P.O.Box 8718, Beijing 100080, China	Journal of Software, Sept. 2006,17(9):1848-1859
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2006 by *Journal of Software*

基于机器学习的文本分类技术研究进展

苏金树, 张博锋, 徐 昕

苏金树1, 张博锋1, 徐 昕1,2
1(国防科学技术大学 计算机学院,湖南 长沙 410073)
2(国防科学技术大学 机电工程与自动化学院,湖南 长沙 410073)
作者简介: 苏金树(1962－),男,福建莆田人,博士,教授,博士生导师,CCF高级会员,主要研究领域为计算机网络,信息安全.张博锋(1978－),男,博士生,主要研究领域为信息安全,互联网内容信息分类.徐昕(1974－),男,博士,副教授,主要研究领域为机器学习,信息安全,自主计算.
联系人: 张博锋 Phn: +86-731-4513504, E-mail: bfzhang@nudt.edu.cn

Received 2005-12-15; Accepted 2006-04-03

Abstract
In recent years, there have been extensive studies and rapid progresses in automatic text categorization, which is one of the hotspots and key techniques in the information retrieval and data mining field. Highlighting the state-of-art challenging issues and research trends for content information processing of Internet and other complex applications, this paper presents a survey on the up-to-date development in text categorization based on machine learning, including model, algorithm and evaluation. It is pointed out that problems such as nonlinearity, skewed data distribution, labeling bottleneck, hierarchical categorization, scalability of algorithms and categorization of Web pages are the key problems to the study of text categorization. Possible solutions to these problems are also discussed respectively. Finally, some future directions of research are given.

Su JS, Zhang BF, Xu X. Advances in machine learning based text categorization. *Journal of Software*, 2006,17(9):1848-1859.
DOI: 10.1360/jos171848
http://www.jos.org.cn/1000-9825/17/1848.htm

摘要
文本自动分类是信息检索与数据挖掘领域的研究热点与核心技术,近年来得到了广泛的关注和快速的发展.提出了基于机器学习的文本分类技术所面临的互联网内容信息处理等复杂应用的挑战,从模型、算法和评测等方面对其研究进展进行综述评论.认为非线性、数据集偏斜、标注瓶颈、多层分类、算法的扩展性及Web页分类等问题是目前文本分类研究的关键问题,并讨论了这些问题可能采取的方法.最后对研究的方向进行了展望.

基金项目：Supported by the National Natural Science Foundation of China under Grant Nos.90604006, 60303012 (国家自然科学基金); the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20049998027 (国家教育部高校博士点基金)

References:

[1] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002,34(1):1-47.

[2] Debole F, Sebastiani F. Supervised term weighting for automated text categorization. In: Haddad H, George AP, eds. Proc. of the 18th ACM Symp. on Applied Computing (SAC-03). Melbourne: ACM Press, 2003. 784-788.

[3] Xue D, Sun M. Chinese text categorization based on the binary weighting model with non-binary smoothing. In: Sebastiani F, ed. Proc. of the 25th European Conf. on Information Retrieval (ECIR-03). Pisa: Springer-Verlag, 2003. 408-419.

[4] Lertnattee V, Theeramunkong T. Effect of term distributions on centroid-based text categorization. Information Sciences, 2004, 158 (1):89-115.

[5] Bigi B. Using Kullback-Leibler distance for text categorization. In: Sebastiani F, ed. Proc. of the 25th European Conf. on Information Retrieval (ECIR-03). Pisa: Springer-Verlag, 2003. 305-319.

[6] Nunzio GMD. A bidimensional view of documents for text categorisation. In: McDonald S, Tait J, eds. Proc. of the 26th European Conf. on Information Retrieval Research (ECIR-04). Sunderland: Springer-Verlag, 2004. 112-126.

[7] Moschitti A, Basili R. Complex linguistic features for text classification: A comprehensive study. In: McDonald S, Tait J, eds. Proc. of the 26th European Conf. on Information Retrieval Research (ECIR-04). Sunderland: Springer-Verlag, 2004. 181-196.

[8] Kehagias A, Petridis V, Kaburlasos VG, Fragkou P. A comparison of word- and sense-based text categorization using several classification algorithms. Journal of Intelligent Information Systems, 2003,21(3):227-247.

[9] Forman G. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 2003,3 (1):1533-7928.

[10] Chen W, Chang X, Wang H, Zhu J, Tianshun Y. Automatic word clustering for text categorization using global information. In: Myaeng SH, Zhou M, Wong KF, Zhang H, eds. Proc. of the Information Retrieval Technology, Asia Information Retrieval Symp. (AIRS 2004). Beijing: Springer-Verlag, 2004. 1-11.

[11] Chen L, Tokuda N, Nagai A. A new differential LSI space-based probabilistic document classifier. Information Processing Letters, 2003,88(5):203-212.

[12] Kim H, Howland P, Park H. Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research, 2005,6(1):37-53.

[13] Rogati M, Yang Y. High-Performing feature selection for text classification. In: David G, Kalpakis K, Sajda Q, Han D, Len S, eds. Proc. of the 11th ACM Int'l Conf. on Information and Knowledge Management (CIKM-02). McLean: ACM Press, 2002. 659-661.

[14] Makrehchi M, Kamel MS. Text classification using small number of features. In: Perner P, Imiya A, eds. Proc. of the 4th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition: (MLDM 2005). 2005. 580-589.

[15] Mladenic D, Brank J, Grobelnik M, Milic-Frayling N. Feature selection using linear classifier weights: Interaction with classification models. In: Jarvelin K, Allan J, Bruza P, Sanderson M, eds. Proc. of the 27th ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR-04). Sheffield: ACM Press, 2004. 234-241.

[16] Fernandez J, Montanes E, Diaz I, Ranilla J, Combarro EF. Text categorization by a machine-learning-based term selection. In: Galindo F, Takizawa R, Traunmuller R, eds. Proc. of the Database and Expert Systems Applications (DEXA-04). Zaragoza: Springer-Verlag, 2004. 253-262.

[17] Chua S, Kulathuramaiyer N. Semantic feature selection using WordNet. In: Yao J, Vijay VR, Wang GY, eds. Proc. of the IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI 2004). Beijing: IEEE Computer Society, 2004. 166-172.

[18] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: Fisher DH, ed. Proc. of the 14th Int'l Conf. on Machine Learning (ICML-97). Nashville: Morgan Kaufmann Publishers, 1997. 412-420.

[19] Gabrilovich E, Markovitch S. Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. In: Brodley CE, ed. Proc. of the 21st Int'l Conf. on Machine Learning (ICML-04). Banff: Morgan Kaufmann Publishers, 2004. 41.

[20] Bekkerman R, Yaniv RE, Tishby N, Winter Y. Distributional word clusters vs. words for text categorization. Journal of Machine Learning Research, 2003,3(2):1183-1208.

[21] Soucy P, Mineau GW. Feature selection strategies for text categorization. In: Xiang Y, Chaib-Draa B, eds. Proc. of the 16th Conf. of the Canadian Society for Computational Studies of Intelligence (CSCSI-03). Halifax: Springer-Verlag, 2003. 505-509.

[22] Yang Y, Zhang J, Kisiel B. A scalability analysis of classifiers in text categorization. In: Callan J, Cormack G, Clarke C, Hawking D, Smeaton A, eds. Proc. of the 26th ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR-03). Toronto: ACM Press, 2003. 96-103.

[23] Liu TY, Yang Y, Wan H, Zhou Q, Gao B, Zeng HJ, Chen Z, Ma WY. An experimental study on large-scale web categorization. In: Ellis A, Hagino T, eds. Proc. of the 14th Int'l World Wide Web Conf (WWW-05). Chiba: ACM Press, 2005. 1106-1107.

[24] Chakrabarti S, Roy S, Soundalgekar M. Fast and accurate text classification via multiple linear discriminant projections. Int'l Journal on Very Large Data Bases, 2003,12(2):170-185.

[25] Wu H, Phang TH, Liu B, Li X. A refinement approach to handling model misfit in text categorization. In: Davis H, Daniel K, Raymoind N, eds. Proc. of the 8th ACM Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD-02). Edmonton: ACM Press, 2002. 207-216.

[26] Wang J, Wang H, Zhang S, Hu Y. A simple and efficient algorithm to classify a large scale of text. Journal of Computer Research and Development, 2005,42(1):85-93 (in Chinese with English abstract).

[27] Tan S, Cheng X, Wang B, Xu H, Ghanem MM, Guo Y. Using dragpushing to refine centroid text classifiers. In: Ricardo ABY, Nivio Z, Gary M, Alistair M, John T, eds. Proc. of the ACM SIGIR-05. Salvador: ACM Press, 2005. 653-654.

[28] Debole F, Sebastiani F. An analysis of the relative hardness of reuters-21578 subsets. Journal of the American Society for Information Science and Technology, 2004,56(6):584-596.

[29] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: Nedellec C, Rouveirol C, eds. Proc. of the 10th European Conf. on Machine Learning (ECML-98). Chemnitz: Springer-Verlag, 1998. 137-142.

[30] Yang Y, Liu X. A re-examination of text categorization methods. In: Gey F, Hearst M, Rong R, eds. Proc. of the 22nd ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR-99). Berkeley: ACM Press, 1999. 42-49.

[31] Lewis DD, Li F, Rose T, Yang Y. RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, 2004,5(3):361-397.

[32] Forman G, Cohen I. Learning from little: Comparison of classifiers given little training. In: Jean FB, Floriana E, Fosca G, Dino P, eds. Proc. of the 8th European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD-04). Pisa: Springer-Verlag, 2004. 161-172.

[33] Kazama J, Tsujii J. Maximum entropy models with inequality constraints: A case study on text categorization. Machine Learning, 2005,60(1-3):159-194.

[34] Li R, Wang J, Chen X, Tao X, Hu Y. Using maximum entropy model for Chinese text categorization. Journal of Computer Research and Development, 2005,42(1):94-101 (in Chinese with English abstract).

[35] Liu WY, Song N. A fuzzy approach to classification of text documents. Journal of Computer Science and Technology, 2003,18(5): 640-647.

[36] Widyantoro DH, Yen J. A fuzzy similarity approach in text classification task. In: Proc. of the 9th IEEE Int'l Conf. on Fuzzy Systems (Fuzz-IEEE 2000), Vol.s 1 and 2. San Antonio: IEEE Computer Society, 2000. 653-658. http://citeseer.ist.psu.edu/692028.html

[37] Lam W, Lai KY. Automatic textual document categorization based on generalized instance sets and a metamodel. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(5):628-633.

[38] Tsay JJ, Wang JD. Improving linear classifier for Chinese text categorization. Information Processing and Management, 2004,40(2): 223-237.

[39] Tan S, Cheng X, Ghanem MM, Wang B, Xu H. A novel refinement approach for text categorization. In: Otthein H, Hans JS, Norbert F, Abdur C, Wilfried T, eds. Proc. of the 14th ACM Conf. on Information and Knowledge Management (CIKM-05). Bremen: ACM Press, 2005. 469-476.

[40] Wei YG, Tsay JJ. A study of multiple classifier systems in automated text categorization [PH.D. Thesis]. Chiayi: College of Engineering National Chung Cheng University, 2002.

[41] Bennett PN, Dumais ST, Horvitz E. The combination of text classifiers using reliability indicators. Information Retrieval, 2005,8(1): 67-100.

[42] Xu X, Zhang B, Zhong Q. Text categorization using SVMs with Rocchio ensemble for internet information classification. In: Lu X, Zhao W, eds. Proc of the3rd Int'l Conf on Networking and Mobile Computing (ICCNMC-05). Springer-Verlag, 2005. 1022-1031.

[43] Aas K, Eikvil L. Text categorization: A survey. Technical Report, NR 941, Oslo: Norwegian Computing Center, 1999.

[44] Schapire RE, Singer Y. BoosTexter: A boosting-based system for text categorization. Machine Learning, 2000,39(2-3):135-168.

[45] Li F, Yang Y. A loss function analysis for classification methods in text categorization. In: Fawcett T, Mishra N, eds. Proc. of the ICML 2003. Washington: AAAI Press, 2003. 472-479.

[46] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2002. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[47] Joachims T. Making large-scale SVM learning practical. In: Scholkopf B, Burges C, Smola A, eds. Advances in Kernel Methods—Support Vector Learning. Cambridge: MIT Press, 1999. 169-184.

[48] Cristianini N, Shawe-Taylor J, Lodhi H. Latent semantic kernels. In: Brodley C, Danyluk A, eds. Proc. of the 18th Int'l Conf. on Machine Learning (ICML-01). Williams College: Morgan Kaufmann Publishers, 2001. 66-73.

[49] Cancedda N, Gaussier E, Goutte C, Renders JM. Word sequence kernels. Journal of Machine Learning Research, 2003,3(6): 1059-1082.

[50] Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C. Text classification using string kernels. Journal of Machine Learning Research, 2002,2(2):419-444.

[51] Leslie C, Kuang R. Fast kernels for inexact string matching. In: Scholkopf B, Warmuth MK, eds. Proc. of the 16th Annual Conf. on Learning Theory and 7th Kernel Workshop (COLT/Kernel 2003). Washington: Springer-Verlag, 2003. 114-128.

[52] Fawcett T. ROC graphs: Notes and practical considerations for researchers. Technical Report, HPL-2003-4, Palo Alto: HP Laboratories, 2003.

[53] Yu K, Yu S, Tresp V. Multilabel informed latent semantic indexing. In: Proc. of the ACM SIGIR-05. Salvador: ACM Press, 2005. 258-265.

[54] Lachiche N, Flach P. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In: Fawcett T, Mishra N, eds. Proc. of the 20th Int'l Conf. on Machine Learning (ICML-01). Washington: AAAI Press, 2003. 416-423.

[55] Lewis DD. Reuters-21578 text categorization test collection. Distribution 1.0. 1997. http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt

[56] Muller KR, Mika S, Ratsh G, Tsuda K, Scholkopf B. An introduction to kernel-based learning algorithms. IEEE Trans. on Neural Networks, 2001,12(2):181-202.

[57] Zaragoza HH, Ralf. The perceptron meets Reuters. In: Proc. of the NIPS 2001 Machine Learning for Text and Images Workshop. 2001. http://citeseer.ist.psu.edu/456556.html

[58] Joachims T, Cristianini N, Shawe-Taylor J. Composite kernels for hypertext categorisation. In: Brodley C, Danyluk A, eds. Proc. of the 18th Int'l Conf. on Machine Learning (ICML-01). Williams College: Morgan Kaufmann Publishers, 2001. 250-257.

[59] Chawla NV, Japkowicz N, Kotcz A. Editorial: Special issue on learning from imbalanced data sets. Sigkdd Explorations Newsletters, 2004,6(1):1-6.

[60] Estabrooks A, Jo TH, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Computational Intelligence, 2004,20(1):18-36.

[61] Manevitz LM, Yousef M. One-Class SVMs for document classification. Journal of Machine Learning Research, 2001, 2(1):139-154.

[62] Brank J, Grobelnik M. Training text classifiers with SVM on very few positive examples. Technical Report, MSR-TR-2003-34, Redmond: Microsoft Research, 2003.

[63] Tan S. Neighbor-Weighted k-Nearest neighbor for unbalanced text corpus. Expert Systems with Applications, 2005,28(4):667-671.

[64] Castillo MDd, Serrano JI. A multistrategy approach for digital text categorization from imbalanced documents. SIGKDD Explorations Newsletter, 2004,6(1):70-79.

[65] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data. SIGKDD Explorations, 2004,6(1):80-89.

[66] Forman G. A pitfall and solution in multi-class feature selection for text classification. In: Brodley CE, ed. Proc. of the 21st Int'l Conf. on Machine Learning (ICML-04). Banff: Morgan Kaufmann Publishers, 2004. 38.

[67] Nigam K. Using unlabeled data to improve text classification [Ph.D. Thesis]. Pittsburgh: Carnegie Mellon University, 2001.

[68] Joachims T. Transductive inference for text classification using support vector machines. In: Bratko I, Dzeroski S, eds. Proc. of the 16th Int'l Conf. on Machine Learning (ICML-99). Bled: Morgan Kaufmann Publishers, 1999. 200-209.

[69] Chen YS, Wang GP, Dong SH. A progressive transductive inference algorithm based on support vector machine. Journal of Software, 2003,14(3):451-460 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/451.htm

[70] Taira H, Haruno M. Text categorization using transductive boosting. In: Raedt LD, Flach PA, eds. Proc. of the 12th European Conf. on Machine Learning (ECML-01). Freiburg: Springer-Verlag, 2001. 454-465.

[71] Park SB, Zhang BT. Co-Trained support vector machines for large scale unstructured document classification using unlabeled data and syntactic information. Information Processing and Management, 2004,40(3):421-439.

[72] Kiritchenko S, Matwin S. Email classification with co-training. In: Stewart DA, Johnson JH, eds. Proc. of the 2001 Conf. of the Centre for Advanced Studies on Collaborative Research. Toronto: IBM Press, 2001. 8.

[73] Liu B, Dai Y, Li X, Lee WS, Yu PS. Building text classifiers using positive and unlabeled examples. In: Proc. of the 3rd IEEE Int'l Conf. on Data Mining. Melbourne (ICDM-03). IEEE Computer Society, 2003. 179-188.

[74] Tong S, Koller D. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2001,2(1):45-66.

[75] Ruiz M. Combining machine learning and hierarchical structures for text categorization [Ph.D. Thesis]. Ames: Graduate College of University of Iowa, 2001.

[76] Ruiz M, Srinivasan P. Hierarchical text classification using neural networks. Information Retrieval, 2002,5(1):87-118.

[77] Sun A, Lim EP, Ng WK. Hierarchical text classification methods and their specification. In: Chan AT, Chan SC, Leong HV, Ng VTY, eds. Cooperative Internet Computing. Dordrecht: Kluwer Academic Publishers, 2003. 236-256.

[78] Sun A, Lim EP. Hierarchical text classification and evaluation. In: Cercone N, Lin TY, Wu X, eds. Proc. of the 1st IEEE Int'l Conf. on Data Mining (ICDM-01). San Jose: IEEE Computer Society, 2001. 521-528.

[79] Sun A, Lim EP, Ng WK. Performance measurement framework for hierarchical text classification. Journal of the American Society for Information Science and Technology, 2003,54(11):1014-1028.

[80] Zhou S, Fan Y, Hua J, Yu F, Hu Y. Hierachically classifying Chinese Web documents without dictionary support and segmentation

procedure. In: Lu H, Zhou A, eds. Proc. of the 1st Int'l Conf. on Web-Age Information Management (WAIM-00). Shanghai: Springer-Verlag, 2000. 215-226.

[81] Ceci M, Malerba D. Hierarchical classification of HTML documents with WebClassII. In: Sebastiani F, ed. Proc. of the 25th European Conf. on Information Retrieval (ECIR-03). Pisa: Springer-Verlag, 2003. 57-72.

[82] Huang CC, Chuang SL, Chien LF. LiveClassifier: Creating hierarchical text classifiers through Web corpora. In: Proc. of the 13th Int'l World Wide Web Conf. New York: ACM Press, 2004. 184 -192.

[83] Sun A, Lim EP, Ng WK, Srivastava J. Blocking reduction strategies in hierarchical text classification. IEEE Trans. on Knowledge and Data Engineering, 2004,16(10):1305-1308.

[84] Liu TY, Yang Y, Wan H, Zeng HJ, Chen Z, Ma WY. Support vector machines classification with a very large-scale taxonomy. SIGKDD Explor. Newsl., 2005,7(1):36-43.

[85] Oh HJ, Myaeng SH, Lee MH. A practical hypertext categorization method using links and incrementally available class information. In: Belkin NJ, Ingwersen P, Leong MK, eds. Proc. of the 23rd ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR-00). Athens: ACM Press, 2000. 264-271.

[86] Yang Y, Slattery S, Ghani R. A study of approaches to hypertext categorization. Journal of Intelligent Information Systems, 2002, 18(2-3):219-241.

[87] Glover EJ, Tsioutsiouliklis K, Lawrence S, Pennock DM, Flake GW. Using web structure for classifying and describing Web pages. In: Proc. of the Int'l Conf. on the World Wide Web (WWW-2002). Honolulu: ACM Press, 2002. 562-569.

[88] Furnkranz J. Exploiting structural information for text classification on the WWW. In: Hand DJ, Kok JN, Berthold MR, eds. Proc. of the Advances in Intelligent Data Analysis. Springer-Verlag, 1999. 487-497.

[89] Kan MY, Thi HON. Fast Webpage classification using URL features. In: Otthein H, Hans JS, Norbert F, Abdur C, Wilfried T, eds. Proc. of the 14th ACM Conf. on Information and Knowledge Management (CIKM-05). Bremen: ACM Press, 2005. 325-326.

[90] Shih LK, Karger DR. Using URLs and table layout for Web classification tasks. In: Feldman SI, Uretsky M, Najork M, Wills CE, eds. Proc. of the 13th Int'l Conf. on the World Wide Web (WWW-2004). New York: ACM Press, 2004. 193-202.

[91] Chakrabarti S, Dom BE, Indyk P. Enhanced hypertext categorization using hyperlinks. In: Haas LM, Tiwary A, eds. Proc. of the ACM Int'l Conf. on Management of Data (SIGMOD-98). Seattle: ACM Press, 1998. 307-318.

附中文参考文献:

[26] 王建会,王洪伟,申展,胡运发.一种实用高效的文本分类算法.计算机研究与发展,2005,42(1):85?93.

[34] 李陆荣,王建会,陈晓芸,陶晓鹏,胡运发.使用最大熵模型进行中文文本分类.计算机研究与发展,2005,42(1):94?101.

[69] 陈毅松,汪国平,董士海.基于支持向量机的渐进直推式分类学习.软件学报,2003,14(3):451?460. http://www.jos.org.cn/1000-9825/14/451.htm