Web Semantics: Science, Services and Agents on the World Wide Web 37–38 (2016) 152–169

Contents lists available at ScienceDirect

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

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

Learning the semantics of structured data sources

Mohsen Taheriyan *, Craig A. Knoblock, Pedro Szekely, José Luis Ambite

University of Southern California, Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292, USA

ARTICLE INFO

Article history:
Received 8 April 2015
Received in revised form
24 September 2015
Accepted 24 December 2015
Available online 11 January 2016

Keywords:
Knowledge graph
Semantic model
Semantic labeling
Semantic web
Ontology
Linked data

ABSTRACT

Information sources such as relational databases, spreadsheets, XML, JSON, and Web APIs contain a tremendous amount of structured data that can be leveraged to build and augment knowledge graphs. However, they rarely provide a semantic model to describe their contents. Semantic models of data sources represent the implicit meaning of the data by specifying the concepts and the relationships within the data. Such models are the key ingredients to automatically publish the data into knowledge graphs. Manually modeling the semantics of data sources requires significant effort and expertise, and although desirable, building these models automatically is a challenging problem. Most of the related work focuses on semantic annotation of the data fields (source attributes). However, constructing a semantic model that explicitly describes the relationships between the attributes in addition to their semantic types is critical.

We present a novel approach that exploits the knowledge from a domain ontology and the semantic models of previously modeled sources to automatically learn a rich semantic model for a new source. This model represents the semantics of the new source in terms of the concepts and relationships defined by the domain ontology. Given some sample data from the new source, we leverage the knowledge in the domain ontology and the known semantic models to construct a weighted graph that represents the space of plausible semantic models for the new source. Then, we compute the top k candidate semantic models and suggest to the user a ranked list of the semantic models for the new source. The approach takes into account user corrections to learn more accurate semantic models on future data sources. Our evaluation shows that our method generates expressive semantic models for data sources and services with minimal user input. These precise models make it possible to automatically integrate the data across sources and provide rich support for source discovery and service composition. They also make it possible to automatically publish semantic data into knowledge graphs.

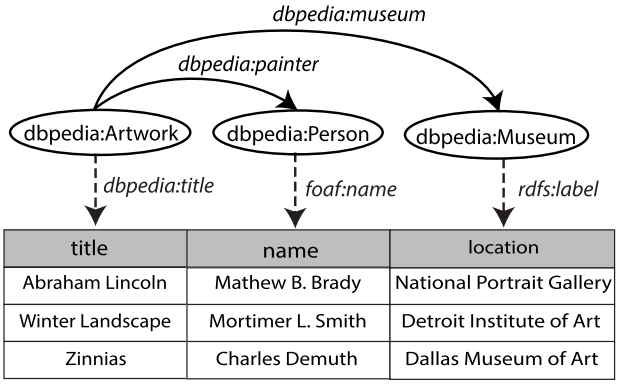
1. Introduction

Knowledge graphs have recently emerged as a rich and flexible representation of domain knowledge. Nodes in this graph represent the entities and edges show the relationships between the entities. Large companies such as Google and Microsoft employ knowledge graphs as a complement for their traditional search methods to enhance the search results with semantic-search information. Linked Open Data (LOD) is an ongoing effort in the Semantic Web community to build a massive public knowledge graph. The goal is to extend the Web by publishing various open datasets as RDF on the Web and then linking data items to other useful information from different data sources. With linked data, starting from a certain point in the graph, a person or machine can explore the graph to find other related data. The focus of this work is the first step of publishing linked data, automatically publishing datasets as RDF using a common domain ontology.

A large amount of data in LOD comes from structured sources such as relational databases and spreadsheets. Publishing these sources into LOD involves constructing *source descriptions* that represent the intended meaning of the data by specifying mappings between the sources and the *domain ontology* [1]. A domain ontology is a formal model that represents the concepts within a domain and the properties and interrelationships of those concepts. In this context, what is meant by a source description is a schema mapping from the source to an ontology. We can represent this mapping as a *semantic network* with ontology classes as the nodes and ontology properties as the links between the nodes. This network, also called a *semantic model*, describes the source in terms of the concepts and relationships defined by the domain ontology. Fig. 1 depicts a semantic model for a sample data source including

* Corresponding author.
E-mail addresses: mohsen@isi.edu (M. Taheriyan), knoblock@isi.edu (C.A. Knoblock), pszekely@isi.edu (P. Szekely), ambite@isi.edu (J.L. Ambite).

http://dx.doi.org/10.1016/j.websem.2015.12.003
1570-8268/



title	name	location
Abraham Lincoln	Mathew B. Brady	National Portrait Gallery
Winter Landscape	Mortimer L. Smith	Detroit Institute of Art
Zinnias	Charles Demuth	Dallas Museum of Art

Fig. 1. The semantic model of a sample data source containing information about paintings.

information about some paintings. This model explicitly represents the meaning of the data by mapping the source to the DBpedia[1] and FOAF[2] ontologies. Knowing this semantic model enables us to publish the data in the table into the LOD knowledge graph.

One step in building a semantic model for a data source is *semantic labeling*, determining the *semantic types* of its data fields, or *source attributes*. That is, each source attribute is labeled with a class and/or a data property of the domain ontology. In our example in Fig. 1, the semantic types of the first, second, and third columns are *title* of *Artwork*, *name* of *Person*, and *label* of *Museum* respectively. However, simply annotating the attributes is not sufficient. Unless the relationships between the columns are explicitly specified, we will not have a precise model of the data. In our example, a *Person* could be the *owner*, *painter*, or *sculptor* of an *Artwork*, but in the context of the given source, only *painter* correctly interprets the relationship between *Artwork* and *Person*. In the correct semantic model, *Museum* is connected to *Artwork* through the link *museum*. Other models may connect *Museum* to *Person* instead of *Artwork*. For instance, *Person* could be *president*, *owner*, or *founder* of *Museum*, or *Museum* could be *employer* or *workplace* of *Person*. To build a semantic model that fully recovers the semantics of the data, we need a second step that determines the relationships between the source attributes in terms of the properties in the ontology.

Manually constructing semantic models requires significant effort and expertise. Although desirable, generating these models automatically is a challenging problem. In Semantic Web research, there is much work on mapping data sources to ontologies [2–13], but most focus on semantic labeling or are very limited in automatically inferring the relationships. Our goal is to construct semantic models that not only include the semantic types of the source attributes, but also describe the relationships between them.

In this paper, we present a novel approach that exploits the knowledge from a domain ontology and *known semantic models* of sources in the same domain to automatically learn a rich semantic model for a new source. The work is inspired by the idea that different sources in the same domain often provide similar or overlapping data and have similar semantic models. Given sample data from the new source, we use a labeling technique [14] to annotate each source attribute with a set of candidate semantic types from the ontology. Next, we build a weighted directed graph from the known semantic models, learned semantic types, and the domain ontology. This graph models the space of plausible semantic models. Then, we find the most promising mappings from the source attributes to the nodes of the graph, and for each mapping, we generate a candidate model by computing the minimal tree that connects the mapped nodes. Finally, we score the candidate models to prefer the ones formed with more coherent and frequent patterns.

This work builds on top of our previous work on learning semantic models of sources [15,16]. The central data structure of our approach to learn a semantic model for a new source is a graph built on top of the known semantic models. In the previous work, we add a new component to the graph for each known semantic model. If two semantic models are very similar to each other and they only differ in one link, for example, we will still have two different components for them in the graph. The graph grows as the number of known semantic models grows, which makes computing the semantic models inefficient if we have a large set of known semantic models. In this paper, we extend our previous work to make it scale to a large number of semantic models. We present a new algorithm that constructs a much more compact graph by merging overlapping segments of the known semantic models. The new technique significantly reduces the size of the graph in terms of the number of nodes and links. Consequently, it considerably decreases the number of possible mappings from the source attributes to the nodes of the graph. It also makes computing the minimal tree that connects the nodes of the candidate mappings in the graph more efficient. The new method to build the graph changes our algorithms to compute and rank the candidate semantic models.

The main contribution of this paper is a scalable approach that exploits the structure of the domain ontology and the known semantic models to build semantic models of new sources. We evaluated our approach on a set of museum data sources modeled using two well-known *data models* in the cultural heritage domain: Europeana Data Model (EDM) [17], and CIDOC Conceptual Reference Model (CIDOC-CRM) [18]. A data model standardizes how to map the data elements in a domain to a set of domain ontologies. The evaluation shows that our approach automatically generates high-quality semantic models that would have required significant user effort to create manually. It also shows that the semantic models learned using both the domain ontology and the known models are approximately 70% more accurate than the models learned with the domain ontology as the only background knowledge. The generated semantic models are the key ingredients to automate tasks such as source discovery, information integration, and service composition. They can also be formalized using mapping languages such as R2RML[19], which can be used for converting data sources into RDF and publishing them into the Linked Open Data (LOD) cloud or any other knowledge graph.

We have implemented our approach in Karma [20], our data modeling and integration framework.[3] Users can import data from a variety of sources including relational databases, spreadsheet, XML files and JSON files into Karma. They can also import the domain ontologies they want to use for modeling the data. The system then automatically suggests a semantic model for the loaded source. Karma provides an easy to use graphical user interface to let users interactively refine the learned semantic models if needed. Once a semantic model is created for the new source, users can publish the data as RDF by clicking a single button. Szekely et al. [21] used Karma to model the data from Smithsonian American Art Museum[4] and then publish it into the

1 http://dbpedia.org/ontology.
2 http://xmlns.com/foaf/spec.
3 http://karma.isi.edu.
4 http://americanart.si.edu.