Expert Systems with Applications 40 (2013) 1393–1399

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A semantic similarity method based on information content exploiting multiple ontologies

David Sánchez *, Montserrat Batet

Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Avda. Països Catalans, 26, 43007 Tarragona, Spain

ARTICLE INFO

Keywords:
Information content
Semantic similarity
Ontologies
MeSH
SNOMED CT

ABSTRACT

The quantification of the semantic similarity between terms is an important research area that configures a valuable tool for text understanding. Among the different paradigms used by related works to compute semantic similarity, in recent years, information theoretic approaches have shown promising results by computing the information content (IC) of concepts from the knowledge provided by ontologies. These approaches, however, are hampered by the coverage offered by the single input ontology. In this paper, we propose extending IC-based similarity measures by considering multiple ontologies in an integrated way. Several strategies are proposed according to which ontology the evaluated terms belong. Our proposal has been evaluated by means of a widely used benchmark of medical terms and MeSH and SNOMED CT as ontologies. Results show an improvement in the similarity assessment accuracy when multiple ontologies are considered.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

The estimation of the semantic similarity between terms contributes to the better understanding of textual resources. As a result, it has been applied in many different tasks such as word-sense disambiguation (Resnik, 1999), document categorization or clustering (Batet, 2011; Cilibrasi & Vitányi, 2006; Luo, Chen, & Xiong, 2011), word spelling correction (Budanitsky & Hirst, 2006), automatic language translation (Cilibrasi & Vitányi, 2006), ontology learning (Sánchez, 2010; Sánchez & Moreno, 2008a, 2008b; Sánchez, Moreno, & Vasto, 2012), semantic annotation (Sánchez, Isern, & Millán, 2011), information extraction (Atkinson, Ferreira, & Aravena, 2009; Sánchez & Isern, 2011), information retrieval (Al-Mubaid & Nguyen, 2006; Budanitsky & Hirst, 2006) or anonymisation of textual documents (Martínez, Sánchez, & Valls, 2012; Martínez, Sánchez, Valls, & Batet, 2012).

Semantic similarity is understood as the degree of taxonomic proximity between terms. Similarity measures assess a numerical score that quantifies this proximity as a function of the semantic evidence observed in one or several knowledge sources. Usually, those resources consist on taxonomies and more general ontologies, which provide a formal and machine-readable way to express a shared conceptualisation by means of a unified terminology and semantic inter-relations from which semantic similarity can be assessed. In the last years, general purpose ontologies have been developed (such as WordNet) but also domain-dependant one (such as MeSH or SNOMED CT for the biomedical domain).

According to the theoretical principles and the way in which ontologies are analysed to estimate similarity, different families of methods can be identified. In a nutshell, *edge-counting* measures base the similarity assessment on the number of taxonomical links of the *minimum path* separating two concepts contained in a given ontology (Leacock & Chodorow, 1998; Li, Bandar, & McLean, 2003; Rada, Mili, Bichnell, & Blettner, 1989; Wu & Palmer, 1994). Due to their simplicity, these approaches offer a limited accuracy due to ontologies model a large amount of taxonomical knowledge that is not considered during the evaluation of the minimum path (Batet, Sánchez, & Valls, 2011). *Feature-based* approaches estimate similarity according to the weighted sum of the amount of common and non-common features (Sánchez, Batet, Isern, & Valls, 2012). By features, authors usually consider taxonomic and non-taxonomic information modelled in an ontology, in addition to concept descriptions (e.g., glosses) retrieved from dictionaries (Petrakis, Varelas, Hliaoutakis, & Raftopoulou, 2006; Rodríguez & Egenhofer, 2003; Tversky, 1977). Due to the additional semantic evidences considered during the assessment, they potentially improve edge-counting approaches. However, they usually rely on non-taxonomic features that are rarely found in ontologies (Ding et al., 2004) and require fine tuning of weighting parameters in order to integrate heterogeneous semantic evidences (Petrakis et al., 2006).

Finally, *information content-based* approaches, which are the focus of this work, assess the similarity between concepts as a function of the information content (IC) that both concepts have

* Corresponding author. Tel.: +34 977559657; fax: +34 977 559710.
E-mail address: david.sanchez@urv.cat (D. Sánchez).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.08.049

in common in a given ontology. In the past, IC was typically computed from concept distribution in tagged textual corpora (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995). However, this introduces a dependency on corpora availability and manual tagging that hampered their accuracy and applicability due to data sparseness (Sánchez, Batet, Valls, & Gibert, 2010). To overcome this problem, in recent years, several authors have proposed ways to infer IC of concepts in an intrinsic manner from the knowledge structure modelled in an ontology (Seco, Veale, & Hayes, 2004; Sánchez & Batet, 2011; Sánchez, Batet, & Isern, 2011; Zhou, Wang, & Gu, 2008). However, the fact that intrinsic IC-based measures only rely on ontological knowledge is also a drawback because they completely depend on the degree of coverage and detail of the unique input ontology. This limitation could be overcome computing concept's IC and estimating semantic similarity from *multiple ontologies*. As stated in Al-Mubaid and Nguyen (2009) the exploitation of multiple ontologies provides additional knowledge that can improve the similarity estimation and solve cases in which terms are not represented in an individual ontology. This is especially interesting in domains such as the biomedical one, in which several big and detailed ontologies are available, offering overlapping and complementary knowledge about the same topics.

As it will be discussed in Section 2, few works propose similarity methods supporting more than one ontology, being all of them framed in the context of edge-counting and feature-based paradigms. In this paper we present a method to extend IC-based semantic similarity measures when multiple ontologies are available. As far as we know, no similarity methods based on IC have been proposed in the past considering more than one input ontology. The method relies on a state of the art approach to compute concept's IC from an ontology in an intrinsic manner (Sánchez et al., 2011). On one hand, our method permits estimating the similarity when a term or a term pair is missing in a certain ontology but it is found in another one. On the other hand, in case of overlapping knowledge (i.e., ontologies covering the same terms), our approach increases the accuracy by selecting the most reliable IC and similarity estimation from those computed from each individual ontology. The method has been evaluated by means of a widely used benchmark of biomedical terms and the above-mentioned biomedical ontologies. Results show that intrinsic IC measures are able to improve other similarity computation paradigms. Moreover, the exploitation of several complementary and/or overlapping ontologies during the similarity assessment was able to improve the accuracy with respect to the mono-ontology scenario.

The rest of the paper is organised as follows. Section 2 introduces related works proposing methods for semantic similarity assessment from multiple ontologies. Section 3 analyses different approaches for computing the IC of a concept, focusing on ontology-based methods. Afterwards, classic IC-based similarity measures are presented. Section 4 describes our method to exploit multiple ontologies for similarity assessment, detailing the strategies proposed to tackle the problem according to which ontology the evaluated terms belong. Section 5 evaluates our approach, comparing it to a mono-ontology scenario. The final section contains the conclusions and some lines of future research.

2. Related work

Semantic similarity estimation methods supporting multiple ontologies are based on the edge-counting and feature-based paradigms.

In Rodríguez and Egenhofer (2003), the similarity is computed as the weighted sum of similarities between synonym sets, features (e.g., meronyms, attributes, etc.) and neighbour concepts (those linked via semantic pointers) of evaluated terms. Petrakis et al. (2006) extended the previous approach relying on the matching between synonym sets and concept glosses (i.e., term definitions). They considered that two terms are similar if their synonyms and glosses and those of the concepts in their neighbourhood (following semantic relations) are lexically similar. In both approaches, when the evaluated term pair belongs to different ontologies, authors connect ontologies by a new imaginary root node that subsumes the root nodes of each ontology. Then, the similarity is computed from the resulting knowledge structure.

A problem of these approaches is the reliance on many ontological features that are rarely found in ontologies. In fact, an investigation of the structure of existing ontologies (Ding et al., 2004) has shown that ontologies very occasionally model non-taxonomic knowledge. Another problem for Rodriguez and Egenhofer's approach is its dependency on the weighting parameters that balance the contribution of each feature. These parameters should be tuned according to the nature of the ontology and the evaluated terms. This hampers the applicability as a general purpose solution. Petrakis et al.'s method does not depend on weighting parameters, because the maximum similarity provided by each feature alone is taken. Even though this adapts the behaviour of the measure to the characteristics of the ontology, the contribution of other features is omitted because only the maximum value is considered.

A more elaborated approach is presented in Sánchez, Solé-Ribalta, Batet, and Serratosa (2012). This work complements the strict matching of subsumers according to their labels with a structural similarity function that aims at discovering similar but not necessarily terminologically identical subsumers. Since only one subsumer pair is matched, the method can only be applied to path-based similarity measures.

With respect to the multi-ontology scenario, the above methods do not consider the case in which a term pair is found in several ontologies at the same time. In consequence, they omit the problem of selecting the most appropriate assessment and/or to evaluate overlapping sources of information.

A more general approach by Al-Mubaid and Nguyen (2009) propose a methodology to exploit biomedical sources (such as SNOMED CT or MeSH) using a similarity measure defined in Al-Mubaid and Nguyen (2006). This measure combines, in a weighted manner, the features *path length* and *common specificity* of the compared concepts. Authors quantify the common specificity of two concepts by subtracting the depth of their least common subsumer (LCS) from the depth of the taxonomic branch to which they belong. In this manner, concepts at a lower level of the taxonomy are considered to be more similar those located at a higher level. In Al-Mubaid and Nguyen (2009), they extended this measure when multiple input ontologies are available. In their approach, the user must select a *primary* ontology (the rest are considered as *secondary*) that acts as the master in cases in which concepts belong to several ontologies. The *primary* ontology is also used as the base to normalise similarity values. Different heuristics are proposed according to which ontologies the compared concepts belong. If both concepts appear in the *primary* ontology, the similarity is computed exclusively from that source (even if they also appear in a *secondary* ontology). When concepts appear in several *secondary* ontologies, authors evaluate the degree of overlapping with respect to the *primary* ontology and the degree of taxonomic detail (granularity). The *secondary* ontology with the highest alikeness to the *primary* one is chosen. Finally, if a concept appears in an ontology and the other concept is found in another ontology, they "connect" both ontologies by finding "common nodes" (i.e., a subsumers representing the same concepts in any of the ontologies).

A problem faced by the authors is the fact that, due to their measure is based on absolute path lengths between concepts, the similarity computed for each term pair from a different ontology will

Download English Version:

https://daneshyari.com/en/article/383045

Download Persian Version:

https://daneshyari.com/article/383045

Daneshyari.com