Analyzing Statistics with Background Knowledge from Linked Open Data

Petar Ristoski and Heiko Paulheim

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

Abstract. Background knowledge from Linked Open Data sources, such as DBpedia, Eurostat, and GADM, can be used to create both interpretations and advanced visualizations of statistical data. In this paper, we discuss methods of linking statistical data to Linked Open Data sources and the use of the Explain-a-LOD toolkit. The paper further shows exemplary findings and visualizations created by combining the statistics datasets with Linked Open Data.

1 Introduction

Linked Open Data (LOD) provide a wide variety of information – ranging from general purpose datasets to government and life science data [1] All that information can potentially serve as background knowledge when interpreting statistics.

This paper discusses means to automatically enrich statistical datasets with information from Linked Open Data, and shows how that background knowledge can serve as a means to create possible interpretations as well as advanced visualization of the statistical datasets.

2 Datasets and Linking

For our challenge contribution, we have used both the *France* and the *Australia* dataset.[2] For France, we use data on region level and department level, for Australia, we use only data on state level (see below). In all three cases, we use the aggregated unemployment rate (over all sexes and age groups) as a target variable, as well as look into sex-specific and youth unemployment rates.

As data sources for background knowledge, we use four datasets in Linked Open Data:

- *DBpedia*[3] as a general purpose dataset
- *Eurostat*[4] as a government dataset (only for regions in France dataset)

[1] http://lod-cloud.net/
[2] http://www.datalift.org/en/event/semstats2013/challenge
[3] http://dbpedia.org/
[4] http://wifo5-03.informatik.uni-mannheim.de/eurostat/

– *GADM*[5] as a dataset providing geographical shape data
– *LinkedGeoData*[6] as a dataset providing information on geographical objects

We link the geographic entities (i.e., departments, regions, etc.) in the statistical datasets to the corresponding entities in the LOD datasets as a first step to enrich the statistical datasets with background knowledge.

2.1 Linking to DBpedia

For linking the datasets to DBpedia, we use the *DBpedia Lookup Service*,[7] which provides a keyword search interface to DBpedia. We search for the label given for the geographic entity in the statistical dataset. First, a list of all possible n-grams is generated, using the label's tokens. Then, for each n-gram we use the DBpedia keyword search service to retrieve all DBpedia entities that contain the n-gram, restricting the search results to instances of type *Place* and *AdministrativeRegion*. As a final result, we choose the DBpedia entity where the edit distance between the entity's label and the original label is minimum.

2.2 Linking to Eurostat

Eurostat does not provide a search interface. Hence, we use SPARQL queries for finding geographic entities that have the label given in the statistical datasets. Since this approach, unlike the DBpedia Lookup Service based search, is not tolerant to alternate spellings, we search for individual tokens in the label and compare the resulting entity's label with the original label. To do so, a list of all possible n-grams is generated, using the label's tokens. Then, we issue a SPARQL query listing all regions whose label contain the respective n-gram. As a final result, we choose the Eurostat entity where the edit distance between the entity's label and the original label is minimum.

2.3 Linking to GADM

While links to GADM are provided in DBpedia, they are quite scarce (for the France dataset, only one entity was linked to GADM). Thus, using the DBpedia links directly is not feasible.

GADM provides a search interface which allows searching for entities by their coordinates as well as by their names. Since we found that searching by name was error-prone due to alternate spellings and ambiguous names, we use the following three search approaches:

Simple search uses the coordinates given in DBpedia to find entities in GADM, and compares their names to the labels given in the statistical dataset

Average search uses the average of coordinates of all objects linked to a DBpedia entity, and compares their names to the labels given in the statistical dataset

[5] `http://gadm.geovocab.org/`
[6] `http://www.linkedgeodata.org`
[7] `http://lookup.dbpedia.org`

Table 1. Quality of links between the statistical and the LOD datasets. The table reports recall (r), precision (p), and F-measure (f).

Dataset/method	France Regions			France Departments			Australia States		
	r	p	f	r	p	f	r	p	f
DBpedia	100.0%	92.6%	96.2%	100.0%	97.0%	98.5%	100.0%	100.0%	100.0%
Eurostat	100.0%	100.0%	100.0%	/	/	/	/	/	/
GADM									
Simple	96.3%	100.0%	98.1%	84.2%	98.8%	90.9%	88.9%	100.0%	94.1%
Average	92.6%	100.0%	96.2%	97.0%	98.0%	97.5%	77.7%	100.0%	87.5%
Combined	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	88.9%	100.0%	94.1%

Combined uses simple search with average search as a fallback strategy if simple search does not return a result

The rationale for average search is that many of the DBpedia entities used in the statistical datasets had either missing or wrong coordinates. The average of linked entities (e.g., all cities that are located in a department) provides an alternative guessing for the coordinates of a larger geographical entity. Table 1 shows the link quality for all datasets. For the evaluation, gold standards were created manually. The numbers of linked geographic entities to the corresponding entities in the LOD datasets, for each of the statistical datasets are as follows:

- France Regions (#27): 27 linked to DBpedia, 26 linked to Eurostat, 27 linked to GADM
- France Departments (#101): 101 linked to DBpedia, 101 linked to GADM
- Australian States (#9): 8 linked to DBpedia, 9 linked to GADM

In contrast to the France dataset and the Australian states, linking the Australian SA3 and SA4 regions to DBpedia and GADM entities is not trivial, because of their ambiguous and descriptive naming conventions.[8] The regions are named according to the areas they represent: city names; combination of city names and employment centres; combination of regional areas; combination of state/territory names and regional areas. Thus, we have decided to discard the SA3 and SA4 data.

2.4 Linking to LinkedGeoData

LinkedGeoData contains data about a variety of singular geographic objects (restaurants, stores, factories, etc.). Since larger geographic units, such as departments or regions, do not exist in LinkedGeoData, we cannot link the data directly. Instead, we use a two stage process for extracting data from LinkedGeoData:[9]

1. Retrieve all objects that are inside the minimal rectangle surrounding the polygon shape retrieved from GADM

[8] `http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/B01A5912123E8D2BCA257801000C64F2?opendocument`

[9] Since queries for objects that are inside a polygon other than a rectangle are not possible in LinkedGeoData

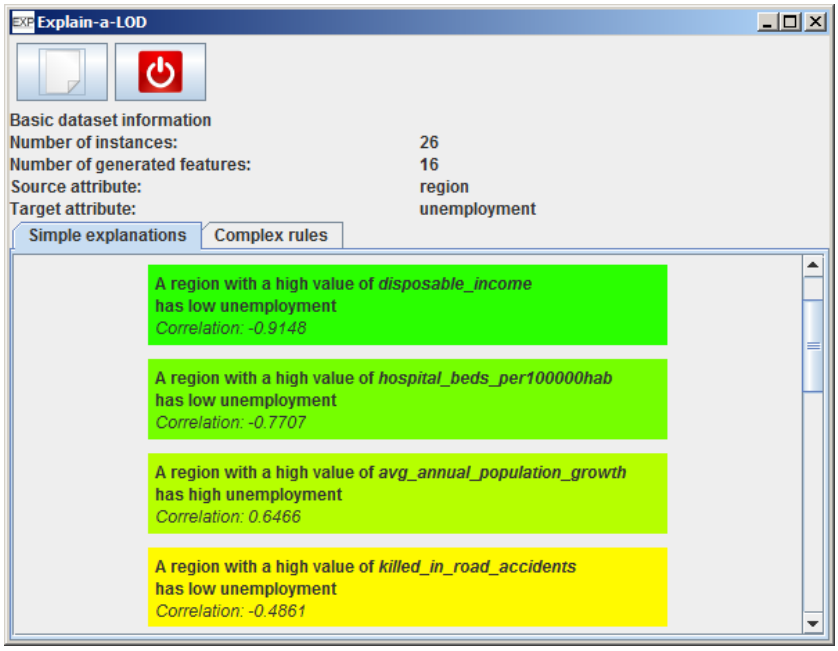


Fig. 1. Screenshot of the Explain-a-LOD tool

2. Remove all objects that are not inside the polygon shape retrieved from GADM

From those objects, we create features that count the number of instances of each type (e.g., number of restaurants) inside a geographic area.

3 Analysis with FeGeLOD and Explain-a-LOD

FeGeLOD [2] is a tool that takes a data table (in our case: a table with statistical regions and unemployment rates) as input and enriches the table with additional features (i.e., data columns) from Linked Open Data. The framework performs entity linking (i.e., identifying corresponding entities in a LOD dataset, as discussed above), feature generation (i.e., reading data from LOD and writing it into the data table), as well as filtering of useless features to deal with noise in LOD data sources. The feature generation step also allows for performing aggregations, e.g., counting the number of resources of a certain type connected to a given entity (e.g., the number of companies located in a region).

Explain-a-LOD [1] uses the FeGeLOD framework in order to create explanations out of statistics, building on correlation analysis and rule learning on the features generated by FeGeLOD. The tool encapsulates the FeGeLOD generation steps and can process a statistics file fully automatically. Fig. 1 shows a screenshot of Explain-a-LOD.

4 Example Results

For the Australian data, it was difficult to obtain meaningful results. As discussed above, we only analyzed the Australian data on state level, where the data turned out to be too scarce (only nine states) for drawing meaningful conclusions. Therefore, this section concentrates on findings on the France datasets (region and department level).

4.1 Explanations

From all the datasets, the best results are obtained from linking the French regions dataset to Eurostat. Besides obvious results (unemployment is negatively correlated with disposable income or GDP of a region), there are also some more interesting correlations for the unemployment rate, such as number of hospital beds per inhabitants (negative), RnD spendings (negative), electrical energy consumption (negative), or population growth (positive). Furthermore, some correlations are found that are rather surprising, e.g., number of casualties in traffic accidents (negative).[10]

With knowledge from DBpedia, we observe that the main findings refer to high unemployment in the overseas departments of France. For example, correlations are found with the types *Outermost regions of the EU*, *African Islands*, and *Islands in the Indian Ocean*, and also in the negative correlation of latitude and unemployment rate. Another example is the positive correlation of the number of footballer players born in a region and the unemployment rate – a fact that may be explained with many of the French top football players having their origins in French overseas departments.[11] Furthermore, this effect mainly applies to male unemployment, while female unemployment is rather on an average level in the overseas departments.

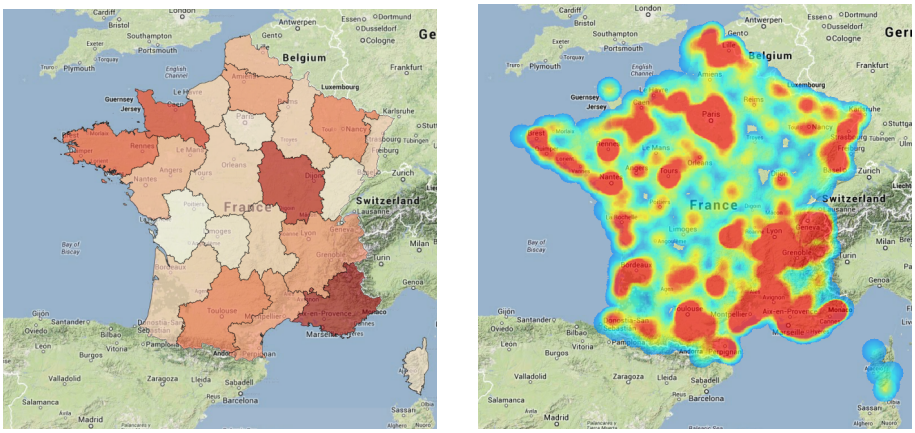
Adding knowledge from LinkedGeoData provides additional insights. For example, there is a high (positive) correlation of unemployment with Police stations and Fastfood restaurants.

4.2 Visualizations

With the help of the polygon shape data in GADM, we are able to visualize the original data geographically, as well as visualize the explanations found, in order to inspect them visually. Fig. 2 depicts the visualization of the original unemployment data by region, created using the polygon data from GADM, as well as the police stations in France, which were found to weakly correlate with the unemployment rate.

[10] A possible explanation is that with more people commuting to work, more traffic accidents occur.

[11] `http://en.wikipedia.org/wiki/France\_national\_football\_team\#Representing\_multi-ethnic\_France`



(a) Unemployment by region (b) Heat map of police stations

Fig. 2. Maps produced for unemployment by region (left) from the input dataset, and number of police stations (right) from LinkedGeoData. The high concentration of police stations in south-eastern France can be observed, which correlates with the high unemployment rate in that region.

5 Conclusion and Outlook

In this paper, we have shown how external knowledge from Linked Open Data can help providing both possible explanations as well as advanced visualizations for interpreting statistical data.

As future work, we aim at more sophisticated feature generation algorithms, which at the moment often cause problems due to scalability. For example, computing quotients (e.g., number of industry companies per inhabitants) could provide more meaningful insights, however, with several thousand features extracted upfront, require strategies to be performed in a scalable way.

For providing the visualizations, mappings from DBpedia to GADM were generated. These mappings have been computed on the whole set of populated places in DBpedia and will become part of the DBpedia 3.9 release, pushing the number of links from DBpedia to GADM from 2,000 to 39,000. Furthermore, additional linking operators, which have been described above, will become part of the next FeGeLOD version.

References

1. Heiko Paulheim. Generating possible interpretations for statistics from linked open data. In *9th Extended Semantic Web Conference (ESWC)*, volume 7295 of *LNCS*, pages 560–574, 2012.
2. Heiko Paulheim and Johannes Fürnkranz. Unsupervised generation of data mining features from linked open data. In *International Conference on Web Intelligence and Semantics (WIMS'12)*, 2012.