A New Model to Compute the Information Content of Concepts from Taxonomic Knowledge

David Sánchez, Universitat Rovira i Virgili, Spain
Montserrat Batet, Universitat Rovira i Virgili, Spain

ABSTRACT

The Information Content (IC) of a concept quantifies the amount of information it provides when appearing in a context. In the past, IC used to be computed as a function of concept appearance probabilities in corpora, but corpora-dependency and data sparseness hampered results. Recently, some other authors tried to overcome previous approaches, estimating IC from the knowledge modeled in an ontology. In this paper, the authors develop this idea, by proposing a new model to compute the IC of a concept exploiting the taxonomic knowledge modeled in an ontology. In comparison with related works, their proposal aims to better capture semantic evidences found in the ontology. To test the authors’ approach, they have applied it to well-known semantic similarity measures, which were evaluated using standard benchmarks. Results show that the use of the authors’ model produces, in most cases, more accurate similarity estimations than related works.

Keywords: Computational Linguistics, Information Content, Knowledge Management, Ontologies, Semantic Similarity

INTRODUCTION

The growth of the so-called Information Society has produced an enormous amount of textual electronic information. Contrary to numerical data, which can be directly managed by means of mathematical operators, the coherent interpretation of textual data is challenging. Since textual resources consist of words or noun phrases, which refer to concepts that represent their meanings, tools and techniques are needed to enable a semantic interpretation of textual terms.

Within the area of semantics, Information Theory has played an important role. Many authors (Jiang & Conrath, 1997; Lin, 1998; Pirrò, 2009; Resnik, 1995; Sánchez & Batet, 2011; Sánchez, Batet, Valls, & Gibert, 2010) have relied on the notion of Information Content (IC) of term conceptualizations to quantify term semantics. Since the IC of a concept states the amount of information provided by the associated term, it provides a numerical measure that enables quantitative comparisons of terms, according to their underlying semantics. The assessment of semantic similarity is one of the most interesting applications: it quantifies the taxonomic resemblance between compared
terms (e.g., wrestling and boxing are similar because both represent concepts that are specializations of contact sports). Since semantic similarity is a fundamental principle by which humans organize objects, it has been extensively used in the past in a variety of natural language applications such as automatic spelling error detection and correction (Budanitsky & Hirst, 2001), word-sense disambiguation (Patwardhan, Banerjee, & Pedersen, 2003), question answering (Tapeh & Rahgozar, 2008) or synonym detection (Lin, 1998). In the field of knowledge management, assessment of semantic similarity has aided in tasks such as information extraction (Sánchez & Isern, 2011; Sánchez, Isern, & Millán, 2011), thesauri generation (Curran, 2002; Hwang & Kim, 2009), knowledge integration (Shakya, Takeda, & Wuwongse, 2009), information retrieval (Fanizzi, d’Amato, & Esposito, 2009; Hliaoutakis, Varelas, Voutsakis, Petrakis, & Milios, 2006) or ontology learning (Ramezani, 2011; Sánchez, 2010). Practical applications, such as privacy-preservation of textual data (Martínez, Sánchez, & Valls, 2012; Martínez, Sánchez, Valls, & Batet, 2012), can also be found.

From the variety of similarity paradigms available in the literature, IC-based measures are some of the most accurate and commonly used ones (Formica, 2008; Jiang & Conrath, 1997; Li, Bandar, & McLean, 2003; Resnik, 1995). These calculate the degree of resemblance between terms, as a function of the amount shared and non-shared information (i.e., IC) of their conceptualizations. Hence, a proper estimation of IC of concepts corresponding to the compared terms is crucial to enable accurate similarity assessments.

In the past, authors used to base the IC calculus on concept appearance probabilities in textual corpora (Resnik, 1995). Corpora-dependency introduced a series of problems such as data sparseness, when limited corpora was available, or scalability issues, due to the need of manual pre-processing of text (Sánchez et al., 2010). To overcome these issues, several authors have proposed to base the IC calculus on the knowledge modeled in an ontology (Sánchez, Batet, & Isern, 2011; Seco, Veale, & Hayes, 2004; Zhou, Wang, & Gu, 2008). In these works, the knowledge structure that ontologies provide have been exploited to estimate the degree of generality/concreteness of concept and, hence, to quantify their IC.

This paper reviews related works based on this last principle and proposes a new ontology-based model to estimate the IC of concepts. Compared to state of the art methods, our proposal aims to better capture the semantic features modeled in the ontology for evaluated concepts, enabling a more accurate quantification of their meaning. Our method has been applied to three classic IC-based similarity measures (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995), and it has been compared against other IC calculus models for different benchmarks. Similarity results have also been compared with other ontology-based paradigms of similarity computation (i.e., edge-counting and feature-based). Evaluation figures show that, on the one hand, our model improves similarity assessments of other IC calculus methods based either on corpora or on ontologies; on the other hand, our model enables similarity assessments that are more accurate than those achieved by other paradigms of similarity computation.

The rest of the paper is organized as follows. First, we review IC calculus models. Then, we present and formalize our proposal. Afterward we go into the details of the evaluation results, by describing the benchmarks, similarity measures and knowledge bases used to test and compare our method. A discussion of these results is also provided. The final section contains the conclusions and proposes some lines of future work.

### INFORMATION CONTENT AND CONCEPT SEMANTICS

Resnik was the first to apply the definition of IC in a semantic context. He computed the IC of a concept $c$ as the inverse of its probability of occurrence in a corpus (Resnik, 1995):

$$IC(c) = -\log p(c)$$ (1)
Related Content

Inductive Classification of Semantically Annotated Resources through Reduced Coulomb Energy Networks
www.igi-global.com/chapter/inductive-classification-semantically-annotated-resources/55051?camid=4v1a

Enabling Scalable Semantic Reasoning for Mobile Services
www.igi-global.com/chapter/enabling-scalable-semantic-reasoning-mobile/55045?camid=4v1a
Ontology Maintenance Through Semantic Text Mining: An Application for IT Governance Domain
[www.igi-global.com/chapter/ontology-maintenance-through-semantic-text-mining/196445?camid=4v1a](www.igi-global.com/chapter/ontology-maintenance-through-semantic-text-mining/196445?camid=4v1a)

Web-Based Collaborative Systems and Harvesting the Collective Intelligence in Business Organizations