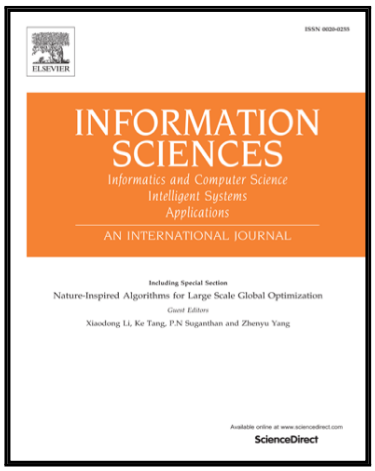

Mathematical properties of Soft Cardinality: Enhancing Jaccard, Dice and cosine similarity measures with element-wise distance

Sergio Jimenez, Fabio A. Gonzalez, Alexander Gelbukh

PII: S0020-0255(16)30437-6
DOI: 10.1016/j.ins.2016.06.012
Reference: INS 12286

To appear in: *Information Sciences*

Received date: 7 June 2015
Revised date: 29 May 2016
Accepted date: 13 June 2016

Please cite this article as: Sergio Jimenez, Fabio A. Gonzalez, Alexander Gelbukh, Mathematical properties of Soft Cardinality: Enhancing Jaccard, Dice and cosine similarity measures with element-wise distance, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.06.012

Mathematical properties of Soft Cardinality: Enhancing Jaccard, Dice and cosine similarity measures with element-wise distance

Sergio Jimenez[a], Fabio A. Gonzalez[a], Alexander Gelbukh[b]

[a] *Universidad Nacional de Colombia, Ciudad Universitaria, Bogotá D.C., Colombia*
[b] *CIC-IPN, Mexico*

Abstract

The soft cardinality function generalizes the concept of counting measure of the classic cardinality of sets. This function provides an intuitive measure of the amount of elements in a collection (i.e. a set or a bag) exploiting the similarities among them. Although soft cardinality was first proposed in an ad-hoc way, it has been successfully used in various tasks in the field of natural language processing. In this paper, a formal definition of soft cardinality is proposed together with an analysis of its boundaries, monotonicity property and a method for constructing similarity functions. Additionally, an empirical evaluation of the model was carried out using synthetic data.

Keywords: Soft cardinality, Jaccard's index, Dice's index, cosine similarity, cardinality-based similarity measures, diversity-based similarity functions

1. Introduction

Similarity measures such as Jaccard ($sim(a,b) = {}^{|A\cap B|}/_{|A\cup B|}$) [16], Dice ($sim(a,b) = {}^{2\cdot|A\cap B|}/_{|A|+|B|}$) [12] and cosine ($sim(a,b) = {}^{|A\cap B|}/_{\sqrt{|A|\cdot|B|}}$) [32] coefficients are defined as arithmetic expressions of the cardinality of the sets $|A|$, $|B|$, their union $|A \cup B|$ and intersection $|A \cap B|$. These coefficients are among the most used similarity measures in science and are subject of active research in fields such as approximate reasoning [17, 26], computer vision [41], decision making [8, 5, 38] and fuzzy sets [14]. However, this approach fails to capture important characteristics of many data such as the fact there could be a continuous degree of similarity between set's elements. Consider two texts having no words in common but sharing the same meaning by using similar words. In this case, when representing texts as sets of words, similarity measures based on standard set operations do not reflect the similarity between the texts, since these operations fail to represent word similarity. In this paper, this approach is formally extended by making it able to handle this and other scenarios where the similarities between elements induce similarities between collections that contain them.

The soft cardinality model was initially proposed (in 2008) as a convenient way to represent text similarity ([19] in Eq. 3.10, and [23]), and its effectiveness was validated by the success of the approach for addressing different natural language processing (NLP) tasks in SemEval competitions in 2012 [3, 31], 2013 [4, 30, 13], 2014 [1, 28, 25] and 2016 [2]. A survey of the approaches using soft cardinality for addressing different NLP problems in open competition can be found in [24]. Recently, a new approach called *polarized soft cardinality* extended the original model to support negative similarity values between elements for modeling opposition such as anonymy between words [20]. In this paper, the formal derivation and definition of soft cardinality are presented along with a review of some of its mathematical properties and an experimental validation over synthetic data. This provides a more sound and formal basis for a model that has proven effective in practice.

In the spirit of soft cardinality but applied to vectors, Sidorov et al. [37] proposed the *soft-cosine* measure, which is particularly similar to our approach. Moreover, Leinster and Cobbold [27] (2012) recently proposed a generalized

Email addresses: sgjimenezv@unal.edu.co (Sergio Jimenez), fagonzalezo@unal.edu.co (Fabio A. Gonzalez), gelbukh@gelbukh.com (Alexander Gelbukh)