Explain-a-LOD: Using Linked Open Data for Interpreting Statistics

Heiko Paulheim
Technische Universität Darmstadt
Hochschulstrasse 10, 64283 Darmstadt
paulheim@ke.tu-darmstadt.de

ABSTRACT
While statistics are omnipresent, e.g., depicting the corruption in different countries, it is often not trivial to find the explanation for a statistical effect, e.g., *why* the corruption is higher in some countries than in others. The necessary facts that can explain a statistic are often not contained in the statistics file itself. This demo shows *Explain-a-LOD*, a tool for generating possible explanations for statistics from Linked Open Data. The tool accepts statistical data as input, and it automatically retrieves data from the Linked Open Data cloud and generates possible explanations.

ACM Classification Keywords
H.2.8 Database Management: Database Applications—*Data Mining, Statistical Databases*; I.2.1 Artificial Intelligence: Applications and Expert Systems

General Terms
Algorithms, Design

Author Keywords
Linked Open Data, Statistics, Semantic Web, Data Analysis

INTRODUCTION

Statistical data is very wide-spread in the media. In many cases, those statistics only contain a few attributes, e.g., a country's name and its corruption index. When asking for an explanation for the phenomenon depicted by the statistics, such as *why* is the corruption higher in country A than in country B, additional information is needed.

Explanations for such phenomena are usually found based by analyzing the correlation of the target attribute with other attributes: for example, the corruption in a country may be negatively correlated with the GDP (gross domestic product) per capita. Such correlations can be found with statistical methods. However, those methods can only be applied if the corresponding attribute (in this case: GDP per capita) is contained in the statistics, which is often not the case. Especially when there are no initial hypotheses, it is difficult to select relevant data for performing the analysis.

In this paper, we introduce *Explain-a-LOD*[1], a tool for automatically generating hypotheses for interpreting statistics. Given a statistic, which may consist only of two columns, such as a country and an indicator of interest, it automatically retrieves information about that country from Linked Open Data (LOD) [1], a large collection of semantically annotated data sets, and creates hypotheses based on that information.

PROTOTYPE

Figure 1 depicts a high-level view of Explain-a-LOD. The statistical data to examine, usually a tabular CSV (comma separated values) file, is loaded into the feature generation toolkit *FeGeLOD* [2]. That toolkit performs three steps:

1. Recognizing corresponding entities in the LOD cloud. For example, an entry about Germany would be tagged with `http://dbpedia.org/resource/Germany`, which serves as an entry point to the LOD cloud for retrieving further information about Germany.

2. Generating attributes (or *features*, as they are commonly called in data mining) for the corresponding entities. FeGeLOD has six different generation strategies that may be combined for creating features: using type information, data properties, and qualified as well as unqualified incoming and outgoing relations.

3. For being able to process the data further in reasonable time, a simple pre-selection of features is performed to reduce the complexity of the data set.

The result of running FeGeLOD on the data set is an enriched data set which contains more attributes. That data set is then processed in two different ways. A simple correlation analysis is performed to find simple correlations of the generated features and the target value under examination. Rule learning is used for discovering more complex explanation patterns that involve more than one feature.

Both the correlated attributes and the discovered rules are presented to the user in the user interface of Explain-a-LOD in a verbalized way. In order to allow the user to quickly grasp the significance of an explanation (the absolute correlation or the rule learner's confidence in a rule, respectively), the

IUI'12, February 14–17, 2012, Lisbon, Portugal.
ACM 978-1-4503-1048-2/12/02.

[1] `http://www.ke.tu-darmstadt.de/resources/explain-a-lod`

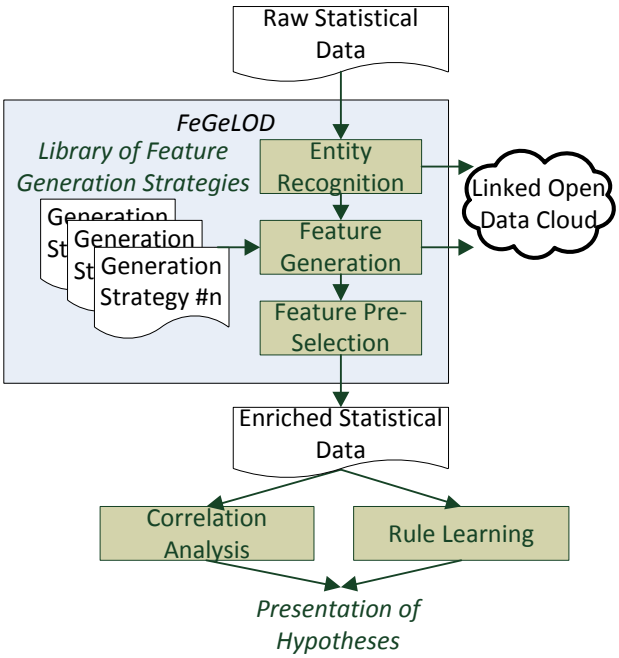

Figure 1. Architecture of Explain-a-LOD. The tool enriches a raw statistics file, using different generators extracting information from Linked Open Data. Correlation analysis and rule learning are employed on the enriched statistics file for creating possible explanations for the statistics.

hypotheses are color coded, using green for high-confidence hypotheses, going over to red for low-confidence hypotheses, as shown Figure 2.

EXPERIMENTS

We have tested our approach on different datasets, such as the 1999 Mercer quality of living survey[2] and the 2010 Corruption Perceptions Index by Transparency International[3]. The first observation is the that the hypotheses that are ranked highest are always useful concepts. Examples for such hypotheses include: *Member states of the EU (European Union) have a low corruption index*, or *Capitals of Africa have a low perceived quality of living*. It is important to notice that the information that a country is an EU member state or that a city is a capital of Africa is not included in the input dataset, but extracted from Linked Open Data.

Some observations from our first set of experiments hint to a set of current limitations of our approach. First, the data quality in LOD is not always good, and information may be inequally distributed. For example, in the hypotheses stating *Cities which are the headquarter of many companies have a high quality of living*, the condition should be more carefully read as *Cities for which the information is stored that they are the headquarter of many companies that have an entry in the LOD cloud*. Furthermore, many data sets in LOD have regional (mostly US-centric) bias, which may be propagated into the generated hypotheses.

[2] http://across.co.nz/qualityofliving.htm
[3] http://www.transparency.org/policy_research/surveys_indices/cpi/2010/results

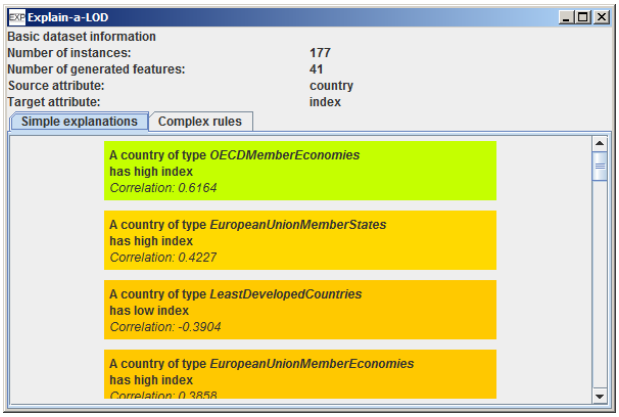

Figure 2. Screenshot of the Explain-a-LOD tool, showing possible explanations for a statistics coded with different colors.

Second, the quality of labels in Linked Open Data sometimes makes it hard to verbalize the hypotheses. Although Linked Open Data follows the structure of subject, predicate, and object, the labels for predicates are often not proper verbs.

Third, the approach can only present correlations to the user, but it is not capable of telling about the causality. In the example shown above, it is not known whether the low corruption of a country was a prerequisite for becoming an EU member state, if its low corruption is a consequence of the country being an EU member state, or if both have a common cause.

CONCLUSION AND FUTURE WORK

With this demo, we have introduced a first prototype of *Explain-a-LOD*, a tool for finding possible explanations for statistical phenomena. For a given statistics, we extract additional information from LOD, and we use statistical as well as rule learning methods for producing hypotheses.

We plan to improve both the generation of hypotheses as well as the presentation of those hypotheses. On the generation side, we aim at using more sophisticated algorithms within the feature generation component *FeGeLOD*, such as the use of different rule learning algorithms for creating interesting hypotheses, as well as the exploitation of more data sets in Linked Open Data. On the presentation side, we aim at visual clustering of hypotheses, allowing for interactive refinements of explanations, as well as more intuitive, non-textual presentation of rules.

ACKNOWLEDGEMENTS

The work presented in this paper was supported by the German Science Foundation (DFG) project FU 580/2 “Towards a Synthesis of Local and Global Pattern Induction (GLocSyn)”.

REFERENCES

1. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.
2. H. Paulheim and J. Fürnkranz. Unsupervised Feature Generation from Linked Open Data. Technical Report TUD-KE-2011-2, Technische Universität Darmstadt, 2011.