Information Sciences 283 (2014) 197–210

Contents lists available at ScienceDirect

Information Sciences

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

An information theoretic approach to improve semantic similarity assessments across multiple ontologies

Montserrat Batet [a,*], Sébastien Harispe [b], Sylvie Ranwez [b], David Sánchez [a], Vincent Ranwez [c]

[a] Department d'Enginyeria Informàtica i Matemàtiques, Univeritat Rovira i Virgili, Av. Paisos Catalans, 26, 43007 Tarragona, Spain
[b] LGI2P/ENSMA Research Centre, Site EERIE, Parc scientifique G. Besse, 30035 Nîmes cedex 1, France
[c] Montpellier SupAgro, UMR AGAP, 2 place Pierre Viala, 34060 Montpellier cedex 1, France

ARTICLE INFO

Article history:
Received 4 June 2013
Received in revised form 31 January 2014
Accepted 26 June 2014
Available online 8 July 2014

Keywords:
Semantic similarity
Information Theory
Ontology
MeSH
SNOMED-CT

ABSTRACT

Semantic similarity has become, in recent years, the backbone of numerous knowledge-based applications dealing with textual data. From the different methods and paradigms proposed to assess semantic similarity, ontology-based measures and, more specifically, those based on quantifying the Information Content (IC) of concepts are the most widespread solutions due to their high accuracy. However, these measures were designed to exploit a single ontology. They thus cannot be leveraged in many contexts in which multiple knowledge bases are considered. In this paper, we propose a new approach to achieve accurate IC-based similarity assessments for concept pairs spread throughout several ontologies. Based on Information Theory, our method defines a strategy to accurately measure the degree of commonality between concepts belonging to different ontologies—this is the cornerstone for estimating their semantic similarity. Our approach therefore enables classic IC-based measures to be directly applied in a multiple ontology setting. An empirical evaluation, based on well-established benchmarks and ontologies related to the biomedical domain, illustrates the accuracy of our approach, and demonstrates that similarity estimations provided by our approach are significantly more correlated with human ratings of similarity than those obtained via related works.

1. Introduction

Semantic similarity is a pillar of text understanding since it quantifies the degree of resemblance between the meanings of textual terms. In recent years, due to the marked increase in electronically available textual data, considerable effort has been focused on defining semantic measures, which have been extensively applied in various contexts such as information retrieval [54], information extraction [56], word sense disambiguation [32], data clustering [4,27], data privacy [6,45–47] and biomedicine (e.g. protein classification and interaction [19,57], chemical entity identification [18], identification of the communalities between brain data and linguistic data [15], etc.) to cite a few. Different knowledge sources have been used to facilitate the semantic similarity calculus, including: measures relying solely on textual corpora to estimate similarity from the degree of co-occurrence of terms [11], and measures involving structured knowledge bases such as ontologies [36,37,44], whereby the similarity calculus is based on the analysis of semantic relationships modelled between concepts. Compared to corpora-based approaches, ontology-based measures have a dual benefit: concept meanings can be

* Corresponding author. Tel.: +34 977559657; fax: +34 977 559710.
E-mail address: montserrat.batet@urv.cat (M. Batet).

http://dx.doi.org/10.1016/j.ins.2014.06.039
0020-0255/

unambiguously retrieved from ontologies and similarities can be assessed from structured knowledge that has been explicitly formalised by human experts.

A plethora of ontology-based measures have been proposed to estimate the similarity of two concepts belonging to a *single* ontology. They can be roughly classified as:

(i) *Edge-based measures.* These measures consider an ontology as a directed graph in which concepts are interrelated by means of semantic links. They estimate similarity according to the number of semantic links separating concept pairs [24,36,58]. However, the fact that they only consider the minimum number of taxonomic links between concept pairs limits their accuracy, because much of the knowledge modelled in the ontology is omitted (e.g. taxonomic links resulting from multiple inheritance).

(ii) *Feature-based measures.* These methods try to overcome the limitations of edge-based measures by building sets of features that describe the concepts. They estimate similarity as a function of the amount of overlapping and non-overlapping knowledge features (e.g. taxonomic ancestors, concept descriptions, etc.) between the compared concepts [34,44].

(iii) *Information Content-based measures.* These measures complement the taxonomical knowledge provided by an ontology with the quantification of the amount of information (i.e. Information Content, IC) that concepts have in common [22,25,37]. The IC of a concept is usually computed as the inverse of the probability of occurrence of that concept in a given corpus. However, in order to avoid textual ambiguity and data sparseness, this IC calculus requires from a large, heterogeneous and tagged corpora, which is not usually available. Because of these limitations, some authors intrinsically derive IC values from an ontology according the number of taxonomic descendants and/or ancestors of concepts [41,43,52,59]. These latter approaches, which constitute the focus of our work, have proved successful in mimicking human judgements of semantic similarity [17,40,43].

Numerous ontologies are currently available due to the widespread adoption of the Semantic Web paradigm. However, because of the complicated maintenance of large ontologies, knowledge modellers tend to spread knowledge through multiple interlinked domain ontologies (e.g. BioPortal, a portal dedicated to biomedical ontologies, provides up to 200 ontologies [30]). An increasing number of applications thus require multiple knowledge bases and ontologies to be taken into account. As an example, documents annotated by multiple ontologies may need to be queried in pluridisciplinary projects (e.g. biodiversity protection requires simultaneous consideration of economical, geographical and biological information) [21]. In order to use an information retrieval system for this task (e.g. OBIRS [54]), the semantic similarity between the query and the document annotations must be evaluated because associated knowledge may be spread in different ontologies. However, due to their monolithic design principles, classic similarity measures cannot be applied when concepts belong to *different* ontologies [2,7].

Such multi-ontology scenarios are common when dealing with cross-domain data (e.g. social and computer sciences) and can also occur in specific fields, such as biomedicine, in which concepts are modelled in several knowledge sources, e.g. SNOMED-CT (Systemized Nomenclature of Medical Clinical Terms) [53] and MeSH (Medical Subject Headings) [29] are knowledge bases with different scopes and purposes, but both model biomedical concepts. Indeed, many ontologies share some common knowledge even if they were initially designed for unrelated applications. They hence model complementary and overlapping aspects of a complex reality that has been split for convenience or historic reasons. Their overlaps are cornerstones for designing the multi-ontology semantic measures required in these scenarios.

Similarity measures coping with multiple ontologies have been seldom considered in the literature [2,7,38,42,51]. In the context of IC-based measures, the identification of the *Most Informative Common Ancestor* (or MICA), which represents the commonality between compared concepts, is essential for similarity assessments. Existing works based on IC [42,51] retrieve the MICA of a pair of concepts belonging to different ontologies by looking for equivalences of concept ancestors sharing the same linguistic labels (i.e. terminological matchings). These approaches are hampered by the fact that ontologies rarely model concepts in the same way or refer to them using the same label (e.g. due to synonymy) [50]. Indeed, in many cases, they either select too abstract ancestors as MICA or they cannot discover any equivalence at all because they miss suitable concepts sharing similar meanings but referred with different labels (e.g. *cancer/neoplasm*). In both cases, the concept similarity is largely underestimated.

In this paper, we propose a method that overcomes the limitations of strict terminological matching between concept ancestors. Based on the Information Theory and solely exploiting ontological knowledge, our approach measures the degree of semantic equivalence between concept ancestors belonging to different ontologies to select a MICA that is more suitable than that obtained via terminological matchings. An empirical evaluation carried on several widespread similarity benchmarks, ontologies and classic IC-based measures shows that, by means of our method, similarity assessments obtained in a multi-ontology setting are more accurate than those obtained from related works.

The rest of the paper is organised as follows. Section 2 introduces similarity measures based on IC and details different IC computational models. Section 3 reviews related works on similarity measures handling multiple ontologies and discusses their main limitations. Section 4 presents and formalises our approach. Section 5 details the evaluation protocol and discusses the results obtained for several benchmarks, ontologies and measures. The last section gives the conclusions of this work.