Classification of Chinese-To-English Translated Social Network Timelines using Naïve Bayes

Xiang-Ru Yu*, Zhong-Liang Xiang*, Dae-Ki Kang**

*Computer Software Institute, Weifang University of Science & Technology, Shouguang city, Shandong Province, China

**Division of Computer and Information Engineering, Dongseo University, Busan city, South Korea

yuxiangru1119@163.com, ugoood@163.com, dkkang@dongseo.ac.kr

***Abstract*—This study proposes a method that classifies Chinese social network positive-negative comments (Weibo) using naïve Bayes algorithm trained from English social network (Twitter) corpus. We train our text classifier using Twitter corpus (in English language), and use this classifier to classify Chinese text. In the previous research, Chinese sentences are processed using Chinese word segmentation algorithms before the application of machine learning algorithm. Chinese word segmentation algorithms split Chinese sentences into a series of words since a Chinese word consists of several Chinese characters unlike English sentences. Therefore, the quality of word segmentation algorithm obviously influences the accuracy of Chinese text categorization problems. In our research, we eliminate Chinese word segmentation stage (a traditional preprocessing stage of Chinese text classification) to avoid the effect on the quality of segmentation algorithms. Instead of Chinese word segmentation processing, we translate Chinese text into English text via Google translator. Based on Twitter corpus, we directly generate a text classifier by using naïve Bayes multinomial algorithm. Finally, the text classifier classifies a new Chinese text (a Weibo text, which has been translated into English by Google translation at preprocessing stage). We conduct an experiment comparing the performance of naïve Bayes multinomial algorithm and C4.5 in terms of accuracy.**

***Keyword*—Text categorization, Classification, Naive Bayes, Multinomial model, Weibo, Micro-blog, Comment**

Xiang-Ru Yu received a science master degree in computer science at Ocean University of China in 2010 and a Bachelor of Science (BS) degree in computer science at Mudanjiang Normal University in 2003. Currently, she is a Lecturer at Weifang University of Science and Technology. Her research interests include data mining and machine learning.

Zhong-Liang Xiang is a candidate Ph.D. student in computer science at Dongseo University in South Korea. He received a science master degree in computer science at Ocean University of China in 2010 and a Bachelor of Science (BS) degree in computer science at Mudanjiang Normal University in 2003. His research interests include data mining and machine learning.

Dae-Ki Kang is a professor at Dongseo University in South Korea. He was a senior member of engineering staff at the attached Institute of Electronics and Telecommunications Research Institute in South Korea. He earned a Ph.D. in computer science from Iowa State University in 2006. His research interests include intrusion detection, security informatics, ontology learning, and relational learning. Prior to joining Iowa State, he worked at a Bay-area startup company and at the Electronics and Telecommunication Research Institute in South Korea. He received a science master degree in computer science at Sogang University in 1994 and a bachelor of engineering (BE) degree in computer science and engineering at Hanyang University in 1992.