Project Outline: ...

Student Project Semantic Web Technologies HWS18

Presented by

Nele Ecker
Alex Lütke
Marvin Messenzehl

Submitted to the
Data and Web Science Group
Prof. Dr. Heiko Paulheim & Sven Hertling
University of Mannheim

October - December 2018

Contents

1 Problem Statement

Everyone who does regular city trips knows the problem. You who visit a foreign city and just dont have the same experience as people who live there. A lot of information sources, like tourist centers, give you a very biased opinion on what you should see and what not. But what if somebody wants to look for something special or has special interests? If somebody really wants to get to know a city and the different districts, this takes a lot of time and preparation.

This problem should be solved within this project through an application that uses semantic web technologies.

2 Application Goal

The overall goal of the application is to give the user a visual help to easily identify the particularities of the different city districts. This should be done with the help of a map as the main user interface where the different districts are marked in different colors in the form of rectangles.
Examples of this would be categories like *museums, science, students* and a lot more. The detailed categories depend on the districts and offer of the respective city.

In the end, the user can visit the web application where he/she sees a map of a city where different special districts are marked. Therefore trips can be planned more individual and efficient.

3 Techniques

From a user's perspective the application goal can be realized in two different ways. (1) Regional districts are getting clustered by their type of public institutions or (2) the user searches for a generic type of institutions and an algorithm searches on-the-fly for regions with a high density of those. Technically, both scenarios rely on similar computations. The main idea is to divide a given region into static squares and calculate for each square the density of public institution types. In the end, a square is labeled with the most occuring institution-type. In more detail, two questions must be clarified beforehand: (1) What is the best size of a square? (2) Due to the non-unique naming assumption, similar institution types might be called similarly. For question (1) experimental setups will clarify the best square size. Mannheim will be used as a reference for evaluating the adequacy of

the classification. For question (2), reasoning techniques will help finding enough overlap. The assumption is that public institutions are categorized in the datasets. Those categories follow some kind of an ontology. For instance the University of Mannheim is of type ËducationalInstitution, Üniversityänd Örganisationïn the DBPedia-ontology. For the labeling, each type in the entire ontology must be considered and wherever the largest overlap within a square is found, is then outputted. However, very general types will be very frequent. So the algorithm has to penalize general types like organization and prefer to label squares with deeper nested labels like Üniversity. This preference is a threshold that must be experimentally optimized as well. In a later stage of the project, one could replace the predefined squares with arbitrarily shaped clusters using a DBScan algorithm. However, that is just a nice-to-have addition. In the end the outcome will always be small regions with labels depending on the type of institutions. With that information foreign travellers can easily find their way through cities and districts.

4 Datasets

The focus is set on two datasets, namely DBPedia and LinkedGeoData. Both offer information about a specific area, though they have a different level of detail. DBPedia is used to have a first overview about the most important points of interests within a specific city. These points of interests have information about longitude and latitude, what makes it easy to include them into the map on the website. LinkedGeoData obtained all its data from the OpenStreetMap, so it also has information about shops, restaurants, etc. This dataset also provides a useful way to distinguish the kind of place that is shown on the map, as each point has an addition attribute type, which specifically defines what the user can expect to find there. For example the University of Mannheim is of type *amenity:university*, while a nearby bookshop is of type *shop:books* and the Kunsthalle Mannheim is of type *tourism:museum*. Next to this attribute also information about longitude and latitude is added, so each point can be included on the map.The information about longitude and latitude is also crucial to divide the map into different areas and color them accordingly.

As described in the section 3 one possible realisation is that an area is clustered by the type of the institution. In this case the given type, or a part of the type - as it consists of a more general and more detailed part - can be used as label for the dataset.

5 Evaluation

The evaluation for this use case is quite subjective. That is why there is no atomic performance measurement available. The assessment of this project will be done in two ways. First of all, manual spot checks for the region of Mannheim and Heidelberg will be done. The project team is quite familiar to the characteristics of districts in Mannheim and Heidelberg. So some plausibility checks can be done that way. In order to objective the evaluation a bit further, a fitness measure based on the deepness of the cluster-labels in the DBPedia and LinkedOpenData ontology will be introduced. This measure is meant to penalize very general labels like institution. This measure can be further combined with a penalization for very heterogeneous or homogeneous regions, where each square is labeled differently or similar. An example fitness function can look like this:

$$d_i = deepness\ of\ current\ cluster\ label$$

$$d = deepest\ label\ in\ ontology$$

$$o_i = current\ cluster\ size$$

$$o = optimal\ cluster\ size$$

$$I = set\ of\ squares$$

$$\sum_{i=0}^{|I|} 0.5 * \frac{d_i}{d} + 0.5 * \frac{1}{o - o_i}$$