A Clustering Based Approach for Type Discovery in RDF Data Sources

Kenza Kellou-Menouer*, Zoubida Kedad*

*PRISM - University of Versailles Saint-Quentin-en-Yvelines,
45 avenue des Etats-Unis , Versailles, France
kenza.menouer@prism.uvsq.fr, zoubida.kedad@prism.uvsq.fr

Querying and exploiting RDF(S)/OWL data sources requires information about the resources and properties they contain. Without a description of the data set, it is difficult to target the relevant properties and resources, and browsing data sets in order to understand their content can be a tedious process. One important feature of RDF(S)/OWL data sources is that they are not organized according to any predefined schema. They are structureless by nature and the languages used to describe data on the Web do not impose any constraints or restrictions on the properties describing resources.

Our goal is to discover missing type definitions in a RDF(S)/OWL data set. We propose a clustering based approach where entities are grouped according to their similarity. The similarity between two given entities is evaluated considering their respective sets of properties using Jaccard similarity.

Our requirements for type discovery are the followings: firstly, the number of types is not known in advance, and secondly, the data sets are evolving, large and may contain noise. The most suitable grouping approach is density-based clustering, introduced by Ester et al. (1996), because it is robust to noise, deterministic and it finds classes of arbitrary shape, which is useful for data sets where resources are described with heterogeneous property sets. In addition, unlike the algorithms based on k-means and k-medoid, the number of classes is not required.

In order to speed up the clustering process, and especially to perform successive executions with different parameters values (the maximum radius of neighborhood ε and the minimum number of neighbors for an entity $MinPts$), we perform once and for all the calculation of the nearest neighbors of each entity. To this end, we index the data and we order the entities according to their similarity. We store a neighborhood matrix containing for each entity the ordered list of its neighbors, as well as the distance between this entity list and the number of the line representing the index of an entity. Thanks to the indexing of the data and the ordering of the neighbors according to the distance, it is not necessary to go through the entire matrix to find the neighbors of a given entity, and the complexity of neighbors search becomes linear o(n).

We have performed some experiments on existing data sets to assess the quality of the inferred types. We have extracted the existing type definitions from our data sets and considered them as a gold standard. Then we have run our algorithm on the data sets without the type definitions and evaluated for each of the inferred classes precision and recall. We have annotated each inferred class with the most frequent type definition of its entities. For each type label T_r

of size n_r in the data set and each class C_i of size n_i inferred by our approach, such that T_r is the label of C_i, the quality metrics are evaluated for each C_i with respect to T_r as follows: n_{ri} being the number of instances in the class C_i that belong to T_r, the precision $P(T_r, C_i)$ is defined as n_{ri}/n_i, the recall $R(T_r, C_i)$ is defined as n_{ri}/n_r. In addition to achieve good precision and recall even when the sets of properties describing entities are very heterogeneous, the approach enabled to infer type definitions which were not specified in the data set.

Some works in the literature have addressed the problem of inferring type of a semi-structured data set. Wang et al. (2000) propose an approximate DataGuide based on COBWEB, but it is not deterministic, expensive and not very suitable for large data sets. The proposed algorithm in Nestorov et al. (1998), uses bottom-up grouping. However, the method requires a threshold of similarity, and the number of classes. Christodoulou et al. (2013) use ascending hierarchical clustering to deduce structural summaries of linked data. Unlike our approach, only outgoing properties are considered and the hierarchical clustering is very expensive. SD-Type (Paulheim and Bizer (2013)) enriches an entity by several types using RDFS inference rules, and computes the confidence of a type for an entity. The contribution is more to evaluate the relevance of inferred types for an entity rather than to find its type, as RDFS inference rules are used for this.

In future works, we will address the generation of fuzzy classes, to allow multiple types for an entity. We will also tackle the annotation of the extracted classes, as well as the discovery of possible links between them in order to produce a complete schema description.

References

Christodoulou, K., N. W. Paton, and A. A. Fernandes (2013). Structure inference for linked data sources using clustering. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, pp. 60–67. ACM.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, Volume 96, pp. 226–231.

Nestorov, S., S. Abiteboul, and R. Motwani (1998). Extracting schema from semistructured data. In *ACM SIGMOD Record*, Volume 27, pp. 295–306. ACM.

Paulheim, H. and C. Bizer (2013). Type inference on noisy rdf data. In *The Semantic Web–ISWC 2013*, pp. 510–525. Springer.

Wang, Q. Y., J. X. Yu, and K.-F. Wong (2000). Approximate graph schema extraction for semi-structured data. In *Advances in Database Technology-EDBT 2000*, pp. 302–316. Springer.

Résumé

RDF(S)/OWL data sources are not organized according to a predefined schema, as they are structureless by nature. This lack of schema limits their use to express queries or to understand their content. Our work is a contribution towards the inference of the structure of RDF(S)/OWL data sources. We present an approach relying on density-based clustering to discover the types describing the entities of possibly incomplete and noisy data sets.