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. As more linked statistical datasets become available, a fundamental question on statistical data comparability arises: To what extent can arbitrary statistical datasets be faithfully compared? Our research focuses in studying whether statistical and semantic relationships influence each other, by comparing the correlation of statistical data with their semantic similarity. The ongoing research problem is to investigate how machines can reveal meaningful correlations or establishing non-coincidental connection between variables in statistical datasets. 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 • Semantic Similarity

1 Introduction

While computers can assist us in discovering strong correlations in large amounts of statistical datasets, whether by chance or through sophisticated methods, humans still need to be critical about the results and interpret them appropriately. This implies that we are still very much involved in the process in discovering meaningful correlations by filtering through everything that is presented to us.

If machines can 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].

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.

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

H_1: Semantic similarity is a good predictor of meaningful correlations.

H_0: 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.

2.1 Linked Statistical Data and Semantic Similarity

We are interested in the interplay of statistical correlation of LSD and their semantic similarity. Do certain semantic linkage imply the existence of correlation? 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.

2.2 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.

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. 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.

3 Experiment

Two experiments were conducted using the same workflow, differing 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 (World Bank's education topic), thereby making it possible to compare whether a control over a topic can be significant for semantic similarity of the dataset titles.

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 was to remain within a consistent classification space to measure semantic similarity.

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*, by joining each dataset pair by their reference area, and using their measure values for the correlation coefficient (Kendall). We computed and stored the correlations for dataset pairs with a sample size, n>10, resulting in 2126912 correlation values.

Semantic similarity for each dataset pair: We first took an unique list of the dataset identifiers (2200) so that their similarity is only in relation to those datasets, as opposed to the complete set of datasets which we originally retrieved. The similarity was measured based on dataset titles. They are in short sentences e.g., "Mortality rate, infant (per 1,000 live births)". The semantic similarity algorithm is based on LSA where each title is placed in a cluster.

Correlation analysis with variables semantic similarity and correlation of dataset: We then took the absolute values for both variables; `|similarity|`, `|correlation|`. The final correlation coefficient (Kendall) and scatter plot was generated by joining the similarity and correlation tables.

4 Results

Experiment results are at the LSD Sense repository, and can be reproduced. Table [Experiment results], Figures [1] and [2] illustrates our findings:

Experiment Results

	All topics	One topic (*education*)
Correlation	0.182	0.227
p-value	< 2.2e-16	< 2.2e-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.

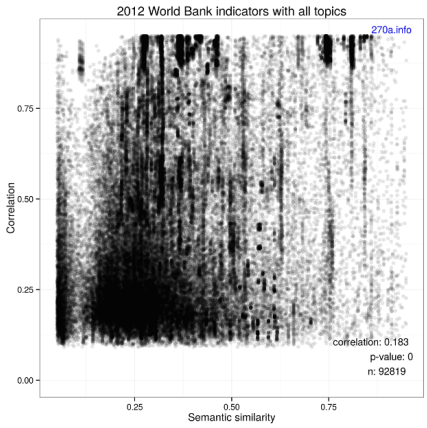

Figure 1: 2012 World Bank indicators with all topics

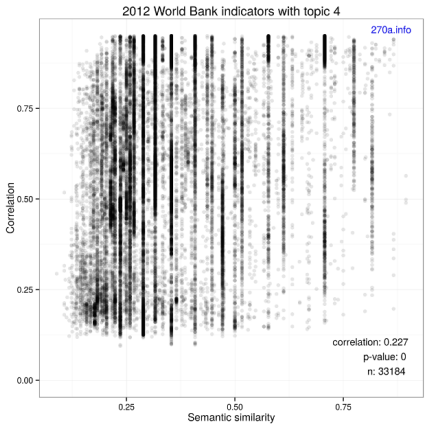

Figure 2: 2012 World Bank indicators with topic education.

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 Conclusions and Future Work

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.

6 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. 270a.info, http://270a.info/
2. LSD Sense code at GitHub, https://github.com/csarven/lsd-sense
3. 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
4. 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
5. gensim: Topic modeling for humans, http://radimrehurek.com/gensim/index.html
6. SemanticCorrelation code at GitHub, https://github.com/albertmeronyo/SemanticCorrelation
7. World Bank Linked Dataspace, http://worldbank.270a.info/