Knowledge Graphs for Impactful Data Science

Victor de Boer[1]

[1]Vrije Universiteit Amsterdam, the Netherlands

Abstract

In this invited talk I will argue that to build scalable, transparent and explainable AI in various domains where heterogeneous data is available, we need to collaborate with domain experts to develop relevant and high-quality knowledge graphs as well as appropriate data science and Machine Learning methods to constantly enrich and analyse these graphs. I give examples in the Digital Humanities and Internet of Things.

In many modern statistical approaches to AI, raw data is the preferred input for (Machine Learning) models. In some areas and in some cases, however, we struggle to find this raw form of data. One such area involves heterogeneous knowledge: entities, their attributes and internal relations. The Semantic Web community has invested decades of work on just this problem: how to use graphs to represent knowledge, in various domains, in as raw and as usable a form as possible, satisfying many use cases. To build scalable, transparent and explainable AI in various domains where such heterogeneous data and knowledge is available, we need to collaborate with domain experts to develop a) relevant and high-quality knowledge graphs as well as b) appropriate data science and ML methods to constantly enrich and analyse these graphs[1].

In the domain of Digital Humanities (DH), a large amount of heterogeneity of data and knowledge exists. Digitized datasets can derive from centuries-old sources and multiple views on history and heritage should be represented. The capacity of the Knowledge Graph to capture such heterogeneity makes this an ideal model to represent, share and combine data sources to allow for new types of analyses. In the domain, Machine Learning and other Data Science methods are more and more looked at to identify patterns in the data, establish new links or categorize entities. Transparency and explainability are key requirements for such methods to be used in serious scholarly analysis.

Although the domain of Internet of Things (IoT) and Smart Homes differs in many ways from that of DH, here too do we find datasets of varying sources, combined to allow for new types of applications and analysis[2]. In smart home scenarios, methods that combine data into knowledge graphs for further analysis or applications will need to be privacy-aware, transparent and explainable. Using ontologies such as SAREF[3], we can achieve this interoperability. Using re-usable (python) notebooks we can establish a Data Science pipeline.

SEMANTICS 2022 EU: 18th International Conference on Semantic Systems, September 13-15, 2022, Vienna, Austria

v.de.boer@vu.nl (V. d. Boer)

http://victordeboer.com/ (V. d. Boer)

0000-0001-9079-039X (V. d. Boer)

References

[1] X. Wilcke, P. Bloem, V. De Boer, The knowledge graph as the default data model for learning on heterogeneous knowledge, Data Science 1 (2017) 39–57.

[2] R. van der Weerdt, V. de Boer, L. Daniele, B. Nouwt, Validating saref in a smart home environment, in: Research Conference on Metadata and Semantics Research, Springer, 2020, pp. 35–46.

[3] L. Daniele, F. d. Hartog, J. Roes, Created in close interaction with the industry: the smart appliances reference (saref) ontology, in: International Workshop Formal Ontologies Meet Industries, Springer, 2015, pp. 100–112.