Review of Doctoral Thesis

Thesis: Modelling and Management of Multi-Model Data

Author: Pavel Koupil
Department of Software Engineering
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

Reviewer: Michal Krátký
Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB – Technical University of Ostrava
Czech Republic

Review

This thesis covers the following areas of multi-modal data management:

1. Conceptual data modelling,
2. Data transformation,
3. Inference of multi-model schemas,
4. Multi-model schema and data evolution.

In more details:

1. A conceptual modelling language is introduced, see Chapter 2 and the following article:

   M. Svoboda, P. Čontoš (Koupil), and I. Holubová. *Categorical Modeling of Multi-Model Data: One Model to Rule Them All.* The 10th International Conference on Model and Data Engineering, MEDI 2021. Tallinn, Estonia, June 2021. doi: 10.1007/978-3-030-78428-7 15 (CORE C).

2. Algorithms for data transformation are introduced, see Chapter 3 and the article:

   P. Koupil and I. Holubová. *A unified representation and transformation of multi-model data using category theory.* Journal of Big Data 9, 61 (2022). doi: 10.1186/s40537-022-00613-3 (Q1, IF: 14.57, SJR: 2.592)

3. An algorithm for inference of multi-model schemas is introduced, see Chapter 4 and the article:

   P. Koupil, S. Hricko, and I. Holubová. *A Universal Approach for Multi-Model Schema Inference.* Journal of Big Data 9, 97 (2022). https://doi.org/10.1186/s40537-022-00645-9 (Q1, IF: 14.57, SJR: 2.592)

4. A tool for multi-model schema and data evolution is introduced, see Chapter 5 and the article:

   P. Koupil, J. Bártík, and I. Holubová. MM-evocat: *A Tool for Modelling and Evolution Management of Multi-Model Data.* In a review process.

In Chapter 0, we can read a cover text for the rest of the thesis. In each chapter, there are a state of art, own approach and a comparison of the approach with other existing methods. Consequently, it seems like a standard scientific work. On the other hand, I distinguish the following issues which should be explained:

1. The list of own articles includes 14 items and some of them seem to be relevant to the topics of the thesis. However, only the above depicted 4 articles are indicated as relevant to the thesis. Can you explain this issue?
2. Is it possible to try out a tool, for example MM-evocat? I tried to follow `https://www.ksi.mff.cuni.cz/~koupil/mm-evocat/index.html`, however it seems that it is not available. Similarly, I tried to follow `http://nosql.ms.mff.cuni.cz/mmcat/`, however without any result.
3. Page 22: Can you explain a difference between GOOD and CGOOD?
4. Page 44: In particular, we focus on five selected schema inference approaches, namely:
   - Why these schema inference approaches have been selected?
5. Page 50: At the same time, schema inference approaches over the collections of JSON documents scale very well and are capable of handling Big Data
   - Is it possible to quantify the Big Data size?

Other notices:

1. Sometimes, references are missing, for example:
   (a) Page 1:
      i. However, nowadays, most of the data consists of very large (volume) and varied (variety) (un)structured data, which in addition are rapidly generated (velocity) and changed (variability).
      ii. Besides, in real-world applications the logical models are often combined, overlapped, and linked by references.

   (b) Page 45: Sevilla Ruiz et al., the approach of Baazizi et al.

Conclusion

There are 4 articles related to individual chapters of the thesis (2 journal articles, one article at a CORE C conference, and one article in a review process), which is acceptable for a PhD thesis. However, author should explain the above depicted issues. In this case I recommend accepting this work as a PhD thesis.

In Ostrava, September 12, 2022

....
Michal Krátký