

Review of Doctoral Thesis

Thesis Title: XML Data in (Object-)Relational Databases

Thesis Author: RNDr. Irena Mlýnková
Faculty of Mathematics and Physics
Charles University
Prague

Review Author: Doc. Ing. Michal Krátký, Ph.D.
Department of Computer Science
VŠB Technical University of Ostrava
17. listopadu 15
708 33 Ostrava-Poruba
Czech Republic

Review

Introduction

A develop of the storage for XML data is up-to-date research problem. This work tries to extend knowledges in the storage of XML data in (object-)relational databases. The thesis includes seven chapters, each chapter develops author's publications.

We can distinguish these publications:

- 1 article published by IEEE Computer Society Press.
- 1 chapter in a book published by Springer-Verlag.
- 1 article in a journal without impact factor.
- 4 articles in international conferences with local proceedings.
- 5 technical reports.

It seems that the two most important publications include rather the state-of-the-art. Publications with novel methods were published at international conferences with local proceedings.

Notices

In the first chapter, author proposes the state-of-the-art for the research area. She depicts a classification of XML-to-relational mapping, we distinguish: generic, schema driven, and user-defined methods.

Question:

Can author give a reference that Oracle RDBMS is included in the third set, see page 2? If we consider the following reference, there is no such information.

M. Krishnaprasad, Z. H. Liu, A. Manikutty, J. W. Warner, V. Arora, and S. Kotsovolos: Query Rewrite for XML in Oracle XML DB. In *Proceedings of 30th International Conference on Very Large Data Bases (VLDB 2004)*. Morgan Kaufmann Publishers, 2004. ■

Question:

There is a kind of inconsistency in this chapter. Can author describe differences between the user-defined term introduced at page 2 and user-driven term proposed at page 3? ■

In the second chapter, basic terms are stated. The third chapter includes techniques to be considered as an efficient way of XML processing based on (object)-relational databases.

Question:

Whereas author proposes adaptive and user-driven methods at page 3, cost-driven and user-driven methods are introduced at page 13.

Is it an inconsistency or term overloading? ■

In chapter 4, author depicts own results, a novel algorithm is introduced. In chapter 5, she describes a similarity measure designed primarily for the purpose of enhancing of user-driven XML-to-relational strategy. This chapter is based on the unpublished author paper.

In chapter 6, author introduces a statistical analysis of existing real-world XML data. This chapter is based on two articles published at international conferences with local proceedings. It includes important and interesting results, however, there are some unclear points for me.

Questions:

1. The Inex 1.9 collection includes 16,071 documents. However, author proposes 16,534 documents in all tested collections.
2. What is the factor of the tested XMark collection, 1, 10, 100?
3. Figure 6.2 includes XML nodes with the level up to 400. In general, the average level of a node in an XML document is 6. Can you show a document with the depth?

■

The seventh chapter is very incomprehensible for me. I waited results of a query evaluation of XML data stored in relations according to proposed algorithms. However, no result is given in this chapter. Therefore, the result of the work is unclear for me. In my opinion, the most important reason for the research in this area is a develop of more efficient storage of XML data. It means, the storage for more efficient evaluation of an XML query. However, this chapter does not include DAC or time of a query processing.

There are many articles where authors have been introduced their methods for the storage of XML data in relational databases. We can select the one article:

Z. Chen, G.J. Korn, F. Koudas, N. Shanmugasundaram, and J.D. Srivastava: Index Structures for Matching XML Twigs Using Relational Query Processors. In *Proceedings of 21st International Conference on Data Engineering (ICDE 2005)*. IEEE Computer Society, 2005. ■

Can author explain the goal of her work? Is her result more efficient than this method (or another one)? This work does not answer the most important question. It is unclear point of this work.

Conclusion

The language of the work is very good. Moreover, this work includes both own theoretical and experimental issues. However, main issues of the storage – the efficiency of querying XML data is missing out and, therefore, the goal of the work is rather unclear. Since some articles were published by IEEE Computer Society Press and Springer-Verlag, I recommend this work for the defense. Author proved the ability of the stand-alone work.

In Ostrava, 20th August 2007

