TermPicker: Recommending Vocabulary Terms for Reuse When Modeling Linked Open Data

Johann Schaible
GESIS - Leibniz Institute for the Social Sciences
Germany
johann.schaible@gesis.org

ABSTRACT
Linked Open Data (LOD) refers to data published on the Web in a way that it is machine-readable, its meaning is explicitly defined, and it is linked to other data sets. So-called Resource Description Framework (RDF) vocabularies are employed for LOD modeling. An RDF vocabulary is a collection of unique vocabulary terms comprising classes, which describe the type of a data entity, and properties, which describe relationships between data entities. When modeling data as LOD, it is considered best practice to reuse such vocabulary terms from already existing RDF vocabularies before inventing new ones. This reduces the heterogeneity in data representation and makes it easier for others to understand and consume the data. Unfortunately, this simple guidance is too unspecific, as deciding which vocabulary terms to reuse is far from trivial. First, even when data engineers are focused on a specific domain, there are more and more vocabularies to choose from. Second, various vocabularies contain multiple terms that seem to represent similar semantics. Data engineers are thereby uncertain about which term is the more appropriate option. Thus, without additional guidance, it is very difficult to reuse appropriate vocabulary terms for representing data as LOD. Using inappropriate vocabulary terms decreases the data quality. As a consequence, others will less likely use the data.

This work proposes and evaluates TermPicker: a novel approach alleviating this situation by recommending vocabulary terms based on the information how other data providers modeled their data as LOD. TermPicker gathers such information and represents it via so-called schema-level patterns (SLPs), which are used to calculate a ranked list of RDF vocabulary term recommendations. The ranking of the recommendations is based either on the machine learning approach “Learning To Rank” (L2R) or on the data mining approach “Association Rule” mining (AR). TermPicker is evaluated in a two-fold way. First, an automated cross-validation evaluates the ranking of TermPicker’s recommendations by comparing them to term recommendations based on the most typical strategies (currently used by LOD engineers) to reuse vocabularies. These most typical reuse strategies are elaborated with the help of an online survey, in which LOD experts are asked to rank given LOD models from “best” to “worst” regarding the utilized vocabulary terms. For the cross-validation, the Mean Average Precision (MAP) as well as the Mean Reciprocal Rank at the first five positions (MRR@5) are used as evaluation metrics, since these measures assess the quality of an ordered list. Second, a user study examines which of the recommendation methods (L2R vs. AR) aids real users more to reuse RDF vocabulary terms in a practical setting. To this end, both approaches are integrated into the UI-based data modeling tool Karma. The participants, i.e., TermPicker’s potential users, are asked to reuse vocabulary terms while modeling three data sets as LOD, but they receive either L2R-based recommendations, AR-based recommendation, or no recommendations.

The results of the cross-validation show that using SLPs, TermPicker achieves 35% higher MAP and MRR@5 values compared to using solely the features based on the typical reuse strategies. Both the L2R-based and the AR-based recommendation methods were able to calculate lists of recommendations with MAP \( \approx 0.75 \) and MRR@5 \( \approx 0.80 \). However, the results of the user study show that the majority of the participants favor the AR-based recommendations. When using Association Rule mining, the participants accept significantly more recommendations, they need less time to complete the modeling tasks, and the quality of the LOD representations as well as the participants’ satisfaction is significantly higher compared to using recommendations based on Learning To Rank.

The outcome of this work illustrates a novel and validated method to find vocabulary terms for reuse. TermPicker alleviates the situation of searching for classes and properties used by other data providers on the LOD cloud for representing similar data. Besides Karma, the recommendation approach can also be integrated in further LOD modeling tools, such as Neologism, allowing for next generation vocabulary reuse when modeling Linked Open Data.