FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Master thesis review

Thesis author	Felipe Vianna
Thesis Title	Expert Classification and Retrieval
Submission year	2020
Study program	Computer Science
Study branch	Artificial Intelligence

Reviewer	Mgr. Marta Vomlelová, Ph.D.	**Role**	oponent
Affiliation	Department of Theoretical Computer Science and Mathematical Logic Faculty of Mathematics and Physics Charles University		

Review:

The thesis presents a data mining analysis. It aggregates multiple data sources: *Internal publications*, *Previous work assignments*, *'Topic I know' in profiles*, and the *Taxonomy.* The goal of the thesis is to create an expert profiling model. Topics from the taxonomy are combined with the keywords and n-grams from other sources. Term similarities are identified by their word embedding comparison. These topics relate both to documents and persons (experts). Documents are related to topics by BM25 metrics. In the training data, the experts a related to topics by terms in their publications and hours spend on their assignments. In the target data, the topics listed as 'I know' by the expert are taken. Several models are learned and evaluated.

Used algorithms are described in the thesis, explanatory figures are provided. The preprocessing setting is analyzed on a set of graphs. Different model settings with and without Flair encoding and a time decay are evaluated.

The organization of the thesis can be improved. The information is often spread through several sections and not always consistent. One issue is the evaluation function. The average precision AP and AP_5 is defined in Section 1.2.1. $F_{0.5}$ measure and the tradeoff precision and recall is described in Section 4.9. Furthermore, the use of bagging (Section 4.9) is not clear: is it used or is the train-test split $S_1,\ldots, S_4$ fixed in advance?

Neural network models are compared with a baseline. Unfortunately, the Baseline model does not use all information available for the neural networks. The 'Previous work assignments' should be incorporated also in the Baseline model.

The topics from expert profiling are based on taxonomy and combined with terms from other sources. I would appreciate a summary overview over the topics: How many come from the taxonomy, from *Topic I know* keywords and from *Internal publications* dataset? Do you topics with high similarity as a list of 'synonyms'? Specifically, what is the maximum of the function `baseline_eval.precision_at_5`? If a recommendation covers several topics it may be more than 1.(?) The exclusion of topics with less than 20 experts p.73 is questionable. The key point in expert finding is to find an expert for a rare topic. Together, the comparison with TREC Enterprise is biased.

The author has shown its ability to combine available data sources, to apply language preprocessing, and to build a suitable model for expert profiling.

I recommend to accept the work as a master thesis.

Prague, January 24^{th}, 2020

Mgr. Marta Vomlelová, Ph.D.