CHARLES UNIVERSITY PRAGUE
faculty of mathematics and physics

Spectabilis,

Jakub Fischer

Dean,

Faculty of Informatics and Statistics

University of Economics, Prague

Report on habilitation thesis

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

By Tomas Kliegr

Main motivations of this thesis and important real world challenges are twofold: first, use entity based text representation to improve text processing and second, for data analysis use rule-based models to increase interpretability (intuitiveness, human understandability, self explainability) of models. As I understand, results should help an untrained user to find relevant information (in ideal case without additional human assistance).

Considering improving text mining by using entity-based representation – main contribution is the way, how these entities are obtained. Previously, these were mainly from info boxes and article categories (as human edited semi structured input in Wikipedia, often missing and available only in dumps available with considerable delay). Author extends this and gets "type of" relation extracted from free text of articles from live Wikipedia (usually from the first sentence of article).This approach has several advantages and enlarged several language editions of DBPedia substantially. Main result is a journal paper J1, Appendix A, [45].

Second main topic is rule mining. There is an apparent contradiction in comparison with "black-box-models". In 18^{7-10} author states "The advantages

Prof. RNDr. Peter Vojtáš, DrSc.
Mobil: +420 739 822 406
vojtas@ksi.mff.cuni.cz
tp://www.ksi.mff.cuni.cz/~vojtas/

partment of Software Engineering
ılostranské nám. 25, 118 00 Praha 1
Phone: +420 22191 4239
fax: +420 22191 4323

of rule-based classifiers as opposed to other commonly used classifier representations, such as random forests or Artificial Neural Networks (ANNs), include typically faster learning times (particularly as opposed to neural networks), and better interpretability". Whereas 19^{5-10} we can read "While rules in general are well suited to provide such meaningful explanation to the end user, one of the limiting factors for the use of association rules is a high number of frequent item sets that can often be discovered even for very small datasets due to combinatorial explosion [15]. The high number of discovered item sets, and consequently rules, does not only have computational costs, but also severely impedes interpretability of the model. While a single association rule is typically easily interpretable, how does one interpret one million rules?".

Author builds upon previous works of colleagues from department, adds new ideas which are tested as "proof of concept". By my opinion, these results are less convincing (when compared to entity part). I think this is mainly due to subjectivity of human understanding, which is vague, can depend on context and can change in time. Author restrict to reducing the number of rules. I did not find any clue for improving complexity.

Both these topics are intertwined in some publications and do not stand only separately (although some papers are "single topic").

My question to discussion concerns authors opinion on how can interpretability be verified on real world users, and/or rephrased "why should a user trust you?".

Thesis is a commented collection of five journal papers and two conference / workshop papers (where comments go far beyond these seven papers – about 30 bibliographical sources co-authored by the adept are listed). Readability could be improved by a back and forth indexing (author uses three types of identifiers, one from tables 4.1 and 4.2, second as appendix from A to G and finally [1234] notation in bibliography).

To conclude. I like this thesis, it evoked a lot of questions and my comments are rather expression of my interest. I can repeat: the results of this thesis are important, published in good journals and deal with hot topic hard problems well aware of up to date state of research. More over author showed ability to motivate and lead younger colleagues – I am sure he will be a good supervisor and teacher.

I approve the work as a habilitation thesis.

Prague, January 23rd 2019

Peter Vojtáš