Semantic Similarity and Correlation of Linked Statistical Data Analysis

Sarven Capadisli[1,3], Albert Meroño-Peñuela[2], Sören Auer[3], Reinhard Riedl[1]

[1]Bern University of Applied Sciences, E-Government-Institute, Bern, Switzerland, [2]VU University Amsterdam, Department of Computer Science, Amsterdam, Netherlands, [3]University of Bonn, Enterprise Information Systems Department, Bonn, Germany
info@csarven.ca, albert.merono@vu.nl, auer@cs.uni-bonn.de, reinhard.riedl@bfh.ch

Document ID: http://csarven.ca/sense-of-lsd-analysis

Abstract. Statistical data is increasingly made available in the form of Linked Data on the Web. As more and more statistical datasets become available, a fundamental question on statistical data comparability arises: To what extent can arbitrary statistical datasets be faithfully compared? Besides a purely statistical comparability, we are interested in the role that semantics plays in the data to be compared. Our hypothesis is that semantic relationships between different components of statistical datasets might have a relationship with their statistical correlation. Our research focuses in studying whether these statistical and semantic relationships influence each other, by comparing the correlation of statistical data with their semantic similarity. The ongoing research problem is, hence, to investigate why machines have a difficulty in revealing meaningful correlations or establishing non-coincidental connection between variables in statistical datasets. We describe a fully reproducible pipeline to compare statistical correlation with semantic similarity in arbitrary Linked Statistical Data. We present a use case using World Bank data expressed as RDF Data Cube, and we highlight whether dataset titles can help predict strong correlations.

Keywords: Linked Data • Statistics • Statistical database • Semantic Similarity • Correlation

1 Introduction

"There was this American who was afraid of a heart attack and he found out that the Japanese ate very little fat and almost did not drink wine but they had much less heart attacks than the Americans. But on the other hand he also found out that the French eat as much fat as the Americans and they drink much more wine but they also have less heart attacks. So he concluded that what kills you is speaking English" [1]. While computers can assist us to discover strong correlations in large amounts of statistical datasets, whether by

chance or through sophisticated methods, humans (or sometimes also known as *domain experts*) still need to be critical about the results and interpret them appropriately. This implies that we are still very much involved in the process to discover meaningful correlations by filtering through everything that is presented to us.

If we could however improve the situation slightly by having machines present us with only *useful* correlations from a random mass of correlations, then we can give more of our attention to what is interesting. Hence, our goal is to set a path towards identifying why some variables have a semantic link between them. Before we establish that, our ongoing approach (as outlined in this research and afterwards) will be to refute or cancel out things which may be in disguise for semantic similarity.

Therefore, we set our investigation with a workflow to experiment with Linked Statistical Datasets in the 270a Cloud [2]. We have first set our hypothesis to uncover the possibility that *semantically similar* variables or datasets need to incorporate semantically rich information in order to find thought-provoking correlations. Then, the question is, what do exceptional or intriguing linkages for semantic similarity look like? We start with our null hypothesis by checking to see whether the dataset titles in World Bank indicators can help indicate strong correlations. Our results show that dataset titles by themselves or within a particular topic area is not a good indicator to predict correlations.

2 Methodology

We first state our research design and hypothesis, then discuss how we employed Linked Statistical Data (LSD) and Semantic Similarity approaches for a workflow in our LSD Sense [3] implementation.

2.1 Research design

Research problem: Why do machines have difficulty in revealing meaningful correlations or establishing non-coincidental connection between variables in statistical datasets? Put another way: How can machines uncover interesting correlations?

Over this ongoing investigation, we want to uncover some of the fundamental components for measuring and declaring semantic similarity between datasets, in order to better predict relevant strong relationships. Can semantic relatedness between datasets imply statistical correlation of the related data points in the datasets?

2.2 Hypothesis

Given our research question, we would like to propose a viable research hypothesis, followed by our investigation with the null hypothesis:

H1: If the absence of semantically rich connection between datasets is inadequate to distinguish meaningful relationships, then making relevant information about dataset connectivity will improve predicting dataset correlations by observing their semantic similarity.

Ho: There exists a significant relationship between the semantic similarity of statistical dataset titles and the correlation among those datasets, because dataset titles can indicate rich connectivity.

We set the significance level to 5% probability.

2.3 Linked Statistical Data and Semantic Similarity

The RDF Data Cube vocabulary does not only allow to express statistical data in a Web exchangeable format, but also to represent the (semantic) links within those statistical data. This ability poses some new interesting research questions around the relationship between the statistical and semantic relatedness of datasets. We are interested in the interplay of statistical correlation of LSD and their semantic similarity, in order to answer questions like: Does correlation between statistical datasets imply some kind of semantic relation? Do certain semantic links imply the existence of correlation? We propose a generic workflow for studying whether or not this relation between correlation and similarity holds for arbitrary LSD. We aim at generic correlation and similarity measures, and our workflow enables the use of any correlation and similarity indicators. For the specific goal of this paper, though, we stick to the use of Kendall's correlation coefficient and Latent Semantic Analysis (LSA) similarity.

2.4 Workflow

Based on preliminary experimentation from data acquisition to analysis, we have created the LSD Sense workflow:

1. Create hypothesis
2. Determine datasets and configurations
3. Get metadata of datasets.
4. Get each dataset's observations.
5. Create correlations and other analysis for each dataset pair combination.
6. Create dataset metadata subset for semantic similarity.
7. Create semantic similarity for each dataset pair combination.
8. Create correlation and other analysis using variables semantic similarity and correlation of LSD.
9. Test and verify hypothesis.
10. Analysis.

2.5 Implementation

We have an implementation of the LSD Sense workflow which can be used to both, reproduce our experiments, as well as run it on new input datasets. With the exception of determining which datasets to inspect, and the system configuration, LSD Sense is automated.

Semantic Correlation. The semantic similarity algorithm is based on a Latent Semantic Index (LSI) [4]. We use the dataset titles to check for their similarity.

Essentially, LSI puts each dataset title into a cluster. The number of clusters can be adjusted (default to 200). It remains as an open research question as to what it should be. Generally, research has demonstrated that optimal values depend on the size and nature of the dataset [5]. We use gensim [6] in our Semantic Correlation [7] implementation for LSD Sense.

Concerning the quality of the dataset titles, it is possible to come across datasets that differ only by one word e.g., "male", "female". This potentially lowers the accuracy to differentiate datasets. As mentioned earlier, we removed the attribute information from the dataset titles with the assumption that it reduced noise.

3 Experiment

Two experiments were conducted using the same workflow. Experiments differed only by their input data. In the first experiment, the analysis was done for a particular reference year over all available datasets. In the second experiment, however, we restricted the data further for only a particular dataset domain (topic), thereby making it possible to compare whether a control over a topic can be significant for semantic similarity of the dataset titles.

3.1 Data

We decided to conduct our experiment on a simple dataset structure, containing two dimensions; *reference area*, and *reference period*, and one measure *value* for its observations, where the World Bank indicators was a good candidate from the 270a Cloud. The rationale for using only one dataspace (at this time) was to remain within a consistent classification space to measure semantic similarity. We fixed the reference period to 2012, and datasets that are part of World Bank's education topic. We have identified one downside concerning the data quality i.e., the attribute/unit information was incorporated as part of the dataset title, usually as a suffix within brackets. We dealt with this by removing the attribute information from the titles as part of preprocessing in the semantic similarity phase.

3.2 World Bank Indicators workflow

The workflow of our experiment is summarized as follows:

Correlations for each dataset pair. We retrieved the 2012 World Bank Indicators datasets, 3267 in total, via SPARQL queries from the World Bank Linked Dataspace [8]. The correlations were computed using R, the statistical software, by joining each dataset pair by their reference area (one of the dimensions of the dataset structure), and using their measure values for the correlation coefficient. Based on preliminary inspection for normality distribution on sample datasets, we noted that observations did not come from a bivariate normal distribution. Hence, we computed Kendall's rank correlation coefficient in our analysis. Initially we computed and stored the correlations for dataset pairs with a sample size, $n>10$, resulting in 2126912 correlation values.

The information on the analysis we generated consisted of the following headers: `datasetX`, `datasetY`, `correlation`, `pValue`, `n`, where `datasetX` and `datasetY` are the identifiers for each dataset pair that is being compared. We later filtered sample values, n<50, for our threshold for significance. The population size i.e., the number of potential reference areas that can have an observation, is 260. That is the number of reference area codes in the World Bank classification, however, it is not known as to which reference areas may occur in a given dataset before hand. We retained majority of the computations in any case, giving us the possibility to do better pruning in the future, in light of more information.

Semantic similarity for each dataset pair. Before doing the semantic similarity, we first took an unique list of the dataset identifiers from `datasetX` and `datasetY` so that what is to be checked for their similarity is only in relation to those datasets, as opposed to the complete set of datasets which we originally retrieved. At this point, we have 2200 unique datasets. The similarity was measured based on dataset titles. They are in short sentences e.g., "Mortality rate, infant (per 1,000 live births)". After minor preprocessing e.g., removal of the text pertaining the unit within brackets, it was left with "Mortality rate, infant". The semantic similarity algorithm is based on LSA. Essentially LSA puts each dataset title into a cluster (default number is 200). The resulting headers of the output was: `datasetX`, `datasetY`, `similarity`.

Correlation analysis with variables semantic similarity and correlation of dataset. We then took the absolute values for both variables; `|similarity|`, `|correlation|` (caring only for the strength of the relationships as opposed to their directionality). We then filtered both similarity and correlation values <0.05 and >0.95, as well as correlation values with p-value>0.05, for reasons to exclude potential outliers, or misleading perfect relations, as well as to exclude insignificant correlations. The final correlation and scatter plot was generated by joining the similarity and correlation tables on datasetX and datasetY columns. Finally the correlation of the final data table was conducted using the Kendall method as the data had a non-normal distribution and we were not interested in modeling (line fitting).

The second experiment followed the same procedure for the analysis, but considering only the datasets associated with the topic education for the same reference period.

4 Results

All of the experiment results are available at the LSD Sense GitHub repository, and can be reproduced. Table [Experiment results] provides our findings, with Figures [1] and [2]:

Experiment Results

	All topics	**One topic (*education*)**
Correlation	0.182	0.227
***p*-value**	$< 2.2e\text{-}16$	$< 2.2e\text{-}16$
n	92819	33184

Datasets are from 2012 World Bank indicators. n is the number of dataset pairs with semantic similarity and correlation as variables.

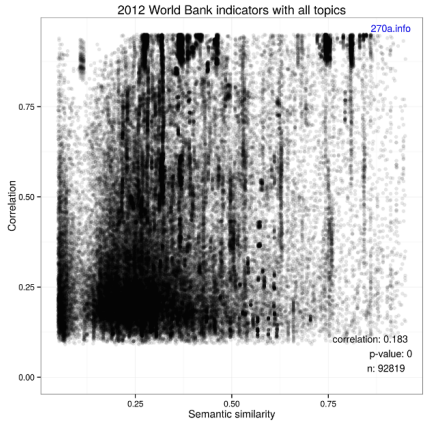

Figure 1: 2012 World Bank indicators with all topics

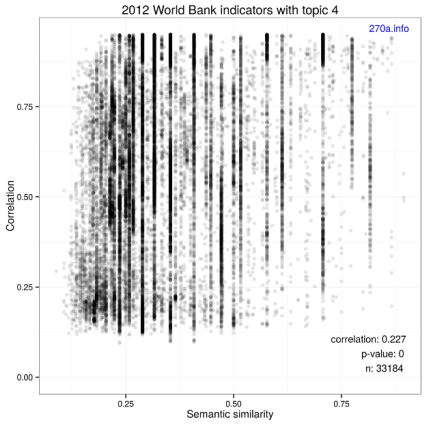

Figure 2: 2012 World Bank indicators with topic education.

Given that both experiments resulted in p-values that are statistically significant and that the strength of the correlation values are weak, we reject our null hypothesis. For extra measure, we can also verify the meaninglessness by looking at the plots. There is **nothing** interesting **to see here**. We will **move along** with our alternative hypothesis.

5 Related Work

Linked Statistical Data Analysis [9], explores a way to reuse statistical linked dataspaces, federated queries, and generation of statistical analyses e.g., regression, for humans and machines. The stats.270a.info [10] service stores computed analysis, and makes it possible for future discovery.

Towards Next Generation Health Data Exploration: A Data Cube-based Investigation into Population Statistics for Tobacco [11], presents the qb.js [12] tool to explore data that is expressed as RDF Data Cubes. It is designed to formulate and explore hypotheses. Under the hood, it makes a SPARQL query to an endpoint which contains the data that it analyzes.

Generating Possible Interpretations for Statistics from Linked Open Data [13] talks about the Explain-a-LOD tool which focuses on generating hypotheses that explain statistics. It has a configuration to compare two variables, and then

provides possible interpretations of the correlation analysis for users to review.

Using Linked Data to Evaluate the Impact of Research and Development in Europe: A Structural Equation Model [14], presents the feasibility of combining different LOD sources to assess the impact of one variable over others.

Spurious Correlations [15] reveals correlations that are not genuine for practical use. In other words, the correlations are type I errors. It emphasizes on the importance for humans to be critical of random correlations, and to investigate whether there is a direct relation between the variables.

Ontology Matching [16] is perhaps the most mature field in the Semantic Web dealing with the general problem of finding semantically related entities of ontologies and Linked Data, although resources like WordNet and DBpedia are also related.

These studies and engineering efforts have created, inspected, and hypothesized possible correlations. However, the missing gap in research is that there is no integrated study on how semantic relatedness between datasets may enhance the detection of meaningful or useful correlations in statistical data. Our contribution is the investigation of highly probably elements which would lead to better prediction of interesting correlations by employing linked statistical datasets and semantic analysis.

6 Conclusions and Future Work

We believe that the presented work here and the prior Linked Statistical Data Analysis effort contributes towards strengthening the relationship between Semantic Web and statistical research. What we have set out to investigate was to minimize human involvement for discovering useful correlations in statistical data. We have implemented a workflow in which we can automate the analysis process, from data retrieval to outputting analysis results for candidate semantic linkages in Linked Statistical Data.

We have evaluated our results by testing and verifying the null hypothesis which we have put forward. While it turned out that the semantic similarity between datasets titles were not useful to determine strong and meaningful correlations — which is a useful finding, in any case — it left us with the remaining alternative hypothesis that can be used in future research.

Possibly fruitful future work might want to run a similar experiment with the semantic similarity of dataset descriptions, test manually configured useful relations for a controlled set of datasets, or looking into interlinked topic domains across linked dataspaces.

Where is *interestingness* hidden?

7 Acknowledgements

This work was supported by a STSM Grant from the COST Action TD1210. Many thanks to colleagues whom helped one way or another during the course of this work (not implying any endorsement); in no particular order: Amber van den Bos (Dakiroa), Michael Mosimann (BFS), Anton Heijs (Treparel b.v.), Frank van Harmelen (VU Amsterdam).

References

1. Rosling, H., Marmot, M.: The Joy Of Stats: Meaningless and meaningful correlations, http://www.open.edu/openlearn/science-maths-technology /mathematics-and-statistics/statistics/the-joy-stats-meaningless-and-meaningful-correlations

2. 270a.info, http://270a.info/

3. LSD Sense code at GitHub, https://github.com/csarven/lsd-sense

4. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, 41(6), pp.391–407 (1990), http://www.cs.bham.ac.uk /~pxt/IDA/lsa_ind.pdf

5. Bradford, R.: An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications, Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp.153–162 (2008), http://dl.acm.org/citation.cfm?id=1458105

6. gensim: Topic modeling for humans, http://radimrehurek.com/gensim /index.html

7. SemanticCorrelation code at GitHub, https://github.com/albertmeronyo /SemanticCorrelation

8. World Bank Linked Dataspace, http://worldbank.270a.info/

9. Capadisli, S., Auer, S. Riedl, R.: Linked Statistical Data Analysis, ISWC SemStats (2013), http://csarven.ca/linked-statistical-data-analysis

10. stats.270a.info, http://stats.270a.info/

11. McCusker, J. P., McGuinness, D. L., Lee, J., Thomas, C., Courtney, P., Tatalovich, Z., Contractor, N., Morgan, G., Shaikh, A.: Towards Next Generation Health Data Exploration: A Data Cube-based Investigation into Population Statistics for Tobacco, Hawaii International Conference on System Sciences (2012), http://www.hicss.hawaii.edu/hicss_46/bp46/hc6.pdf

12. qb.js, http://orion.tw.rpi.edu/~jimmccusker/qb.js/

13. Paulheim, H.: Generating Possible Interpretations for Statistics from Linked Open Data, ESWC (2012), http://www.ke.tu-darmstadt.de/bibtex/attachments /single/310

14. Zaveri, A., Vissoci, J. R. N., Daraio, C., Pietrobon, R.: Using Linked Data to Evaluate the Impact of Research and Development in Europe: A Structural Equation Model, pp.244–259, ISWC (2013), http://svn.aksw.org/papers /2013/LODSEM/ISWC2013_AZ_LODSEM_public.pdf

15. Vigen, T.: Spurious Correlations, http://tylervigen.com/

16. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering (2013), http://disi.unitn.it/~p2p/RelatedWork/Matching/SurveyOMtkde_SE.pdf