ARTICLE IN PRESS

Web Semantics: Science, Services and Agents on the World Wide Web ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

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

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

Overview of YAM++—(not) Yet Another Matcher for ontology alignment task

DuyHoa Ngo, Zohra Bellahsene *

University Montpellier, LIRMM, 161 rue Ada, 34095, Montpellier, France

ARTICLE INFO

Article history:
Received 24 May 2016
Accepted 22 September 2016
Available online xxxx

Keywords:
Ontology matching
Similarity measure
Matcher combination
Similarity propagation
Mapping selection
Large scale ontology matching

ABSTRACT

Several challenges to the field of ontology matching have been outlined in recent research. The selection of the appropriate similarity measures as well as the configuration tuning of their combination are known as fundamental issues the community should deal with. Verifying the semantic coherence of the discovered alignment is also known as a crucial task. As the challenging issues are both in basic matching techniques and in their combination, our approach is aimed to provide improvement at the basic matcher level and also at the level of framework. Matching large scale ontologies is currently one of the most challenging issues in ontology matching field. The main reason is that large ontologies are highly heterogeneous both at terminological and conceptual levels. Furthermore, matching very large ontologies entails exploring a very large searching space to discover correspondences. It may also require a huge amount of main memory to maintain the temporary results at each computational step. These factors strongly impact the effectiveness and efficiency of any ontology matching tool. To overcome these issues, we have developed a disk-based ontology matching approach. The underlying idea of our approach is that the complexity and therefore the cost of the matching algorithms are reduced thanks to the indexing data structures by avoiding exhaustive pair-wise comparisons. Indeed, we extensively used indexing techniques in many places. For example, we defined a bitmap encoding the structural information of an ontology. This indexing structure will be exploited for accelerating similarity propagation. Moreover, our approach uses a disk-based mechanism to store temporary data. This allows to perform any ontology matching task on a simple PC or laptop instead of a powerful server. In this paper, we describe YAM++, an ontology matching tool, aimed at solving these issues. We evaluated the efficiency of YAM++ in various OAEI 2012 and OAEI 2013 tracks. YAM++ was one of the best ontology matching systems in terms of *F*-measure. Most notably, the current version of YAM++ has passed all scalability and large scale ontology matching tests and obtained high matching quality results.

1. Introduction

In recent years, ontologies have attracted a lot of attention in Computer Science, especially in the Semantic Web field. They serve as explicit conceptual knowledge models and provide the semantic vocabulary that make domain knowledge available to be exchanged and interpreted among information systems. Hence, they open new opportunities for developing a new line of semantic applications such as semantic search [1,2], semantic portal [3–5], semantic information integration [6–8], intelligent advisory systems [9,10], semantic middleware [11,12], semantic software engineering [13], etc. However, one of the most difficult issues is how to deal with heterogeneity of ontologies [14,15]. Due to the decentralized nature of the semantic web, an explosion in the number of ontologies is expected. Many of them may describe similar domains, but they are very different because they have been designed and developed independently by different ontology engineers following diverse modeling principles and patterns.

For example, within a collection of ontologies describing the domain of organizing conferences [16]. People attending to the conference can be conceptualized with different names such as `conference_Participant`, `attendee`, `participant`, `delegate`, `listener`.[1] The heterogeneity of ontologies mainly causes problems of variation in meaning or ambiguity in entity interpretation and, consequently, it prevents information systems

* Corresponding author.
E-mail addresses: Hoa.Ngo@csiro.au (D. Ngo), bella@lirmm.fr (Z. Bellahsene).

[1] (in conference dataset: confOf.owl, ekaw.owl, edas.owl, iasted.owl, sigkdd.owl)

http://dx.doi.org/10.1016/j.websem.2016.09.002
1570-8268/

from sharing their own domain knowledge to the community. Therefore, without knowing the semantic mappings between entities of ontologies, information systems cannot perform interaction, communication and collaboration with each other.

According to [17], ontology matching is a key solution to the semantic heterogeneity problem. It discovers correspondences between semantically related entities of ontologies. Ontology matching can be done either by hand or by using (semi) automatic tools. Discovering manually mappings is tedious, error-prone, and impractical due to the number, size and heterogeneity of ontologies. Hence, the development of fully or semi automatic ontology matching tools becomes crucial to the success of the semantic information systems and applications. In the last decade, through the annual campaign OAEI,[2] many ontology matching systems/tools have been proposed. These state-of-the-art approaches have made a significant progress in the ontology matching field, but none of them gained a clear success in terms of matching quality for all the matching scenarios [18]. In [19–22], challenging issues in ontology matching have been described in detail. Among these challenges, selecting the appropriate similarity measures as well as tuning the configuration of their combination are the toughest fundamental problems of all matching systems. Matching scenarios may require to combine the outcome of the used similarity measures in a different way. Furthermore, the difficulty of the problem grows with the size of the ontologies. Indeed, matching large scale ontologies is one of the most difficult problems in ontology matching field. In particular, the size of ontologies being matched strongly impacts the performance, i.e., effectiveness and efficiency of any ontology matching system. The main reasons are: (i) large ontologies usually lead to a high conceptual heterogeneity and (ii) The complexity of matching is usually proportional to the size of the input ontologies. Furthermore, discovery mappings in a huge space is very time consuming especially if multiple matchers need to be evaluated and combined. Thus, the efficiency of the matching system will be degraded.

To deal with large-scale ontology matching, several techniques have been proposed. The most promising approaches are: filtering-based methods, partitioning methods and background-based ones. The main idea behind these techniques of filtering methods is to reduce the search space by heuristically eliminating less promising candidate mappings. For example, in Eff2Match [23], the heuristic to select candidate mappings for each entity in the source ontology is taken by performing the top-K entities algorithm in the target ontology according to their context (Virtual Document) similarity. More sophisticated heuristics strategies based on different extracted features such as label, hierarchy, neighbors, etc. are applied in each iteration to select the promising mappings [24], ServOMap [25].

While in partitioning-based methods, two large ontologies are firstly divided into sub-ontologies according to their structural information. Then the alignment process is performed between entities of pairs of sub-ontologies. In order to avoid exhaustive pair-wise comparisons, only the high relevant pairs of sub-ontologies will be passed to the matching process. These methods can be found in Falcon-AO [26] and COMA++ [27].

A sub-class of this category is known as anchor-based partitioning methods. These methods are a modified version of the algorithms above, which partition to-be-matched ontologies are done according to the set of anchors. In short, an anchor is a pair of entities mapping determined by a similarity measure. A fragment or sub-ontology is constructed by collecting neighbors entities of the chosen anchors. Then, the alignment process will be performed for each pair of related sub-ontologies. These methods can be found in Anchor-Prompt [28], AnchorFlood [29], Lily [30], TaxoMap [31].

The underlying idea of our approach is that the annotation, the structural and the contextual information of entities are indexed in order to improve the whole matching process both in terms of matching quality and time performance. Unlike those of related work, our filtering methods make use of the annotation-based indexes in order to accelerate the filtering process. Furthermore, the structural indexes are also exploited to check the coherence of the resulting mappings. Indeed, in addition, verifying the semantic coherence of the discovered alignment is known as a challenging issue in large scale ontology matching because almost all reasoning systems fail or cannot completely classify large ontologies [32]. We have implemented a new inconsistency removing algorithm based on Clarkson algorithm for the weighted minimum vertex cover problem. The details of this contribution can be found in [33]. In this paper, we highlight the main contributions and techniques that have been implemented in YAM++ and that have made it one of the best of ontology matching tools. These contributions are the following:

- Effective and efficient filtering methods to deal with large scale ontology matching.
- A heuristic-based label similarity measure which integrates a strict heuristic filter with the label similarity measure, which is aimed at detecting of *informative words*.
- A machine learning-based method to combine terminological similarity measures without the effort of manual setting.
- An information retrieval-based similarity measure to improve the matching quality and to deal with terminological heterogeneity. This new similarity measure takes into account not only syntactic similarity but also information content of words. This measure constitutes an alternative to machine learning method when training data are not available or in large scale setting.
- A bitmap encoding the structural information of an ontology that is exploited for accelerating the similarity propagation. This method is stable and reliable because it exploits and uses all the structural information of an ontology for discovering mappings.
- A dynamic weighted sum method to combine the mappings resulting from the element matcher and structure matcher. The benefit is that it automatically assigns weights to each matcher for a given matching scenario. Moreover, it also automatically determines the filter's threshold value to produce the final mappings.
- A fast semantic filtering method to detect the inconsistent mappings when matching large ontologies.
- The experimental results demonstrate that YAM++ is both effective in terms of quality of the alignments, and efficient in terms of time performance and scalability.

YAM++ was one of the very best system in OAEI competitions from 2011 to 2013. Particularly, thanks to the contributions on dealing with terminological heterogeneity (i.e., machine learning, information retrieval methods), structural heterogeneity (i.e., propagation method), YAM++ achieved the best results in the series of `Systematic Benchmark` tracks in years 2012 and 2013; in the `Conference` track in years 2011, 2012 and 2013; especially, in the *multilingual* `Multifarm` tracks in years 2011.5, 2012 and 2013. Additionally, thanks to the fast semantic indexing and the inconsistency filtering method that has been devoted to large scale ontologies matching, YAM++ was one of the best systems on the `Anatomy` track, `Library` track and `Large Biomedical Ontologies` tracks in years 2012 and 2013.

The rest of the paper is organized as follows:
Section 2 provides the basic notions and definitions used in this paper as well as the evolution and the architecture of

[2] http://oaei.ontologymatching.org/.

Please cite this article in press as: D. Ngo, Z. Bellahsene, Overview of YAM++—(not) Yet Another Matcher for ontology alignment task, Web Semantics: Science, Services and Agents on the World Wide Web (2016), http://dx.doi.org/10.1016/j.websem.2016.09.002