Knowledge-Based Systems 39 (2013) 173–184

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

A unified approach to matching semantic data on the Web

Zhichun Wang [a,b,*], Juanzi Li [a], Yue Zhao [c], Rossi Setchi [d], Jie Tang [a]

[a] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[b] College of Information Science and Technology, Beijing Normal University, Beijing, China
[c] Department of Computer Science, Columbia University, USA
[d] School of Engineering, Cardiff University, Cardiff, Wales, UK

ARTICLE INFO

Article history:
Received 16 September 2011
Received in revised form 29 September 2012
Accepted 23 October 2012
Available online 2 November 2012

Keywords:
Semantic Web
Ontology
Linked data
Ontology matching
Instance matching

ABSTRACT

In recent years, the Web has evolved from a global information space of linked documents to a space where data are linked as well. The Linking Open Data (LOD) project has enabled a large number of semantic datasets to be published on the Web. Due to the open and distributed nature of the Web, both the schema (ontology classes and properties) and instances of the published datasets may have heterogeneity problems. In this context, the matching of entities from different datasets is important for the integration of information from different data sources. Recently, much work has been conducted on ontology matching to resolve the schema heterogeneity problem in the semantic datasets. However, there is no unified framework for matching both schema entities and instances. This paper presents a unified matching approach to finding equivalent entities in ontologies and LOD datasets on the Web. The approach first combines multiple lexical matching strategies using a novel voting-based aggregation method; then it utilizes the structural information and the already found correspondences to discover additional ones. We evaluated our approach using datasets from both OAEI and LOD. The results show that the voting-based aggregation method provides highly accurate matching results, and that the structural propagation procedure effectively improves the recall of the results.

1. Introduction

In recent years, the Web has evolved from a global information space of linked documents to a space where data are linked as well. Many Linking Open Data (LOD) datasets have been published on the Web, leading to the creation of the Web of Data, a global data space containing billions of assertions [1]. Linked data on the Web are machine-readable, and their meaning is explicitly defined by using ontology classes and properties. Datasets are linked to other external datasets and can in turn be linked by external datasets. Although several rules guide the publication of LOD data, two important issues still require further investigation. The first issue is the schema heterogeneity problem. Although the use of common vocabularies/ontologies such as FOAF[1], SIOC[2] and SKOS[3] is encouraged to simplify the processing of LOD data by client applications [2], existing datasets often employ their own schemas, which can be quite different because they are often defined by different organizations. Schema heterogeneity hinders data sharing and data integration. Second, there are fewer established RDF links that connect data than real links between those data. For example, the DBLP and ACM libraries are two datasets in the LOD that contain information about scientific publications. Both datasets contain many duplicate authors and papers, most of which are not linked to each other. RDF links allow client applications to navigate between data sources, discover additional data and combine data on the same individuals that are stored in different locations. Although tools, such as D2RQ[4] and Openlink Virtuoso[5] have been developed to publish LOD data, they do not provide functions for discovering the links between different data sources [3]. Few tools can build semantic correspondences between entities of different datasets at both the schema and instance level. Therefore, the LOD provides a new environment for investigating ontology matching techniques and also raises some challenges for ontology matching.

The problem of ontology matching has been extensively studied in the last decade [4–6]. Most matching strategies calculate the semantic similarity between any two ontology entities and find the correspondences between them. Different information can be

* Corresponding author at: Department of Computer Science and Technology, Tsinghua University, Beijing, China. Tel.: +86 010 62773618; fax: +86 010 62781461.
E-mail addresses: zcwang@bnu.edu.cn (Z. Wang), ljz@keg.cs.tsinghua.edu.cn (J. Li), zhaoy1030@gmail.com (Y. Zhao), setchi@cardiff.ac.uk (R. Setchi), jietang@tsinghua.edu.cn (J. Tang).

[1] http://xmlns.com/foaf/spec/.
[2] http://rdfs.org/sioc/spec/.
[3] http://www.w3.org/TR/skos-reference.
[4] http://www4.wiwiss.fu-berlin.de/bizer/d2rq/.
[5] http://virtuoso.openlinksw.com/.

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.10.015

exploited to assess the similarity between ontology entities, e.g., entity name, entity description, taxonomy structure and instance. To perform well, most matching systems, such as RiMOM [7], Falcon-AO [8] and ASMOV [9], combine multiple strategies based on different ontology information. Recently, ontology instance matching has attracted much research interest because many LOD datasets are published on the Web. Instance matching attempts to evaluate the degree of similarity between different descriptions of real objects across heterogeneous data sources and determine whether two object descriptions refer to the same real object in a given domain. Existing systems for instance matching, such as Silk [10] and KonFuss [11] also use multiple strategies to build links between instances.

An important problem in ontology matching is how to combine the results of the multiple matching strategies that are used. Although combining multiple strategies can improve the overall matching results in a set of matching tasks, it cannot guarantee better results in every matching task. Multi-strategy approaches do not outperform their single-strategy alternatives in certain circumstances. Peukert et al. [12] evaluated several similarity combination methods and found that no single strategy returns the best results in all test cases. A good matching result depends on choosing the most appropriate matching and combination method. Furthermore, combination strategies that perform well in ontology schema matching may fail to achieve good results in instance matching because more fields are compared when matching instances. Therefore, an aggregation strategy that performs well in both ontology schema matching and instance matching tasks is needed. Another important issue is how to combine both lexical and structural information to accurately predict the matching results. Some existing approaches use a vector space model to represent merged lexical information on entities and their neighbors and then compute the cosine similarity between the vectors of entities. Other approaches utilize the structural information to propagate the similarities between entities. However, combining the information on neighbors or propagating similarity cannot always improve the quality of matching results because this technique may degrade the performance of matching algorithms. Certain methods should be used to control the use of structural information. For instance, RiMOM [7] defines two factors to determine whether to use structural information in a certain matching task, while Duan et al. [13] proposed a supervised learning approach to determine the degree to which the similarity should be propagated through the structural information.

In this paper, we propose a unified approach for discovering matching correspondences of both schema entities and instances. Given two sets of ontology entities, our approach uses lexical and structural information separately, in two steps, to identify matches between the entities. In the first step, a set of initial matching correspondences are identified using strategies based on the lexical information on the entities; different strategies are combined using a voting-based method to obtain a set of accurate matching results. Additional semantic correspondences are then identified by utilizing the initial matching results and the internal links between the entities. This paper advances the state-of-the-art in this area by making the following contributions:

(1) Development of a novel voting-based aggregation method to combine the matching results of multiple strategies. Instead of extracting matching correspondences after aggregating similarity values, we generate matching results for each individual strategy and then merge them to produce the final result. A voting scheme is used to refine the results by eliminating less possible correspondences. The method generates a set of matching results with very high precision.

(2) Development of a structural matching strategy that only utilizes confirmed matching results. Our approach uses the lexical and structural information separately in two steps. In the first step, a set of initial matching results is identified based on lexical information, and a structural strategy is then used to match entities by comparing the entities that were already matched in their neighborhood. This two-step approach operates effectively and efficiently to identify matching correspondences. The lexical matching strategies produce high-accuracy results, and the structural matching strategy improves the recall of the results.

The rest of this paper is organized as follows. Section 2 presents the background of our work. Section 3 describes our new combination approach and the structure propagation method. Section 4 presents an analysis of the experimental results, and Section 5 discusses some related work. The conclusion is provided in Section 6.

2. Background

2.1. RDF and linked data

The Resource Description Framework (RDF)[6] is a family of W3C[7] specifications that has been widely used as a general method for the conceptual description or modeling of information in Web resources. The RDF describes resources in the form of subject–predicate–object expressions. These expressions are called triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and object.

The RDF Vocabulary Definition Language (RDFS) and the Web Ontology Language (OWL) provide a basis for the creation of vocabularies that can be used to describe entities and their relationships [14]. Vocabularies are collections of classes and properties. Vocabularies are expressed in the RDF using terms from RDFS and OWL that provide varying degrees of expressivity in modeling domains of interest. Anyone can publish vocabularies in the Web of Data, which in turn can be connected by RDF triples that link classes and properties in one vocabulary to those in another, thereby defining mappings between related vocabularies [1].

The term "Linked Data" refers to a set of best practices for publishing and connecting structured data on the Web [1]. Berners-Lee [15] outlined four basic rules for publishing LOD data on the Web. Fig. 1 shows an example of LOD data in the form of a graph that represents three RDF links taken from Tim Berners-Lees FOAF profile. The terms con:assistant, con:phone, rdf:label and foaf:person are schema information in the graph; they are entities from the Contact[8], RDF[9] and FOAF namespaces, respectively, which specify the meaning of the links. The two URIs in Fig. 1 represent two person instances, namely Tim Berners-Lee and his assistant. Therefore, publishing LOD data can simply be considered as the use of certain ontologies to create typed links between things that are represented by URIs.

2.2. Ontology and ontology matching

In computer science, an ontology is a formal, explicit specification of a shared conceptualization [16,17]. Ontologies can be expressed in several standard languages, including the Web Ontology Language (OWL) [18]. OWL is an ontology language that is recommended by the W3C and provides vocabularies used to de-

[6] http://www.w3.org/RDF/.
[7] http://www.w3.org/.
[8] http://www.w3.org/2000/10/swap/pim/contact.
[9] http://www.w3.org/1999/02/22-rdf-syntax-ns.