IIM> Vol.4 No.6, November 2012

OPEN ACCESS

Low-Power Themes Classifier (LPTC): A Human-Expert-Based Approach for Classification of Scientific Papers/Theses with Low-Power Theme

PDF (Size:3949KB) PP. 364-382 DOI: 10.4236/iim.2012.46041

Author(s)

Mohsen Abasi, Mohammad Bagher Ghaznavi-Ghoushchi

ABSTRACT

Document classification is widely applied in many scientific areas and academic environments, using NLP techniques and term extraction algorithms like CValue, TfIdf, TermEx, GlossEx, Weirdness and the others like. Nevertheless, they mainly have weaknesses in extracting most important terms when input text has not been rectified grammatically, or even has non-alphabetic methodical and math or chemical notations, and cross-domain inference of terms and phrases. In this paper, we propose a novel Text-Categorization and Term-Extraction method based on human-expert choice of classified categories. Papers are the training phase substances of the proposed algorithm. They have been already labeled with some scientific pre-defined field specific categories, by a human expert, especially one with high experiences and researches and surveys in the field. Our approach thereafter extracts (concept) terms of the labeled papers of each category and assigns all to the category. Categorization of test papers is then applied based on their extracted terms and further comparing with each category ' s terms. Besides, our approach will produce semantic enabled outputs that are useful for many goals such as knowledge bases and data sets complement of the Linked Data cloud and for semantic querying of them by some languages such as SparQL. Besides, further finding classified papers ' gained topic or class will be easy by using URIs contained in the ontological outputs. The experimental results, comparing LPTC with five well-known term extraction algorithms by measuring precision and recall, show that categorization effectiveness can be achieved using our approach. In other words, the method LPTC is significantly superior to CValue, TfIdf, TermEx, GlossEx and Weirdness in the target study. As well, we conclude that higher number of papers for training, even higher precision we have.

KEYWORDS

Natural Language Processing (NLP); Semantic Web; Term Extraction; Text Categorization; Resource Description Framework (RDF); Low-Power Theme

Cite this paper

M. Abasi and M. Ghaznavi-Ghoushchi, "Low-Power Themes Classifier (LPTC): A Human-Expert-Based Approach for Classification of Scientific Papers/Theses with Low-Power Theme," *Intelligent Information Management*, Vol. 4 No. 6, 2012, pp. 364-382. doi: 10.4236/iim.2012.46041.

References

[1] A. P. Chandrakasan and R. W. Brodersen, "Low Power Digital CMOS Design," Norwell, Kluwer, 1995. doi:10.1007/978-1-4615-2325-3

[2] S. Aseervatham and Y. Bennani, "Semi-Structured Document Categorization with a Semantic Kernel," Pattern Recognition, Vol. 42, No. 9, 2009, pp. 2067-2076. doi:10.1016/j.patcog.2008.10.024

[3] N. Fuhr, S. Hartmann, G. Knorz, G. Lustig, M. Schwantner and K. Tzeras, "AIR/X-a Rule Based Multistage Indexing System for Large Subject Fields," Proceedings of Riao ' 91, 1991, pp. 606-623.

[4] Y. Yang and C. G. Chute, "An Example-Based Mapping Method for Text Categorization and Retrieval," ACM Transactions on Information Systems (TOIS), Vol. 12, No. 3, 1994, pp. 252-277.

• Open Special Issues
• Published Special Issues
• Special Issues Guideline

IIM Subscription

Most popular papers in IIM

About IIM News

Frequently Asked Questions

Recommend to Peers

Recommend to Library

Contact Us

doi:10.1145/183422.183424

[5] D. D. Lewis and M. Ringuette, "A Comparison of Two Learning Algorithms for Text Categorization," Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, 11-13 April 1994, pp. 81-93.

[6] K. Tzeras and S. Hartmann, "Automatic Indexing Based on Bayesian Inference Networks," ACM Press, New York City, 1993, pp. 22-35.

[7] E. Wiener, J. Pedersen and A. Weigend, "A Neural Network Approach to Topic Spotting," Proceedings of SDAIR- 95, 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, 1995, pp. 317-332.

[8] W. W. Cohen and Y. Singer, "Context-Sensitive Learning Methods for Text Categorization," ACM Transactions on Information Systems (TOIS), Vol. 17, No. 2, 1999, pp. 141-173. doi:10.1145/306686.306688

[9] D. D. Lewis, R. E. Schapire, J. P. Callan and R. Papka, "Training Algorithms for Linear Text Classifiers," ACM Press, New York City, 1996, pp. 298-306.

[10] C. Apté, F. Damerau and S. M. Weiss, "Towards Language Independent Automated Learning of Text Categorization Models," Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Springer-Verlag, New York, 1994, pp. 23-30.

[11] I. Moulinier, G. Raskinis and J. G. Ganascia, "Text Categorization: A Symbolic Approach," Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, April 1996, pp. 87-99.

[12] R. H. Creecy, B. M. Masand, S. J. Smith and D. L. Waltz, "Trading MIPS and Memory for Knowledge Engineering," Communications of the ACM, Vol. 35, No. 8, 1992, pp. 48-64. doi:10.1145/135226.135228

[13] Y. Yang, "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval," Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Springer-Verlag, New York, 1994, pp. 13-22.

[14] C. W. Hsu and C. J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," IEEE Transactions on Neural Networks, Vol. 13, No. 2, 2002, pp. 415-425. doi:10.1109/72.991427

[15] C. Staelin, "Parameter Selection for Support Vector Machines," Hewlett-Packard Company, Palo Alto, 2003.

[16] L. H. Lee, C. H. Wan, R. Rajkumar and D. Isa, "An Enhanced Support Vector Machine Classification Framework by Using Euclidean Distance Function for Text Document Categorization," Applied Intelligence, Vol. 37, 2012, pp. 80-99.

[17] T. Y. Wang and H. M. Chiang, "One-Against-One Fuzzy Support Vector Machine Classifier: An Approach to Text Categorization," Expert Systems with Applications, Vol. 36, No. 6, 2009 pp. 10030-10034. doi:10.1016/j.eswa.2009.01.025

[18] Q. Luo, E. Chen and H. Xiong, "A Semantic Term Weighting Scheme for Text Categorization," Expert Systems with Applications, Vol. 38, No. 10, 2011, pp. 12708- 12716. doi:10.1016/j.eswa.2011.04.058

[19] Z. Li, Z. Xiong, Y. Zhang, C. Liu and K. Li, "Fast Text Categorization Using Concise Semantic Analysis," Pattern Recognition Letters, Vol. 32, No. 3, 2011, pp. 441- 448. doi:10.1016/j.patrec.2010.11.001

[20] B. Yu, Z.-B. Xu and C.-H. Li, "Latent Semantic Analysis for Text Categorization Using Neural Network," Knowledge-Based Systems, Vol. 21, No. 8, 2008, pp. 900-904. doi:10.1016/j.knosys.2008.03.045

[21] W. J. Wilbur and K. Sirotkin, "The Automatic Identification of Stop Words," Journal of Information Science, Vol. 18, No. 1, 1992, pp. 45-55. doi:10.1177/016555159201800106

[22] F. Zarrinkalam and M. Kahani, "A New Metric for Measuring Relatedness of Scientific Papers Based on Non-Textual Features," Intelligent Information Management, Vol. 4, No. 4, 2012, pp. 99-107. doi:10.4236/iim.2012.44016

[23] D. M. Bikel, S. Miller, R. Schwartz and R. Weischedel, “Nymble: A High-Performance Learning Name-Finder,” Proceedings of the 5th Conference on Applied Natural Language Processing, Washington DC, April 1997, pp. 194-201.

[24] A. Borthwick, J. Sterling, E. Agichtein and R. Grishman, “Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition,” Proceedings of the 6th Workshop on very Large Corpora, 1998, pp. 152-160.

[25] E. Miller, “An Introduction to the Resource Description Framework,” Bulletin of the American Society for Information Science and Technology, Vol. 25, No. 1, 1998, pp. 15-19. doi:10.1002/bult.105

[26] J. Golbeck and M. Rothstein, “Linking Social Networks on the Web with Foaf: A Semantic Web Case Study,” Proceedings of the 23rd national Conference on Artificial Intelligence—Vol. 2, AAAI Press, Chicago, 2008, pp. 1138- 1143.

[27] T. R. Gruber, “A Translation Approach to Portable Ontology Specifications,” Knowledge Acquisition, Vol. 5, No. 2, 1993, pp. 199-220. doi:10.1006/knac.1993.1008

[28] M. D ’ Aquin and N. F. Noy, “Where to Publish and Find Ontologies? A Survey of Ontology Libraries,” Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 11, 2011, pp. 96-111. doi:10.1016/j.websem.2011.08.005

[29] M. Ley, “The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives String Processing and Information Retrieval,” Springer, Berlin and Heidelberg, 2002.

[30] K. Frantzi, S. Ananiadou and H. Mima, “Automatic Recognition of Multi-Word Terms: The C-value/NC-Value Method,” International Journal on Digital Libraries, Vol. 3, No. 2, 2000, pp. 115-130. doi:10.1007/s007999900023

[31] D. A. Evans and R. G. Lefferts, “Clarit-Trec Experiments,” Information Processing & Management, Vol. 31, No. 3, 1995, pp. 385-395. doi:10.1016/0306-4573(94)00054-7