A general fuzzy-based framework for text representation and its application to text categorization

Doan S., Ha Q.-T., Horiguchi S.

Graduate School of Information Science, Tohoku University, Aoba 09, Sendai, 980-8579, Japan; College of Technology, Vietnam National University, Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam

Abstract: In this paper we develop the general framework for text representation based on fuzzy set theory. This work is extended from our original ideas [5],[4], in which a document is represented by a set of fuzzy concepts. The importance degree of these fuzzy concepts characterize the semantics of documents and can be calculated by a specified aggregation function of index terms. Based on this representation, a general framework is proposed and applied to text categorization problem. An algorithm is given in detail for choosing fuzzy concepts. Experiments on the real-world data set show that the proposed method is superior to the conventional method for text representation in text categorization. ?? Springer-Verlag Berlin Heidelberg 2006.

Index Keywords: Algorithms; Computational methods; Data structures; Functions; Fuzzy sets; Semantics; Data sets; Documents; Text categorization; Text processing

Year: 2006

Source title: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

Volume: 4223 LNAI

Page : 611-620

Link: Scorpus Link

Correspondence Address: Doan, S.; Graduate School of Information Science, Tohoku University, Aoba 09, Sendai, 980-8579, Japan; email: s-doan@ecei.tohoku.ac.jp

Conference name: 3rd International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2006

Conference date: 24 September 2006 through 28 September 2006

Conference location: Xi'an

Conference code: 68440

ISSN: 3029743

ISBN: 3540459162; 9783540459163

Language of Original Document: English

Abbreviated Source Title: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

Document Type: Conference Paper

Source: Scopus

Authors with affiliations:

1. Doan, S., Graduate School of Information Science, Tohoku University, Aoba 09, Sendai, 980-8579, Japan

2. Ha, Q.-T., College of Technology, Vietnam National University, Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam
3. Horiguchi, S., Graduate School of Information Science, Tohoku University, Aoba 09, Sendai, 980-8579, Japan

References:

1. Billhardt, H., Bonajo, D., Maojo, V., A context vector model for information retrieval (2002) Journal of the American Society for Information Science and Technology (JASIST), 53 (3), pp. 236-249
2. Buell, D.A., An analysys of some fuzzy subsets application to information retrieval systems (1982) Fuzzy Sets and Systems, 7 (1), pp. 35-42
3. Deerwester, S., Furnas, G.W., Dumais, S., Landauer, T.K., Indexing by latent semantic indexing (1990) Journal of the American Society for Information Science and Technology (JASIST), 41 (6), pp. 391-407
4. Doan, S., A fuzzy-based approach to text representation in text categorization (2005) Proceeding of 14th IEEE Int'l Conference Onn Fuzzy Systems - FUZZ-IEEE 2005, pp. 1008-1013. , Nevada, U.S
5. Doan, S., Horiguchi, S., A new text representation using fuzzy concepts in text categorization (2002) Proceeding of 1st International Conference on Fuzzy Set and Knowledge Discovery (FSKD), 2, pp. 514-518. , Singapore
6. 20newsgroups Dataset, , http://www.cs.cmu.edu/~textlearning
7. Joachims, T., Text categorization with support vector machines: Learning with many relevant features (1998) Proceedings 10th European Conference on Machine Learning(ECML), pp. 137-142
8. Lewis, D., (1991) Representation and Learning in Information Retrieval, , PhD thesis, Graduate School of the University of Massachusetts
9. Lucarella, D., Marara, R., First: Fuzzy informatioon retrieval system (1991) Journal of Information Science, 17 (2), pp. 81-91
10. Manning, C.D., Schutze, H., (1999) Foundations of Statistical Natural Language Processing, , The MIT Press
11. Miyamoto, S., (1990) Fuzzy Sets in Information Retrieval and Cluster Analysis, , Kluwer Academic Publishers
12. Molinari, A., Pasi, G., A fuzzy representation of html document for information retrieval system (1996) Proceeding of 5th IEEE Int't Conference on Fuzzy Systems, pp. 107-112
13. Moulinier, I., A framework for comparing text categorization approaches (1996) AAAI Symposium on Machine Learning and Information Access, , Stanford University
14. Moulinier, I., Ganascia, J.G., Applying an existing machine learning algorithm to text categorization (1996) Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing, pp. 343-354. , S. Wermter, E. Riloff, and G. Schaler, editors. Springer-Verlag, Heidelbeg
15. Murai, T., Miyakoshi, M., Shimbo, M., A fuzzy document retrieval method based on two-valued indexing (1989) Fuzzy Sets and Systems, 30 (2), pp. 103-120
16. Rocchio, J., Relevance feedback in information retrieval (1971) The SMART Retrieval System: Experiments on Automatic Document Processing, pp. 313-323. , G. Salton, editor, chapter 14. Prentice Hall
17. Salton, G., Buckley, C., Term weighting approaches in automatic text retrieval (1988) Information Processing and Management, 24 (5), pp. 513-523
18. Salton, G., Wong, A., Yang, C.S., A vector space model for automatic indexing (1975) Communications of the ACM, 18 (11), pp. 613-620
19. Sebastiani, F., Machine learning in automated text categorization (2002) ACM Computing Survey, 34 (1), pp. 1-47
20. Sparck-Jones, K.A., A statistical interpretation of term specifility and its application in retrieval (1972) Journal of Documentation, 28 (1), pp. 11-20
21. Witte, R., Bergler, S., Fuzzy coreference resolution for summarization (2003) Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS), pp. 43-50. , http://rene-

witte.net, Venice, Italy, June 23-24 Universit? Ca' Foscari

22. Yang, Y., An evaluation of statistical approaches to text categorization (1999) Information Retrieval Journal, 1, pp. 69-90
23. Yang, Y., Pedersen, J.O., A comparative study on feature selection in text categorization (1997) Proceeding of the 14th International Conference on Machine Learning (ICML97), pp. 412-420
24. Zadeh, L.A., Fuzzy sets (1965) Information Control, 8, pp. 338-353