Author profiling for Vietnamese blogs

Pham D.D., Tran G.B., Pham S.B.

Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam

Abstract: This paper presents the first work in the task of author profiling for Vietnamese blogs. This task is important in threat identification and marketing intelligence. We have developed a Vietnamese Blog Profiling framework to automatically predict age, gender, geographic origin and occupation of weblogs' authors purely based on language use. The experiments on the blogs corpus we collected show very promising results with accuracy of around 80% across all traits. ?? 2009 IEEE.

Index Keywords: Geographic origins; Threat identification; Weblogs; Blogs; Internet; Linguistics

Year: 2009

Source title: 2009 International Conference on Asian Language Processing: Recent Advances in Asian Language Processing, IALP 2009

Art. No.: 5380763

Page : 190-194

Link: Scorpus Link

Correspondence Address: Pham, D. D.; Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam; email: dangpd@vnu.edu.vn

Conference name: 2009 International Conference on Asian Language Processing: Recent Advances in Asian Language Processing, IALP 2009

Conference date: 7 December 2009 through 9 December 2009

Conference location: Singapore

Conference code: 79727

ISBN: 9.78E+12

DOI: 10.1109/IALP.2009.47

Language of Original Document: English

Abbreviated Source Title: 2009 International Conference on Asian Language Processing: Recent Advances in Asian Language Processing, IALP 2009

Document Type: Conference Paper

Source: Scopus

Authors with affiliations:

1. Pham, D.D., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam
2. Tran, G.B., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam
3. Pham, S.B., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam

References:

1. Abbasi, A., Chen, H., Applying authorship authorship to extremist group web forum messages (2005) Homeland Security, , IEEE Intelligence System

2. Argamon, S., Koppel, M., Fine, J., Shimoni, A., Gender, genre, and writing style in formal written texts (2003) Text, 23 (3)
3. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B., Author Profiling for English Emails 10th Conference of the Pacific Association for Computational Linguistics (PACLING, 2007), 2007
4. Gill, A., Harrison, A., Oberlander, J., Interpersonality: Individual differences and interpersonal priming (2005) Proceedings of the 26th Annual Conference of the Cognitive Science Society, pp. 464-469. , Hillsdale, NJ: Lawrence Erlbaum Associates
5. Gill, A.J., (2004) Personality and Language: The Projection and Perception of Personality in Computer-mediated Communication, , Doctoral Thesis, University of Edinburgh
6. Groom, C.J., Pennebaker, J.W., (2005) The Language of Love: Sex, Sexual Orientation, and Language Use in Online Personal
7. Koppel, M., Argamon, S., Shimoni, A.R., Automatically categorizing written texts by author gender (2002) Literary and Linguistic Computing, 17 (4), pp. 401-412
8. Nowson, S., (2006) The Language of Weblogs: A Study of Genre and Individual Differences, , Doctoral Thesis, University of Edinburgh
9. Oberlander, J., Gill, A., Individual difference and implicit language: Personality, parts-of-speech and pervasiveness (2004) Proceedings of the 26th Annual Conference of the Cognitive Science Society, pp. 1035-1040. , Hillsdale, NJ: LEA
10. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G., Psychological Aspects of Natural Language Use: Our Words, Our Selves (2003) Annual Review of Psychology, 54, pp. 547-577
11. Schler, J., Koppel, M., Argamon, S., Penebaker, J., Effects of Age and Gender on Blogging (2006) AAAI Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), , AAAI Technical report SS-06-103
12. Zheng, R., Qin, Y., Huang, Z., Chen, H., Authorship ananlysis in Cybercrime Investigation. Intelligence and Security Informatics (2003) Proceedings of the IEEE International Conference on Intelligence and Security Informatics, pp. 59-73. , IEEE
13. Witten, I.H., Frank, E., (2005) Data Mining: Practical Machine Learning Tools and Techniques, , Morgan Kaufmann, San Francisco, second edition
14. Pham, D.D., Tran, B.G., Pham, S.B., A Hybrid Approach to Vietnamese Word Segmentation using Part of Speech tags IEEE International Conference on Knowledge System Engineering, Vietnam, 2009
15. Nguyen, T.M.H., Vu, X.L., Le, H.P., Using QTAG POS tagging for Vietnamese documents ICT.rda'03, Vietnam, 2003