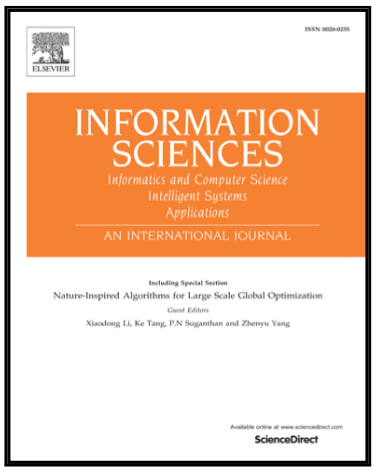

A Scalable Framework for Cross-lingual Authorship Identification

Raheem Sarwar, Qing Li, Thanawin Rakthanmanon,
Sarana Nutanong

PII: S0020-0255(18)30523-1
DOI: 10.1016/j.ins.2018.07.009
Reference: INS 13776

To appear in: *Information Sciences*

Received date: 18 September 2017
Revised date: 20 May 2018
Accepted date: 7 July 2018

Please cite this article as: Raheem Sarwar, Qing Li, Thanawin Rakthanmanon, Sarana Nutanong, A Scalable Framework for Cross-lingual Authorship Identification, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.07.009

A Scalable Framework for Cross-lingual Authorship Identification

Raheem Sarwar[a], Qing Li[a], Thanawin Rakthanmanon[b,c], Sarana Nutanong[a,*]

[a]*Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, HKSAR, China*
[b]*School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology, Thailand*
[c]*Department of Computer Engineering, Kasetsart University, Thailand*

Abstract

Cross-lingual authorship identification aims at finding the author of an anonymous document written in one language by using labeled documents written in other languages. The main challenge of cross-lingual authorship identification is that the stylistic markers (features) used in one language may not be applicable to other languages in the corpus. Existing methods overcome this challenge by using external resources such as machine translation and part-of-speech tagging. However, such solutions are not applicable to languages with poor external resources (known as low resource languages). They also fail to scale as the number of candidate authors and/or the number of languages in the corpus increases. In this investigation, we analyze different types of stylometric features and identify 10 high-performance language-independent features for cross-lingual stylometric analysis tasks. Based on these stylometric features, we propose a cross-lingual authorship identification solution that can accurately handle a large number of authors. Specifically, we partition the documents into fragments where each fragment is further decomposed into fixed size chunks. Using a multilingual corpus of 400 authors with 825 documents written in 6 different languages, we show that our method can achieve an accuracy level of

*Corresponding author
Email addresses: rsarwar2-c@my.cityu.edu.hk (Raheem Sarwar), itqli@cityu.edu.hk (Qing Li), thanawin.r@ku.ac.th (Thanawin Rakthanmanon), snutanon@cityu.edu.hk (Sarana Nutanong)