Knowledge-Based Systems 73 (2015) 199–211

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies

Mohammed Maree [a], Mohammed Belkhatir [b,*]

[a] Department of Multimedia Technology, Faculty of Engineering and Information Technology, The Arab American University, Palestine
[b] University of Lyon, Campus de la Doua, France

ARTICLE INFO

Article history:
Received 15 January 2013
Received in revised form 28 September 2014
Accepted 2 October 2014
Available online 12 October 2014

Keywords:
World Wide Web
Semantic heterogeneity
Ontologies
Knowledge base assisted merging
Missing background knowledge

ABSTRACT

With the development of the Semantic Web (SW), the creation of ontologies to formally conceptualize our understanding of various domains has widely increased in number. However, the conceptual and terminological differences (a.k.a *semantic heterogeneity problem*) between ontologies form a major limiting factor towards their use/reuse and full adoption in practical settings. A key solution to addressing this problem can be through identifying semantic correspondences between the entities (including concepts, relations, and instances) of heterogeneous ontologies, and consequently achieving interoperability between them. This process is also known as *ontology alignment*. The output of this process can be further exploited to merge ontologies into a single coherent ontology. Indeed, this is widely regarded as a crucial, yet difficult task, specifically when dealing with heavyweight ontologies that consist of hundreds of thousands of concepts. To address this issue, various ontology merging approaches have been proposed. These approaches can be classified into three categories: single-strategy-based approaches, multiple-strategy-based approaches, and approaches based on exploiting external semantic resources. In this paper, we first discuss the strengths and limitations of each of these approaches, and then present our framework for addressing the semantic heterogeneity problem through merging domain-specific ontologies based on multiple external semantic resources. The novelty of the proposed approach is mainly based on employing knowledge represented by multiple external resources (knowledge bases in our work) to make aggregated decisions on the semantic correspondences between the entities of heterogeneous ontologies. Other important issues that we attempt to tackle in the proposed framework are: (i) Identifying and handling inconsistency of semantic relations between the ontology concepts and, (ii) Handling the issue of missing background knowledge (such as concepts and instances) in the exploited knowledge bases by utilizing an integrated statistical and semantic technique. Additionally, the proposed solution soundly enriches the knowledge bases with missing background knowledge, and thus enables the reuse of the newly obtained knowledge in future ontology merging tasks. To validate our proposal, we tested the framework using the OAEI 2009 benchmark and compared the produced results with state-of-the-art syntactic and semantic based systems. In addition, we utilized the proposed techniques to merge three heavyweight ontologies from the environmental domain.

1. Introduction

The incorporation of semantic technology in information systems is regarded as an important issue, particularly with the development of Web 3.0. The semantics are captured in domain-specific ontologies, which aim at providing a formal, explicit and shared conceptualization and understanding of common domains between different communities [19,38]. With the advent of the Internet, which has enabled the development of an ever-increasing number of ontologies with different terminologies, it has become difficult to make use of this vast and heterogeneous source of knowledge. The difficulty of this task is due to the decentralized nature of ontology development and the differences between the viewpoints of ontology engineers. This has resulted in the so called "semantic heterogeneity" problem, which constitutes the major obstacle against achieving interoperability between ontologies.

Solving the semantic heterogeneity problem can be achieved through merging two or more ontologies from the same domain

* Corresponding author.
E-mail addresses: mohammad.maree@aauj.edu (M. Maree), belkhatir@univ-lyon1.fr (M. Belkhatir).

http://dx.doi.org/10.1016/j.knosys.2014.10.001
0950-7051/

into a single coherent ontology [34]. Several automatic and semi-automatic ontology merging approaches have been proposed. Details on the types of strategies that are used by state-of-the-art ontology merging systems are listed below:

- Strategies based on linguistic matching a.k.a. Name-based strategies: these approaches compute distances between the strings of the concepts (e.g. Jaro-Winkler distance function [43]) from the source ontologies to obtain correspondences between them [5,8]. However, they do not take the semantic aspects of the compared strings into account and therefore, errors like considering the concepts "car" and "care" as "equivalent" concepts, or considering the concepts "car" and "automobile" as "not equivalent" concepts would be produced.
- Structure-based similarity strategies: these approaches rely on the structure of the source ontologies to merge them [1,10,12,33]. Typically, the graph structure of both ontologies is provided (through is-a or other relations) and the similarity is computed for each concept based on its neighboring concepts. In this context, the neighbors of each concept are its parents (ancestors) and/or its children (descendants). However, when the number of concepts is huge such as in heavyweight ontologies, these approaches suffer from resource consumption problems as they utilize in-memory structures to merge both ontologies [45].
- Combination-based strategies: other approaches rely on the combination of name-based and structure-based similarity measures such as in the systems proposed by the authors of [10,29,30,42]. Although these techniques may produce good results, there is a considerable number of cases where they fail to discover semantic correspondences between the source ontologies due to the drawbacks of each individual strategy discussed above.
- Strategies based on external auxiliary resources: these approaches propose to integrate an external resource or knowledge base to support the task of finding semantic correspondences between heterogeneous ontologies [7,17,27,41]. However, these approaches are subjected to the limitations of the exploited knowledge base. For example, Aleksovski et al. use a background knowledge based paradigm in the medical domain where the DICE ontology acts as a semantic bridge between the matched ontologies [2]. Also, Sabou et.al. provide an approach to ontology matching which can exploit multiple heterogeneous ontologies obtained from the Web [37]. Another example is the system proposed by the authors of [46]. This system employs knowledge represented by an external resource (WordNet [32] and Web pattern-based queries to derive the semantic aspects of the entities of the source ontologies. However, the most successful and widely employed knowledge bases (e.g. WordNet, Cyc, OpenCyc, SUMO, etc.) are man-made; they suffer from low coverage, high assembly cost and fast aging whereby they do not know the latest Windows version or latest soccer stars [40]. Typically, background information such as concepts and instances are missing. For example, we find that concepts such as "Corporate Body" or instances such as "Monash University" are missing when using WordNet as an external knowledge base.

In this paper, we propose an ontology merging framework that takes two domain-specific ontologies as input, finds semantic correspondences (alignments) between both ontologies and produces a single merged ontology as output. In our approach, decisions on the semantic correspondences between the entities of semantically heterogeneous ontologies are made by considering multiple knowledge bases. When the exploited knowledge bases have missing background knowledge such as concepts or instances, we utilize other techniques to capture their semantics. To do this, we employ a process integrating name-based and coupled statistical and semantic based techniques respectively based on the Jaro-Winkler distance and the Normalized Retrieval Distance (NRD) functions. It is important to mention that other string edit distance techniques can be used; however, we found the Jaro-Winkler function among the best techniques to compute the string distances between the labels of the ontologies' entities. Moreover, the implementation of this function is publically available and can be easily integrated to any framework that deals with string processing. This step has three major benefits. First, it tackles the issue of the single use of string distance to obtain correspondences between ontology concepts, such as in name-based approaches, by considering an additional coupled statistical and semantic technique. Then, it enables the reuse of newly obtained knowledge in subsequent ontology merging. Finally, it eliminates the concern of manually defining the relations between the missing concepts or instances and other concepts of the knowledge bases. Accordingly, we summarize the contributions of the paper as follows:

- Prioritizing the ontology merging techniques into semantic-based, name-based, and statistical-based techniques respectively.
- Exploiting multiple knowledge bases to make aggregated mapping decisions for merging heterogeneous ontologies.
- Addressing the issue of missing background knowledge in the exploited knowledge bases by utilizing a process integrating name-based and coupled statistical and semantic-based techniques.
- Enriching knowledge bases based on the obtained information. It is important to mention that although some of the approaches use external resources such as knowledge bases to support the merging task, they use only one knowledge base and do not attempt to enrich the knowledge base with missing background knowledge as we do in our framework.

The rest of this paper is organized as follows. Section 2 gives a general overview of the proposed framework. Background information related to the proposed framework is presented in Section 3. We detail the steps of inconsistency checking and resolution between heterogeneous domain specific ontologies and knowledge bases in Section 4.1. The multiple knowledge base assisted merging process is developed in Section 4.2 while dealing with missing background knowledge is discussed in Section 4.3. We develop knowledge base enrichment in Section 4.4. Section 5 presents the experimental results carried out to evaluate the effectiveness of the employed methods and techniques in the proposed framework. The final section presents the conclusions and outlines the future work.

2. General overview

The first technique that we utilize to find semantic correspondences between the entities of different ontologies is based on employing multiple knowledge bases. Each knowledge base represents a repository of facts about entities and their relationships that exist in different domains. In this context, a fact is a triple consisting of an *entity-relation-entity* structure, where entities are related through different types of semantic and taxonomic relations. These relations are automatically extracted from multiple heterogeneous data sources such as plain texts, image and video captions and online ontologies. It is important to mention that