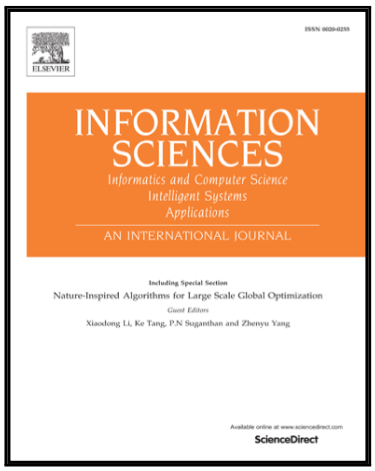

On the selection of the correct number of terms for profile construction: theoretical and empirical analysis

Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete

PII: S0020-0255(17)31098-8
DOI: 10.1016/j.ins.2017.11.034
Reference: INS 13262

To appear in: *Information Sciences*

Received date: 20 July 2017
Revised date: 2 October 2017
Accepted date: 17 November 2017

Please cite this article as: Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, On the selection of the correct number of terms for profile construction: theoretical and empirical analysis, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.11.034

On the selection of the correct number of terms for profile construction: theoretical and empirical analysis

Luis M. de Campos[a], Juan M. Fernández-Luna[a], Juan F. Huete[a,*]

[a] *Departamento de Ciencias de la Computación e Inteligencia Artificial, ETSI Informática y de Telecomunicación, CITIC-UGR, Universidad de Granada, 18071, Granada, Spain*

Abstract

In this paper, we examine the problem of building a user profile from a set of documents. This profile will consist of a subset of the most representative terms in the documents that best represent user preferences or interests. Inspired by the discrete concentration theory we have conducted an axiomatic study of seven properties that a selection function should fulfill: the minimum and maximum uncertainty principle, invariant to adding zeros, invariant to scale transformations, principle of nominal increase, transfer principle and the richest get richer inequality. We also present a novel selection function based on the use of similarity metrics, and more specifically the cosine measure which is commonly used in information retrieval, and demonstrate that this verifies six of the properties in addition to a weaker variant of the transfer principle, thereby representing a good selection approach.

The theoretical study was complemented with an empirical study to compare the performance of different selection criteria (weight- and unweight-based) using real data in a parliamentary setting. In this study, we analyze the performance of the different functions focusing on the two main factors affecting the selection process: profile size (number of terms) and weight distribution. These profiles are then used in a document filtering task to show that our similarity-based approach performs well in terms not only of recommendation accuracy but also efficiency (we obtain smaller profiles and consequently faster recommendations).

Keywords: Content analysis, Term selection, Document-based profiles, Expert search

*Corresponding author

Email addresses: lci@decsai.ugr.es (Luis M. de Campos), jmfluna@decsai.ugr.es (Juan M. Fernández-Luna), jhg@decsai.ugr.es (Juan F. Huete)