Vietnamese spam detection based on language classification

Nguyen T.A., Tran Q.A., Nguyen N.B.
Library and Information Network Center, Hanoi University of Technology, Hanoi, Viet Nam; Faculty of Information Technology, Hanoi University, Hanoi, Viet Nam; College of Technology, Vietnam National University Hanoi, Hanoi, Viet Nam

Abstract: Language classification is the process of identifying the disposition of a presented text, such as classifying an email or a text document into a particular category. Classifying text can involve determining the genre of a book, categorizing a document, or in our case deciding whether an email is spam. The idea behind language classification is to teach the computer to be a filing clerk. Spam filters using a Bayesian combination of the spam probabilities of individual words that employ language classification read and filter your email by learning your personal email behavior (what you think is and isn't spam). There are many spam filters written based on this technology and applied effectively for English and other languages. But they got a low effect when applied directly at Vietnamese spam. Because the token segmentation of the Bayesian filters is not suitable for Vietnamese specific characteristics. We, therefore, propose a Vietnamese segmentation for using token selection for building a Vietnamese spam filter based on language classification and Bayesian combination to sufficiently support Vietnamese. The result is very satisfactory. Thanks to this technique, our filter for Vietnamese spam is 9% more accurate when compared to other filters which use other segmentation technical. ©2008 IEEE.
Author Keywords: Bayesian antispam; Language classification; Spam; Vietnamese segmentation
Index Keywords: Bayesian networks; Classification (of information); Computer networks; Electronic mail; Information retrieval systems; Linguistics; Query languages; Security of data; Text processing; Wave filters; Bayesian antispam; Language classification; Spam; Spam filtering; Vietnamese segmentation; Internet

Year: 2008
Source title: HUT-ICCE 2008 - 2nd International Conference on Communications and Electronics
Art. No.: 4578936
Page : 74-79
Link: Scorpus Link
Correspondence Address: Nguyen, T. A.; Library and Information Network Center, Hanoi University of Technology, Hanoi, Viet Nam; email: anhnt-linc@mail.hut.edu.vn
Conference name: HUT-ICCE 2008 - 2nd International Conference on Communications and Electronics
Conference date: 4 June 2008 through 6 June 2008
Conference location: Hoi an
Conference code: 73510
ISBN: 9.78E+12
Language of Original Document: English
Abbreviated Source Title: HUT-ICCE 2008 - 2nd International Conference on Communications and Electronics

Document Type: Conference Paper

Source: Scopus

Authors with affiliations:

- Nguyen, T.A., Library and Information Network Center, Hanoi University of Technology, Hanoi, Viet Nam
- Tran, Q.A., Faculty of Information Technology, Hanoi University, Hanoi, Viet Nam
- Nguyen, N.B., College of Technology, Vietnam National University Hanoi, Hanoi, Viet Nam

References:

- Graham, P., (2002) A plan for spam, , http://www.paulgraham.com/spam.html, Web document, URL
- Zdziarski, J.A., (2005) Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification, , No Starch Press
- Salton, G., Wong, A., Yang, C.S., A Vector Space Model for Automatic Indexing (1975) Communications of the ACM, 18 (11), pp. 613-620
- Tran, Q.-A., Duan, H., Li, X., Real-time statistical rules for spam detection (2006) IJCSNS International Journal of Computer Science and Network Security, 6 (2 B), pp. 178-184. , February
- Gary Robinson, A., Statistical Approach to the Spam Problem (2003) Linux Journal, 2003 (107). , http://www.linuxjournal.com/article/6467, March, URL
- Androutsopoulos I., Koutsias, J., Chandrinos, K.V., Paliouras, G., Spyropoulos, C.D., 2000. An evaluation of Naive Bayesian anti-Spam filtering. Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona, SpainCormac O'Brien & Carl Vogel,. 2003. Spam Filters: Bayes vs. Chisquared
- Letters vs. Words. International Symposium on Information and Communication Technologies. Markus Aleksy, et al. (Eds). pp. pp. 298-303Mehran Sahami, Susan Dumais, David Heckerman, Eric Horvitz (1998),9A Bayesian Approach to Filtering Junk E-Mail, In Learning For Text Categorization , Papers form AAAI Workshop. Madison Wisconsin. AAAI Technical Report WS-98-05Dien, D., Kiem, H., Van Toan, N., Vietnamese Word Segmentation (2001) The Sixth Natural Language Processing Pacific Rim Symposium, pp. 749-756. , Tokyo, Japan, pp
- Phan, T.-L., (2004) Tolerance Rough Set Model in Vietnamese text processing, , Bachelor's Thesis. Faculty of Information Technology, Ha Noi University of Technology
- Yerazunis, B., (2002) Better Than Human, , http://www.paulgraham.com/wsy.html, Web document, URL
- Anti Spam SMTP Proxy (ASSP), , http://assp.sourceforge.net