LA TROBE UNIVERSITY

Dept. of Computer Science & Computer Engineering
School of Engineering & Mathematical Sciences
Faculty of Science, Technology & Engineering

Mailing address
Victoria 3086 Australia

T +61 3 9479 1107
E info@cs.latrobe.edu.au
F +61 3 9479 3060
latrobe.edu.au/cs

MELBOURNE CAMPUSES
Bundoora
Collins Street CBD
Franklin Street CBD

REGIONAL CAMPUSES
Bendigo
Albury-Wodonga
Mildura
Shepparton

Melbourne, 27 February 2015

Reviewer's Report

Thesis: Correction of Invalid Trees with Respect to Regular Tree Grammar
Author: Martin Svoboda

In many situations, when an XML document does not comply with its schema, the document will be discarded because the applications using the document simply cannot process it. This is not always desirable because information in the document might be required for a particular purpose. Therefore, there is a need for a mechanism to correct the document following its schema specifications. This is the original motivation of this research, which is to find mechanisms to efficiently correct XML documents so they comply with the associated schema represented in XML schema languages such as DTD and XSD. Instead of only focusing on XML, Svoboda in his thesis takes further step by considering correction of any invalid tree with respect to regular tree grammar.

In **Chapter 1**, the clear motivation of the work is presented. How the thesis work is positioned within the whole range of project undertaken by the author in his research group is clearly defined. The research has strong connection to other research area such as data analytics, linked data processing and graph data management. This chapter also briefly summarised the latest existing work that is closely related or comparable to his work.

In **Chapter 2**, the author provides foundation knowledge required to fully understand his proposed approaches. Since the goal of the thesis is to determine the correction approaches of XML document represented in data tree abides its corresponding DTD or XSD represented in a regular tree grammar, the author explains the concept and states the definition on regular expression, finite automata and regular tree grammars. The scope and context are stated clearly (for example, the scope of data trees used only for element corrections and the rationale of Glushkov automata as the selected finite automata formalization). Overall, this section is complete to provide basic knowledge before readers comprehend the actual proposal.

In **Chapter 3**, the author formally explains the correction model including its capabilities, features and principles. In general, the model is described sequentially into four steps: (i) the identification of *edit operations* that data trees can employ; (ii) the identification of *corrections intents* required by the data trees against allowed grammar; (iii) the formation of *correction multigraphs* from the sequence of corrections identified earlier and (iv) the computation of the final *intent repairs* using the shortest correction paths. Each aforementioned step is formally presented and where appropriate, there are numerous examples that explain the process more clearly.

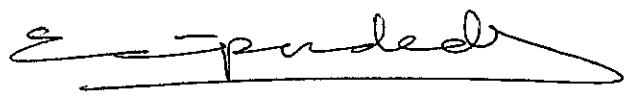
To carry out the correction model in the previous chapter, in **Chapter 4**, the author proposes the algorithms. The algorithms show how to search the shortest correction paths and associate correction costs for every edge in the correction multigraphs. The algorithms are configured as a combination of three possible options: three correction strategies (default, exploring and refinement), five execution approaches (nesting single, invoking single, invoking multiple, forwarding single and forwarding multiple) and two signature modes (disabled and enabled). They are formally presented with sufficient explanation. The interconnection between each component of the algorithms is clear and logically structured.

In **Chapter 5**, the experiments of the correction algorithms are implemented and demonstrated. The implementation is developed using Java in an integrated system called *Corrector*, which is publicly available if needed for reproducibility purpose. As the inputs, Corrector takes XML documents and regular tree grammar, while the output is a list of edit operations required from the corrections. The experimental setting is explained briefly but complete. The set of experiments provided by the author shows that the correction algorithms work successfully and the performance implication (such as execution times and scalability factors) is also discussed.

In **Chapter 6**, the author concisely summarises his overall research and lists the contributions.

As a conclusion, this thesis contributes to the area of correction of invalid trees with respect to regular tree grammars. The theoretical contribution is presented in sufficient manner and proof of concepts in terms of rigorous experimentation is provided. All components of this research have been published by the authors in one reputable peer-reviewed journal and several peer-reviewed conference proceedings. The examiner believes that **the author of the thesis has proven to have ability for creative scientific work.** The reviewer thus recommends the thesis to PhD defense.

Kind Regards,



Dr Eric Pardede
Department of Computer Science and Information Technology
College of Science, Health and Engineering
La Trobe University, Melbourne AUSTRALIA 3083
Email: E.Pardede@latrobe.edu.au
Ph: +61 3 9479 3459 – Fax: +61 3 9479 3060