Journal of Biomedical Informatics 48 (2014) 38–53

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain

Sébastien Harispe [a,*], David Sánchez [b], Sylvie Ranwez [a], Stefan Janaqi [a], Jacky Montmain [a]

[a] LGI2P/EMA Research Centre, Site de Nîmes, Parc scientifique G. Besse, 30035 Nîmes cedex 1, France
[b] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Spain

ARTICLE INFO

Article history:
Received 16 July 2013
Accepted 9 November 2013
Available online 21 November 2013

Keywords:
Ontologies
Semantic similarity measures
Unifying framework
SNOMED-CT
Biomedical ontologies

ABSTRACT

Ontologies are widely adopted in the biomedical domain to characterize various resources (e.g. diseases, drugs, scientific publications) with non-ambiguous meanings. By exploiting the structured knowledge that ontologies provide, a plethora of ad hoc and domain-specific semantic similarity measures have been defined over the last years. Nevertheless, some critical questions remain: which measure should be defined/chosen for a concrete application? Are some of the, a priori different, measures indeed equivalent? In order to bring some light to these questions, we perform an in-depth analysis of existing ontology-based measures to identify the core elements of semantic similarity assessment. As a result, this paper presents a unifying framework that aims to improve the understanding of semantic measures, to highlight their equivalences and to propose bridges between their theoretical bases. By demonstrating that groups of measures are just particular instantiations of parameterized functions, we unify a large number of state-of-the-art semantic similarity measures through common expressions. The application of the proposed framework and its practical usefulness is underlined by an empirical analysis of hundreds of semantic measures in a biomedical context.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Over the last decade, considerable efforts have been made to standardize our understanding of various fields by means of ontologies, i.e. formal and explicit specifications of shared conceptualizations [1]. Ontologies enable modelling domains through sets of concepts and semantic relationships established between them. Due to the importance of knowledge representation and terminology in biology and medicine, the biomedical domain has been very prone to the definition of structured thesauri or ontologies (e.g. UMLS, SNOMED-CT, MeSH). They enable characterizing medical resources such as clinical records, diseases, genes, or even scientific articles, through unambiguous conceptualizations. To take advantage of this valuable knowledge for information retrieval and knowledge discovery, *semantic similarity measures* are used to estimate the similarity of concepts defined in ontologies and, hence, to assess the semantic proximity of the resources indexed by them.

Ontology-based semantic similarity measures compare how similar the meanings of concepts are according to the taxonomical evidences modelled in the ontology. They are used in a wide array of applications: to design information retrieval algorithms [2,3], to disambiguate texts [4,5], to suggest drug repositioning [6] and to cluster genes according to their molecular function [7], to cite a few. Semantic similarity measures are indeed critical components of many knowledge-based systems [6,8,9]. Moreover, they are nowadays receiving more attention due to the growing adoption of both Semantic Web and Linked Data paradigms [10].

A plethora of measures have been proposed over the last decades (see surveys [7,9,11]). Although some context-independent semantic similarity measures have been proposed [12–15], most measures were designed in an ad hoc manner and were expressed on the basis of domain-specific or application-oriented formalisms [8]. Therefore, most proposals related to those measures target a specific audience and fail to benefit other communities. In this way, a non-specialist can only interpret the large diversity of state-of-the-art proposals as an extensive list of measures. As a consequence, the selection of an appropriate measure for a specific usage context is a challenging task. Actually, no extensive studies enabled characterizing the large diversity of proposals, even though few seminal contributions focusing on theoretical aspects of ontology-based semantic similarity measures exist [8,16,17].

Despite the large number of contributions related to ontology-based semantic similarity measures, the understanding of their foundations is nowadays limited. For a designer/practitioner, some fundamental questions remain: Why does a measure work better than another one? How does one choose or design a measure? Is it possible to distinguish families of measures sharing specific

* Corresponding author.
E-mail address: sebastien.harispe@mines-ales.fr (S. Harispe).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.11.006

properties? How can one identify the most appropriate measures according to particular criteria?

To fill these gaps, this paper proposes an extensive study of ontology-based semantic similarity measures from which a unifying framework decomposing measures through a set of intuitive core elements is proposed.

1.1. Contributions and plan

The framework presented in this paper proposes to model, in a generic and flexible way, the core elements on which most measures available in the literature rely. Thus, particular semantic measures can be properly characterized and can directly be obtained as instantiations of the framework components. This brings new insights for the study of semantic measures:

- *Distinguishing the core elements on which measures rely*. The theoretical characterization of semantic measures helps to understand the different measure paradigms and the large diversity of expressions proposed in the state-of-the-art.
- *Unifying measures through parameterized measures*. Based on the characterization of the core elements of semantic measures, our framework enables the identification of commonalities, bridges and equivalences between exiting measures. Indeed, their design could be unified through abstract expressions, even if many of them are (i) of ad hoc nature, (ii) domain-specific or (iii) based on different theoretical principles. Expressing semantic similarity measures through parameterized expressions can therefore facilitate the detection of their common properties and the analysis of their behaviour in specific applications.
- *Selecting appropriate domain-specific measures*. Such a framework provides a systematic, theoretically-coherent and direct way to define or tune the semantic similarity assessment for particular application scenarios. Semantic similarity measures expressed through parameterized functions could therefore be used to optimize measure tuning in domain-specific applications.
- *Designing new families of semantic measures*. New measures can be easily defined due to the modularity provided by the framework. Their design can take into account (i) the elements that affect the semantic assessment the most (e.g. estimation of concept specificity) and (ii) the particularities of ontology/application to which it will be applied (e.g. the presence of multiple inheritances).
- *Identifying the crucial aspects of semantic similarity assessment*. Empirical studies could be used to highlight the core elements best impacting measures' accuracies. As a result, the framework can be used to guide research efforts towards the aspects that can improve measure performances.

Such an approach will not just benefit a single measure designed for a domain-specific application (which is, to date, the focus of most related works) but will rather result in improvements of a wide set of measures and applications.

The rest of the paper is organized as follows. Section 2 introduces the reader to ontology-based semantic similarity measures, distinguishing the various paradigms proposed for their design. In addition, this section reviews previous works regarding the unification of semantic measures. Section 3 describes the proposed framework from which state-of-the-art measures are unified, and from which new proposals can be derived. Section 4 illustrates the practical application of the framework in which semantic measures' behaviours are analysed in a biomedical scenario. Section 5 provides the conclusions as well as some lines of future work.

2. Ontology-based semantic similarity measures

This section reviews the various paradigms used for the definition of ontology-based semantic similarity measures (SSMs). Each paradigm is illustrated by a selection of proposals emphasizing the essence of the approach. We then introduce the reader to existing contributions related to the unification of SSMs.

2.1. Paradigms for semantic similarity estimation

SSMs aim at estimating the likeness of two concepts considering the taxonomical knowledge modelled in ontologies. We consider approaches measuring taxonomic distance/dissimilarity indistinctly; notice that the latter can be converted to similarities by means of a linear transformation. In this section, we present state-of-the art SSMs organized according to the various paradigms proposed for their definition.

As a running example to illustrate the study, Fig. 1 presents a snapshot of the SNOMED-CT clinical healthcare terminology [18], in which biomedical concepts are organized by taxonomic relationships. The topology of SNOMED-CT defines a partial order $\preceq$ between concepts, e.g. '*Heparin*' $\preceq$ '*Protein*' means that the concept '*Heparin*' is subsumed by the concept '*Protein*', that is, the heparin is a specific class of protein.

2.1.1. Edge-based approaches

Edge-based measures estimate the similarity of two concepts according to the strength of their interlinking in the ontology. The most usual approach considers the similarity as a function of the distance which separates the two concepts in the ontology. For instance, Rada et al. estimate the distance of two concepts u, v as the shortest-path linking them ($sp(u, v)$) [15].

$$Dist_{Rada}(u, v) = sp(u, v) \tag{1}$$

In Fig. 1, the shortest path between the concepts c_5 and c_3 is $c_5 \rightarrow c_4 \rightarrow c_3$. Leacock and Chodorow proposed a non-linear adaptation of Rada's distance to define the similarity measure Sim_{LC} [19]:

$$Sim_{LC}(u, v) = -\log\left(\frac{sp(u, v)}{2 \cdot Max\_depth}\right) \tag{2}$$

Rada's distance is here normalized by the maximal depth of the ontology, *Max_depth*, i.e. the longest of the shortest paths linking a concept to the concept which subsumes all the others (the root of the ontology, c_0 in Fig. 1).

More refined approaches propose to consider variations of the strength of the links between concepts; the deeper two linked concepts are, the stronger their semantic relationship will be

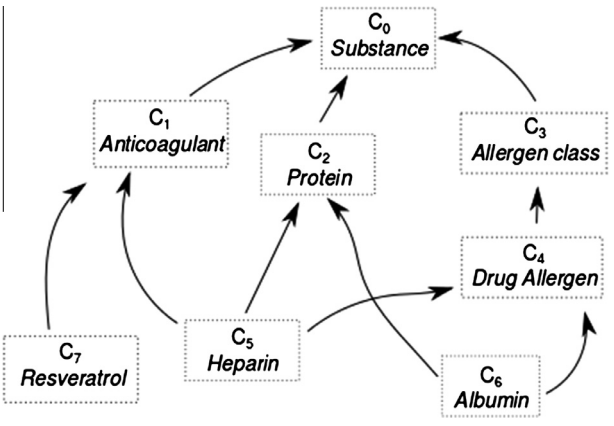

Fig. 1. Snapshot of the taxonomy of concepts defined in the SNOMED-CT.

Download English Version:

https://daneshyari.com/en/article/517112

Download Persian Version:

https://daneshyari.com/article/517112

Daneshyari.com