# Generating Possible Interpretations for Statistics from Linked Open Data

Heiko Paulheim

Technische Universität Darmstadt
Knowledge Engineering Group
paulheim@ke.tu-darmstadt.de

**Abstract.** Statistics are very present in our daily lives. Every day, new statistics are published, showing the perceived quality of living in different cities, the corruption index of different countries, and so on. Interpreting those statistics, on the other hand, is a difficult task. Often, statistics collect only very few attributes, and it is difficult to come up with hypotheses that explain, e.g., *why* the perceived quality of living in one city is higher than in another. In this paper, we introduce *Explain-a-LOD*, an approach which uses data from Linked Open Data for generating hypotheses that explain statistics. We show an implemented prototype and compare different approaches for generating hypotheses by analyzing the perceived quality of those hypotheses in a user study.

## 1 Introduction

Statistical data plays an important role in our daily lives. Every day, a new statistic is published, telling about, e.g., the perceived quality of living in different cities (used as a running example throughout the following sections), the corruption in different countries, or the box office revenue of films. While it is often possible to retrieve a statistic on a certain topic quite easily, *interpreting* that statistic is a much more difficult task. The raw data of a statistic often consists only of a few attributes, collected; in the extreme case, it may only comprise a source and a target attribute, such as a city and its score. Therefore, formulating hypotheses, e.g., *why* the perceived quality of living is higher in some cities than in others is not easy and requires additional *background information*.

While there are tools for discovering correlations in statistics, those tools require that the respective background information is already contained in the statistic. For example, the quality of living in a city may depend on the population size, the weather, or the presence of cultural institutions such as cinemas and theaters. For discovering those correlations, the respective data has to be contained in the dataset. For creating useful hypotheses, the dataset should contain a larger number of attributes, which makes the compilation of such a dataset a large amount of manual work.

More severely, the selection of attributes for inclusion in a statistical dataset introduces a bias: attributes are selected since the person creating the dataset

already assumes a possible correlation. For *disovering new and unexpected hypotheses*, this turns out to be a chicken-and-egg problem: we have to know what we are looking for to include the respective attribute in the dataset. For example, if we assume that the cultural live in a city influences the quality of living, we will include background information about theaters and festivals in our dataset.

For many common statistical datasets (e.g. datasets which relate real-world entities of a common class with one or more target variables), there is background information available in Linked Open Data [2]. In the quality of living example, information about all major cities in the world can be retrieved from the semantic web, including information about the population and size, the weather, and facilities that are present in that city. Thus, Linked Open Data appear to be an ideal candidate for generating attributes to enhance statistical datasets, so that new hypotheses for interpreting the statistic can be found.

In this paper, we introduce *Explain-a-LOD*, a prototype for automatically generating hypotheses for explaining statistics by using Linked Open Data. Our prototype implementation can import arbitrary statistics files (such as CSV files), and uses *DBpedia* [3] for generating attributes in a fully automatic fashion. While our main focus is on enhancing statistical datasets with background information, we have implemented the full processing chain in our prototype, using correlation analysis and rule learning for producing hypotheses which are presented to the user.

The rest of this paper is structured as follows. In Sect. 2, we introduce our approach, show a proof-of-concept prototype, and discuss the underlying algorithms. Section 3 discusses the validity of the approach and the individual algorithms with the help of a user study. In Sect. 4, we review related approaches. We conclude with a summary and an outlook on future research directions.

## 2 Approach

We have developed an approach for using Linked Open Data in a way that new hypotheses for interpreting statistics can be generated. The approach starts with a plain statistic, e.g., a CSV file, and comes up with hypotheses, which can be output in a user interface. To that end, three basic steps are performed: first, the statistical data is enhanced such that additional data from Linked Open Data is added, second, hypotheses are sought in this enhanced data set by means of correlation analysis and rule learning, and third, the hypotheses that are found are presented to the user. The basic workflow of our approach is depicted in Fig. 1. We have implemented that approach in a proof-of-concept prototype.

### 2.1 Data Preparation

In the first step, the statistical data is prepared using our feature generation toolkit *FeGeLOD* [15]. FeGeLOD itself performs three steps: entity recognition, feature generation, and feature selection, as depicted in Fig. 2.

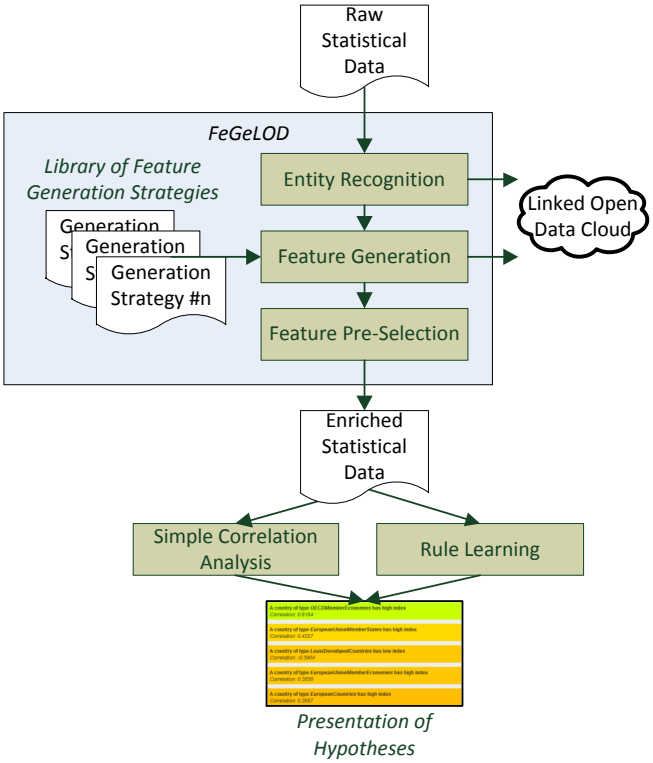


**Fig. 1.** Basic prototype of the Explain-a-LOD prototype

In the first step, the entities that the statistic is about – cities in the quality of living example – have to be mapped to corresponding URIs in Linked Open Data, so that additional information about those entities can be retrieved. For the first prototype of FeGeLOD, we have used a very basic mechanism for entity recognition: it retrieves all possible matching resources, e.g., such as `http://dbpedia.org/resource/Vancouver` for the city name *Vancouver*, and performs an optional type check, e.g. for `dbpedia-owl:City`. If the landing page of the first step is a disambiguation page, all disambiguated entities are followed, and the first one matching the type checks is used.

### 2.2 Generation of Hypotheses

Once an entity is recognized, attributes (or *features*, as they are called in data mining) can be generated for that entity as a second step. In the prototype, FeGeLOD supports six different generation strategies:

- *Simple datatype properties*, such as the population of a city.
- *Class information*. For example, a city can be of type `dbpedia-owl:City`, among others. Since DBpedia also uses *YAGO* types [17], there are also a lot

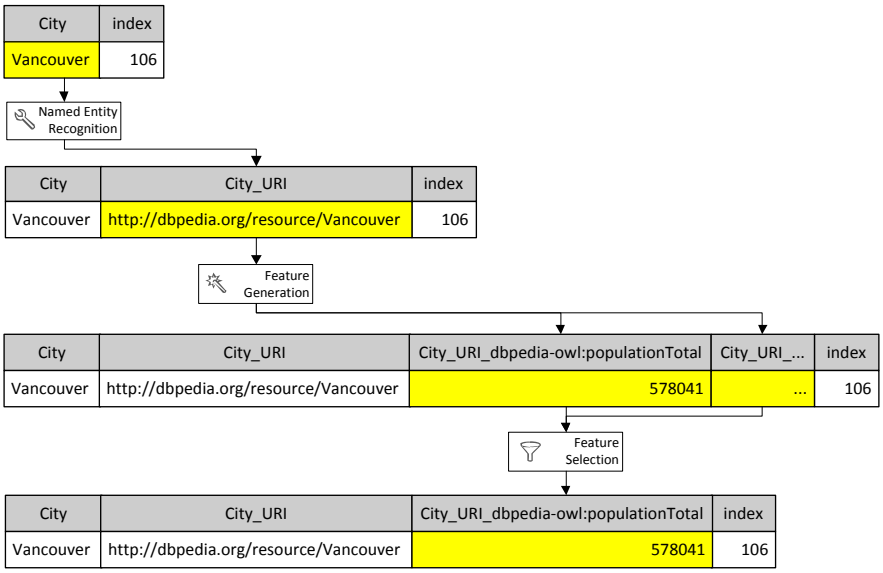


**Fig. 2.** The three steps performed at the data preparation stage

of very specific types that can be used as features, such as `yago:PopulatedCoastalPlacesInCanada`.

- *Unqualified relations.* Features are generated for incoming and outgoing relations without any information about the related entity. For example, a city may have incoming relations of type `dbpedia-owl:foundationPlace`. Those features can be generated as boolean (incoming/outgoing relations of the specified type exist or not) or numeric (counting the related entities) features[1].
- *Qualified relations.* Unlike unqualified relations, boolean or numeric features are generated including the type of the related entity, e.g., the presence or number of entities of type `dbpedia-owl:Company` which have a `dbpedia-owl:foundationPlace` relation to a city. The detailed YAGO typing system leads to a lot of very specific features, such as *number of airlines that are founded in 2000 that are located in a city.*

To illustrate how the individual strategies work, Fig. 3 shows an excerpt of DBpedia, depicting some information about the city of Darmstadt. Table 1 shows the features that are generated in this example by the individual generators.

[1] We are aware that the creation of such attributes neglects the two central semantic principles of Linked Open Data, i.e., the open world assumption and the non-unique name assumption. For example, the actual number of companies and organizations founded in a city will probably be higher than that in DBpedia. However, re-interpreting that feature, e.g., as *approximate number of* important *companies/organizations founded in a city* fixes these issues, and still serves as a useful feature for analysing the statistic.

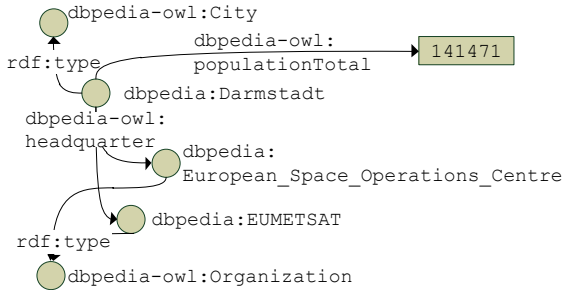


**Fig. 3.** An excerpt from DBpedia, showing data about Darmstadt

Not all of the features generated by the different strategies are equally helpful. For example, the generator for class information may generate a feature for the classes `dbpedia-owl:City` or even `owl:Thing`, which are `true` for all entities. Likewise, qualified relations may yield a large number of features which are not useful, such as the number of entities of type `yago:ArtSchoolsInParis` which are located in a city: this attribute will have a non-zero value only for one entity, i.e., Paris.

Since those features are very unlikely to produce useful hypotheses, we apply a simple heuristic to filter them out before processing the dataset in order to improve the runtime behavior of the remaining processing steps. Given a threshold $p, 0 \leq p \leq 1$, we discard all features that have a ratio of more than $p$ unknown, equal, or different values (different values, however, are not discarded for numeric features). In our previous experiments, values of $p$ between 0.95 and 0.99 have proved to produce data sets of reasonable size without reducing the results' quality significantly [15].

The result of the data preparation step is a table with many additional attributes. That table can then be further analyzed to generate possible hypotheses. Currently, we pursue two strategies for creating hypotheses:

- The correlation of each attribute with the respective target attribute is analyzed. Attributes that are highly correlated (positively or negatively) lead to a hypothesis such as "Cities with a high value of population have a low quality of living".
- Rule learning is used to produce more complex hypotheses which may take more than one feature into account. We have used the standard machine learning library *Weka* [4] for rule learning. Possible algorithms are class association rule mining [1], the use of separate-and-conquer rule learners [6], where in the latter case, only the first, i.e., most general rules are used, as the subsequent rules are often not valid on the whole data set.

### 2.3 Presentation of Hypotheses

After importing and processing a statistics file, the hypotheses found are presented to the user in a user interface, as depicted in Fig. 4. To that end, all

**Table 1.** Features generated for the example shown in Fig. 3

| Generator | Feature Name | Feature Value |
|---|---|---|
| Data properties | `dbpedia-owl:populationTotal` | 141471 |
| Types | `type_dbpedia-owl:City` | true |
| Unqualified relations boolean | `dbpedia-owl:headquarter_boolean` | true |
| Unqualified relations numeric | `dbpedia-owl:headquarter_numeric` | 2 |
| Qualified relations boolean | `dbpedia-owl:headquarter_type _dbpedia-owl:Organization_out_boolean` | true |
| Qualified relations numeric | `dbpedia-owl:headquarter_type _dbpedia-owl:Organization_out_numeric` | 2 |

hypotheses are verbalized. For example, a positive correlation between the type `yago:EuropeanCapitals` and the quality of living is turned into a sentence such as *In cities of type European Capitals, the quality of living is high.* Likewise, learned rules are verbalized.

All hypotheses have a quality measure. For simple correlations, it is the correlation coefficient itself. Rules learned by a rule learning algorithm also come with a confidence or accuracy measure provided by the algorithm. Therefore, the hypotheses may be sorted, presenting the most likely ones on top. Furthermore, to improve the usability, we use a color coding schema, depicting the best rated hypotheses in green, going over to red for the worst rated hypotheses.

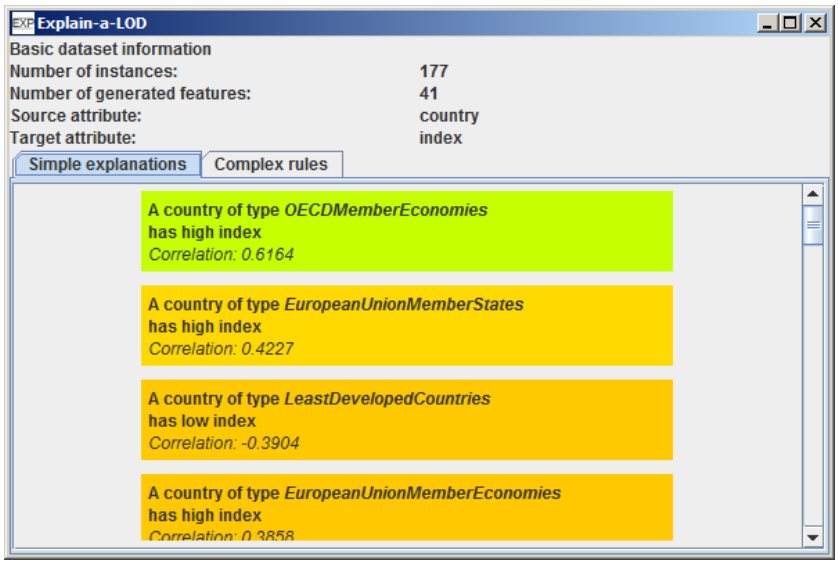


**Fig. 4.** Screenshot of the Explain-a-LOD User Interface

**Table 2.** Number of features generated for the two data sets used in the study. This table shows the numbers without any post-processing feature selection. The boolean and numerical variants of relations and qualified relations produce an equal number of features.

| | Mercer | Transparency International |
|---|---|---|
| Data | 1,205 | 614 |
| Types | 622 | 237 |
| Relations | 2,414 | 1,523 |
| Qualified Relations | 48,441 | 34,302 |

## 3 Experimental Evaluation

In order to examine the quality of the hypotheses generated by Explain-a-LOD, also with respect to the different feature generation strategies, we have asked a number of users to evaluate those hypotheses. To that end, users were presented a list of hypotheses generated by Explain-a-LOD, and they were asked to rate those hypotheses by the perceived plausibility. Furthermore, all participants were asked a number of general questions on the approach in the end.

### 3.1 Setup

We have conducted the user study with 18 voluntary participants, who were undergraduate and graduate students as well as researchers at Technische Universität Darmstadt. The participants were between 24 and 45 years old, 15 of them were male, 3 female.

For the evaluation, we have used two statistics datasets: the already mentioned Mercer quality of living survey with data[2], which comprises 218 cities, and the corruption perception dataset by Transparency International[3], which comprises 178 countries. With our entity recognition approach, we could map 97.7% of the cities and 99.4% of the countries to the corresponding URIs in DBpedia.

For each data set, we have generated hypotheses with the approaches discussed above, using the different feature generation algorithms, and used the top three hypotheses from both the simple correlation analysis and the rule learning approach. Table 2 depicts the number of features generated and used in each dataset.

For the rule learning approach, we also used a joint set of all feature generators, so that rules involving features from different generators could also be found. As the joint set of features cannot produce any new hypotheses when only regarding correlations of single features, that dataset was only used with the rule learning approach. After removing duplicates (a hypothesis with only

[2] Data available at `http://across.co.nz/qualityofliving.htm`

[3] Data available at `http://www.transparency.org/policy_research/surveys_indices/cpi/2010/results`

one feature can be found by both approaches), we had 37 hypotheses for the Mercer dataset and 38 hypotheses for the Transparency International dataset. Each participant was asked to evaluate all 75 hypotheses[4].

From those hypotheses, we have constructed a questionnaire listing those hypotheses for both datasets in random order, and asking for the plausibility of each hypothesis on a scale from 1 (worst) to 5 (best).

At the end of the questionnaire, the users were asked to which degree they feel that the hypotheses in total are useful, surprising, non-trivial, and trustworthy. Filling out a questionnaire took the participants between 15 and 20 minutes.

### 3.2 Results

The first goal was to understand which strategies for feature generation and for creating the hypotheses work well, also in conjunction. To that end, we analyzed the ratings of the respective hypotheses. Figure 5 shows the results for the Mercer dataset, Figure 6 shows the respective results for the Transparency International dataset. The intra-class correlation (i.e., the agreement score of the participants) was 0.9044 and 0.8977, respectively.

The first basic observation is that the evaluations for both datasets are very different. For the Mercer dataset, simple correlations produce the more plausible hypotheses, while for the Transparency International dataset, rule learning is significantly better in some cases. In both cases, the best rated hypotheses are produced when using the type features. In both cases, joining all the attributes in a common dataset did not lead to significantly better rules.

For the Mercer dataset, the best rated hypotheses were *Cities in which many events take place have a high quality of living* (found with correlation analysis from unqualified relations, average rating 3.94), and *Cities that are European Capitals of Culture have a high quality of living* (generated from a type feature with type `yago:EuropeanCapitalsOfCulture`, found both by correlation analysis and with a rule learner, rating 3.89). The worst rated hypotheses were *Cities where at least one music record was made and where at least 22 companies or organizations are located have a high quality of living* (generated with a rule learner from unqualified relations, rating 1.5) and *Cities that are the hometown of at least 18 bands, but the headquarter of at most one airline founded in 2000, have a high quality of living* (generated with a rule learner from qualified relations, rating also 1.5).

For the Transparency International Dataset, the two best rated hypotheses are *Countries of type Least Developed Countries have a high corruption index* (generated by correlation analysis from a type feature with type `yago:LeastDevelopedCountries`, rating 4.29), and *Countries where no military conflict is carried out and where no schools and radio stations are located have a high corruption index* (generated by rule learning from three different qualified relation features, rating 4.24). The two worst rated hypotheses are *Countries with many*

[4] The hypotheses used in the evaluation are listed at `http://www.ke.tu-darmstadt.de/resources/explain-a-lod/user-study`

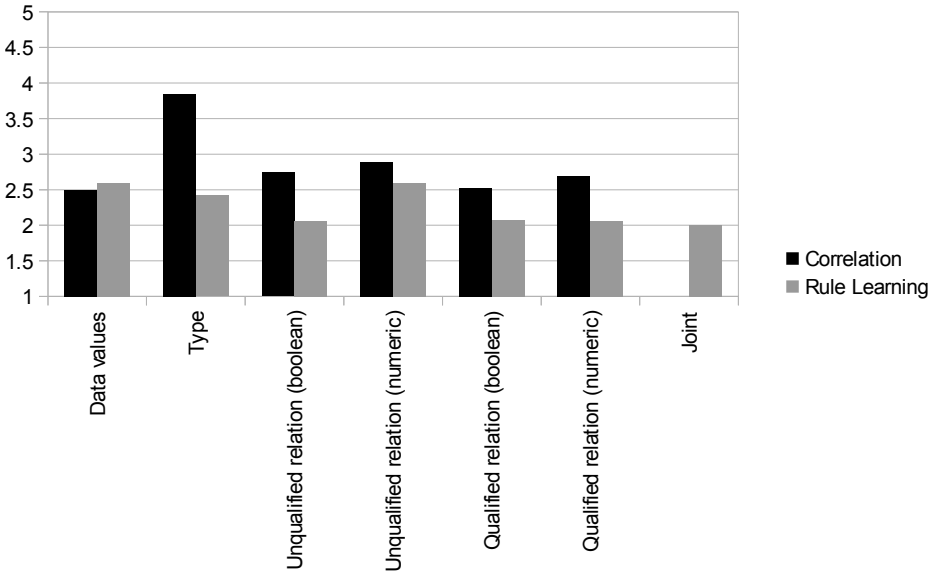


**Fig. 5.** Average user ratings of the hypotheses generated for the Mercer dataset, analyzed by feature generation and hypothesis generation strategy

*mountains have a low corruption index* and *Countries where no music groups that have been disbanded in 2008 come from have a high corruption index* (both generated by correlation analysis from a qualified relation feature, ratings 1.39 and 1.28, respectively).

There are some hypotheses that are rated badly, because the explanations they hint at are not trivial to see. For example, one hypothesis generated for the Mercer dataset is *Cities with a high longitude value have a high quality of living* (average rating 1.52). When looking at a map, this hypothesis becomes plausible: it separates cities in, e.g., North America, Australia, and Japan, from those in, e.g., Africa and India. Interestingly enough, a corresponding hypothesis concerning the latitude (which essentially separates cities in the third world from those in the rest of the world) was rated significantly ($p < 0.05$) higher (rating 3.15). Another example for an hypothesis that is not trivial to interpret is the following: *Countries with an international calling code greater than 221 have a high corruption index* (rating 1.69). Those calling codes mostly identify African countries. On the other hand, the following hypothesis is rated significantly higher (rating 4.0): *Countries in Africa have a high corruption index.*

The second goal of the user study was to get an impression of how the overall usefulness of the tool is perceived. Figure 7 shows the results of the general questions. The hypotheses got positive results on three of the four scales, i.e., the users stated that the results were at least moderately useful, surprising, and non-trivial. The latter two are significantly better than the average value of three with $p < 0.05$. The trustworthiness of the results, on the other hand,

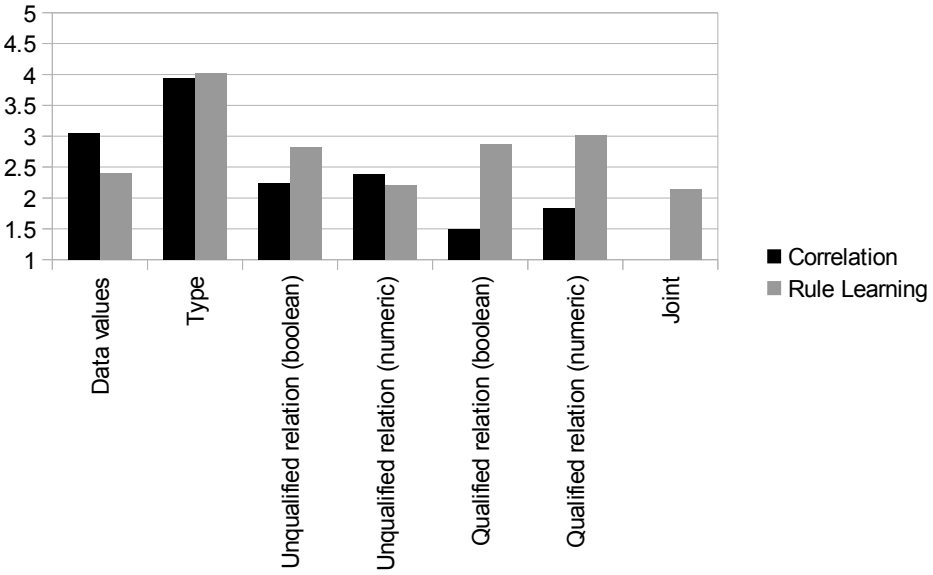


**Fig. 6.** Average user ratings of the hypotheses generated for the Transparency International dataset, analyzed by feature generation and hypothesis generation strategy

was not rated well ($p < 0.01$). These results show that the tool is well suited for generating *hypotheses*, but these hypotheses always need a human judging whether these hypotheses are valid explanations or not.

At the end of the questionnaire, users were asked to give some additional comments. One user was asking for detail information on certain explanations, e.g., showing the average corruption of African and non-African states for a hypothesis such as *Countries in Africa have a high corruption index*. Another user remarked that some rules are hard to comprehend without background knowledge (such as those involving latitude/longitude values, as discussed above).

Another user remarked that longer hypotheses were in general less plausible. This may partly explain the bad performance of the rule-based approaches on the Mercer dataset. Rule learning approaches most often seek to find rules that have a good coverage and accuracy, i.e., split the dataset into positive and negative examples as good as possible. Since rule learning algorithms may choose combinations of arbitrary features for that, it may happen that an unusual combination of features leads to a good separation of the example space, but that the resulting rule is not perceived as a very plausible one.

One example is the following rule, which was among the worst rated hypotheses (average rating of 1.5): *Cities which are the hometown of at least 18 bands, but are the headquarter of at most one airline founded in 2000, have a high quality of life*. While the second condition may increase the rule's accuracy by some percent, it decreases the perceived plausibility of the rule, mostly since there is no obvious coherence between bands and airlines. In contrast, the following rule

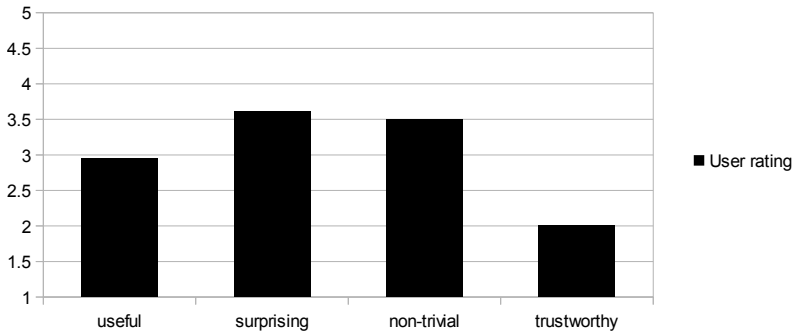


**Fig. 7.** Results of the overall rating of the Explain-a-LOD hypotheses

received a significantly ($p < 0.01$) higher average rating (2.72): Cities that are the origin of at least 33 artists and bands have a high quality of life. On the other hand, the first rule has an accuracy of 98.0%, while the second rule has an accuracy of only 88.6%. This shows that, while the accuracy of rules may increase with additional, non-related features, this does not necessarily imply an increase in the perceived plausibility. A similar observation can be made for correlation analysis: the best-rated hypothesis for the Transparency International dataset, *Countries of type Least Developed Countries have a high corruption index* is actually the one in the set of hypotheses with the lowest correlation (Pearson's correlation coefficient 0.39, rating 4.33).

Another observation made is that rule-based approaches are capable of finding very exact conditions, i.e., they find the value which separates best between positive and negative examples. One example are the following two corresponding rules: *Countries with a high HDI have a low corruption index* (average rating: 4.0, found with correlation analysis), and *Countries with a HDI less than 0.712 have a high corruption index* (average rating: 3.39, found with rule learner). While both hypotheses express the same finding, the second one, which is formulated in a more specific way, is rated significantly ($p < 0.05$) lower. These examples show that very accurate rules are not always perceived as plausible at the same time.

## 4 Related Work

There is a vast body of work that is concerned with the analysis of statistical data [14]. Given a statistic, there are various methods to find out correlations and interrelations of the attributes contained in those statistics. Highly developed toolkits such as *R* [9] can be used for performing such analyses.

Those methods always assume that all the possible attributes are known, and thus, they are only capable of finding correlations between attributes that are included in the statistic. The work presented in this paper can be seen as a complement to those approaches, as it enhances a dataset by a multitude of

attributes that can then be examined by such statistical analysis algorithms and tools.

One of the works closest to Explain-a-LOD is proposed by Zapilko et al. [20]. The authors propose a method for publishing statistical data as linked data, which allows for combining different of such data sets. Kämpgen and Harth suggest a similar approach for analyzing statistical linked data with online analytical processing (OLAP) tools [11]. They discuss a common schema for such data and present various case studies. While OLAP allows for asking for specific correlations (i.e., the user has to come up with the hypotheses by himself upfront), our approach *generates* hypotheses automatically. Furthermore, while we are able to exploit any arbitrary, general-purpose datasets, such as DBpedia, the authors of the two approaches are restricted to specialized statistical datasets, following a specific schema. Nevertheless, including such specific statistical linked data sets in our approach may help increasing the quality of our hypotheses significantly.

*g-SEGS* [13] uses ontologies as background knowledge in data mining tasks. Ontologies are used as additional taxonomic descriptions for nominal attributes. For example, a nominal attribute with the values *Student*, *Apprentice*, *Employee*, *Self-employed*, and *Unemployed* may be augmented with a taxonomy of those values. Thus, regularities that hold for all people in education (regardless of whether they are students or apprentices) may be found better. In contrast to our approach, g-SEGS uses T-Box information, while we use A-Box information. Furthermore, in g-SEGS, the ontology has to be known in advance and mapped to the dataset manually. This makes it difficult to discover *new* hypotheses, since the designer of the ontology can be tempted to model only those facts in the ontology that are considered relevant for the mining problem at hand.

SPARQL-ML [10] is an approach that foresees the extension of the SPARQL query language [18] with a specialized statement to learn a model for a specific concept or numeric attribute in an RDF dataset. Such models can be seen as explanations in the way we use them in *Explain-a-LOD*. However, the approach requires support of the endpoint in question, e.g., DBpedia, to support the SPARQL-ML language extensions. In contrast, our approach works with any arbitrary SPARQL endpoint providing Linked Open Data.

Mulwad et al. have proposed an approach for annotating tables on the web [12]. The authors try to automatically generate links to DBpedia both for entities in the table as well as for column names, which are linked to classes in ontologies. Unlike the approach presented in this paper, the authors are not concerned with creating hypotheses. Since tables are typical ways to present statistical data on the web, their approach could be a useful complement to the Explain-a-LOD for generating hypotheses on arbitrary tabular statistical data found on the web.

## 5 Conclusion and Future Work

In this paper, we have introduced *Explain-a-LOD*, an approach for using Linked Open Data as a means to interpret statistics. Given a "plain" statistics file, i.e., containing only a source and a target variable, such as a city and a numerical

indicator for that city, we map the values of the source variable to entities in Linked Open Data, gather additional attributes from those Linked Open Data entities, and use those attributes to generate hypotheses for explaining the statistic using correlation analysis as well as rule learning. The whole process from loading the statistic to presenting the hypotheses can be performed in a fully automatic manner.

We have conducted a user study, in which we asked people to rate hypotheses generated by Explain-a-LOD for two different datasets, as well as to give a general impression of the tool. The hypotheses received mixed ratings: while some approaches produce hypotheses of high value (especially those exploiting the types in DBpedia), others are not suitable for producing good hypotheses. In the overall rating, the study participants stated that the hypotheses generated by Explain-a-LOD are useful, surprising, and non-trivial.

Although often useful, the hypotheses generated by Explain-a-LOD should be handled with care. The data preparation algorithm disrespects some essential fundamentals of Linked Open Data, such as the open world assumption, when generating attributes such as *number of organizations with headquarter in that city*. Furthermore, there might be cultural biases in the Linked Open Data sets used. When generating a feature such as *number of (famous) persons born in that city*, a larger amount of information on, e.g., US celebrities may introduce a cultural bias [5]. Likewise, we have observed a slight bias in DBpedia towards facts from popular culture, since many of our hypotheses were concerned with bands and music records.

Additionally, Explain-a-LOD cannot distinguish correlations from causal relations: from an explanation such as *countries where many companies have their headquarter are less corrupt*, we cannot tell whether companies tend to choose such countries with low corruption as headquarters, or whether a flourishing economy leads to a lower corruption.

In the future, we want to extend our approach to other Linked Open Data sets, such as Freebase, and compare the quality of hypotheses that can be obtained with data preprocessing using such different datasets. Since many datasets are already linked to DBpedia, drawing background information from those datasets is not difficult once the entities in the statistic are mapped to DBpedia. It may also be useful to produce deeper features, such as *number of companies in that country that have more than X Mio.$ turnover*, which do not only take direct neighbors of the entity into account, but also further information about those entities. However, the explosion of the search space has to be taken care of in that case.

We have implemented Explain-a-LOD as a proof of concept prototype with a set of simple algorithms and toolkits. That prototype can be improved with respect to many aspects. The entity recognition step can be enhanced by using frameworks such as the DBpedia lookup service, or by adapting the algorithm by Mulwad et al. discussed above. The generation of hypotheses can be enhanced by adding further mechanisms, such as subgroup discovery [19], and more sophisticated algorithms for feature selection [8].

We have also recognized that some of the hypotheses generated by rule learners are not considered plausible if the conditions are not semantically coherent, such as in *Cities which are the hometown of at least 18 bands, but are the headquarter of at most one airline founded in 2000, have a high quality of life*, where bands and airlines are not semantically close, which lowers the total plausibility. A rule with two conditions involving, e.g., bands and TV stars, or airlines and logistics companies, would probably be perceived more plausible, since the semantic distance between the conditions is lower. Therefore, an interesting research direction would be finding accurate, but semantically coherent rules.

Concerning the presentation of the hypotheses, several improvements can be thought of. The sorting of hypotheses by their rating is essential to the user, since the best hypotheses are expected to be on top. However, our user study showed that the natural ratings (such as the correlation coefficient for simple attributes) do not always reflect the perceived plausibility. In future user studies, we want to explore the impact of different rating measures for hypotheses. Furthermore, the verbalization of hypotheses does not always work too well because of mixed quality of labels used in the datasets [7]. Here, we aim for more intuitive and readable verbalizations, such as proposed in [16].

Finally, an interactive user interface would be helpful, where the user can mark implausible hypotheses (such as a correlation between the number of mountains in a country and the country's corruption index) and receive an explanation and/or an alternative hypothesis. Taking the informal feedback from the user study into account, it would also be helpful to provide evidence for hypothesis, e.g., list those instances that fulfill a certain condition. Such a functionality might also help to improve the trust in the hypotheses generated by Explain-a-LOD, which was not perceived very high in our user study.

In summary, we have introduced an approach and an implemented prototype that demonstrates how Linked Open Data can help in generating hypotheses for interpreting statistics. The evaluation of the user study show that the approach is valid and produces useful results.

**Acknowledgements**

This work was supported by the German Science Foundation (DFG) project FU 580/2 "Towards a Synthesis of Local and Global Pattern Induction (GLoc-Syn)". The author would like to thank Sebastian Döweling, Aristotelis Hadjakos, and Axel Schulz for their valuable advice on designing the evaluation, and everybody who took the time to participate in the user study.

## References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. (1993) 207–216
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1–22

3. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics - Science Services and Agents on the World Wide Web **7**(3) (2009) 154–165
4. Bouckaert, R.R., Frank, E., Hall, M., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: WEKA — Experiences with a Java open-source project. Journal of Machine Learning Research **11** (September 2010) 2533–2541
5. Callahan, E.S., Herring, S.C.: Cultural bias in wikipedia content on famous persons. Journal of the American Society for Information Science and Technology **62**(10) (2011) 1899–1915
6. Cohen, W.W.: Fast effective rule induction. In: Twelfth International Conference on Machine Learning, Morgan Kaufmann (1995) 115–123
7. Ell, B., Vrandečić, D., Simperl, E.: Labels in the Web of Data. In: 10th International Semantic Web Conference (ISWC 2011). (2011)
8. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A., eds.: Feature Extraction – Foundations and Applications. Springer (2006)
9. Ihaka, R.: R: Past and future history. In: Proceedings of the 30th Symposium on the Interface. (1998)
10. Kiefer, C., Bernstein, A., , Locher, A.: Adding data mining support to sparql via statistical relational learning methods. In: 5th European Semantic Web Conference (ESWC 2008). (2008) 478–492
11. Kämpgen, B., Harth, A.: Transforming statistical linked data for use in olap systems. In: 7th International Conference on Semantic Systems (I-SEMANTICS 2011). (2011)
12. Mulwad, V., Finin, T., Syed, Z., Joshi, A.: Using linked data to interpret tables. In: Proceedings of the First International Workshop on Consuming Linked Data (COLD2010). (2010)
13. Novak, P.K., Vavpetič, A., Trajkovski, I., Lavrač, N.: Towards semantic data mining with g-segs. In: Proceedings of the 11th International Multiconference Information Society (IS 2009). (2009)
14. Ott, R.L., Longnecker, M.: Introduction to Statistical Methods and Data Analysis. Brooks/Cole (2006)
15. Paulheim, H., Fürnkranz, J.: Unsupervised Feature Generation from Linked Open Data. In: International Conference on Web Intelligence, Mining, and Semantics (WIMS'12). (2012)
16. Piccinini, H., Casanova, M.A., Furtado, A.L., Nunes, B.P.: Verbalization of rdf triples with applications. In: ISWC 2011 – Outrageous Ideas track. (2011)
17. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. WWW '07, ACM (2007) 697–706
18. W3C: SPARQL Query Language for RDF. `http://www.w3.org/TR/rdf-sparql-query/` (2008)
19. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Symposium on Pattern Discovery in Databases (PKDD'97). (1997)
20. Zapilko, B., Harth, A., Mathiak, B.: Enriching and analysing statistics with linked open data. In: Conference on New Techniques and Technologies for Statistics (NTTS). (2011)