Review of Habilitation Thesis

# Interpretable Data Analysis with Entity-based Text Representations and Rule-based Models

by

**Tomáš Kliegr**

The thesis primarily collects author's journal and conference papers dealing with two distinct topics of interpretable data analytics – entity-based semantic extension of textual resources and association rules extracted from structured data. The text is supplied with an introductory part which motivates the work, summarizes the current state of the art in the two mentioned areas and stresses author's contributions to them. Tomáš Kliegr clearly defines research challenges tackled by his research and describes developed methods addressing the challenges. The quality of papers, the reprints of which were selected for inclusion, demonstrates the high level of knowledge and experience Tomáš has gained in recent years.

New entity-based text representation methods based on hypernym identification as well as domain-knowledge pruning techniques applied in the association rule learning form the strongest part of the presented research work. The author shows that the proposed approach for extracting types of entities represented by the first hypernym in a Wikipedia-like defining sentence brings a significant value in the quality of extraction with numerous applications. Similarly, author's contribution to the CBA (Classification by Associations) approach and rule pruning integrating user feedback is substantial. Another strong aspect of the presented research lies in the wide range of areas the author applied his results in. For example, the CBA algorithm was adopted for a submission to the CLEF NewsREEL 2017 Challenge by a team from the Czech Technical University. Tomáš Kliegr has also actively cooperated with key persons behind the knowledge processing part of the DBpedia initiative and participated in relevant international knowledge-processing competitions and challenges. I also acknowledge the attention the author pays to integration and evaluation of the research results in applications, particularly in the domain of entity classification. It links the work to real-world problems and proves that the results can be directly employed in current knowledge engineering practice. Tomáš also contributed to various benchmarking resources, for example, through the collection of datasets for evaluating word similarity computation algorithms according to paradigmatic association. Last but not least, I appreciate the clarity in delimiting author's contribution to each particular research direction and publication specified in the thesis.

To point out also less clear parts and, potentially, to foster discussion during the thesis defence, it can be stated that some aspects of the presented work dealing with interpretability of machine learning methods do not fully reflect up-to-date trends in the field. Indeed, knowledge mining is a very active area and new methods are published every month. The brief survey of the state of the art in the entity-based text representation does not pay sufficient attention to advanced learning methods (for example, the recent DeepType system based on deep learning) and interpretability of quantitative models (typically, neural networks) by means of various attention mechanisms and visualisa-

tion tools. Similarly, the thesis does not deal with contextual word relatedness models that have been recently popularized by systems such as BERT.

Despite the mentioned minor shortcomings, I can ascertain that the reviewed habilitation thesis proves author's excellent research track in the knowledge engineering field and clearly demonstrates his potential for future scientific work in the area. I propose to accept it for the habilitation at the University of Economics, Prague.

Brno, January 2, 2019 Pavel Smrž