Visual Analysis of Statistical Data on Maps Using Linked Open Data

Petar Ristoski(✉) and Heiko Paulheim

University of Mannheim, Research Group Data and Web Science B6 26,
Mannheim, Germany
{petar.ristoski,heiko}@informatik.uni-mannheim.de

Abstract. When analyzing statistical data, one of the most basic and at the same time widely used techniques is analyzing correlations. As shown in previous works, Linked Open Data is a rich resource for discovering such correlations. In this demo, we show how statistical analysis and visualization on maps can be combined to facilitate a deeper understanding of the statistical findings.

Keywords: Linked Open Data · Visualization · Correlations · Statistical data

1 Introduction

Statistical datasets are widely spread and published on the Web. However, many users' information need is not the mere consumption of statistical data as such, but the search for patterns and explanations. As shown in previous works [3], information from the Linked Open Data (LOD) cloud can serve as background knowledge for interpreting statistical data, as it covers various domains, ranging from general purpose datasets to government and life science data [7].

In this paper, we present the Web-based tool *ViCoMap*[1], which allows automatic correlation analysis and visualizing statistical data on maps using Linked Open Data. The tool automatically enriches statistical datasets, imported from LOD, RDF datacubes, or local datasets, with information from Linked Open Data, and uses that background knowledge as a means to create possible interpretations as well as advanced map visualization of the statistical datasets. To visualize geospatial entities on a map, we use GADM[2], a LOD database of polygon shapes of the world's administrative areas.

So far, many tools for visualization of LOD and statistical data have been developed [4]. In particular, for RDF data cubes[3] exposing statistical data, different browsers have been developed, such as *CubeViz*[4] or *Payola*[5] [1]. The *CODE*

[1] The tool is available at http://vicomap.informatik.uni-mannheim.de/. Note: response times may vary according to the availability of the LOD sources used.

[2] http://gadm.geovocab.org/.

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

[4] http://aksw.org/Projects/CubeViz.html.

[5] http://live.payola.cz/.

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9341, pp. 138–143, 2015.
DOI: 10.1007/978-3-319-25639-9_27

Visualisation Wizard[6] [2] also features different chart and map based visualizations. Tialhun et al. [8] have developed a LOD-based visualization system for healthcare data. The system visualizes different healthcare indicators per country on a map, and is able to perform correlation analysis between selected indicators, which can later be visualized as a chart.

The direct predecessor of *ViCoMap* is *Explain-a-LOD* [3], which is one of the first approaches for automatically generating hypothesis for explaining statistics using LOD. The tool enhances statistical datasets with background information from DBpedia[7], and uses correlation analysis and rule learning for producing hypothesis which are presented to the user.

ViCoMap, presented in this demo, combines map-based visualizations on the one hand side, and mining for correlations using background knowledge from LOD on the other. As such, it opens new ways of interpreting statistical data.

2 The ViCoMap Tool

The architecture of the ViCoMap tool consists of three main components, as shown in Fig. 1. The base component of the tool is the *RapidMiner Linked Open Data extension* [5]. The extension hooks into the powerful data mining platform *RapidMiner*[8], and offers operators for accessing LOD in RapidMiner, allowing for using it in sophisticated data analysis workflows using data from LOD. 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 redundant data found in dfferent datasets, and wraps additional services such as *DBpedia Lookup*[9] and *DBpedia Spotlight*[10].

All processes built within the RapidMiner platform can be exposed as Web services through the RapidMiner Server, which can be consumed in a user Web

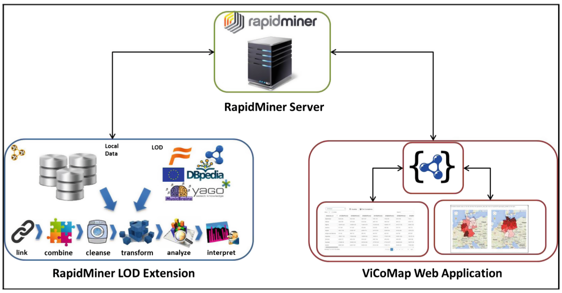


Fig. 1. ViCoMap architecture

[6] http://code.know-center.tugraz.at/search.
[7] http://dbpedia.org/.
[8] http://www.rapidminer.com.
[9] http://lookup.dbpedia.org.
[10] http://lookup.dbpedia.org.

application. We use such a setup to integrate the functionalities of the RapidMiner LOD extension, as well as the functionalities of RapidMiner built-in operators, in the ViCoMap Web application. The ViCoMap Web application offers three main functionalities to the end-user: *Data Import*, *Correlation Analysis*, and *Visualization on Maps*.

2.1 Data Import

There are three options to import data, i.e.,import a dataset published using the RDF Data Cube vocabulary, import data from a SPARQL endpoint, and import data from a local file.

Import RDF Data Cubes: To import a dataset published using the RDF Data Cube vocabulary, the user first needs to select the data publisher source, and a dataset that will be explored. Currently, we provide a static list of most used RDF Data Cube publishers, like WorldBank[11]. After selecting a data cube and the dimensions to be analyzed, the dataset is loaded into RapidMiner by the LOD extension.

SPARQL Data Import: To import data from a SPARQL endpoint, the user first needs to select a SPARQL endpoint and to provide a SPARQL query. The tool offers a SPARQL query builder assistant, which helps the user formulate queries such as *Select the number of universities per federal state in Germany*, by selecting a set of spatial entities (states in Germany) and a subject entity (universities).

Local Dataset Import: The user can import data from a local CSV file.

2.2 Correlation Analysis

Once the data is loaded, the user can select a column that will be used for correlation analysis. The user can choose the LOD sources that will be explored to find interesting factors that correlate with the target value at hand. To link the data at hand to remote LOD datasets, the tool exploits existing `owl:sameAs` links, and it automatically creates additional links, e.g., via *DBpedia Lookup* for non-linked datasets, such as CSV files. From the data retrieved from the additional LOD sources, a simple correlation analysis is performed to find simple correlations of the generated features and the target value under examination. The discovered correlations are sorted by confidence and presented to the user.

Visualization on Maps. After the correlation analysis is completed, the user can visualize any correlation on a map, using the Google Maps API[12] and displaying two maps for the correlated values side by side. The shape data of the

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

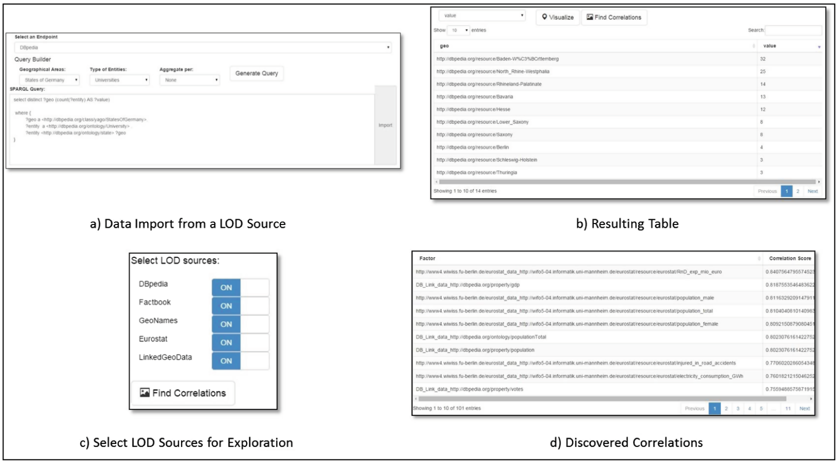
[12] https://developers.google.com/maps/.

geographical entities is retrieved from GADM. DBpedia provides external links to the GADM dataset, which were created using different heuristics based on the label and coordinates of geographical entities [6]. DBpedia 2014 contains 65, 616 links to the GADM dataset, for entities on different administration level, e.g., municipalities, regions, states, departments, countries, etc. Such links allow us to visualize spatial entities on any administrative level.

3 Use Case: Number of Universities per State in Germany

In this use case, we analyze which factors correlate with the number of universities per state in Germany[13]. To import the initial data we use the query builder assistant from the SPARQL data import tab (Fig. 2a). After executing the query, the data is presented in a table with two columns, i.e., the DBpedia URI for each state, and the number of universities per state (Fig. 2b). By pressing the button *Find Correlations*, we can select the LOD sources that will be included in the correlation analysis (Fig. 2c). Next, the discovered correlations are presented in a new table with two columns, i.e., a column with the factor label, and the a column with the correlation confidence (Fig. 2d).

We can see that in this case, as shown in Fig. 3, the highest positive correlation is the RnD expenses of the states (+0.84), which is retrieved from Eurostat. The highest negative correlation is the latitude of the states (−0.73), which is retrieved from GeoNames, which reflects the north-south gradient of the wealth distribution in Germany.[14]



```
select distinct ?geo (count(?entity) AS ?value)

 where {
        ?geo a <http://dbpedia.org/class/yago/StatesOfGermany>.
        ?entity  a <http://dbpedia.org/ontology/University> .
        ?entity <http://dbpedia.org/ontology/state> ?geo
}
```

a) Data Import from a LOD Source

geo	value
http://dbpedia.org/resource/Baden-W%C3%BCrttemberg	32
http://dbpedia.org/resource/North_Rhine-Westphalia	25
http://dbpedia.org/resource/Rhineland-Palatinate	14
http://dbpedia.org/resource/Bavaria	13
http://dbpedia.org/resource/Hesse	12
http://dbpedia.org/resource/Lower_Saxony	8
http://dbpedia.org/resource/Saxony	8
http://dbpedia.org/resource/Berlin	4
http://dbpedia.org/resource/Schleswig-Holstein	3
http://dbpedia.org/resource/Thuringia	3

Showing 1 to 10 of 14 entries

b) Resulting Table

Select LOD sources:

DBpedia	ON
Factbook	ON
GeoNames	ON
Eurostat	ON
LinkedGeoData	ON

c) Select LOD Sources for Exploration

Factor	Correlation Score
http://www4.wiwiss.fu-berlin.de/eurostat_data_http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/RnD_exp_mio_euro	0.8407564795574523
DB_Link_data_http://dbpedia.org/property/gdp	0.8187553546483622
http://www4.wiwiss.fu-berlin.de/eurostat_data_http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/population_male	0.8116329209147911
http://www4.wiwiss.fu-berlin.de/eurostat_data_http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/population_total	0.8104040810140983
http://www4.wiwiss.fu-berlin.de/eurostat_data_http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/population_female	0.8092150879080451
DB_Link_data_http://dbpedia.org/ontology/populationTotal	0.8023076161422752
DB_Link_data_http://dbpedia.org/property/population	0.8023076161422752
http://www4.wiwiss.fu-berlin.de/eurostat_data_http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/injured_in_road_accidents	0.7706020286054348
http://www4.wiwiss.fu-berlin.de/eurostat_data_http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/electricity_consumption_GWh	0.7601821215046252
DB_Link_data_http://dbpedia.org/property/votes	0.7559488575871915

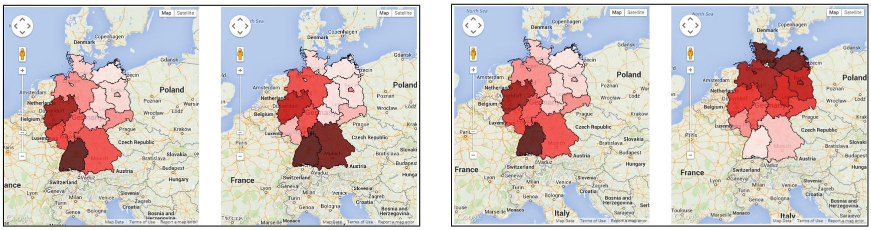
Showing 1 to 10 of 101 entries

d) Discovered Correlations

Fig. 2. German states use case workflow

[13] States of Germany: http://en.wikipedia.org/wiki/States_of_Germany.

[14] http://www.bundesbank.de/Redaktion/EN/Topics/2013/2013_07_10_to_save_or_not_to_save_private_wealth_in_germany.html.



(a) Positive correlation between #universities (left) and RnD expenses (right) per state

(b) Negative correlation between #universities (left) and latitude (right) per state

Fig. 3. Correlations visualized on a map using GADM geographical shape data

4 Conclusion

With this demo, we have introduced the web-based ViCoMap tool, which allows the users to analyze statistical data, and visualize it on maps using external knowledge from Linked Open Data. While at the moment, we use only literal data properties and types for finding correlations, we aim at a more intelligent exploration of the feature space, e.g., by automatically finding meaningful aggregations of different measures.

Acknowledgements. The work presented in this paper has been partly funded by the German Research Foundation (DFG) under grant number PA 2373/1-1 (Mine@LOD).

References

1. Klímek, J., Helmich, J., Neasky, M.: Application of the linked data visualization model on real world data from the Czech LOD cloud. In: 6th International Workshop on the Linked Data on the Web (LDOW 2014) (2014)
2. Mutlu, B., Hoefler, P., Tschinkel, G., Veas, E.E., Sabol, V., Stegmaier, F., Granitzer, M.: Suggesting visualisations for published data. In: Proceedings of IVAPP, pp. 267–275 (2014)
3. Paulheim, H.: Generating possible interpretations for statistics from linked open data. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 560–574. Springer, Heidelberg (2012)
4. Peña, O., Aguilera, U., López-de Ipiña, D.: Linked open data visualization revisited: a survey. Under Review at Semant. Web J. (2015)
5. Ristoski, P., Bizer, C., Paulheim, H.: Mining the web of linked data with rapidminer. In: Semantic Web Challenge at ISWC (2014)
6. Ristoski, P., Paulheim, H.: Analyzing statistics with background knowledge from linked open data. In: Workshop on Semantic Statistics (2013)

7. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: Mika, P., et al. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 245–260. Springer, Heidelberg (2014)
8. Tilahun, B., Kauppinen, T., Keßler, C., Fritz, F.: Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation. JMIR Med. Inf. **2**(2), 196–208 (2014)