999), pp. 352–356. ACM Press, New York, US.

Agrawal, R., Bayardo, R. J., and Srikant, R. 2000. Athena: Mining-based interactive management of text databases. In C. Zaniolo, P. C. Lockemann, M. H. Scholl, and T. Grust Eds., *Proceedings of EDBT-00, 7th International Conference on Extending Database Technulogy* (Konstanz, DE, 2000), pp. 365–379. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1777.

Aizawa, A. 2000. The feature quantity: an information-theoretic perspective of tfidf-like measures. In N. J. Belkin, P. Ingwersen, and M.-K. Leong Eds., *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* (Athens, GR, 2000), pp. 104–111. ACM Press, New York, US.

Aizawa, A. 2001. Linguistic techniques to improve the performance of automatic text categorization. In *Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium* (Tokyo, JP, 2001), pp. 307–314.

Al-Kofahi, K., Tyrrell, A., Vachher, A., Travers, T., and Jackson, P. 2001. Combining multiple classifiers for text categorization. In H. Paques, L. Liu, and D. Grossman Eds., *Proceedings of CIKM-01, 10th ACM International Conference on Information and Knowledge Management* (Atlanta, US, 2001), pp. 97–104. ACM Press, New York, US.

Amati, G. and Crestani, F. 1999. Probabilistic learning for selective dissemination of information. *Information Processing and Management 35*, 5, 633–654.

Amati, G., Crestani, F., and Ubaldini, F. 1997. A learning system for selective dissemination of information. In M. E. Pollack Ed., *Proceedings of IJCAI-97, 15th International Joint Conference on Artificial Intelligence* (Nagoya, JP, 1997), pp. 764–769. Morgan Kaufmann Publishers, San Francisco, US.

Amati, G., Crestani, F., Ubaldini, F., and Nardis, S. D. 1997. Probabilistic learning for information filtering. In L. Devroye and C. Chrisment Eds., *Proceedings of RIAO-97, 1st International Conference "Recherche d'Information Assistee par Ordinateur"* (Montreal, CA, 1997), pp. 513–530. An extended version appears as [Amati and Crestani 1999].

Amati, G., D'Aloisi, D., Giannini, V., and Ubaldini, F. 1996. An integrated system for filtering news and managing distributed data. In *Proceedings of PAKM-96, 1st International Conference on Practical Aspects of Knowledge Management* (Basel, CH, 1996). An extended version appears as [Amati et al. 1997].

Amati, G., D'Aloisi, D., Giannini, V., and Ubaldini, F. 1997. A framework for filtering news and managing distributed data. *Journal of Universal Computer Science 3*, 8, 1007–1021.

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., and Spyropoulos, C. D. 2000. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In N. J. Belkin, P. Ingwersen, and M.-K. Leong Eds., *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* (Athens, GR, 2000), pp. 160–167. ACM Press, New York, US.

Appiani, E., Cesarini, F., Colla, A., Diligenti, M., Gori, M., Marinai, S., and Soda, G. 2001. Automatic document classification and indexing in high-volume applications. *International Journal on Document Analysis and Recognition 4*, 2, 69–83.

Apté, C., Damerau, F. J., and Weiss, S. M. 1994a. Towards language-independent automated learning of text categorization models. In W. B. Croft and C. J. van Rijsbergen Eds., *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval* (Dublin, IE, 1994), pp. 23–30. Springer Verlag, Heidelberg, DE. An extended version appears as [Apté et al. 1994b].

Apté, C. d., Damerau, F. J., and Weiss, S. M. 1994b. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems 12*, 3, 233–251.

Attardi, G., Di Marco, S., and Salvi, D. 1998. Categorization by context. *Journal of Universal Computer Science 4*, 9, 719–736.

Attardi, G., Gullí, A., and Sebastiani, F. 1999. Automatic Web page categorization

by link and context analysis. In C. Hutchison and G. Lanzarone Eds., *Proceedings of THAI-99, 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence* (Varese, IT, 1999), pp. 105–119.

Baker, L. D. and McCallum, A. K. 1998. Distributional clustering of words for text classification. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel Eds., *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval* (Melbourne, AU, 1998), pp. 96–103. ACM Press, New York, US.

Basili, R., Moschitti, A., and Pazienza, M. T. 2000. Language-sensitive text classification. In *Proceeding of RIAO-00, 6th International Conference "Recherche d'Information Assistee par Ordinateur"* (Paris, FR, 2000), pp. 331–343.

Basili, R., Moschitti, A., and Pazienza, M. T. 2001a. An hybrid approach to optimize feature selection process in text classification. In F. Esposito Ed., *Proceedings of AI*IA-01, 7th Congress of the Italian Association for Artificial Intelligence* (Bari, IT, 2001), pp. 320–325. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2175.

Basili, R., Moschitti, A., and Pazienza, M. T. 2001b. NLP-driven IR: Evaluating performances over a text classification task. In B. Nebel Ed., *Proceeding of IJCAI-01, 17th International Joint Conference on Artificial Intelligence* (Seattle, US, 2001), pp. 1286–1291.

Bayer, T., Kressel, U., Mogg-Schneider, H., and Renz, I. 1998. Categorizing paper documents. a generic system for domain and language independent text categorization. *Computer Vision and Image Understanding 70*, 3, 299–306.

Bekkerman, R., El-Yaniv, R., Tishby, N., and Winter, Y. 2001. On feature distributional clustering for text categorization. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel Eds., *Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval* (New Orleans, US, 2001), pp. 146–153. ACM Press, New York, US.

Benkhalifa, M., Bensaid, A., and Mouradi, A. 1999. Text categorization using the semi-supervised fuzzy c-means algorithm. In *Proceedings of NAFIPS-99, 18th International Conference of the North American Fuzzy Information Processing Society* (New York, US, 1999), pp. 561–565.

Benkhalifa, M., Mouradi, A., and Bouyakhf, H. 2001a. Integrating external knowledge to supplement training data in semi-supervised learning for text categorization. *Information Retrieval 4*, 2, 91–113.

Benkhalifa, M., Mouradi, A., and Bouyakhf, H. 2001b. Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization. *International Journal of Intelligent Systems 16*, 8, 929–947.

Biebricher, P., Fuhr, N., Knorz, G., Lustig, G., and Schwantner, M. 1988. The automatic indexing system AIR/PHYS. From research to application. In Y. Chiaramella Ed., *Proceedings of SIGIR-88, 11th ACM International Conference on Research and Development in Information Retrieval* (Grenoble, FR, 1988), pp. 333–342. ACM Press, New York, US. Reprinted in Karen Sparck Jones and Peter Willett (eds.), "Readings in Information Retrieval", Morgan Kaufmann, San Francisco, US, 1997, pp. 513–517.

Bloedorn, E. and Michalski, R. S. 1998. Data-driven constructive induction. *IEEE Intelligent Systems 13*, 2, 30–37.

Blosseville, M., Hebrail, G., Montell, M., and Penot, N. 1992. Automatic document classification: natural langage processing and expert system techniques used together. In N. J. Belkin, P. Ingwersen, and A. M. Pejtersen Eds., *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval* (Kobenhavn, DK, 1992), pp. 51–57. ACM Press, New York, US.

Borko, H. and Bernick, M. 1963. Automatic document classification. *Journal of the Association for Computing Machinery 10*, 2, 151–161.

Borko, H. and Bernick, M. 1964. Automatic document classification. part ii: additional experiments. *Journal of the Association for Computing Machinery 11*, 2, 138–151.

Bruckner, T. 1997. The text categorization system TEKLIS at TREC-6. In E. M. Voorhees and D. K. Harman Eds., *Proceedings of TREC-6, 6th Text Retrieval Conference* (Gaithersburg, US, 1997), pp. 619–621. National Institute of Standards and Technology, Gaithersburg, US.

Carbonell, J., Cohen, W. W., and Yang, Y. 2000. Guest editors' introduction to the special issue on machine learning and information retrieval. *Machine Learning 39*, 2/3, 99–101.

Caropreso, M. F., Matwin, S., and Sebastiani, F. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In A. G. Chin Ed., *Text Databases and Document Management: Theory and Practice*, pp. 78–102. Hershey, US: Idea Group Publishing.

Carreras, X. and Márquez, L. 2001. Boosting trees for anti-spam email filtering. In *Proceedings of RANLP-01, 4th International Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, BG, 2001).

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

Cerny, B. A., Okseniuk, A., and Lawrence, J. D. 1983. A fuzzy measure of agreement between machine and manual assignment of documents to subject categories. In R. F. Vondran, A. Caputo, C. Wasserman, and R. A. Diener Eds., *Proceedings of ASIS-83, 46th Annual Meeting of the American Society for Information Science* (Washington, US, 1983), pp. 265. American Society for Information Science, Washington, US.

Chakrabarti, S., Dom, B. E., Agrawal, R., and Raghavan, P. 1997. Using taxonomy, discriminants, and signatures for navigating in text databases. In M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld Eds., *Proceedings of VLDB-97, 23rd International Conference on Very Large Data Bases* (Athens, GR, 1997), pp. 446–455. Morgan Kaufmann Publishers, San Francisco, US. An extended version appears as [Chakrabarti et al. 1998].

Chakrabarti, S., Dom, B. E., Agrawal, R., and Raghavan, P. 1998. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *Journal of Very Large Data Bases 7*, 3, 163–178.

Chakrabarti, S., Dom, B. E., and Indyk, P. 1998. Enhanced hypertext categorization using hyperlinks. In L. M. Haas and A. Tiwary Eds., *Proceedings of SIGMOD-98, ACM International Conference on Management of Data* (Seattle, US, 1998), pp. 307–318. ACM Press, New York, US.

Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., and Kleinberg, J. 1999. Mining the Web's link structure. *Computer 32*, 8, 60–67.

Chandrinos, K. V., Androutsopoulos, I., Paliouras, G., and Spyropoulos, C. D. 2000. Automatic Web rating: Filtering obscene content on the Web. In J. L. Borbinha and T. Baker Eds., *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries* (Lisbon, PT, 2000), pp. 403–406. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1923.

Chen, C. C., Chang Chen, M., and Sun, Y. 2002. PVA: A self-adaptive personal view agent. *Journal of Intelligent Information Systems 18*, 2/3, 173–194. Special Issue on Automated Text Categorization.

Chen, H. and Dumais, S. T. 2000. Bringing order to the Web: automatically categorizing search results. In *Proceedings of CHI-00, ACM International Conference on Human Factors in Computing Systems* (Den Haag, NL, 2000), pp. 145–152. ACM Press, New York, US.

Chen, H. and Ho, T. K. 2000. Evaluation of decision forests on text categorization. In D. P. Lopresti and J. Zhou Eds., *Proceedings of the 7th SPIE Conference on Document Recognition and Retrieval* (San Jose, US, 2000), pp. 191–199. SPIE - The International Society for Optical Engineering.

Cheng, C.-h., Tang, J., Wai-chee, A., and King, I. 2001. Hierarchical classification of documents with error control. In D. Cheung, Q. Li, and G. Williams Eds., *Proceedings of PAKDD-01, 5th Pacific-Asia Conferenece on Knowledge Discovery and Data Mining* (Hong Kong, CN, 2001), pp. 433–443. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2035.

Chouchoulas, A. and Shen, Q. 2001. Rough set-aided keyword reduction for text categorization. *Applied Artificial Intelligence 15*, 9, 843–873.

Chuang, W. T., Tiyyagura, A., Yang, J., and Giuffrida, G. 2000. A fast algorithm for hierarchical text classification. In Y. Kambayashi, M. Mohania, and A. Tjoa Eds., *Proceedings of DaWaK-00, 2nd International Conference on Data Warehousing and Knowledge Discovery* (London, UK, 2000), pp. 409–418. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1874.

Ciravegna, F., Lavelli, A., Mana, N., Matiasek, J., Gilardoni, L., Mazza, S., Black, W. J., and Rinaldi, F. 1999. FACILE: Classifying texts integrating pattern matching and information extraction. In T. Dean Ed., *Proceedings of IJCAI-99, 16th International Joint Conference on Artificial Intelligence* (Stockholm, SE, 1999), pp. 890–895. Morgan Kaufmann Publishers, San Francisco, US.

Clack, C., Farringdon, J., Lidwell, P., and Yu, T. 1997. Autonomous document classification for business. In W. L. Johnson Ed., *Proceedings of the 1st International Conference on Autonomous Agents* (Marina del Rey, US, 1997), pp. 201–208. ACM Press, New York, US.

Cohen, W. W. 1995a. Learning to classify English text with ILP methods. In L. De Raedt Ed., *Advances in inductive logic programming*, pp. 124–143. Amsterdam, NL: IOS Press.

Cohen, W. W. 1995b. Text categorization and relational learning. In A. Prieditis and S. J. Russell Eds., *Proceedings of ICML-95, 12th International Conference on Machine Learning* (Lake Tahoe, US, 1995), pp. 124–132. Morgan Kaufmann Publishers, San Francisco, US.

Cohen, W. W. and Hirsh, H. 1998. Joins that generalize: text classification using Whirl. In R. Agrawal, P. E. Stolorz, and G. Piatetsky-Shapiro Eds., *Proceedings of KDD-98, 4th International Conference on Knowledge Discovery and Data Mining* (New York, US, 1998), pp. 169–173. AAAI Press, Menlo Park, US.

Cohen, W. W. and Singer, Y. 1996. Context-sensitive learning methods for text categorization. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson Eds., *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval* (Zürich, CH, 1996), pp. 307–315. ACM Press, New York, US. An extended version appears as [Cohen and Singer 1999].

Cohen, W. W. and Singer, Y. 1999. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems 17*, 2, 141–173.

Crammer, K. and Singer, Y. 2002. A new family of online algorithms for category ranking. In *Proceedings of SIGIR-02, 25th ACM International Conference on Research and Development in Information Retrieval* (Tampere, FI, 2002). ACM Press, New York, US.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A. K., Mitchell, T. M., Nigam, K., and Slattery, S. 1998. Learning to extract symbolic knowledge from the World Wide Web. In *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence* (Madison, US, 1998), pp. 509–516. AAAI Press, Menlo Park, US. An extended version appears as [Craven et al. 2000].

Craven, M., DiPasquo, D., Freitag, D., McCallum, A. K., Mitchell, T. M., Nigam, K., and Slattery, S. 2000. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence 118*, 1/2, 69–113.

Craven, M. and Slattery, S. 2001. Relational learning with statistical predicate invention: Better models for hypertext. *Machine Learning 43*, 1/2, 97–119.

Creecy, R. M., Masand, B. M., Smith, S. J., and Waltz, D. L. 1992. Trading MIPS and memory for knowledge engineering: classifying census returns on the Connection Machine. *Communications of the ACM 35*, 8, 48–63.

Cristianini, N., Shawe-Taylor, J., and Lodhi, H. 2001. Latent semantic kernels. In C. Brodley and A. Danyluk Eds., *Proceedings of ICML-01, 18th International Conference on Machine Learning* (Williams College, US, 2001), pp. 66–73. Morgan Kaufmann Publishers, San Francisco, US.

Cristianini, N., Shawe-Taylor, J., and Lodhi, H. 2002. Latent semantic kernels. *Journal of Intelligent Information Systems 18*, 2/3, 127–152. Special Issue on Automated Text Categorization.

Cunningham, S. J., Witten, I. H., and Littin, J. 1999. Applications of machine learning in information retrieval. *Annual Review of Information Science 34*, 341–384.

Dagan, I., Feldman, R., and Hirsh, H. 1996. Keyword-based browsing and analysis of large document sets. In *Proceedings of SDAIR-96, 5th Annual Symposium on Document Analysis and Information Retrieval* (Las Vegas, US, 1996), pp. 191–207.

Dagan, I., Karov, Y., and Roth, D. 1997. Mistake-driven learning in text categorization. In C. Cardie and R. Weischedel Eds., *Proceedings of EMNLP-97, 2nd Conference on Empirical Methods in Natural Language Processing* (Providence, US, 1997), pp. 55–63. Association for Computational Linguistics, Morristown, US.

D'Alessio, S., Murray, K., Schiaffino, R., and Kershenbaum, A. 1998. Category levels in hierarchical text categorization. In *Proceedings of EMNLP-98, 3rd Conference on Empirical Methods in Natural Language Processing* (Granada, ES, 1998). Association for Computational Linguistics, Morristown, US.

D'Alessio, S., Murray, K., Schiaffino, R., and Kershenbaum, A. 2000. The effect of using hierarchical classifiers in text categorization. In *Proceeding of RIAO-00, 6th International Conference "Recherche d'Information Assistee par Ordinateur"* (Paris, FR, 2000), pp. 302–313.

Damashek, M. 1995. Gauging similarity with N-grams: Language-independent categorization of text. *Science 267*, 5199, 843–848.

de Buenaga Rodríguez, M., Gómez-Hidalgo, J. M., and Díaz-Agudo, B. 1997. Using WordNet to complement training information in text categorization. In R. Milkov, N. Nicolov, and N. Nikolov Eds., *Proceedings of RANLP-97, 2nd International Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, BL, 1997).

de Lima, L. R., Laender, A. H., and Ribeiro-Neto, B. A. 1998. A hierarchical approach to the automatic categorization of medical documents. In G. Gardarin, J. C. French, N. Pissinou, K. Makki, and L. Bouganim Eds., *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management* (Bethesda, US, 1998), pp. 132–139. ACM Press, New York, US.

Denoyer, L., Zaragoza, H., and Gallinari, P. 2001. HMM-based passage models for document classification and ranking. In *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research* (Darmstadt, DE, 2001), pp. 126–135.

Diao, Y., Lu, H., and Wu, D. 2000. A comparative study of classification-based personal e-mail filtering. In T. Terano, H. Liu, and A. L. Chen Eds., *Proceedings of PAKDD-00, 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Kyoto, JP, 2000), pp. 408–419. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1805.

Díaz Esteban, A., de Buenaga Rodríguez, M., Ureña López, L. A., and García Vega, M. 1998. Integrating linguistic resources in an uniform way for text classification tasks. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada Eds., *Proceedings of LREC-98, 1st International Conference on Language Resources and Evaluation* (Grenada, ES, 1998), pp. 1197–1204.

Dörre, J., Gerstl, P., and Seiffert, R. 1999. Text mining: finding nuggets in mountains of textual data. In *Proceedings of KDD-99, 5th ACM International Conference on Knowledge Discovery and Data Mining* (San Diego, US, 1999), pp. 398–401. ACM Press, New York, US.

Doyle, L. B. 1965. Is automatic classification a reasonable application of statistical analysis of text? *Journal of the ACM 12*, 4, 473–489.

Drucker, H., Vapnik, V., and Wu, D. 1999. Automatic text categorization and its applications to text retrieval. *IEEE Transactions on Neural Networks 10*, 5, 1048–1054.

Dumais, S. T. and Chen, H. 2000. Hierarchical classification of Web content. In N. J. Belkin, P. Ingwersen, and M.-K. Leong Eds., *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* (Athens, GR, 2000), pp. 256–263. ACM Press, New York, US.

Dumais, S. T., Platt, J., Heckerman, D., and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In G. Gardarin, J. C. French, N. Pissinou, K. Makki, and L. Bouganim Eds., *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management* (Bethesda, US, 1998), pp. 148–155. ACM Press, New York, US.

El-Yaniv, R. and Souroujon, O. 2001. Iterative double clustering for unsupervised and semi-supervised learning. In L. D. Raedt and P. A. Flach Eds., *Proceedings of ECML-01, 12th European Conference on Machine Learning* (Freiburg, DE, 2001), pp. 121–132. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2167.

Escudero, G., Màrquez, L., and Rigau, G. 2000. Boosting applied to word sense disambiguation. In R. L. de Mántaras and E. Plaza Eds., *Proceedings of ECML-00, 11th European Conference on Machine Learning* (Barcelona, ES, 2000), pp. 129–141. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1810.

Fangmeyer, H. and Lustig, G. 1968. The EURATOM automatic indexing project. In *Proceedings of the IFIP Congress (Booklet J)* (Edinburgh, UK, 1968), pp. 66–70.

Fangmeyer, H. and Lustig, G. 1970. Experiments with the CETIS automated indexing system. In *Proceedings of the Symposium on the Handling of Nuclear Information* (1970), pp. 557–567. International Atomic Energy Agency.

Ferilli, S., Fanizzi, N., and Semeraro, G. 2001. Learning logic models for automated text categorization. In F. Esposito Ed., *Proceedings of AI*IA-01, 7th Congress of the Italian Association for Artificial Intelligence* (Bari, IT, 2001), pp. 81–86. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2175.

Field, B. 1975. Towards automatic indexing: automatic assignment of controlled-language indexing and classification from free indexing. *Journal of Documentation 31*, 4, 246–265.

Finn, A., Kushmerick, N., and Smyth, B. 2002. Genre classification and domain transfer for information filtering. In F. Crestani, M. Girolami, and C. J. van Rijsbergen Eds., *Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research* (Glasgow, UK, 2002), pp. 353–362. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2291.

Forsyth, R. S. 1999. New directions in text categorization. In A. Gammerman Ed., *Causal models and intelligent data management*, pp. 151–185. Heidelberg, DE: Springer Verlag.

Frank, E., Chui, C., and Witten, I. H. 2000. Text categorization using compression models. In J. A. Storer and M. Cohn Eds., *Proceedings of DCC-00, IEEE Data Compression Conference* (Snowbird, US, 2000), pp. 200–209. IEEE Computer Society Press, Los Alamitos, US.

Frasconi, P., Soda, G., and Vullo, A. 2001. Text categorization for multi-page documents: A hybrid naive Bayes HMM approach. In *Proceedings of JCDL, 1st ACM-IEEE Joint Conference on Digital Libraries* (Roanoke, US, 2001), pp. 11–20. IEEE Computer Society Press, Los Alamitos, US.

Frasconi, P., Soda, G., and Vullo, A. 2002. Text categorization for multi-page documents: A hybrid naive Bayes HMM approach. *Journal of Intelligent Information Systems 18*, 2/3, 195–217. Special Issue on Automated Text Categorization.

Frommholz, I. 2001. Categorizing Web documents in hierarchical catalogues. In *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research* (Darmstadt, DE, 2001).

Fuhr, N. 1985. A probabilistic model of dictionary-based automatic indexing. In *Proceedings of RIAO-85, 1st International Conference "Recherche d'Information Assistee par Ordinateur"* (Grenoble, FR, 1985), pp. 207–216.

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

Fuhr, N. and Knorz, G. 1984. Retrieval test evaluation of a rule-based automated indexing (AIR/PHYS). In C. J. van Rijsbergen Ed., *Proceedings of SIGIR-84, 7th ACM International Conference on Research and Development in Information Retrieval* (Cambridge, UK, 1984), pp. 391–408. Cambridge University Press.

Fuhr, N. and Pfeifer, U. 1991. Combining model-oriented and description-oriented approaches for probabilistic indexing. In A. Bookstein, Y. Chiaramella, G. Salton, and V. V. Raghavan Eds., *Proceedings of SIGIR-91, 14th ACM International Conference on Research and Development in Information Retrieval* (Chicago, US, 1991), pp. 46–56. ACM Press, New York, US. An extended version appears as [Fuhr and Pfeifer 1994].

Fuhr, N. and Pfeifer, U. 1994. Probabilistic information retrieval as combination of abstraction inductive learning and probabilistic assumptions. *ACM Transactions on Information Systems 12*, 1, 92–115.

Fürnkranz, J. 1999. Exploiting structural information for text classification on the WWW. In D. J. Hand, J. N. Kok, and M. R. Berthold Eds., *Proceedings of IDA-99, 3rd Symposium on Intelligent Data Analysis* (Amsterdam, NL, 1999), pp. 487–497. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1642.

Galavotti, L., Sebastiani, F., and Simi, M. 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. In J. L. Borbinha and T. Baker Eds., *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries* (Lisbon, PT, 2000), pp. 59–68. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1923.

Gale, W. A., Church, K. W., and Yarowsky, D. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities 26*, 5, 415–439.

Gaussier, É., Goutte, C., Popat, K., and Chen, F. 2002. A hierarchical model for clustering and categorising documents. In F. Crestani, M. Girolami, and C. J. van Rijsbergen Eds., *Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research* (Glasgow, UK, 2002), pp. 229–247. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2291.

Gentili, G., Marinilli, M., Micarelli, A., and Sciarrone, F. 2001. Text categorization in an intelligent agent for filtering information on the Web. *International Journal of Pattern Recognition and Artificial Intelligence 15*, 3, 527–549.

Geutner, P., Bodenhausen, U., and Waibel, A. 1993. Flexibility through incremental learning: Neural networks for text categorization. In *Proceedings of WCNN-93, World Congress on Neural Networks* (Portland, US, 1993), pp. 24–27.

Ghani, R. 2000. Using error-correcting codes for text classification. In P. Langley Ed., *Proceedings of ICML-00, 17th International Conference on Machine Learning* (Stanford, US, 2000), pp. 303–310. Morgan Kaufmann Publishers, San Francisco, US.

Ghani, R. 2001. Combining labeled and unlabeled data for text classification with a large number of categories. In N. Cercone, T. Y. Lin, and X. Wu Eds., *Proceedings of the IEEE International Conference on Data Mining* (San Jose, US, 2001), pp. 597–598. IEEE Computer Society, Los Alamitos, US.

Ghani, R., Slattery, S., and Yang, Y. 2001. Hypertext categorization using hyperlink patterns and meta data. In C. Brodley and A. Danyluk Eds., *Proceedings of ICML-01, 18th International Conference on Machine Learning* (Williams College, US, 2001), pp. 178–185. Morgan Kaufmann Publishers, San Francisco, US.

Glover, E. J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D. M., and Flake, G. W. 2002. Using Web structure for classifying and describing Web pages. In *Proceedings of WWW-02, International Conference on the World Wide Web* (Honolulu, US, 2002). Forthcoming.

Goldberg, J. L. 1995. CDM: an approach to learning in text categorization. In *Proceedings of ICTAI-95, 7th International Conference on Tools with Artificial Intelligence* (Herndon, US, 1995), pp. 258–265. IEEE Computer Society Press, Los Alamitos, US. An extended version appears as [Goldberg 1996].

Goldberg, J. L. 1996. CDM: an approach to learning in text categorization. *International Journal on Artificial Intelligence Tools 5*, 1/2, 229–253.

Gómez-Hidalgo, J. M. 2002. Evaluating cost-sensitive unsolicited bulk email categorization. In *Proceedings of SAC-02, 17th ACM Symposium on Applied Computing* (Madrid, ES, 2002).

Gómez-Hidalgo, J. M., de Buenaga Rodríguez, J. M., Urea López, L. A., Martín Valdivia, M. T., and García Vega, M. 2002. Integrating lexical knowledge in learning-based text categorization. In *Proceedings of JADT-02, 6th International Conference on the Statistical Analysis of Textual Data* (St-Malo, FR, 2002).

Goodman, M. 1990. Prism: a case-based telex classifier. In A. Rappaport and R. Smith Eds., *Proceedings of IAAI-90, 2nd Conference on Innovative Applications of Artificial Intelligence* (1990), pp. 25–37. AAAI Press, Menlo Park, US.

Gövert, N., Lalmas, M., and Fuhr, N. 1999. A probabilistic description-oriented approach for categorising Web documents. In *Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management* (Kansas City, US, 1999), pp. 475–482. ACM Press, New York, US.

Gray, W. A. and Harley, A. J. 1971. Computer-assisted indexing. *Information Storage and Retrieval 7*, 4, 167–174.

Guthrie, L., Guthrie, J. A., and Leistensnider, J. 1999. Document classification and routing. In T. Strzalkowski Ed., *Natural language information retrieval*, pp. 289–310. Dordrecht, NL: Kluwer Academic Publishers.

Guthrie, L., Walker, E., and Guthrie, J. A. 1994. Document classification by machine: theory and practice. In *Proceedings of COLING-94, 15th International Conference on Computational Linguistics* (Kyoto, JP, 1994), pp. 1059–1063.

Hadjarian, A., Bala, J., and Pachowicz, P. 2001. Text categorization through multistrategy learning and visualization. In A. Gelbukh Ed., *Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing* (Mexico City, ME, 2001), pp. 423–436. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes for Computer Science" series, number 2004.

Hamill, K. A. and Zamora, A. 1978. An automatic document classification system using pattern recognition techniques. In E. H. Brenner Ed., *Proceedings of ASIS-78, 41st Annual Meeting of the American Society for Information Science* (New York, US, 1978), pp. 152–155. American Society for Information Science, Washington, US.

Hamill, K. A. and Zamora, A. 1980. The use of titles for automatic document classification. *Journal of the American Society for Information Science 33*, 6, 396–402.

Han, E.-H., Karypis, G., and Kumar, V. 2001. Text categorization using weight-adjusted k-nearest neighbor classification. In D. Cheung, Q. Li, and G. Williams Eds., *Proceedings of PAKDD-01, 5th Pacific-Asia Conferenece on Knowledge Discovery and Data Mining* (Hong Kong, CN, 2001), pp. 53–65. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2035.

Hayes, P. J., Andersen, P. M., Nirenburg, I. B., and Schmandt, L. M. 1990. Tcs: a shell for content-based text categorization. In *Proceedings of CAIA-90, 6th IEEE Conference on Artificial Intelligence Applications* (Santa Barbara, US, 1990), pp. 320–326. IEEE Computer Society Press, Los Alamitos, US.

Hayes, P. J., Knecht, L. E., and Cellio, M. J. 1988. A news story categorization system. In *Proceedings of ANLP-88, 2nd Conference on Applied Natural Language Processing*

(Austin, US, 1988), pp. 9–17. Association for Computational Linguistics, Morristown, US. Reprinted in Karen Sparck Jones and Peter Willett (eds.), "Readings in Information Retrieval", Morgan Kaufmann, San Francisco, US, 1997, pp. 518–526.

Hayes, P. J. and Weinstein, S. P. 1990. Construe/Tis: a system for content-based indexing of a database of news stories. In A. Rappaport and R. Smith Eds., *Proceedings of IAAI-90, 2nd Conference on Innovative Applications of Artificial Intelligence* (1990), pp. 49–66. AAAI Press, Menlo Park, US.

Heaps, H. 1973. A theory of relevance for automatic document classification. *Information and Control 22*, 3, 268–278.

Hearst, M. A. 1991. Noun homograph disambiguation using local context in large corpora. In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary* (Oxford, UK, 1991), pp. 1–22.

Hearst, M. A. and Hirsh, H. Eds. 1996. *Machine Learning in Information Access. Papers from the 1996 AAAI Spring Symposium* (Stanford, US, 1996). Available as Technical Report SS-96-05.

Hersh, W., Buckley, C., Leone, T., and Hickman, D. 1994. Ohsumed: an interactive retrieval evaluation and new large text collection for research. In W. B. Croft and C. J. van Rijsbergen Eds., *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval* (Dublin, IE, 1994), pp. 192–201. Springer Verlag, Heidelberg, DE.

Hoashi, K., Matsumoto, K., Inoue, N., and Hashimoto, K. 2000. Document filtering methods using non-relevant information profile. In N. J. Belkin, P. Ingwersen, and M.-K. Leong Eds., *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* (Athens, GR, 2000), pp. 176–183. ACM Press, New York, US.

Hoch, R. 1994. Using IR techniques for text classification in document analysis. In W. B. Croft and C. J. van Rijsbergen Eds., *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval* (Dublin, IE, 1994), pp. 31–40. Springer Verlag, Heidelberg, DE.

Hoyle, W. 1973. Automatic indexing and generation of classification by algorithm. *Information Storage and Retrieval 9*, 4, 233–242.

Hsu, W.-L. and Lang, S.-D. 1999a. Classification algorithms for NETNEWS articles. In *Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management* (Kansas City, US, 1999), pp. 114–121. ACM Press, New York, US.

Hsu, W.-L. and Lang, S.-D. 1999b. Feature reduction and database maintenance in NETNEWS classification. In *Proceedings of IDEAS-99, 1999 International Database Engineering and Applications Symposium* (Montreal, CA, 1999), pp. 137–144. IEEE Computer Society Press, Los Alamitos, US.

Huffman, S. 1995. Acquaintance: Language-independent document categorization by n-grams. In D. K. Harman and E. M. Voorhees Eds., *Proceedings of TREC-4, 4th Text Retrieval Conference* (Gaithersburg, US, 1995), pp. 359–371. National Institute of Standards and Technology, Gaithersburg, US.

Huffman, S. and Damashek, M. 1994. Acquaintance: A novel vector-space n-gram technique for document categorization. In D. K. Harman Ed., *Proceedings of TREC-3, 3rd Text Retrieval Conference* (Gaithersburg, US, 1994), pp. 305–310. National Institute of Standards and Technology, Gaithersburg, US.

Hull, D. A. 1994. Improving text retrieval for the routing problem using latent semantic indexing. In W. B. Croft and C. J. van Rijsbergen Eds., *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval* (Dublin, IE, 1994), pp. 282–289. Springer Verlag, Heidelberg, DE.

Hull, D. A. 1998. The TREC-7 filtering track: description and analysis. In E. M. Voorhees and D. K. Harman Eds., *Proceedings of TREC-7, 7th Text Retrieval Conference* (Gaithersburg, US, 1998), pp. 33–56. National Institute of Standards and Technology, Gaithersburg, US.

Hull, D. A., Pedersen, J. O., and Schütze, H. 1996. Method combination for document filtering. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson Eds., *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval* (Zürich, CH, 1996), pp. 279–288. ACM Press, New York, US.

Ittner, D. J., Lewis, D. D., and Ahn, D. D. 1995. Text categorization of low quality images. In *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval* (Las Vegas, US, 1995), pp. 301–315.

Iwayama, M. and Tokunaga, T. 1994. A probabilistic model for text categorization: Based on a single random variable with multiple values. In *Proceedings of ANLP-94, 4th Conference on Applied Natural Language Processing* (Stuttgart, DE, 1994), pp. 162–167. Association for Computational Linguistics, Morristown, US.

Iwayama, M. and Tokunaga, T. 1995a. Cluster-based text categorization: a comparison of category search strategies. In E. A. Fox, P. Ingwersen, and R. Fidel Eds., *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval* (Seattle, US, 1995), pp. 273–281. ACM Press, New York, US.

Iwayama, M. and Tokunaga, T. 1995b. Hierarchical Bayesian clustering for automatic text classification. In C. E. Mellish Ed., *Proceedings of IJCAI-95, 14th International Joint Conference on Artificial Intelligence* (Montreal, CA, 1995), pp. 1322–1327. Morgan Kaufmann Publishers, San Francisco, US.

Iwazume, M., Takeda, H., and Nishida, T. 1996. Ontology-based information gathering and text categorization from the Internet. In *Proceedings of IEA/AIE-96, 9th International Conference in Industrial and Engineering Applications of Artificial Intelligence and Expert Systems* (Fukuoka, JP, 1996), pp. 305–314.

Iyer, R. D., Lewis, D. D., Schapire, R. E., Singer, Y., and Singhal, A. 2000. Boosting for document routing. In A. Agah, J. Callan, and E. Rundensteiner Eds., *Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management* (McLean, US, 2000), pp. 70–77. ACM Press, New York, US.

Jacobs, P. S. 1992. Joining statistics with NLP for text categorization. In M. Bates and O. Stock Eds., *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing* (Trento, IT, 1992), pp. 178–185. Association for Computational Linguistics, Morristown, US.

Jacobs, P. S. 1993. Using statistical methods to improve knowledge-based news categorization. *IEEE Expert 8*, 2, 13–23.

Jo, T. C. 1999a. News article classification based on categorical points from keywords in backdata. In M. Mohammadian Ed., *Computational Intelligence for Modelling, Control and Automation*, pp. 211–214. Amsterdam, NL: IOS Press.

Jo, T. C. 1999b. News articles classification based on representative keywords of categories. In M. Mohammadian Ed., *Computational Intelligence for Modelling, Control and Automation*, pp. 194–198. Amsterdam, NL: IOS Press.

Jo, T. C. 1999c. Text categorization with the concept of fuzzy set of informative keywords. In *Proceedings of FUZZ-IEEE'99, IEEE International Conference on Fuzzy Systems* (Seoul, KR, 1999), pp. 609–614. IEEE Computer Society Press, Los Alamitos, US.

Joachims, T. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In D. H. Fisher Ed., *Proceedings of ICML-97, 14th International Conference on Machine Learning* (Nashville, US, 1997), pp. 143–151. Morgan Kaufmann Publishers, San Francisco, US.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol Eds., *Proceedings of ECML-98, 10th European Conference on Machine Learning* (Chemnitz, DE, 1998), pp. 137–142. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1398.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In I. Bratko and S. Dzeroski Eds., *Proceedings of ICML-99, 16th International Conference on Machine Learning* (Bled, SL, 1999), pp. 200–209. Morgan Kaufmann Publishers, San Francisco, US.

Joachims, T. 2000. Estimating the generalization performance of a SVM efficiently. In P. Langley Ed., *Proceedings of ICML-00, 17th International Conference on Machine Learning* (Stanford, US, 2000), pp. 431–438. Morgan Kaufmann Publishers, San Francisco, US.

Joachims, T. 2001. A statistical learning model of text classification with support vector machines. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel Eds., *Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval* (New Orleans, US, 2001), pp. 128–136. ACM Press, New York, US.

Joachims, T. 2002. *Learning Text Classifiers with Support Vector Machines*. Kluwer Academic Publishers, Dordrecht, NL. Forthcoming.

Joachims, T., Cristianini, N., and Shawe-Taylor, J. 2001. Composite kernels for hypertext categorisation. In C. Brodley and A. Danyluk Eds., *Proceedings of ICML-01, 18th International Conference on Machine Learning* (Williams College, US, 2001), pp. 250–257. Morgan Kaufmann Publishers, San Francisco, US.

Joachims, T., Freitag, D., and Mitchell, T. M. 1997. WebWatcher: a tour guide for the Word Wide Web. In M. E. Pollack Ed., *Proceedings of IJCAI-97, 15th International Joint Conference on Artificial Intelligence* (Nagoya, JP, 1997), pp. 770–775. Morgan Kaufmann Publishers, San Francisco, US.

Joachims, T. and Sebastiani, F. 2002. Guest editors' introduction to the special issue on automated text categorization. *Journal of Intelligent Information Systems 18*, 2/3, 103–105. Special Issue on Automated Text Categorization.

Junker, M. and Abecker, A. 1997. Exploiting thesaurus knowledge in rule induction for text classification. In R. Milkov, N. Nicolov, and N. Nikolov Eds., *Proceedings of RANLP-97, 2nd International Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, BL, 1997), pp. 202–207.

Junker, M. and Dengel, A. 2001. Preventing overfitting in learning text patterns for document categorization. In S. Singh, N. A. Murshed, and W. Kropatsch Eds., *Proceedings of ICAPR-01, 2nd International Conference on Advances in Pattern Recognition* (Rio de Janeiro, BR, 2001), pp. 137–146. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2013.

Junker, M. and Hoch, R. 1998. An experimental evaluation of OCR text representations for learning document classifiers. *International Journal on Document Analysis and Recognition 1*, 2, 116–122.

Junker, M., Sintek, M., and Rinck, M. 2000. Learning for text categorization and information extraction with ilp. In J. Cussens and S. Dzeroski Eds., *Proceedings of the 1st Workshop on Learning Language in Logic* (Bled, SL, 2000), pp. 247–258. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1925.

Kaban, A. and Girolami, M. 2002. A dynamic probabilistic model to visualise topic evolution in text streams. *Journal of Intelligent Information Systems 18*, 2/3, 107–125. Special Issue on Automated Text Categorization.

Kar, G. and White, L. J. 1978. A distance measure for automated document classification by sequential analysis. *Information Processing and Management 14*, 2, 57–69.

Karypis, G. and Han, E.-H. 2000. Fast supervised dimensionality reduction algorithm with applications to document categorization and retrieval. In A. Agah, J. Callan, and E. Rundensteiner Eds., *Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management* (McLean, US, 2000), pp. 12–19. ACM Press, New York, US.

Kawatani, T. 2002. Topic difference factor extraction between two document sets and its application to text categorization. In *Proceedings of SIGIR-02, 25th ACM International Conference on Research and Development in Information Retrieval* (Tampere, FI, 2002). ACM Press, New York, US.

Kessler, B., Nunberg, G., and Schütze, H. 1997. Automatic detection of text genre. In P. R. Cohen and W. Wahlster Eds., *Proceedings of ACL-97, 35th Annual Meeting*

of the Association for Computational Linguistics (Madrid, ES, 1997), pp. 32–38. Morgan Kaufmann Publishers, San Francisco, US.

Kim, Y.-H., Hahn, S.-Y., and Zhang, B.-T. 2000. Text filtering by boosting naive Bayes classifiers. In N. J. Belkin, P. Ingwersen, and M.-K. Leong Eds., *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* (Athens, GR, 2000), pp. 168–75. ACM Press, New York, US.

Kindermann, J., Diederich, J., Leopold, E., and Paass, G. 2002. Identifying the author of a text with support vector machines. *Applied Intelligence*.

Kindermann, J., Paass, G., and Leopold, E. 2001. Error correcting codes with optimized Kullback-Leibler distances for text categorization. In L. D. Raedt and A. Siebes Eds., *Proceedings of ECML-01, 12th European Conference on Machine Learning* (Freiburg, DE, 2001), pp. 266–275. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2168.

Klas, C.-P. and Fuhr, N. 2000. A new effective approach for categorizing Web documents. In *Proceedings of BCSIRSG-00, the 22nd Annual Colloquium of the British Computer Society Information Retrieval Specialist Group* (Cambridge, UK, 2000).

Klingbiel, P. H. 1973a. Machine-aided indexing of technical literature. *Information Storage and Retrieval 9*, 2, 79–84.

Klingbiel, P. H. 1973b. A technique for machine-aided indexing. *Information Storage and Retrieval 9*, 9, 477–494.

Klinkenberg, R. and Joachims, T. 2000. Detecting concept drift with support vector machines. In P. Langley Ed., *Proceedings of ICML-00, 17th International Conference on Machine Learning* (Stanford, US, 2000), pp. 487–494. Morgan Kaufmann Publishers, San Francisco, US.

Knorz, G. 1982. A decision theory approach to optimal automated indexing. In G. Salton and H.-J. Schneider Eds., *Proceedings of SIGIR-82, 5th ACM International Conference on Research and Development in Information Retrieval* (Berlin, DE, 1982), pp. 174–193. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 146.

Ko, Y. and Seo, J. 2000. Automatic text categorization by unsupervised learning. In *Proceedings of COLING-00, the 18th International Conference on Computational Linguistics* (Saarbrücken, DE, 2000).

Koller, D. and Sahami, M. 1997. Hierarchically classifying documents using very few words. In D. H. Fisher Ed., *Proceedings of ICML-97, 14th International Conference on Machine Learning* (Nashville, US, 1997), pp. 170–178. Morgan Kaufmann Publishers, San Francisco, US.

Kongovi, M., Guzman, J. C., and Dasigi, V. 2002. Text categorization: An experiment using phrases. In F. Crestani, M. Girolami, and C. J. van Rijsbergen Eds., *Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research* (Glasgow, UK, 2002), pp. 213–228. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2291.

Kosmynin, A. and Davidson, I. 1996. Using background contextual knowledge for documents representation. In C. K. Nicholas and D. Wood Eds., *Proceedings of PODP-96, 3rd International Workshop on Principles of Document Processing* (Palo Alto, CA, 1996), pp. 123–133. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1293.

Kwok, J. T. 1998. Automated text categorization using support vector machine. In *Proceedings of ICONIP'98, 5th International Conference on Neural Information Processing* (Kitakyushu, JP, 1998), pp. 347–351.

Kwon, O.-W., Jung, S.-H., Lee, J.-H., and Lee, G. 1999. Evaluation of category features and text structural information on a text categorization using memory based reasoning. In *Proceedings of ICCPOL-99, 18th International Conference on Computer Processing of Oriental Languages* (Tokushima, JP, 1999), pp. 153–158.

Labrou, Y. and Finin, T. 1999. Yahoo! as an ontology: using Yahoo! categories to

describe documents. In *Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management* (Kansas City, US, 1999), pp. 180–187. ACM Press, New York, US.

Lai, K.-Y. and Lam, W. 2001. Meta-learning models for automatic textual document categorization. In D. Cheung, Q. Li, and G. Williams Eds., *Proceedings of PAKDD-01, 5th Pacific-Asia Conferenece on Knowledge Discovery and Data Mining* (Hong Kong, CN, 2001), pp. 78–89. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2035.

Lam, S. L. and Lee, D. L. 1999. Feature reduction for neural network based text categorization. In A. L. Chen and F. H. Lochovsky Eds., *Proceedings of DASFAA-99, 6th IEEE International Conference on Database Advanced Systems for Advanced Application* (Hsinchu, TW, 1999), pp. 195–202. IEEE Computer Society Press, Los Alamitos, US.

Lam, W. and Ho, C. Y. 1998. Using a generalized instance set for automatic text categorization. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel Eds., *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval* (Melbourne, AU, 1998), pp. 81–89. ACM Press, New York, US.

Lam, W. and Lai, K.-Y. 2001. A meta-learning approach for text categorization. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel Eds., *Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval* (New Orleans, US, 2001), pp. 303–309. ACM Press, New York, US.

Lam, W., Low, K. F., and Ho, C. Y. 1997. Using a Bayesian network induction approach for text categorization. In M. E. Pollack Ed., *Proceedings of IJCAI-97, 15th International Joint Conference on Artificial Intelligence* (Nagoya, JP, 1997), pp. 745–750. Morgan Kaufmann Publishers, San Francisco, US.

Lam, W., Ruiz, M. E., and Srinivasan, P. 1999. Automatic text categorization and its applications to text retrieval. *IEEE Transactions on Knowledge and Data Engineering 11*, 6, 865–879.

Lang, K. 1995. NewsWeeder: learning to filter netnews. In A. Prieditis and S. J. Russell Eds., *Proceedings of ICML-95, 12th International Conference on Machine Learning* (Lake Tahoe, US, 1995), pp. 331–339. Morgan Kaufmann Publishers, San Francisco, US.

Larkey, L. S. 1998. Automatic essay grading using text categorization techniques. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel Eds., *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval* (Melbourne, AU, 1998), pp. 90–95. ACM Press, New York, US.

Larkey, L. S. 1999. A patent search and classification system. In E. A. Fox and N. Rowe Eds., *Proceedings of DL-99, 4th ACM Conference on Digital Libraries* (Berkeley, US, 1999), pp. 179–187. ACM Press, New York, US.

Larkey, L. S. and Croft, W. B. 1996. Combining classifiers in text categorization. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson Eds., *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval* (Zürich, CH, 1996), pp. 289–297. ACM Press, New York, US.

Lee, Y.-B. and Myaeng, S. H. 2002. Text genre classification with genre-revealing and subject-revealing features,. In *Proceedings of SIGIR-02, 25th ACM International Conference on Research and Development in Information Retrieval* (Tampere, FI, 2002). ACM Press, New York, US.

Lehnert, W., Soderland, S., Aronow, D., Feng, F., and Shmueli, A. 1994. Inductive text classification for medical applications. *Journal of Experimental and Theoretical Artificial Intelligence 7*, 1, 49–80.

Leopold, E. and Kindermann, J. 2002. Text categorization with support vector machines: How to represent texts in input space? *Machine Learning 46*, 1/3, 423–444.

Leung, C.-H. and Kan, W.-K. 1997. A statistical learning approach to automatic indexing of controlled index terms. *Journal of the American Society for Information Science 48*, 1, 55–67.

Lewis, D. D. 1991. Data extraction as text categorization: An experiment with the MUC-3 corpus. In *Proceedings of MUC-3, 3rd Message Understanding Conference* (San Diego, US, 1991), pp. 245–255. Morgan Kaufmann Publishers, San Francisco, US.

Lewis, D. D. 1992a. An evaluation of phrasal and clustered representations on a text categorization task. In N. J. Belkin, P. Ingwersen, and A. M. Pejtersen Eds., *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval* (Kobenhavn, DK, 1992), pp. 37–50. ACM Press, New York, US.

Lewis, D. D. 1992b. *Representation and learning in information retrieval*. Ph. D. thesis, Department of Computer Science, University of Massachusetts, Amherst, US.

Lewis, D. D. 1995a. Evaluating and optmizing autonomous text classification systems. In E. A. Fox, P. Ingwersen, and R. Fidel Eds., *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval* (Seattle, US, 1995), pp. 246–254. ACM Press, New York, US.

Lewis, D. D. 1995b. A sequential algorithm for training text classifiers: corrigendum and additional data. *SIGIR Forum 29*, 2, 13–19.

Lewis, D. D. 1995c. The TREC-4 filtering track: description and analysis. In D. K. Harman and E. M. Voorhees Eds., *Proceedings of TREC-4, 4th Text Retrieval Conference* (Gaithersburg, US, 1995), pp. 165–180. National Institute of Standards and Technology, Gaithersburg, US.

Lewis, D. D. 1997. Reuters-21578 text categorization test collection. Distribution 1.0. Available as `http://www.research.att.com/~lewis/reuters21578/README.txt`.

Lewis, D. D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec and C. Rouveirol Eds., *Proceedings of ECML-98, 10th European Conference on Machine Learning* (Chemnitz, DE, 1998), pp. 4–15. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1398.

Lewis, D. D. 2000. Machine learning for text categorization: background and characteristics. In M. E. Williams Ed., *Proceedings of the 21st Annual National Online Meeting* (New York, US, 2000), pp. 221–226. Information Today, Medford, USA.

Lewis, D. D. and Catlett, J. 1994. Heterogeneous uncertainty sampling for supervised learning. In W. W. Cohen and H. Hirsh Eds., *Proceedings of ICML-94, 11th International Conference on Machine Learning* (New Brunswick, US, 1994), pp. 148–156. Morgan Kaufmann Publishers, San Francisco, US.

Lewis, D. D. and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In W. B. Croft and C. J. van Rijsbergen Eds., *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval* (Dublin, IE, 1994), pp. 3–12. Springer Verlag, Heidelberg, DE. See also [Lewis 1995b].

Lewis, D. D. and Hayes, P. J. 1994. Guest editors' introduction to the special issue on text categorization. *ACM Transactions on Information Systems 12*, 3, 231.

Lewis, D. D. and Ringuette, M. 1994. A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval* (Las Vegas, US, 1994), pp. 81–93.

Lewis, D. D., Schapire, R. E., Callan, J. P., and Papka, R. 1996. Training algorithms for linear text classifiers. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson Eds., *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval* (Zürich, CH, 1996), pp. 298–306. ACM Press, New York, US.

Lewis, D. D., Stern, D. L., and Singhal, A. 1999. Attics: a software platform for on-line text classification. In M. A. Hearst, F. Gey, and R. Tong Eds., *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval* (Berkeley, US, 1999), pp. 267–268. ACM Press, New York, US.

Li, H. and Yamanishi, K. 1997. Document classification using a finite mixture model. In P. R. Cohen and W. Wahlster Eds., *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics* (Madrid, ES, 1997), pp. 39–47. Morgan

Kaufmann Publishers, San Francisco, US.

Li, H. and Yamanishi, K. 1999. Text classification using ESC-based stochastic decision lists. In *Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management* (Kansas City, US, 1999), pp. 122–130. ACM Press, New York, US.

Li, H. and Yamanishi, K. 2002. Text classification using ESC-based stochastic decision lists. *Information Processing and Management 38*, 3, 343–361.

Li, W., Lee, B., Krausz, F., and Sahin, K. 1991. Text classification by a neural network. In *Proceedings of the 23rd Annual Summer Computer Simulation Conference* (Baltimore, US, 1991), pp. 313–318.

Li, Y. H. and Jain, A. K. 1998. Classification of text documents. *The Computer Journal 41*, 8, 537–546.

Liddy, E. D., Paik, W., and Yu, E. S. 1994. Text categorization for multiple users based on semantic features from a machine-readable dictionary. *ACM Transactions on Information Systems 12*, 3, 278–295.

Liere, R. and Tadepalli, P. 1997. Active learning with committees for text categorization. In *Proceedings of AAAI-97, 14th Conference of the American Association for Artificial Intelligence* (Providence, US, 1997), pp. 591–596. AAAI Press, Menlo Park, US.

Liere, R. and Tadepalli, P. 1998. Active learning with committees: Preliminary results in comparing Winnow and Perceptron in text categorization. In *Proceedings of CONALD-98, 1st Conference on Automated Learning and Discovery* (Pittsburgh, US, 1998). AAAI Press, Menlo Park, US.

Lim, J. H. 1999. Learnable visual keywords for image classification. In E. A. Fox and N. Rowe Eds., *Proceedings of DL-99, 4th ACM Conference on Digital Libraries* (Berkeley, US, 1999), pp. 139–145. ACM Press, New York, US.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. 2002. Text classification using string kernels. *Journal of Machine Learning Research 2*, 419–444.

Lodhi, H., Shawe-Taylor, J., Cristianini, N., and Watkins, C. J. 2001. Discrete kernels for text categorisation. In T. K. Leen, T. G. Dietterich, and V. Tresp Eds., *Advances in Neural Information Processing Systems*, Volume 13, pp. 563–569. MIT Press, Cambridge, MA.

Macskassy, S. A., Hirsh, H., Banerjee, A., and Dayanik, A. A. 2001. Using text classifiers for numerical classification. In B. Nebel Ed., *Proceeding of IJCAI-01, 17th International Joint Conference on Artificial Intelligence* (Seattle, US, 2001), pp. 885–890.

Maderlechner, G., Suda, P., and Bruckner, T. 1997. Classification of documents by form and content. *Pattern Recognition Letters 18*, 11/13, 1225–1231.

Manevitz, L. M. and Yousef, M. 2001. One-class SVMs for document classification. *Journal of Machine Learning Research 2*, 139–154.

Manning, C. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*, Chapter 16: Text Categorization, pp. 575–608. The MIT Press, Cambridge, US.

Maron, M. 1961. Automatic indexing: an experimental inquiry. *Journal of the Association for Computing Machinery 8*, 3, 404–417.

Masand, B. 1994. Optimising confidence of text classification by evolution of symbolic expressions. In K. E. Kinnear Ed., *Advances in genetic programming*, Chapter 21, pp. 459–476. Cambridge, US: The MIT Press.

Masand, B., Linoff, G., and Waltz, D. 1992. Classifying news stories using memory-based reasoning. In N. J. Belkin, P. Ingwersen, and A. M. Pejtersen Eds., *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval* (Kobenhavn, DK, 1992), pp. 59–65. ACM Press, New York, US.

Matsuda, K. and Fukushima, T. 1998. Task-oriented World Wide Web retrieval by document type classification. In G. Gardarin, J. C. French, N. Pissinou, K. Makki, and L. Bouganim Eds., *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management* (Bethesda, US, 1998), pp. 109–113. ACM Press, New York, US.

McCallum, A. K. and Nigam, K. 1998. Employing EM in pool-based active learning for text classification. In J. W. Shavlik Ed., *Proceedings of ICML-98, 15th International Conference on Machine Learning* (Madison, US, 1998), pp. 350–358. Morgan Kaufmann Publishers, San Francisco, US.

McCallum, A. K., Rosenfeld, R., Mitchell, T. M., and Ng, A. Y. 1998. Improving text classification by shrinkage in a hierarchy of classes. In J. W. Shavlik Ed., *Proceedings of ICML-98, 15th International Conference on Machine Learning* (Madison, US, 1998), pp. 359–367. Morgan Kaufmann Publishers, San Francisco, US.

Meretakis, D., Fragoudis, D., Lu, H., and Likothanassis, S. 2000. Scalable association-based text classification. In A. Agah, J. Callan, and E. Rundensteiner Eds., *Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management* (McLean, US, 2000), pp. 373–374. ACM Press, New York, US.

Merkl, D. 1998. Text classification with self-organizing maps: Some lessons learned. *Neurocomputing 21*, 1/3, 61–77.

Mladenić, D. 1998a. Feature subset selection in text learning. In C. Nédellec and C. Rouveirol Eds., *Proceedings of ECML-98, 10th European Conference on Machine Learning* (Chemnitz, DE, 1998), pp. 95–100. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1398.

Mladenić, D. 1998b. *Machine Learning on non-homogeneous, distributed text data*. Ph. D. thesis, J. Stefan Institute, University of Ljubljana, Ljubljana, SL.

Mladenić, D. 1998c. Turning Yahoo! into an automatic Web page classifier. In H. Prade Ed., *Proceedings of ECAI-98, 13th European Conference on Artificial Intelligence* (Brighton, UK, 1998), pp. 473–474. John Wiley and Sons, Chichester, UK.

Mladenić, D. 1999. Text learning and related intelligent agents: a survey. *IEEE Intelligent Systems 14*, 4, 44–54.

Mladenić, D. and Grobelnik, M. 1998. Word sequences as features in text-learning. In *Proceedings of ERK-98, the Seventh Electrotechnical and Computer Science Conference* (Ljubljana, SL, 1998), pp. 145–148.

Mladenić, D. and Grobelnik, M. 1999. Feature selection for unbalanced class distribution and naive bayes. In I. Bratko and S. Dzeroski Eds., *Proceedings of ICML-99, 16th International Conference on Machine Learning* (Bled, SL, 1999), pp. 258–267. Morgan Kaufmann Publishers, San Francisco, US.

Moens, M.-F. and Dumortier, J. 2000. Text categorization: the assignment of subject descriptors to magazine articles. *Information Processing and Management 36*, 6, 841–861.

Mooney, R. J. and Roy, L. 2000. Content-based book recommending using learning for text categorization. In *Proceedings of DL-00, 5th ACM Conference on Digital Libraries* (San Antonio, US, 2000), pp. 195–204. ACM Press, New York, US.

Mostafa, J. and Lam, W. 2000. Automatic classification using supervised learning in a medical document filtering application. *Information Processing and Management 36*, 3, 415–444.

Moulinier, I. 1997. Feature selection: a useful preprocessing step. In J. Furner and D. Harper Eds., *Proceedings of BCSIRSG-97, the 19th Annual Colloquium of the British Computer Society Information Retrieval Specialist Group*, Electronic Workshops in Computing (Aberdeen, UK, 1997). Springer Verlag, Heidelberg, DE.

Moulinier, I. and Ganascia, J.-G. 1996. Applying an existing machine learning algorithm to text categorization. In S. Wermter, E. Riloff, and G. Scheler Eds., *Connectionist, statistical, and symbolic approaches to learning for natural language processing* (1996), pp. 343–354. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1040.

Moulinier, I., Raškinis, G., and Ganascia, J.-G. 1996. Text categorization: a symbolic approach. In *Proceedings of SDAIR-96, 5th Annual Symposium on Document Analysis and Information Retrieval* (Las Vegas, US, 1996), pp. 87–99.

Myers, K., Kearns, M., Singh, S., and Walker, M. A. 2000. A boosting approach to topic spotting on subdialogues. In P. Langley Ed., *Proceedings of ICML-00, 17th*

International Conference on Machine Learning (Stanford, US, 2000), pp. 655–662. Morgan Kaufmann Publishers, San Francisco, US.

NG, H. T., GOH, W. B., AND LOW, K. L. 1997. Feature selection, perceptron learning, and a usability case study for text categorization. In N. J. BELKIN, A. D. NARASIMHALU, AND P. WILLETT Eds., *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval* (Philadelphia, US, 1997), pp. 67–73. ACM Press, New York, US.

NIGAM, K. 2001. *Using Unlabeled Data to Improve Text Classification*. Ph. D. thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, US.

NIGAM, K. AND GHANI, R. 2000. Analyzing the applicability and effectiveness of co-training. In A. AGAH, J. CALLAN, AND E. RUNDENSTEINER Eds., *Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management* (McLean, US, 2000), pp. 86–93. ACM Press, New York, US.

NIGAM, K., MCCALLUM, A. K., THRUN, S., AND MITCHELL, T. M. 1998. Learning to classify text from labeled and unlabeled documents. In *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence* (Madison, US, 1998), pp. 792–799. AAAI Press, Menlo Park, US. An extended version appears as [Nigam et al. 2000].

NIGAM, K., MCCALLUM, A. K., THRUN, S., AND MITCHELL, T. M. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning 39*, 2/3, 103–134.

OH, H.-J., MYAENG, S. H., AND LEE, M.-H. 2000. A practical hypertext categorization method using links and incrementally available class information. In N. J. BELKIN, P. INGWERSEN, AND M.-K. LEONG Eds., *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* (Athens, GR, 2000), pp. 264–271. ACM Press, New York, US.

ONTRUP, J. AND RITTER, H. 2001. Text categorization and semantic browsing with self-organizing maps on non-Euclidean spaces. In L. D. RAEDT AND A. SIEBES Eds., *Proceedings of PKDD-01, 5th European Conference on Principles and Practice of Knowledge Discovery in Databases* (Freiburg, DE, 2001), pp. 338–349. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2168.

PAIJMANS, H. 1999. Text categorization as an information retrieval task. *The South African Computer Journal* 21, 4–15.

PALIOURAS, G., KARKALETSIS, V., AND SPYROPOULOS, C. D. 1999. Learning rules for large vocabulary word sense disambiguation. In T. DEAN Ed., *Proceedings of IJCAI-99, 16th International Joint Conference on Artificial Intelligence* (Stockholm, SE, 1999), pp. 674–679. Morgan Kaufmann Publishers, San Francisco, US.

PETASIS, G., CUCCHIARELLI, A., VELARDI, P., PALIOURAS, G., KARKALETSIS, V., AND SPYROPOULOS, C. D. 2000. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In N. J. BELKIN, P. INGWERSEN, AND M.-K. LEONG Eds., *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* (Athens, GR, 2000), pp. 128–135. ACM Press, New York, US.

PETERS, C. AND KOSTER, C. H. 2002. Uncertainty-based noise reduction and term selection in text categorization. In F. CRESTANI, M. GIROLAMI, AND C. J. VAN RIJSBERGEN Eds., *Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research* (Glasgow, UK, 2002), pp. 248–267. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2291.

RAGAS, H. AND KOSTER, C. H. 1998. Four text classification algorithms compared on a Dutch corpus. In W. B. CROFT, A. MOFFAT, C. J. VAN RIJSBERGEN, R. WILKINSON, AND J. ZOBEL Eds., *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval* (Melbourne, AU, 1998), pp. 369–370. ACM Press, New York, US.

RASKUTTI, B., FERRÁ, H., AND KOWALCZYK, A. 2001. Second order features for maximising text classification performance. In L. D. RAEDT AND P. A. FLACH Eds., *Proceedings of ECML-01, 12th European Conference on Machine Learning* (2001).

Rau, L. F. and Jacobs, P. S. 1991. Creating segmented databases from free text for text retrieval. In A. Bookstein, Y. Chiaramella, G. Salton, and V. V. Raghavan Eds., *Proceedings of SIGIR-91, 14th ACM International Conference on Research and Development in Information Retrieval* (Chicago, US, 1991), pp. 337–346. ACM Press, New York, US.

Rennie, J. and McCallum, A. K. 1999. Using reinforcement learning to spider the Web efficiently. In I. Bratko and S. Dzeroski Eds., *Proceedings of ICML-99, 16th International Conference on Machine Learning* (Bled, SL, 1999), pp. 335–343. Morgan Kaufmann Publishers, San Francisco, US.

Ribeiro-Neto, B., Laender, A. H., and de Lima, L. R. 2001. An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science 52*, 5, 391–401.

Riloff, E. 1993. Using cases to represent context for text classification. In B. Bhargava, T. Finin, and Y. Yesha Eds., *Proceedings of CIKM-93, 2nd International Conference on Information and Knowledge Management* (New York, US, 1993), pp. 105–113. ACM Press, New York, US.

Riloff, E. 1994. *Information Extraction as a Basis for Portable Text Classification Systems*. Ph. D. thesis, Department of Computer Science, University of Massachusetts, Amherst, US.

Riloff, E. 1995. Little words can make a big difference for text classification. In E. A. Fox, P. Ingwersen, and R. Fidel Eds., *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval* (Seattle, US, 1995), pp. 130–136. ACM Press, New York, US.

Riloff, E. 1996. Using learned extraction patterns for text classification. In S. Wermter, E. Riloff, and G. Scheler Eds., *Connectionist, statistical, and symbolic approaches to learning for natural language processing* (1996), pp. 275–289. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1040.

Riloff, E. and Lehnert, W. 1994. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems 12*, 3, 296–333.

Riloff, E. and Lehnert, W. 1998. Classifying texts using relevancy signatures. In *Proceedings of AAAI-92, 10th Conference of the American Association for Artificial Intelligence* (San Jose, US, 1998), pp. 329–334. AAAI Press, Menlo Park, US.

Riloff, E. and Lorenzen, J. 1999. Extraction-based text categorization: Generating domain-specific role relationships. In T. Strzalkowski Ed., *Natural language information retrieval*, pp. 167–196. Dordrecht, NL: Kluwer Academic Publishers.

Robertson, S. E. and Harding, P. 1984. Probabilistic automatic indexing by learning from human indexers. *Journal of Documentation 40*, 4, 264–270.

Roth, D. 1998. Learning to resolve natural language ambiguities: a unified approach. In *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence* (Madison, US, 1998), pp. 806–813. AAAI Press, Menlo Park, US.

Ruiz, M. and Srinivasan, P. 2002. Hierarchical text classification using neural networks. *Information Retrieval 5*, 1, 87–118.

Ruiz, M. E. and Srinivasan, P. 1997. Automatic text categorization using neural networks. In E. Efthimiadis Ed., *Proceedings of the 8th ASIS/SIGCR Workshop on Classification Research* (Washington, US, 1997), pp. 59–72. American Society for Information Science, Washington, US.

Ruiz, M. E. and Srinivasan, P. 1999a. Combining machine learning and hierarchical indexing structures for text categorization. In *Proceedings of the 10th ASIS/SIGCR Workshop on Classification Research* (Washington, US, 1999). American Society for Information Science, Washington, US.

Ruiz, M. E. and Srinivasan, P. 1999b. Hierarchical neural networks for text categorization. In M. A. Hearst, F. Gey, and R. Tong Eds., *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval* (Berkeley, US, 1999), pp. 281–282. ACM Press, New York, US.

Sable, C. and Church, K. 2001. Using bins to empirically estimate term weights for text categorization. In L. Lee and D. Harman Eds., *Proceedings of EMNLP-01, 6th Conference on Empirical Methods in Natural Language Processing* (Pittsburgh, US, 2001), pp. 58–66. Association for Computational Linguistics, Morristown, US.

Sable, C. L. and Hatzivassiloglou, V. 1999. Text-based approaches for the categorization of images. In S. Abiteboul and A.-M. Vercoustre Eds., *Proceedings of ECDL-99, 3rd European Conference on Research and Advanced Technology for Digital Libraries* (Paris, FR, 1999), pp. 19–38. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1696. An extended version appears as [Sable and Hatzivassiloglou 2000].

Sable, C. L. and Hatzivassiloglou, V. 2000. Text-based approaches for non-topical image categorization. *International Journal of Digital Libraries 3*, 3, 261–275.

Sahami, M. Ed. 1998. *Proceedings of the 1998 Workshop on Learning for Text Categorization* (Madison, US, 1998). Available as Technical Report WS-98-05.

Sahami, M., Hearst, M. A., and Saund, E. 1996. Applying the multiple cause mixture model to text categorization. In L. Saitta Ed., *Proceedings of ICML-96, 13th International Conference on Machine Learning* (Bari, IT, 1996), pp. 435–443. Morgan Kaufmann Publishers, San Francisco, US.

Sahami, M., Yusufali, S., and Baldonado, M. Q. 1998. SONIA: a service for organizing networked information autonomously. In I. Witten, R. Akscyn, and F. M. Shipman Eds., *Proceedings of DL-98, 3rd ACM Conference on Digital Libraries* (Pittsburgh, US, 1998), pp. 200–209. ACM Press, New York, US.

Sakakibara, Y., Misue, K., and Koshiba, T. 1996. A machine learning approach to knowledge acquisitions from text databases. *International Journal of Human Computer Interaction 8*, 3, 309–324.

Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C. D., and Stamatopoulos, P. 2001. Stacking classifiers for anti-spam filtering of e-mail. In L. Lee and D. Harman Eds., *Proceedings of EMNLP-01, 6th Conference on Empirical Methods in Natural Language Processing* (Pittsburgh, US, 2001), pp. 44–50. Association for Computational Linguistics, Morristown, US.

Sasaki, M. and Kita, K. 1998a. Automatic text categorization based on hierarchical rules. In *Proceedings of the 5th International Conference on Soft Computing and Information* (Iizuka, JP, 1998), pp. 935–938. World Scientific, Singapore, SN.

Sasaki, M. and Kita, K. 1998b. Rule-based text categorization using hierarchical categories. In *Proceedings of SMC-98, IEEE International Conference on Systems, Man, and Cybernetics* (La Jolla, US, 1998), pp. 2827–2830. IEEE Computer Society Press, Los Alamitos, US.

Schapire, R. E. and Singer, Y. 2000. BoosTexter: a boosting-based system for text categorization. *Machine Learning 39*, 2/3, 135–168.

Schapire, R. E., Singer, Y., and Singhal, A. 1998. Boosting and Rocchio applied to text filtering. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel Eds., *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval* (Melbourne, AU, 1998), pp. 215–223. ACM Press, New York, US.

Scheffer, T. and Joachims, T. 1999. Expected error analysis for model selection. In I. Bratko and S. Dzeroski Eds., *Proceedings of ICML-99, 16th International Conference on Machine Learning* (Bled, SL, 1999), pp. 361–370. Morgan Kaufmann Publishers, San Francisco, US.

Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics 24*, 1, 97–124.

Schütze, H., Hull, D. A., and Pedersen, J. O. 1995. A comparison of classifiers and document representations for the routing problem. In E. A. Fox, P. Ingwersen, and R. Fidel Eds., *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval* (Seattle, US, 1995), pp. 229–237. ACM Press, New York, US.

Scott, S. 1998. Feature engineering for a symbolic approach to text classification. Master's thesis, Computer Science Department, University of Ottawa, Ottawa, CA.

Scott, S. and Matwin, S. 1999. Feature engineering for text classification. In I. Bratko and S. Dzeroski Eds., *Proceedings of ICML-99, 16th International Conference on Machine Learning* (Bled, SL, 1999), pp. 379–388. Morgan Kaufmann Publishers, San Francisco, US.

Sebastiani, F. 1999. A tutorial on automated text categorisation. In A. Amandi and R. Zunino Eds., *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence* (Buenos Aires, AR, 1999), pp. 7–35. An extended version appears as [Sebastiani 2002].

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys 34*, 1, 1–47.

Sebastiani, F., Sperduti, A., and Valdambrini, N. 2000. An improved boosting algorithm and its application to automated text categorization. In A. Agah, J. Callan, and E. Rundensteiner Eds., *Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management* (McLean, US, 2000), pp. 78–85. ACM Press, New York, US.

Shin, C., Doermann, D., and Rosenfeld, A. 2001. Classification of document pages using structure-based features. *International Journal on Document Analysis and Recognition 3*, 4, 232–247.

Siolas, G. and d'Alche Buc, F. 2000. Support vector machines based on a semantic kernel for text categorization. In S.-I. Amari, C. L. Giles, M. Gori, and V. Piuri Eds., *Proceedings of IJCNN-00, International Joint Conference on Neural Networks*, Volume 5 (Como, IT, 2000), pp. 205–209. IEEE Computer Society Press, Los Alamitos, US.

Skarmeta, A., Bensaid, A., and Tazi, N. 2000. Data mining for text categorization with semi-supervised agglomerative hierarchical clustering. *International Journal of Intelligent Systems 15*, 7, 633–646.

Slattery, S. and Craven, M. 1998. Combining statistical and relational methods for learning in hypertext domains. In D. Page Ed., *Proceedings of ILP-98, 8th International Conference on Inductive Logic Programming* (Madison, US, 1998), pp. 38–52. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1446.

Slattery, S. and Craven, M. 2000. Discovering test set regularities in relational domains. In P. Langley Ed., *Proceedings of ICML-00, 17th International Conference on Machine Learning* (Stanford, US, 2000), pp. 895–902. Morgan Kaufmann Publishers, San Francisco, US.

Slonim, N., Friedman, N., and Tishby, N. 2002. Unsupervised document classification using sequential information maximization. In *Proceedings of SIGIR-02, 25th ACM International Conference on Research and Development in Information Retrieval* (Tampere, FI, 2002). ACM Press, New York, US.

Slonim, N. and Tishby, N. 2001. The power of word clusters for text classification. In *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research* (Darmstadt, DE, 2001).

Soucy, P. and Mineau, G. W. 2001. A simple feature selection method for text classification. In B. Nebel Ed., *Proceeding of IJCAI-01, 17th International Joint Conference on Artificial Intelligence* (Seattle, US, 2001), pp. 897–902.

Spitz, L. and Maghbouleh, A. 2000. Text categorization using character shape codes. In D. P. Lopresti and J. Zhou Eds., *Proceedings of the 7th SPIE Conference on Document Recognition and Retrieval* (San Jose, US, 2000), pp. 174–181. SPIE - The International Society for Optical Engineering.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics 26*, 4, 471–495.

Taghva, K., Nartker, T. A., Borsack, J., Lumos, S., Condit, A., and Young, R. 2000. Evaluating text categorization in the presence of ocr errors. In P. B. Kantor, D. P. Lopresti, and J. Zhou Eds., *Proceedings of the 8th SPIE Conference on Document Recog-*

nition and Retrieval (San Jose, US, 2000), pp. 68–74. SPIE, The International Society for Optical Engineering, Washington, US.

Taira, H. and Haruno, M. 1999. Feature selection in SVM text categorization. In *Proceedings of AAAI-99, 16th Conference of the American Association for Artificial Intelligence* (Orlando, US, 1999), pp. 480–486. AAAI Press, Menlo Park, US.

Taira, H. and Haruno, M. 2001. Text categorization using transductive boosting. In L. D. Raedt and P. A. Flach Eds., *Proceedings of ECML-01, 12th European Conference on Machine Learning* (Freiburg, DE, 2001), pp. 454–465. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2167.

Takamura, H. and Matsumoto, Y. 2001. Feature space restructuring for SVMs with application to text categorization. In L. Lee and D. Harman Eds., *Proceedings of EMNLP-01, 6th Conference on Empirical Methods in Natural Language Processing* (Pittsburgh, US, 2001), pp. 51–57. Association for Computational Linguistics, Morristown, US.

Tan, A.-H. 2001. Predictive self-organizing networks for text categorization. In D. Cheung, Q. Li, and G. Williams Eds., *Proceedings of PAKDD-01, 5th Pacific-Asia Conferenece on Knowledge Discovery and Data Mining* (Hong Kong, CN, 2001), pp. 66–77. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2035.

Tan, C.-M. 2000. Finding and using high quality word-pairs for enhanced text categorization. Master's thesis, Department of Computer Science, University of California at Santa Barbara, Santa Barbara, US.

Taskar, B., Segal, E., and Koller, D. 2001. Probabilistic classification and clustering in relational data. In B. Nebel Ed., *Proceeding of IJCAI-01, 17th International Joint Conference on Artificial Intelligence* (Seattle, US, 2001), pp. 870–878.

Tauritz, D. R., Kok, J. N., and Sprinkhuizen-Kuyper, I. G. 2000. Adaptive information filtering using evolutionary computation. *Information Sciences 122*, 2/4, 121–140.

Tauritz, D. R. and Sprinkhuizen-Kuyper, I. G. 1999. Adaptive information filtering algorithms. In D. J. Hand, J. N. Kok, and M. R. Berthold Eds., *Proceedings of IDA-99, 3rd Symposium on Intelligent Data Analysis* (Amsterdam, NL, 1999), pp. 513–524. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1642.

Teahan, W. J. 2000. Text classification and segmentation using minimum cross-entropy. In *Proceeding of RIAO-00, 6th International Conference "Recherche d'Information Assistee par Ordinateur"* (Paris, FR, 2000).

Tong, R., Winkler, A., and Gage, P. 1992. Classification trees for document routing: A report on the TREC experiment. In D. K. Harman Ed., *Proceedings of TREC-1, 1st Text Retrieval Conference* (Gaithersburg, US, 1992), pp. 209–228. National Institute of Standards and Technology, Gaithersburg, US.

Tong, S. and Koller, D. 2000. Support vector machine active learning with applications to text classification. In P. Langley Ed., *Proceedings of ICML-00, 17th International Conference on Machine Learning* (Stanford, US, 2000), pp. 999–1006. Morgan Kaufmann Publishers, San Francisco, US. An extended version appears as [Tong and Koller 2001].

Tong, S. and Koller, D. 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research 2*, 45–66.

Toutanova, K., Chen, F., Popat, K., and Hofmann, T. 2001. Text classification in a hierarchical mixture model for small training sets. In H. Paques, L. Liu, and D. Grossman Eds., *Proceedings of CIKM-01, 10th ACM International Conference on Information and Knowledge Management* (Atlanta, US, 2001), pp. 105–113. ACM Press, New York, US.

Turney, P. D. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval 2*, 4, 303–336.

Tzeras, K. and Hartmann, S. 1993. Automatic indexing based on Bayesian inference networks. In R. Korfhage, E. Rasmussen, and P. Willett Eds., *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval* (Pittsburgh, US, 1993), pp. 22–34. ACM Press, New York, US.

Ureña-López, L. A., Buenaga, M., and Gómez, J. M. 2001. Integrating linguistic resources in TC through WSD. *Computers and the Humanities 35*, 2, 215–230.

Vert, J.-P. 2001. Text categorization using adaptive context trees. In A. Gelbukh Ed., *Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing* (Mexico City, ME, 2001), pp. 423–436. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2004.

Vinokourov, A. and Girolami, M. 2001. Document classification employing the Fisher kernel derived from probabilistic hierarchic corpus representations. In *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research* (Darmstadt, DE, 2001), pp. 24–40.

Vinokourov, A. and Girolami, M. 2002. A probabilistic framework for the hierarchic organisation and classification of document collections. *Journal of Intelligent Information Systems 18*, 2/3, 153–172. Special Issue on Automated Text Categorization.

Wang, H. and Son, N. H. 1999. Text classification using lattice machine. In A. Skowron and Z. W. Raś Eds., *Proceedings of ISMIS-99, 11th International Symposium on Methodologies for Intelligent Systems* (Warsaw, PL, 1999), pp. 235–243. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1609.

Wang, K., Zhou, S., and He, Y. 2001. Hierarchical classification of real life documents. In *Proceedings of the 1st SIAM International Conference on Data Mining* (Chicago, US, 2001).

Wang, K., Zhou, S., and Liew, S. C. 1999. Building hierarchical classifiers using class proximity. In M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik, and M. L. Brodie Eds., *Proceedings of VLDB-99, 25th International Conference on Very Large Data Bases* (Edinburgh, UK, 1999), pp. 363–374. Morgan Kaufmann Publishers, San Francisco, US.

Wang, W., Meng, W., and Yu, C. 2000. Concept hierarchy based text database categorization in a metasearch engine environment. In Q. Li, Z. M. Ozsoyoglu, R. Wagner, Y. Kambayashi, and Y. Zhang Eds., *Proceedings of WISE-00, 1st International Conference on Web Information Systems Engineering*, Volume 1 (Hong Kong, CN, 2000), pp. 283–290. IEEE Computer Society Press, Los Alamitos, US.

Wei, C.-P. and Dong, Y.-X. 2001. A mining-based category evolution approach to managing online document categories. In R. H. Sprague Ed., *Proceedings of HICSS-01, 34th Annual Hawaii International Conference on System Sciences* (Maui, US, 2001). IEEE Computer Society Press, Los Alamitos, US.

Weigend, A. S., Wiener, E. D., and Pedersen, J. O. 1999. Exploiting hierarchy in text categorization. *Information Retrieval 1*, 3, 193–216.

Weiss, S. M., Apté, C., Damerau, F. J., Johnson, D. E., Oles, F. J., Goetz, T., and Hampp, T. 1999. Maximizing text-mining performance. *IEEE Intelligent Systems 14*, 4, 63–69.

Wermter, S. 2000. Neural network agents for learning semantic text classification. *Information Retrieval 3*, 2, 87–103.

Wermter, S., Arevian, G., and Panchev, C. 1999. Recurrent neural network learning for text routing. In *Proceedings of ICANN-99, 9th International Conference on Artificial Neural Networks* (Edinburgh, UK, 1999), pp. 898–903. Institution of Electrical Engineers, London, UK.

Wermter, S., Panchev, C., and Arevian, G. 1999. Hybrid neural plausibility networks for news agents. In *Proceedings of AAAI-99, 16th Conference of the American Association for Artificial Intelligence* (Orlando, US, 1999), pp. 93–98. AAAI Press, Menlo Park, US.

Wiener, E. D. 1995. A neural network approach to topic spotting in text. Master's thesis, Department of Computer Science, University of Colorado at Boulder, Boulder, US.

Wiener, E. D., Pedersen, J. O., and Weigend, A. S. 1995. A neural network approach to topic spotting. In *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval* (Las Vegas, US, 1995), pp. 317–332.

Wong, J. W., Kan, W.-K., and Young, G. H. 1996. Action: automatic classification for

full-text documents. *SIGIR Forum 30*, 1, 26–41.

Yamazaki, T. and Dagan, I. 1997. Mistake-driven learning with thesaurus for text categorization. In *Proceedings of NLPRS-97, the Natural Language Processing Pacific Rim Symposium* (Phuket, TH, 1997), pp. 369–374.

Yang, H.-C. and Lee, C.-H. 2000a. Automatic category generation for text documents by self-organizing maps. In S.-I. Amari, C. L. Giles, M. Gori, and V. Piuri Eds., *Proceedings of IJCNN-00, International Joint Conference on Neural Networks*, Volume 3 (Como, IT, 2000), pp. 581–586. IEEE Computer Society Press, Los Alamitos, US.

Yang, H.-C. and Lee, C.-H. 2000b. Automatic category structure generation and categorization of Chinese text documents. In D. A. Zighed, J. Komorowski, and J. Zytkow Eds., *Proceedings of PKDD-00, 4th European Conference on Principles of Data Mining and Knowledge Discovery* (Lyon, FR, 2000), pp. 581–586. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1910.

Yang, Y. 1994. Expert network: effective and efficient learning from human decisions in text categorisation and retrieval. In W. B. Croft and C. J. van Rijsbergen Eds., *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval* (Dublin, IE, 1994), pp. 13–22. Springer Verlag, Heidelberg, DE.

Yang, Y. 1995. Noise reduction in a statistical approach to text categorization. In E. A. Fox, P. Ingwersen, and R. Fidel Eds., *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval* (Seattle, US, 1995), pp. 256–263. ACM Press, New York, US.

Yang, Y. 1996. An evaluation of statistical approaches to MEDLINE indexing. In J. J. Cimino Ed., *Proceedings of AMIA-96, Fall Symposium of the American Medical Informatics Association* (Washington, US, 1996), pp. 358–362. Hanley and Belfus.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval 1*, 1/2, 69–90.

Yang, Y. 2001. A study on thresholding strategies for text categorization. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel Eds., *Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval* (New Orleans, US, 2001), pp. 137–145. ACM Press, New York, US.

Yang, Y., Ault, T., and Pierce, T. 2000. Combining multiple learning strategies for effective cross-validation. In P. Langley Ed., *Proceedings of ICML-00, 17th International Conference on Machine Learning* (Stanford, US, 2000), pp. 1167–1182. Morgan Kaufmann Publishers, San Francisco, US.

Yang, Y., Ault, T., Pierce, T., and Lattimer, C. W. 2000. Improving text categorization methods for event tracking. In N. J. Belkin, P. Ingwersen, and M.-K. Leong Eds., *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* (Athens, GR, 2000), pp. 65–72. ACM Press, New York, US.

Yang, Y. and Chute, C. G. 1993. An application of Least Squares Fit mapping to text information retrieval. In R. Korfhage, E. Rasmussen, and P. Willett Eds., *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval* (Pittsburgh, US, 1993), pp. 281–290. ACM Press, New York, US. An extended version appears as [Yang and Chute 1994].

Yang, Y. and Chute, C. G. 1994. An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems 12*, 3, 252–277.

Yang, Y. and Liu, X. 1999. A re-examination of text categorization methods. In M. A. Hearst, F. Gey, and R. Tong Eds., *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval* (Berkeley, US, 1999), pp. 42–49. ACM Press, New York, US.

Yang, Y. and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In D. H. Fisher Ed., *Proceedings of ICML-97, 14th International Conference on Machine Learning* (Nashville, US, 1997), pp. 412–420. Morgan Kaufmann Publishers, San Francisco, US.

Yang, Y., Slattery, S., and Ghani, R. 2002. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems 18*, 2/3, 219–241. Special Issue on Automated Text Categorization.

Yang, Y. and Wilbur, J. W. 1996a. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Computers in Biology and Medicine 26*, 3, 209–222.

Yang, Y. and Wilbur, J. W. 1996b. Using corpus statistics to remove redundant words in text categorization. *Journal of the American Society for Information Science 47*, 5, 357–369.

Yavuz, T. and Güvenir, H. A. 1998. Application of k-nearest neighbor on feature projections classifier to text categorization. In U. Gudukbay, T. Dayar, A. Gursoy, and E. Gelenbe Eds., *Proceedings of ISCIS-98, 13th International Symposium on Computer and Information Sciences* (Ankara, TR, 1998), pp. 135–142. IOS Press, Amsterdam, NL.

Yu, E. S. and Liddy, E. D. 1999. Feature selection in text categorization using the Baldwin effect networks. In *Proceedings of IJCNN-99, 10th International Joint Conference on Neural Networks* (Washington, DC, 1999), pp. 2924–2927. IEEE Computer Society Press, Los Alamitos, US.

Yu, K. L. and Lam, W. 1998. A new on-line learning algorithm for adaptive text filtering. In G. Gardarin, J. C. French, N. Pissinou, K. Makki, and L. Bouganim Eds., *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management* (Bethesda, US, 1998), pp. 156–160. ACM Press, New York, US.

Zelikovitz, S. and Hirsh, H. 2000. Improving short text classification using unlabeled background knowledge. In P. Langley Ed., *Proceedings of ICML-00, 17th International Conference on Machine Learning* (Stanford, US, 2000), pp. 1183–1190. Morgan Kaufmann Publishers, San Francisco, US.

Zelikovitz, S. and Hirsh, H. 2001. Using LSI for text classification in the presence of background text. In H. Paques, L. Liu, and D. Grossman Eds., *Proceedings of CIKM-01, 10th ACM International Conference on Information and Knowledge Management* (Atlanta, US, 2001), pp. 113–118. ACM Press, New York, US.

Zhang, T. and Oles, F. J. 2001. Text categorization based on regularized linear classification methods. *Information Retrieval 4*, 1, 5–31.