Review of Doctoral Thesis

Author: RNDr. Martin Svoboda
Title: Correction of Invalid Trees with Respect to Regular Tree Grammars
Reviewer: Doc. RNDr. Irena Holubová, Ph.D. (thesis supervisor)

General Description

The thesis consists of 6 chapters covering 145 pages altogether. The first chapter provides an introduction and motivation for the problems solved in the rest of the thesis and explains which aspects are solved in which parts of the text and respective papers. In the second chapter the author provides the necessary formal background on which the rest of the thesis is based. In the third section the author defines the proposed correction model and discusses its features and advantages. The fourth section describes the proposed correction algorithms and the fifth section their respective experimental evaluation and comparison. The sixth section sums up and concludes.

The structure of the text conforms to principles and requirements on the structure of a scientific thesis. The author has studied and used appropriate number of bibliography sources quoted in the thesis. It is the evidence of the deep theoretical knowledge and very good orientation in the problems discussed in the text.

The proposed approaches are described using a complex formal background which demonstrates the quality of the results and enables one to precisely study and compare the proposed approaches.

The word processing of the thesis is adequate. The usage of different fonts and the structure of the text is proper and helps the reader in better orientation. There are also numerous figures and examples which help the reader to understand the ideas. The thesis fulfils the general formal requirements at a very good level.

In general, the author proved the ability to prepare a sound and explanatory research text describing the solutions of the selected problem with all appropriate parts.

The Topic and Results of the Thesis

The problem of correction of invalid XML (and not only XML) data is a key challenge encountered whenever we want to process real-world data. Probably all existing analyses of real-world data agree on the fact that a significant portion (namely more than 30%) of the crawled data contained various types of errors. Currently there exist several approaches to the correction problem; however, none of them solves it efficiently, completely, and with both formal background and experimentally evaluated implementation.

The author of the thesis has performed a good orientation and a wide knowledge of different parts of the target area from both applied and theoretical point of view. He has studied and solved the target problems from several points of view and within several research teams and projects, namely the *XML and Web Engineering Research Group*, the GACR projects (1) *Processing of XML Data*, (2) *Management of XML Data in (Object-)Relational Databases and Related Issues*, and (3) *Handling XML Data in Heterogeneous and Dynamic Environments*, and partially also the GAUK project *Efficient Processing of Linked Data* (as a principal researcher) and EU FP7 project *LOD2*.

Author’s Contributions

The main author’s contributions are related to the correction algorithms. They result from a very practical need which was a part of the *Analyzer* project devoted to statistical analyses of real-world XML data and operations. It was implemented as a student SW project at the Faculty of Mathematics and Physics of the Charles University in Prague and Martin Svoboda was a member of the team. The basic ideas for corrections of XML data were extended in his Master thesis (which gained the Dean's Award for the Best Master Thesis of 2010) and further in the following papers forming the core of the doctoral thesis.

Despite the practical origin, the algorithms have a strong and sound formal basis. In addition, they are proven to be applicable not only on XML data but their more general superset. And, last but not least, the algorithms outperform any existing similar approach.

Author’s Publications

The publications covered in the thesis involve 1 journal paper with impact factor, 9 international and 2 national refereed proceeding papers. The author has also 1 other impacted journal paper under the review process which confirms that the research is not closed but continues.

Most of the papers directly cover the content of the thesis, other are closely related which confirms that the author does not limit himself to a single narrow topic. The other topics are related to cooperation with his colleagues in the XML and Web Engineering Research Group or topics the author has encountered during his three international internships. All the studied areas however have a common aim of efficient processing of graph data, most of them resulting from the basic topic of analysis of real-world data.

In general, the results are more than sufficient with regard to the respective research level.

Conclusion

In my opinion, the doctoral thesis of RNDr. Martin Svoboda fulfils all the conditions for gaining the Ph.D. degree in Computer Science; therefore I recommend it.

In Prague, February 5, 2015

Doc. RNDr. Irena Holubová, Ph.D.

Department of Software Engineering
Faculty of Mathematics and Physics
Charles University in Prague
Malostranské nám. 25
118 00 Praha 1
Czech Republic