Review of the Doctoral Thesis by Martin Svoboda

Title of the thesis: Correction of Invalid Trees with Respect to Regular Tree Grammars

Author: RNDr. Martin Svoboda

Reviewer: Ing. Radim Bača, Ph.D.

The work "Correction of Invalid Trees with Respect to Regular Tree Grammars" describes an approach allowing to correct a tree with respect to a regular tree grammar. This work has a straightforward motivation in a correction of XML trees which should be valid to some DTD or XML schema.

Major negative comments:

1. I really miss the proof of the approach correctness. Author claims that the approach is "always able to find all the minimal corrections". This sentence is in the contributions as well as in the conclusion. However, such a strong statement needs a proof! I can expect that author suppose that it comes from the model, however, I miss necessary steps for the claim. Author even did not bother to proof that *unfold* finds *all* shortest correction paths in a correction multigraph. If the thesis does not contain a proof of the correctness nor the algorithm complexity, why to introduce such a complicated model in the first place? I had to write down a long list of non-trivial terms in order to be able to read the thesis, however, I do not see the main purpose of them if they are not used in a theorems. I had a feeling that many terms could be described informally.

2. Author give up giving examples in Section 4. I was really glad for examples in Section three and they would be also useful in Section four. I have a feeling that algorithm is often best explained using appropriate examples. Of course a lot of code is quite obvious if I understand the model. However, that was true for a several algorithms at the beginning. Author could give some examples when describing the optimization algorithms. I do not see any reason why not to show the difference between correction strategies and execution approaches on an example.

3. When author speaks about contributions he mentions the polynomial worst-case time complexity of the algorithm. I miss more thorough analysis of the worst-case and the algorithm complexity analysis in general. If I compare the work with other works then the complexity proof is somewhat a standard in this field [1].

4. In the experiments author mention that it doesn't make a sense to compare their approach with an existing one since it "outperform according to features and expected time complexities at the theoretical level". First of all, authors approach complexity is not proven so the author should not refer to it. Moreover, I do not see a reason why not to compare the authors approach processing time with some other work even though the test would work with local tree grammars or single-type tree grammars and there would some other assumptions.

5. The data sets used in experiments can be more diverse. If the author is not comparing directly with other implementations he can use the data sets from other authors. For example if he

would use the National Corpus of Polish project [2] he could at least approximately compare his results with the results of [1].

Major positive comments:

- Author introduce a very promising approach to a tree correction with respect to a regular tree grammars (super set of DTD and XML schema) that allows to process even very large trees. The method is interesting in a way how it transforms the correction problem into a shortest path problem.
- Since a naïve correction algorithm can be quite inefficient author introduces several optimization methods of it. The first optimization based on a dynamic programming avoids repeated computations when possible using signatures. The second optimization avoids unnecessary computations by intelligent combination of building and searching algorithm. The author tests different combinations of his algorithm settings and compares his results.
- The whole solution is exceptional among similar works since it works for a very general type of grammars (regular tree grammars). Moreover, approach can process in real-time trees in order of magnitude larger than the previous works.

Literature

1. Amavi, Joshua, Béatrice Bouchou, and Agata Savary. "On correcting XML documents with respect to a schema." *The Computer Journal* 57.5 (2014): 639-674.
2. Savary, A., Waszczuk, J. and Przepiórkowski, A. (2010) Towards the Annotation of Named Entities in the National Corpus of Polish. Proc. LREC 10, Valletta, Malta, May 17–23. European Language Resources Association

Conclusion:

The author of this thesis seems to be technically skilled and he really loves theoretical description. However, I believe that theory should not be self-supporting only. Theory should be introduced in order to show some important results which is not clearly visible here.

On the other hand the experimental results of this work are quite impressive since it is possible to process in a order of magnitude larger trees than in previous works. The author optimized his algorithm in a very simple, yet, efficient way. He created an efficient tree correction approach for a very general types of schemas (regular tree grammars) which makes the approach exceptional among the related work. Due to this reasons I recommend the thesis to be defended.

In Ostrava, March 4, 2015

Radim Bača

Department of Computer Science

FEECS, VŠB - Technical University of Ostrava

17. listopadu 15, 708 00 Ostrava-Poruba