Web Semantics: Science, Services and Agents on the World Wide Web 8 (2010) 69–88

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

On the discovery of subsumption relations for the alignment of ontologies

Vassilis Spiliopoulos[a,*], George A. Vouros[a,*], Vangelis Karkaletsis[b]

[a] AI Lab, Information and Communication Systems Engineering Department, University of the Aegean, Karlovassi, Samos 83 200, Greece
[b] Institution of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece

ARTICLE INFO

Article history:
Received 15 March 2009
Received in revised form
22 December 2009
Accepted 4 January 2010
Available online 13 January 2010

Keywords:
Ontology alignment
Subsumption
Supervised machine learning

ABSTRACT

For the effective alignment of ontologies, the subsumption mappings between the elements of the source and target ontologies play a crucial role, as much as equivalence mappings do. This paper presents the "Classification-Based Learning of Subsumption Relations" (*CSR*) method for the alignment of ontologies. Given a pair of two ontologies, the objective of *CSR* is to learn patterns of features that provide evidence for the subsumption relation among concepts, and thus, decide whether a pair of concepts from these ontologies is related via a subsumption relation. This is achieved by means of a classification task, using state of the art supervised machine learning methods. The paper describes thoroughly the method, provides experimental results over an extended version of benchmarking series of both artificially created and real world cases, and discusses the potential of the method.

1. Introduction

Despite the fact that ontologies provide a formal and unambiguous representation of domain conceptualizations, it is rather expectable to deal with different ontologies describing the same domain of knowledge, introducing heterogeneity to the conceptualization of the domain and difficulties in integrating information.

Although many efforts [1] aim to the automatic discovery of equivalence mappings between the elements of ontologies, in this paper we conjecture that this is not enough: to deal effectively with the ontologies' alignment problem, we also have to deal with the discovery of non-equivalence mappings among ontology elements. To this end, in this work we investigate the discovery of subsumption mappings. Although the usefulness of subsumption mappings may be known to the ontology alignment community, to the best of own knowledge, no alignment method has thoroughly investigated the computation of such mappings. Therefore, the progress that has been made towards the location of subsumption mappings is not sufficient, in comparison to the progress made to the computation of equivalence mapping relations.

Subsumption mappings are particularly useful when we deal with ontologies whose conceptualizations are at different "granularity levels": in these cases, the elements (concepts or properties) of an ontology are more generic than the corresponding elements of another ontology. Although subsumption mappings between the elements of two ontologies may be deduced by exploiting equivalence mappings between other elements (e.g. a concept C_1 is subsumed by all subsumers of C_2, if C_1 is equivalent to C_2), in the extreme cases where no equivalence mappings exist, or in cases where the assessed/provided equivalences are erroneous, this cannot be done effectively. This paper conjectures that the direct discovery of subsumption relations between elements of different ontologies can enhance the discovery/filtering of equivalence relations, and vise-versa, augmenting the effectiveness of our ontology alignment and merging methods. This is of great importance, since, as it is also stated in the conclusions of the Consensus Track of OAEI 06 [2], current state of the art systems "confuse" subsumption relations with equivalence ones.

To make the above claims more concrete, let us consider the ontologies depicted in Fig. 1. These specify the concept `Citation` in the 1st ontology (which is equivalent to the concept `Reference` in the 2nd ontology), and `Publication` in the 2nd ontology (which is equivalent to the concept `Work` in the 1st ontology). Each of these ontologies elaborate on the specification of different concepts: the second ontology elaborates on the concept `Publication`, defining different kinds of publications, while the first ontology elaborates on the concept `Citation`, defining different kinds of citations. Given these ontologies, the fact that equivalent concepts in the two ontologies do not have the same lexicalization, and that non-equivalent concepts do have the same lexicalization, we may distinguish two cases.

In case that the equivalence mappings between the concepts of the two ontologies are not known, conclusions concerning subsumption mappings between the concepts of the two ontologies cannot be drawn by a reasoning mechanism. This case shows in a very clear way the necessity to discover equivalence and sub-

* Corresponding authors. Tel.: +30 6937363087; fax: +30 2273082229.
E-mail addresses: vspiliop@aegean.gr (V. Spiliopoulos), georgev@aegean.gr (G.A. Vouros), vangelis@iit.demokritos.gr (V. Karkaletsis).

1570-8268/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2010.01.001

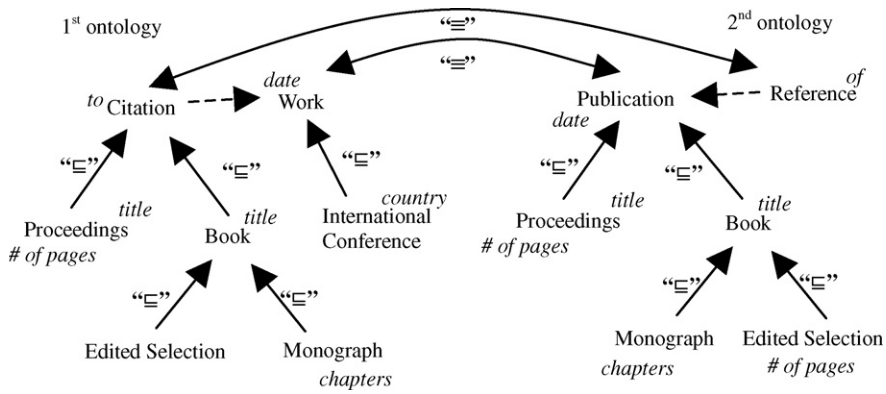

Fig. 1. Example ontologies for assessing the subsumption relation between concepts.

sumption relations between the concepts of the source and target ontologies.

In the second case where equivalence mappings between the concepts of the two ontologies are known, these can be exploited by a reasoning mechanism to deduce subsumption mappings. However, in case that equivalence mappings have been computed by an alignment mechanism, then wrong equivalences shall provide evidence to wrong subsumption mappings. For example, a state of the art alignment tool may wrongly assess that the concept `Monograph` in the 1st ontology is equivalent to the concept `Monograph` in the 2nd ontology, as their ontological features (e.g. labels, defined properties, direct super/sub concept, and depth in the taxonomy) are exactly the same, as far as their surface appearance is concerned. A reasoning mechanism exploiting this equivalence relation would wrongly deduce that the concept `Monograph` in the 1st ontology is subsumed by the concepts `Book` and `Publication` in the 2nd ontology. However, the correct relation is that the concept `Monograph` in the 1st ontology is subsumed by the concept `Reference` in the 2nd ontology.

Furthermore, even if one (human or software entity) can assess that the concept `Work` in the 1st ontology is equivalent to the concept `Publication` in the 2nd ontology, a reasoning mechanism exploiting this knowledge would correctly infer that `International Conference` is subsumed by `Publication`, but it would not be able to place `International Conference` under its direct subsumer (i.e. its correct place in the hierarchy), which in this example is the concept `Proceedings` (this is so since the concept `International Conference` represents publications that appear in the proceedings of international conferences, while `Proceedings` represent publications that appear in any kind of scientific event, e.g. workshops). This example shows that even if we exploit correct equivalences to derive subsumptions, there are cases where the subsumptions found are not sufficient for the merging of the involved ontologies.

The above examples provide evidence towards our conjecture: What is clearly needed is a method that shall discover subsumption relations between concept pairs of two distinct ontologies, separately from subsumptions and equivalences that can be deduced by a reasoning mechanism. In other words, the method should directly pursue the location of subsumption mappings, without necessarily relying on equivalence mappings.

This paper deals with the problem of discovering subsumption mappings between concepts of two distinct ontologies, without relying on known equivalence mappings among them. This is done by using the "Classification-Based Learning of Subsumption Relations" (*CSR*) method for the alignment of ontologies. *CSR* computes subsumption mappings between concept pairs of two ontologies by means of a classification task, using state of the art supervised machine learning methods. Specifically, given a pair of concepts from the source and target ontologies, the classification method "locates" a hypothesis concerning relation of concepts, which best fits to the training examples [3], while generalizing beyond them. The training examples are generated by exploiting both the source and target ontologies, without requiring human intervention (this is thoroughly explained in Section 4). The classification mechanisms proposed exploit features of concepts of different types, for the representation of concept pairs. A detailed description of the classification features used is provided in Section 4.

The basic version of *CSR* has been presented in [4]. The work presented in this article extends the one presented in [4] to the following: (a) we investigate six more different types of classification features (there are two such types in [4]), that improved the efficiency of the method in terms of precision and recall (b) we introduce a new dataset balancing technique based on the semantics of the source and target ontologies, again with a positive impact on the method, and finally (c) we provide a thorough evaluation of *CSR* using three different datasets (in contrast to the one used in [4]).

The machine learning approach has been chosen since (a) there are no evident generic rules that capture *directly* the existence of a subsumption relation between a pair of ontology elements (e.g. by means of their surface appearance, labels/vicinity similarity or dissimilarity), and (b) concept pairs of the same ontology can provide examples for the subsumption relation, making the method self-adapting to the idiosyncrasies of specific domains and conceptualizations provided, and non-dependant to external resources.

The rest of the paper is structured as follows: Section 2 states the problem and presents works that are most closely related to our approach. Section 3 provides necessary background knowledge concerning supervised machine learning and the classification methods used. Furthermore, this section provides background information concerning probabilistic topic models, which are used for the generation of classification features. Section 4 presents the proposed classification-based method for the discovery of subsumption mappings, and discusses specific choices regarding method's alternative configurations. Section 5 presents and thoroughly discusses the experimental settings, as well as the results. Finally, Section 6 concludes the paper by pointing out the main aspects of our method and sketching further work for its enhancement and exploitation.

2. Problem definition and related work

2.1. Problem definition

An ontology is a pair $O=(S, A)$, where S is the ontological signature describing the vocabulary (i.e. the terms that lexicalize ontology elements) and A is a set of ontological axioms, restricting