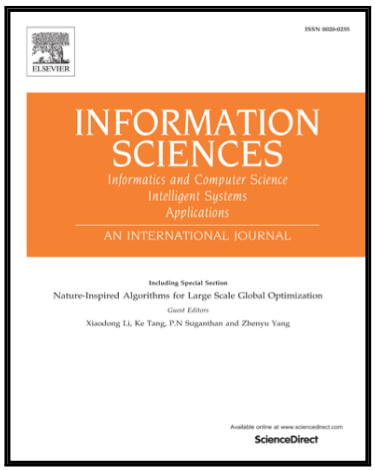

Terms-based Discriminative Information Space for Robust Text Classification

Khurum Nazir Junejo, Asim Karim, Malik Tahir Hassan, Moongu Jeon

PII: S0020-0255(16)30651-X
DOI: 10.1016/j.ins.2016.08.073
Reference: INS 12476

To appear in: *Information Sciences*

Received date: 19 June 2015
Revised date: 19 July 2016
Accepted date: 19 August 2016

Please cite this article as: Khurum Nazir Junejo, Asim Karim, Malik Tahir Hassan, Moongu Jeon, Terms-based Discriminative Information Space for Robust Text Classification, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.08.073

Terms-based Discriminative Information Space for Robust Text Classification

Khurum Nazir Junejo[a,b], Asim Karim[c], Malik Tahir Hassan[d,e], Moongu Jeon[e,*]

[a] *Singapore University of Technology and Design, Singapore*
[b] *Karachi Institute of Economics and Technology, Pakistan*
[c] *Department of Computer Science, SBASSE, Lahore University of Management Sciences, Lahore Pakistan*
[d] *University of Management and Technology, Lahore, Pakistan*
[e] *School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, South Korea*

Abstract

With the popularity of Web 2.0, there has been a phenomenal increase in the utility of text classification in applications like document filtering and sentiment categorization. Many of these applications demand that the classification method be efficient and robust, yet produce accurate categorizations by using the terms in the documents only. In this paper, we propose a novel and efficient method using terms-based discriminative information space for robust text classification. Terms in the documents are assigned weights according to the discrimination information they provide for one category over the others. These weights also serve to partition the terms into category sets. A linear opinion pool is adopted for combining the discrimination information provided by each set of terms to yield a feature space (discriminative information space) having dimensions equal to the number of classes. Subsequently, a discriminant function is learned to categorize the documents in the feature space. This classification methodology relies upon corpus information only, and is robust to distribution shifts and noise. We develop theoretical parallels of our methodology with generative, discriminative, and hybrid classifiers. We evaluate our methodology extensively with five different discriminative term weighting schemes on six data sets from different

*Corresponding author. Tel.: +82-62-715-2406. Email address: mgjeon@gist.ac.kr