Web Semantics: Science, Services and Agents on the World Wide Web 35 (2015) 142–151

Contents lists available at ScienceDirect

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

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

Mining the Web of Linked Data with RapidMiner

Petar Ristoski *, Christian Bizer, Heiko Paulheim

Data and Web Science Group, University of Mannheim, B6, 26, 68159 Mannheim, Germany

ARTICLE INFO

Article history:
Received 30 January 2015
Received in revised form
11 May 2015
Accepted 11 June 2015
Available online 8 July 2015

Keywords:
Linked Open Data
Data mining
RapidMiner

ABSTRACT

Lots of data from different domains are published as Linked Open Data (LOD). While there are quite a few browsers for such data, as well as intelligent tools for particular purposes, a versatile tool for deriving additional knowledge by mining the Web of Linked Data is still missing. In this system paper, we introduce the *RapidMiner Linked Open Data extension*. The extension hooks into the powerful data mining and analysis platform *RapidMiner*, and offers operators for accessing Linked Open Data in RapidMiner, allowing for using it in sophisticated data analysis workflows without the need for expert knowledge in SPARQL or RDF. The extension allows for autonomously exploring the Web of Data by following links, thereby discovering relevant datasets on the fly, as well as for integrating overlapping data found in different datasets. As an example, we show how statistical data from the World Bank on scientific publications, published as an RDF data cube, can be automatically linked to further datasets and analyzed using additional background knowledge from ten different LOD datasets.

1. Introduction

The Web of Linked Data contains a collection of machine processable, interlinked datasets from various domains, ranging from general cross-domain knowledge sources to government, library and media data, which today comprises roughly a thousand datasets [1,2]. While many domain-specific applications use Linked Open Data, general-purpose applications rarely go beyond displaying the mere data, and provide little means of deriving additional knowledge from the data.

At the same time, sophisticated data mining platforms exist, which support the user with finding patterns in data, providing meaningful visualizations, etc. What is missing is a *bridge* between the vast amount of data on the one hand, and intelligent data analysis tools on the other hand. Given a data analysis problem, a data analyst should be able to automatically find suitable data from different relevant data sources, which will then be combined and cleansed, and served to the user for further analysis. This data collection, preparation, and fusion process is an essential part of the data analysis workflow [3], however, it is also one of the most time consuming parts, constituting roughly half of the costs in data analytics projects [4]. Furthermore, since the step is time consuming, a data analyst most often makes a heuristic selection of data sources based on his a priori assumptions, and hence is subject to the selection bias. Despite these issues, automation at that stage of the data processing step is still rarely achieved.

In this paper, we discuss how the Web of Linked Data can be mined using the full functionality of the state of the art data mining environment *RapidMiner*[1] [5]. We introduce an extension to RapidMiner, which allows for bridging the gap between the Web of Data and data mining, and which can be used for carrying out sophisticated analysis tasks on and with Linked Open Data. The extension provides means to automatically connect local data to background knowledge from Linked Open Data, or load data from the desired Linked Open Data source into the RapidMiner platform, which itself provides more than 400 operators for analyzing data, including classification, clustering, and association analysis.

RapidMiner is a programming-free data analysis platform, which allows the user to design data analysis processes in a plug-and-play fashion by wiring *operators*. Furthermore, functionality can be added to RapidMiner by developing extensions, which are made available on the *RapidMiner Marketplace*.[2] The *RapidMiner Linked Open Data extension* adds operators for loading data from datasets within Linked Open Data, as well as autonomously following RDF links to other datasets and gathering additional data from there. Furthermore, the extension supports schema matching for data gathered from different datasets.

* Corresponding author.
E-mail address: petar.ristoski@informatik.uni-mannheim.de (P. Ristoski).

http://dx.doi.org/10.1016/j.websem.2015.06.004
1570-8268/

[1] http://www.rapidminer.com/.
[2] https://marketplace.rapidminer.com/.

As the operators from that extension can be combined with all RapidMiner built-in operators, as well as those from other extensions (e.g., for time series analysis), complex data analysis processes on Linked Open Data can be built. Such processes can automatically combine and integrate data from different datasets and support the user in making sense of the integrated data.

The use case we pursue in this paper starts from a Linked Open Dataset publishing various World Bank indicators. Among many others, this dataset captures the number of scientific journal publications in different countries over a period of more than 25 years. An analyst may be interested in which factors drive a high increase in that indicator. Thus, she needs to first determine the *trend* in the data. Then, additional *background knowledge* about the countries is gathered from the Web of Linked Data, which helps her in identifying relevant factors that may explain a high or low increase in scientific publications. Such factors are obtained, e.g., by running a correlation analysis, and the significant correlations can be visualized for a further analysis, and for determining outliers from the trend.

The rest of this paper is structured as follows. Section 2 describes the functionality of the RapidMiner Linked Open Data extension. In Section 3, we show the example use case of scientific publications in detail, whereas Section 4 briefly showcases other use cases for which the extension has been employed in the past. Section 5 presents evaluations of various aspects of the extension. Section 6 discusses related work, and Section 7 provides an outlook on future directions pursued with the extension.

2. Description

RapidMiner is a data mining platform, in which data mining and analysis processes are designed from elementary building blocks, so called *operators*. Each operator performs a specific action on data, e.g., loading and storing data, transforming data, or inferring a model on data. The user can compose a process from operators by placing them on a canvas and wiring their input and output ports, as shown in Fig. 1.

The *RapidMiner Linked Open Data* extension adds a set of operators to RapidMiner, which can be used in data mining processes and combined with RapidMiner built-in operators, as well as other operators. The operators in the extension fall into different categories: data import, data linking, feature generation, schema matching, and feature subset selection.

2.1. Data import

RapidMiner itself provides import operators for different data formats (e.g., Excel, CSV, XML). The Linked Open Data extension adds two import operators:

- A *SPARQL Importer* lets the user specify a SPARQL endpoint or a local RDF model, and a SPARQL query, and loads the query results table into RapidMiner. For local RDF models, SPARQL queries can also be executed with RDFS and different flavors of OWL inference [6].
- A *Data Cube Importer* can be used for datasets published using the RDF Data Cube vocabulary.[3] Following the Linked Data Cube Explorer (LDCX) prototype described in [7], the importer provides a wizard which lets the user select the dimensions to use, and creates a pivot table with the selected data.

[3] http://www.w3.org/TR/vocab-data-cube/.

2.2. Data linking

In order to combine a local, potentially non-RDF dataset (e.g., data in a CSV file or a database) with data from the LOD cloud, links from the local dataset to remote LOD cloud datasets have to be established first. For that purpose, different linking operators are implemented in the extension:

- The *Pattern-based linker* creates URIs based on a string pattern. If the pattern a dataset uses for constructing its URIs is known, this is the fastest and most accurate way to construct URIs. For example, the *RDF Book Mashup* [8] dataset uses a URI pattern for books which is based on the ISBN.[4]
- The *Label-based linker* searches for resources whose label is similar to an attribute in the local dataset, e.g., the product name. It can only be used on datasets providing a SPARQL interface and is slower than the pattern-based linker, but can also be applied if the link patterns are not known, or cannot be constructed automatically.
- The *Lookup linker* uses a specific search interface[5] for the *DBpedia* dataset [9]. It also finds resources by alternative names (e.g., *NYC* or *NY City* for *New York City*). For DBpedia, it usually provides the best accuracy.
- For processing text, a linker using *DBpedia Spotlight*[6] [10] has also been included, which identifies multiple DBpedia entities in a textual attribute.
- The *SameAs linker* can be used to follow links from one dataset to another. Since many datasets link to DBpedia, a typical setup to link to an arbitrary LOD dataset is a two-step approach: the Lookup linker first establishes links to DBpedia at high accuracy. Then, `owl:sameAs` links between DBpedia and the target dataset are exploited to set the links to the latter.

2.3. Feature generation

For creating new data mining features from Linked Open Data sources, different strategies are implemented in the extension's operators:

- The *Direct Types* generator extracts all types (i.e., objects of `rdf:type`) for a linked resource. For datasets such as YAGO,[7] those types are often very informative, for example, products may have concise types such as *Smartphone* or *AndroidDevice*.
- The *Datatype Properties* generator extracts all datatype properties, i.e., numerical and date information (such as the price and release date of products).
- The *Relations* generator creates a binary or a numeric attribute for each property that connects a resource to other resource. For example, if a dataset contains awards won by products, an *award* attribute would be generated, either as a binary feature stating whether the product won an award or not, or a numerical one stating the number of awards.
- The *Qualified Relations* generator also generates binary or numeric attributes for properties, but takes the type of the related resource into account. For example, for a product linked to a manufacturer of type *GermanCompany*, a feature stating whether the product has been manufactured by a German company or not would be created.

[4] In cases where additional processing is required, such as removing dashes in an ISBN, the operator may be combined with the built-in *Generate Attributes* operator, which can perform such operations.

[5] http://lookup.dbpedia.org/.

[6] http://spotlight.dbpedia.org/.

[7] http://yago-knowledge.org.