Mining the Web of Linked Data with RapidMiner

Petar Ristoski, Christian Bizer, and Heiko Paulheim

University of Mannheim, Germany
Data and Web Science Group
{petar.ristoski,heiko,chris}@informatik.uni-mannheim.de

Abstract. Lots of data from different domains is published as Linked Open Data. While there are quite a few browsers for that 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 challenge entry, we introduce the *RapidMiner Linked Open Data extension*. The extension hooks into the powerful data mining platform *RapidMiner*, and offers operators for accessing Linked Open Data in RapidMiner, allowing for using it in sophisticated data analysis workflows without the need to know SPARQL or RDF. As an example, we show how statistical data on scientific publications, published as an RDF data cube, can be linked to further datasets and analyzed using additional background knowledge from various 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 [8]. While many domain-specific applications use Linked Open Data, general-purpose applications rarely go beyond displaying the mere data, and provide little means of obtaining knowledge from Linked Open 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 pile of knowledge on the one hand, and intelligent data analysis tools on the other hand. Given a problem at hand, such a bridge should be able to automatically find suitable data from different 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 step when analyzing open data, however, it is also one of the most time consuming parts. 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.

This Semantic Web Challenge submission demonstrates how the Web of Linked Data can be mined using the full functionality of the state of the art data mining environment *RapidMiner*. 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 Linked Open Data. The extension provides means to automatically connect local data to background knowledge from Linked Open Data, or simply load data from the desired Linked Open Data source into the RapidMiner platform[1], which itself provides more than 400 operators for analyzing data, including classification, clustering, and association analysis. RapidMiner lets the user design data analysis processes in a plug-and-play fashion by wiring operators. The 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 gather additional data from there. Furthermore, schema matching for data gathered from different datasets is supported.

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 Data set 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 help her in identifying relevant factors, which 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.

2 Description

RapidMiner is a data mining platform which consists of *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. In the following, we provide details on those operators, and describe an example data analysis process making use of data published in the Linked Open Data cloud.

2.1 Operator Overview

The operators in the extension fall into different categories: data import, data linking, feature generation, schema matching, and feature subset selection. Furthermore, there is a meta-operator for exploring the data web.

[1] `http://www.rapidminer.com/`

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.
- A *Data Cube Importer* can be used for datasets published using the RDF Data Cube vocabulary[2]. Following the Linked Data Cube Explorer (LDCX) prototype described in [1], the importer provides a wizard which lets the user select the dimensions to use, and creates a pivot table with the selected data.

Data Linking. In order to combine a local, non-RDF dataset (e.g., data in a CSV 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* dataset uses a URI pattern for books which is based on the ISBN.[3]
- 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 be applied if the link patterns are not known, or cannot be constructed automatically.
- The *Lookup linker* uses a specific search interface[4] for the *DBpedia* dataset. 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*[5] 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, it is typically combined with the Lookup linker, which first establishes links to DBpedia at high accuracy. The RDF links from the identified resources can then be used to discover further resources in other datasets.

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 of a linked resource. For datasets such as YAGO[6], 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).

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

[3] 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.

[4] `http://lookup.dbpedia.org/`

[5] `http://spotlight.dbpedia.org/`

[6] `http://yago-knowledge.org`

- 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.
- The *Qualified Relations* generator also generates binary or numeric attributes for properties, but takes the type of the related resource into account. For example, an attribute stating that the manufacturer of a product is a German company would be created.
- The *Specific Relations* generator creates features for a user-specified relation, such as Wikipedia categories included in DBpedia.

All of those operators can work in three different modes, using a predefined SPARQL endpoint, dereferencing URIs and processing the returned RDF, or using a local RDF model. While the SPARQL-based variant is usually faster, the dereferencing URIs variant is more versatile, as it can also work with datasets not offering a SPARQL endpoint.

Feature Subset Selection. All standard methods for feature subset selection can be used in conjunction with the RapidMiner Linked Open Data extension, as well as operators from the *Feature Subset Selection extension*[7]. Furthermore, the Linked Open Data extension provides the *Simple Hierarchy Filter*, which exploits the schema information of a Linked Open Data source, and often achieves a better compression of the feature set than standard, non-hierarchical operators, without losing valuable features [7].

Exploring Links. The feature generation algorithms above so far use only one input URI, and obtain features from that URI. This means that they are usually restricted to one dataset. For making use of the entire LOD cloud, the extension provides a meta-operator called *Link Explorer*, which follows links of a given type (by default: `owl:sameAs`) to a specified depth, and applies a set of operators to each resource discovered by that. A typical configuration is to use the link explorer in combination with the datatype properties generator, which results in following links from one starting point, and collect all the datatype properties for all linked resources.

Since the datasets that are used by that meta-operator are known a priori, and there is no reliable way of discovering a SPARQL endpoint for a given resource [3], the link explorer only works by derefencing URIs, but not by means of SPARQL queries.

Data Integration. When combining data from different LOD sources, those usually use different schemas. For example, the population of a country can be contained in different datasets, using a different datatype property to denote the information. In order to use that data more effectively, such attributes can be merged into one by applying schema matching. The extension provides the *PARIS LOD Matcher*, which is an adaptation of the PARIS framework [9], and is able to perform alignment of instances, relations and classes. The matching operator outputs all discovered correspondences, which then can be resolved using the *Data Fusion* operator, which offers various conflict resolution strategies.

[7] `http://sourceforge.net/projects/rm-featselext/`

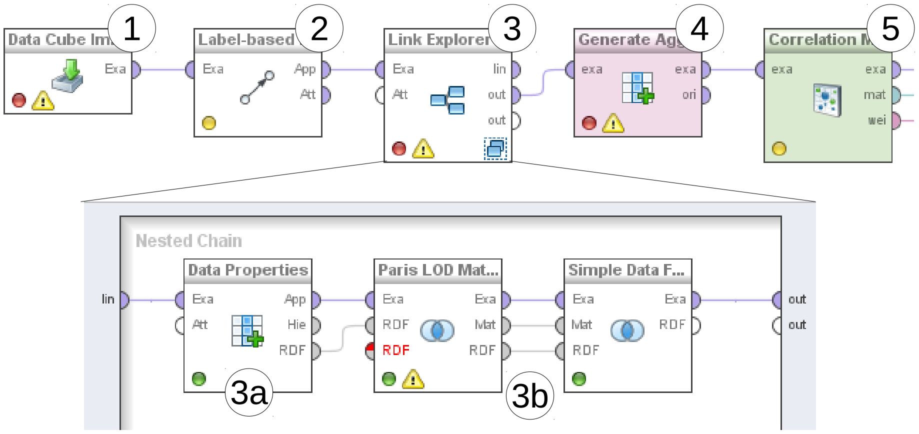

Fig. 1: Overview of the process used in the running example, including the nested subprocess in the link explorer operator

2.2 Example Use Case

In our example use case, we use an RDF data cube with WorldBank economic indicators data[8] as a starting point. The data cube contains time-indexed data for more than 1,000 indicators in over 200 countries. As shown in Fig. 1, the process starts with importing data from that data cube (1). To that end, a wizard is used, which lets the user select the indicator(s) of interest. In our example, we select the indicator "Scientific and technical journal articles", which reports the number of such articles per country and year. The indicator is present for 165 countries, so our resulting data table contains 165 rows, which columns per year, depicting the target value from 1960 to 2011. We are interested in understanding which factors drive a large *increase* in that indicator.[9]

In the next step, we set links of the data imported from the RDF cube to other datasets (2). In our example, we use the label-based linker to find countries in DBpedia which have the same name as the country in the imported slice of the data cube.

The subsequent step is identifying more relevant datasets by following RDF links, and getting the data from there. This is carried out by the link explorer operator (3). Starting from DBpedia, we follow all `owl:sameAs` links to a depth of 2. Inside the link explorer, we collect datatype properties (3a) and also perform the matching (3b). We end up with data from ten different datasets, i.e., DBpedia[10], LinkedGeoData[11],

[8] http://worldbank.270a.info

[9] Note that this example, although similar to the experiments described in our previous works [2,5], goes beyond those experiments by autonomously exploring and gathering data from the Linked Data Cloud, instead of running only on a predefined dataset, such as DBpedia.

[10] http://dbpedia.org

[11] http://linkedgeodata.org

Eurostat[12], Geonames[13], WHO's Global Health Observatory[14], Linked Energy Data[15], OpenCyc[16], World Factbook[17], and YAGO[18].

The initial set of datatype properties extracted has 1,118 attributes, which, after processing by schema matching and conflict resolution, are fused to 818. For example, we find roughly 10 different sources stating the population of countries by following RDF links between datasets. Those are merged into one attribute.

Once all the attributes have been extracted and matched, the actual analysis starts. First, the *Generate Aggregate* operator (4), which is a RapidMiner built-in operator, computes the *increase* in scientific publications from the individual per-year values. Then, a correlation matrix is computed (5), again with a RapidMiner built-in operator, to find interesting factors that explain an increase in scientific publications.

From all the additional attributes we found by following links to different datasets in the Web of Data, we are now able to identify various attributes that explain a strong increase in scientific publications:

- The fragile state index (FSI) and the Human Development Index (HDI) are artificial measures comprised of different social, political and health indicators, and both are good indicators for the growth of scientific publications.
- The GDP per capita is also strongly correlated with the increase in scientific publications. This hints at wealthier countries being able to invest more federal money into science funding.
- For European countries, the number of EU seats shows a significant correlation with the increase in scientific publications. As larger countries have more seats (e.g,. Germany, France, UK), this may hint at an increasing fraction of EU funding for science going being attributed to those countries.
- Additionally, many climate indicators show a strong correlation with the increase in scientific publications: precipitation has a negative correlation with the increase, while hours of sun and temperature averages are positively correlated. This can be explained by an unequal distribution of wealth across different climate zones, with the wealthier nations often located in more moderate climate zones.[19]

So far, these results have concentrated on one specific world bank indicator, while, as stated above, there are more than a thousand. We have conducted similar experiments with other indicators as well, revealing different findings. For example, we looked into the savings of energy consumption over the last years. Here, we can observe, e.g., a correlation with the GDP, showing that wealthier countries can afford putting efforts into saving energy, while less wealthy countries first strive at increasing their economic growth without putting too much of an emphasis on ecology-friendly growth.

[12] `http://eurostat.linked-statistics.org/` and `http://wifo5-03.informatik.uni-mannheim.de/eurostat/`

[13] `http://sws.geonames.org/`

[14] `http://gho.aksw.org/`

[15] `http://en.openei.org/lod/`

[16] `http://sw.opencyc.org/`

[17] `http://wifo5-03.informatik.uni-mannheim.de/factbook/`

[18] `http://yago-knowledge.org`

[19] A more tongue-in-cheek interpretation may be that if the weather is bad, scientists spend more time in the lab writing journal articles.

In summary, the experiment shows that

- we can follow RDF links between datasets, and, by that, gather and combine data from different sources to solve a task at hand, and
- we can use analysis methods that identify the relevant answers to a question.

3 Online Access and Demonstration

The RapidMiner Linked Open Data extension is available online and can be installed from within RapidMiner[20]. The data sets and processes used in this paper are also available online[21]. Furthermore, a detailed documentation for each operator in the extension is provided[22].

4 Appendix

4.1 Mandatory Criteria

The application has to be an end-user application. The application can be used by any person capable of using RapidMiner, i.e., people with basic data mining knowledge. Most of the extension's operators can be used without any prior knowledge of RDF, SPARQL, and the like. The extension is available online, and it already counts more than $3,000$ downloads by users. Furthermore, the extension has commercial potential, as it might serve users from wide range, starting from students using it for educational purposes, to enterprises using it for improving their data analysis processes.

Information sources used. The presented tool can make use of data from the entire LOD cloud by autonomously following links from one dataset to the other. Semantic heterogeneity is handled by schema matching and data fusion methods.

The meaning of data has to play a central role. We use semantically annotated data from the LOD cloud, and we follow `owl:sameAs` links between datasets to navigate the Web of Linked Data and automatically gather data from different sources. The schema information found for the extracted data is taken into account for matching data from different sources.

4.2 Additional Desired Features

The application should be scalable. The proposed tool works as an extension of RapidMiner, which itself is scalable and allows for distributed, parallel processing in a cloud environment.[23]

[20] http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension/

[21] http://data.dws.informatik.uni-mannheim.de/rmlod/swc14/

[22] http://dws.informatik.uni-mannheim.de/fileadmin/lehrstuehle/ki/research/RapidMinerLODExtension/RapidMinerLODExtensionManual.pdf

[23] http://rapidminer.com/news-posts/data-time-rapidminer-cloud-makes-simple-use-data-predict-take-actions-cloud/

Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained. While this paper presents the system from an end user's point of view, in-depth evaluations of the underlying algorithms have been performed, e.g., in [4, 6, 7].

Novelty, in applying semantic technology to a domain or task that have not been considered before Even though there have been proposed several approaches for exploiting Linked Open Data for data mining, the presented LOD extension is the first publicly available tool that exploits Linked Open Data in novel ways in order to support all steps of the knowledge discovery and data analysis process within a powerful data mining platform.

Functionality is different from or goes beyond pure information retrieval. By allowing the analysis of data from Linked Open Data with all sorts of data mining operators, such as classification, clustering, or association analysis, insights can be created that go much beyond a mere information retrieval.

There is a use of dynamic data. The extension performs live access on the Web of Linked Data, which is subject to continuous updates and upgrades over time.

References

1. Benedikt Kämpgen and Andreas Harth. Olap4ld - a framework for building analysis applications over governmental statistics. In *ESWC 2014 Posters & Demo session*. ESWC, Springer, Mai 2014.
2. Heiko Paulheim. Generating possible interpretations for statistics from linked open data. In *9th Extended Semantic Web Conference (ESWC)*, 2012.
3. Heiko Paulheim and Sven Hertling. Disoverability of sparql endpoints in linked open data. In *International Semantic Web Conference, Posters and Demos*, 2013.
4. Heiko Paulheim, Petar Ristoski, Evgeny Mitichkin, and Christian Bizer. Data mining with background knowledge from the web. In *RapidMiner World*, 2014.
5. Petar Ristoski and Heiko Paulheim. Analyzing statistics with background knowledge from linked open data. In *Workshop on Semantic Statistics*, 2013.
6. Petar Ristoski and Heiko Paulheim. A comparison of propositionalization strategies for creating features from linked open data. In *Linked Data for Knowledge Discovery*, 2014.
7. Petar Ristoski and Heiko Paulheim. Feature selection in hierarchical feature spaces. In *Discovery Science*, 2014.
8. Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In *International Semantic Web Conference*, 2014.
9. Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. PARIS: Probabilistic Alignment of Relations, Instances, and Schema. *PVLDB*, 5(3):157–168, 2011.