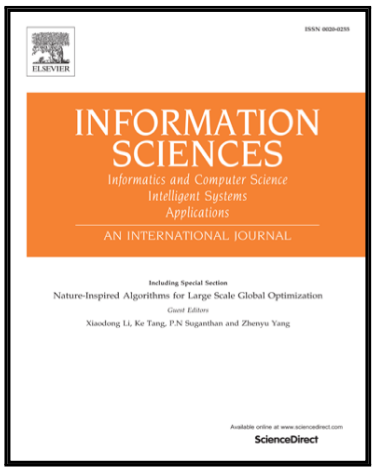

An efficient Wikipedia semantic matching approach to text document classification

Zongda Wu, Hui Zhu, Guiling Li, Zongmin Cui, Hui Huang, Jun Li, Enhong Chen, Guandong Xu

PII: S0020-0255(17)30429-2
DOI: 10.1016/j.ins.2017.02.009
Reference: INS 12732

To appear in: *Information Sciences*

Received date: 28 July 2016
Revised date: 6 January 2017
Accepted date: 3 February 2017

Please cite this article as: Zongda Wu, Hui Zhu, Guiling Li, Zongmin Cui, Hui Huang, Jun Li, Enhong Chen, Guandong Xu, An efficient Wikipedia semantic matching approach to text document classification, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.02.009

An efficient Wikipedia semantic matching approach to text document classification

Zongda Wu[a,*], Hui Zhu[b,*], Guiling Li[c], Zongmin Cui[d], Hui Huang[e], Jun Li[e], Enhong Chen[f], Guandong Xu[g]

[a] *Oujiang College, Wenzhou University, Wenzhou, Zhejiang, China*
[b] *Wenzhou Vocational College of Science and Technology, Wenzhou, Zhejiang, China*
[c] *School of Computer Science, China University of Geosciences, Wuhan, China*
[d] *School of Information Science and Technology, Jiujiang University, Jiangxi, China*
[e] *College of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou, Zhejiang, China*
[f] *School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China*
[g] *Faculty of Engineering and IT, University of Technology, Sydney, Australia*

Abstract

A traditional classification approach based on keyword matching represents each text document as a set of keywords, without considering the semantic information, thereby, reducing the accuracy of classification. To solve this problem, a new classification approach based on Wikipedia matching was proposed, which represents each document as a concept vector in the Wikipedia semantic space so as to understand the text semantics, and has been demonstrated to improve the accuracy of classification. However, the immense Wikipedia semantic space greatly reduces the generation efficiency of a concept vector, resulting in a negative impact on the availability of the approach in an online environment. In this paper, we propose an efficient Wikipedia semantic matching approach to document classification. First, we define several heuristic selection rules to quickly pick out related concepts for a document from the Wikipedia semantic space, making it no longer necessary to match all the concepts in the semantic space, thus greatly improving the generation efficiency of the concept vector. Second, based on the semantic representation of each text document, we compute the similarity between documents so as to accurately classify the documents. Finally, evaluation experiments demonstrate the effectiveness of our approach, i.e., which can improve the classification efficiency of the Wikipedia matching under the precondition of not compromising the classification accuracy.

Keywords: Wikipedia matching, keyword matching, document classification, semantics

1. Introduction

The rapid growth of online documents in the World Wide Web has raised an urgent demand for efficient and effective classification algorithms to help people achieve fast navigation and browsing of online documents [7, 17, 31]. In general, traditional document classification algorithms were developed based on keyword matching [18, 13], whose basic idea is to represent a document as a vector of weighted occurrence frequencies of individual keywords, and then analyze the relevance of keyword vectors to measure the text similarity of documents. However, keyword matching techniques only take into consideration the surface text information, and do not consider the semantic information contained in documents, resulting in problems such as semantic confusion caused by polysemy, and content mismatch caused by synonym, thus reducing the effectiveness of the techniques [12, 33, 5]. To solve this problem, a new technique called Wikipedia matching was proposed [10, 11, 3, 1], whose basic idea is to use the semantic concepts from Wikipedia as an intermediate reference space, upon which a document is mapped from a keyword vector to a concept vector, so as to capture the semantic information contained in the document. As pointed out in [11], compared to other knowledge repositories, Wikipedia has the following advantages: (1) it has broad knowledge coverage about different concepts; (2) its knowledge concepts are always in step with the times; and (3) it contains a lot of new terms that cannot be found in other repositories. All the advantages enable Wikipedia matching to overcome the semantic mismatch problem encountered in keyword matching [1] and as a result improve the accuracy of document similarity computation. Below, we use a simple example to show the superiority of Wikipedia matching over keyword matching.

ID	Document Content
Doc 1	Puma, an American Feline Resembling a Lion.
Doc 2	Puma, a Famous Sports Brand from German.
Doc 3	Welcome to Zoo, an Animal World.

Table 1: Three short text documents

Given three short text documents shown in Table 1, keyword matching would mistakenly think that the similarity between Doc 1 and Doc 2 is higher than that between Doc 1 and Doc 3, because there is the polysemous keyword 'Puma' contained in both Doc 1 and Doc 2. However, in the Wikipedia matching approach, the documents would be mapped into concept vectors in the Wikipedia reference space by using keyword matching.

*Corresponding author
Email address: zongda1983@163.com (Zongda Wu)