Digital Text Categorization Based on Directional Term Structures

Chia-Chuan Wu and Jau-Ji Shen*

Department of Management Information Systems,
National Chung Hsing University,
Taichung 402, Taiwan, ROC
jjshen@nchu.edu.tw

Received 10 July 2011; Revised 15 August 2011; Accepted 10 October 2011

Abstract. A rule learning classifier is capable of identifying potentially interesting patterns from training documents to establish classification rules. There have been a number of related studies for classifiers that utilize the characteristics of readable rules, such as association rule-based techniques and decision trees. These rule-based studies do not consider the structures of the terms in the documents. However, we believe that such structures may help reveal the core themes of the documents. Hence, this paper presents a new concept: Meaningful Inner Link Object (MILO). MILO classifies documents by finding the underlying directional link formed by the terms shared between different paragraphs. Moreover, a hierarchical classification structure, which considers the similarity between categories, is presented to improve the accuracy of the classification. The experimental results show that MILO is quite competitive when compared against other state-of-the-art classification techniques and may even surpass them in certain cases.

Keywords: Automatic text categorization; rule-based; term distribution

References

[1] F. Sebastiani, “Machine Learning in Automated Text categorization,” *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1-47, 2002.

[2] R. J. Quinlan, *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*: Morgan Kaufmann, 1993.

[3] A. McCallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification,” in *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, Madison, Wisconsin, USA, pp. 41-48, 1998.

[4] D. D. Lewis, “Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval,” in *Proceedings of ECML-98*, *Lecture Notes in Computer Science*, Chemnitz, Germany, pp. 4-15, 1998.

[5] T. M. Cover and J. A. Thomas, Elements of Information Theory, Wiley-Interscience, New York, NY, USA, 1991.

[6] E. H. Han, G. Karypis, V. Kumar, “Text Categorization Using Weight Adjusted K-nearest Neighbor Classification,” in *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Hong Kong, China, pp. 53-65, 2001.

[7] R. L. Liu, “Dynamic Category Profiling for Text Filtering and Classification,” *Information Processing & Management*, Vol. 43, No. 1, pp. 154-168, 2007.

[8] G. Escudero, L. Marquez, G. Rigau, “Boosting Applied toe Word Sense Disambiguation,” in *Proceedings of the 11th European Conference on Machine Learning*, Barcelona, Catalonia, Spain, pp. 129-141, 2000.

[9] H. Avancini, A. Lavelli, F. Sebastiani, R. Zanoli, “Automatic Expansion of Domain-specific Lexicons by Term Categorization,” *ACM Transaction on Speech Language and Processing*, Vol. 3, No. 1, pp. 1-30, 2006.

*Correspondence author

[10] C. Liang, L. Guo, Z.J. Xia, F.G. Nie, X.X. Li, L. Su, Z.Y. Yang, "Dictionary-based Text Categorization of Chemical Web Pages," *Information Processing & Management*, Vol. 42, No. 4, pp. 1017-1029, 2006.

[11] Y.M. Chung and Y.H. Noh, "Developing a Specialized Directory System by Automatically Classifying Web Documents," *Journal of Information Science*, Vol. 29, No. 2, pp. 117-126, 2003.

[12] Y. Guo, Z. Shao, N. Hua, "Automatic Text Categorization based on Content Analysis with Cognitive Situation Models," *Information Sciences*, Vol. 180, No. 5, pp. 613-630, 2010.

[13] T. A. van Dijk and W. Kintsch, Strategies of discourse comprehension, New York, Academic Press, 1983.

[14] R. A. Zwaan and G. A. Radvansky, "Situation Models in Language Comprehension and Memory," *Psychological Bulletin*, Vol. 123, No. 2, pp. 162-185, 1998.

[15] A. Glenberg, M. Meyer, K. Lindem, "Mental Models Contribute to Foregrounding during Text Comprehension," *Journal of Memory and Language*, Vol. 26, No. 1, pp. 69-83, 1987.

[16] T. Li, S. Zhu, M. Ogihara, "Text Categorization via Generalized Discriminant Analysis," *Information Processing & Management*, Vol. 44, No. 5, pp. 1684-1697, 2008.

[17] G. Miller and C. Fellbaum, "Semantic Networks of English," *Cognition*, Vol. 41, No. 1-3, pp. 197-229, 1991.

[18] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, "WordNet: An on-line Lexical Database," *International Journal of Lexicography*, Vol. 3, pp. 235-244, 1990.

[19] P. Rullo, V. L. Policicchio, C. Cumbo, S. Iiritano, "Olex: Effective Rule Learning for Text Categorization," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 8, pp. 1118-1132, 2009.

[20] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, pp. 487-499, 1994.

[21] M. L. Antonie and O. R. Zaiane, "Text Document Categorization by Term Association," in *Proceedings of 2002 IEEE International Conference on Data Mining*, Maebashi city, Japan, pp. 19-26, 2002.

[22] J.J. Shen and C.C. Wu, "Meaningful Inner Link Objects for Automatic Text Categorization," in *Proceedings of the 15th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Kyoto, Japan, pp. 266-269, 2009.

[23] D. Lewis, "Reuters-21578 text collection, http://www.daviddlewis.com/resources/testcollections/reuters21578/

[24] W. N. Francis and H. Kučera, "Manual of Information to Accompany a Standard Corpus of Present-day Edited American English, for Use with Digital Computers," *Department of Linguistics*, *Brown University*, 1964.

[25] J. Feng, H. Liu, J. Zou, "SAT-MOD: Moderate Itemset Fittest for Text Classification," in *Proceedings of Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, Chiba, Japan, pp. 1054-1055, 2005.

[26] J. Wang and G. Karypis, "On Mining Instance-Centric Classification Rules," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 11, pp. 1497-1511, 2006.

[27] T. Qian, H. Xiong, Y. Wang, E. Chen, "On the Strength of Hyperclique Patterns for Text Categorization," *Information Sciences*, Vol. 177, No. 19, pp. 4040-4058, 2007.

[28] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features " in *Proceedings of the 10th European Conference on Machine Learning*, Dorint-Parkhotel, Chemnitz, Germany, pp. 137-142, 1998.

[29] C. Fox, "A Stop List for General Text," *SIGIR Forum*, Vol. 24, No. 1-2, pp. 19-21, 1989.

[30] M. F. Porter, “An Algorithm for Suffix Stripping,” Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp. 313-316, 1997.

[31] Y. Yang and J. O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization,” in *Proceedings of the 14th International Conference on Machine Learning*, Nashville, Tennessee, USA, pp. 412-420, 1997.

[32] S. Dumais, J. Platt, D. Heckerman, M. Sahami, “Inductive Learning Algorithms and Representations for Text Categorization,” in *Proceedings of the 7th International Conference on Information and Knowledge Management*, Bethesda, Maryland, USA, pp. 148-155, 1998.

[33] C. Apte, F. Damerau, S. M. Weiss, “Towards Language Independent Automated Learning of Text Categorization Models,” in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, pp. 23-30, 1994.

[34] R. Bekkerman, R. El-Yaniv, Y. Winter, N. Tishby, “On Feature Distributional Clustering for Text Categorization,” in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, USA, pp. 146-153, 2001.

[35] L. Larkey and B. Croft, “Combining Classifiers in Text Categorization,” in *Proceedings of 19th ACM International Conference on Research and Development in Information Retrieval*, New York, USA, pp. 289-297, 1996.

[36] J. Han, J. Pei, Y. Yin, R. Mao, “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach,” *Data Min. Knowl. Discov.*, Vol. 8, No. 1, pp. 53-87, 2004.

[37] X.B. Xue and Z.H. Zhou, “Distributional Features for Text Categorization,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 1, pp. 428-442, 2009.