Using Semantic Domain-Specific Dataset Profiles for Data Analytics

Simon Gottschalk

L3S Research Center
Leibniz University Hannover, Germany
gottschalk@L3S.de

Abstract. The availability of a vast amount of heterogeneous datasets provides means to conduct data analytics in a wide range of applications. However, operations on these datasets demand not only data science expertise, but also knowledge about the structure and semantics behind the data. Semantic data profiles can enable non-expert users to interact with heterogeneous data sources without the need for such expertise.
To support efficient semantic data analytics, a domain-specific data catalog, that describes datasets utilizable in a given application domain, can be used [1]. Precisely, such a data catalog consists of dataset profiles, where each dataset profile semantically describes the characteristics of a dataset. Dataset profile features not only include a set of well-established features (e.g. statistical and provenance features), but also connections to a given semantic domain model. Such a domain model describes concepts and relations in a specific domain and thus helps to automate data processing in a semantic meaningful manner. An example is the mobility domain and the integration of different spatial representations.
Once created, a domain-specific data catalog can support a whole data analytics workflow. This includes, but is not limited to search through the use of semantic concepts (e.g. datasets about street segments), domain-specific feature extraction (e.g. geo-transformations), and machine learning with the help of concept-based type checking. These examples demonstrate that the provision of semantic domain-specific profiles is a valuable step towards data analytics when dealing with heterogenous datasets.

Acknowledgements

This work was partially funded by the Federal Ministry of Education and Research (BMBF), Germany under Simple-ML (01IS18054).

References

1. S. Gottschalk, et al. "Simple-ML: Towards a Framework for Semantic Data Analytics Workflows." SEMANTiCS (2019).