Poznań, 04.02.2006

Dr. Hab. Tadeusz Pankowski
Faculty of Mathematics and Computer Science
Adam Mickiewicz University
ul. Umultowska 87
61-614 Poznań
Poznań/Poland

Referee report on Ph.D. thesis "XPath, XSLT, XQuery: Formal Approach" by Mgr. Pavel Hlouska

Summary

The thesis consists of seven chapters and 129 pages. In Chapter 1 "Introduction" contributions and structure of the thesis are outlined. Chapter 2 "Background" discusses the background of the thesis as well as gives an overview of related work. Notations used throughout the thesis are introduced in Chapter 3 "Notation". The candidate defines notations for basic types of the XML data model and some notations which are essential for the definition of formal semantics of the languages under consideration. A running example is also introduced in this chapter. Chapters 4, 5 and 6 are main parts of the thesis. In Chapter 4 "XPath" some new and essential results concerning the XPath language are presented. Some important results about the XQuery language are presented in Chapter 5 "XQuery". Chapter 6 "XSLT" discuss the XSLT language and gives results about a core of XSLT in the sense that all language constructs can be rewritten using construct of the core. Chapter 7 concludes and summarizes the thesis and shows some directions of the future work. References include 57 bibliographic positions, three of them are written by the candidate. The thesis is written in good English and is generally clear and well structured. The tables and figures are appropriate.

Contribution

The main original contribution of the thesis is included in Chapters 4, 5 and 6. The contribution concerns definition of formal semantics of languages XPath, XQuery and XSLT, as well as identification of the core parts of these languages. The obtained new scientific results can be characterized as follows:

1. The candidate proved that the XPath 2.0 language has expressive power allowing for sorting arbitrary sequences (Theorem 4.1, p. 38). This fact was not noticed in the standard W3C Recommendation, on the contrary, authors of the recommendation claimed that it is infeasible to define sorting semantics in the data model for XPath/XQuery. Consequences of XPath's sorting ability are next used to define formal semantics of sorting for XQuery.
2. In Chapter 5 the candidate proved that the sorting semantics of XQuery is expressible in the used data model. To this aim he used his own results about sorting ability of the XPath 2.0 language. This semantics is based on the notion of tuples. It is worthy to note that in the W3C Recommendation the formal semantics specification does not formally define terms of tuples and tuple types. Formal definitions of these objects introduced in the thesis

enables to express ordering semantics formally. The main achievement of the candidate is proving that there exists a transformation from a general FLWOR expressions with an *order by* clause to the XQuery Core. This result is new and essential since the W3C formal semantics is not able to express semantics of sorting and the authors of the formal specification concluded that sorting is not expressible in the data model of XPath and XQuery. The reason is that there is no formal specification of tuple semantics of FLWOR expressions. The candidate proposed such semantics and was able to obtain new valuable results. Thanks to this, the process of FLWOR normalization, i.e. the transformation an XQuery query to an XQuery Core query, can be now performed without losing information provided by the *order by* clause. The result is summarized in Theorem 5.1 and Collorary 5.1 in page 63. Some examples demonstrate applications of these theorems in Subsection 5.4.4.

3. The next contribution of the candidate is the result obtained from the deep investigation of XSLT language. He has identified the XSLT Core language and defined the formal semantics for it and proved some theorems showing how general constructs of XSLT can be expressed by means of core constructs. To proof these theorems the candidate define his own non-XML syntax for XSLT expressions. In particular, he proved that many of instructions, such as sorting, copying and grouping, may be expressed by other instructions and need not be included into XSLT Core. To proof this, the results obtained in the context of XQuery investigations have been used.
4. The candidate provided a unified framework for defining the formal semantics of all three languages XPath, XQuery and XSLT.

Remarks

The problem of considering XML query languages formally is a very hot topic in computer science and in mathematical foundation of computer science. The three languages investigated by the candidate are the most important and popular in this field. Especially XQuery 1.0 with its sublanguage XPath 2.0 is considered to become the standard query language for XML documents. However, its syntax and semantics continuously evolve – the last version of its formal semantics is dated on November 3, 2005. Nevertheless, popularity of XPath, XQuery and XSLT for querying XML documents is growing rapidly. Thus, I appreciate both the candidate effort and contribution to investigation and formal description of these languages – especially definition of formal semantics and identifying their core components. Contributions of the candidate in this area have been mentioned in the previous paragraph. However, there are some drawbacks in the work, which I address below.

1. There are some lemmas in the thesis which could be omitted because they are either trivial or not original. For example, Lemma 4.3, Lemma 4.4 and Lemma 6.2 seems to be trivial.
2. The Definition 4.2 is misleading since it defines a total ordering relation as the difference between a partial ordering relation R and the set of equal elements under R. I think that some additional assumptions should be stated to make the definition correct.
3. The second paragraph following the Definition 4.5 is slightly mismatched.
4. Considerations in Section 4.5 about quantified expressions are rather not original. Moreover, XPath expression in Definition 4.6 is wrong, because the function `fn:boolean` has illegal argument, i.e. the truth value `fn:true()`, this argument should be the variable $v instead.
5. In the W3C Recommendation the formal semantics for an expression is defined with respect to an evaluation context called *environment*, and the notion of environment is well defined. In the thesis, the term *context* and sometimes also the term *focus* are used. I think

it is not justified and misleading, and I think that the term *environment* should be used instead.

6. In Figure 30 there are two non-terminals ParamDecl and Param, but only one of them should occur.

Importance of results

The obtained results have the following importance and possible applications.

1. They can be treated as a contribution to the theory of functional languages. XPath and XQuery are functional languages, thus the result concerning the formal semantics of sorting has the importance to our knowledge about this class of languages.
2. The definition of formal semantics forms a base for implementation because it clarifies the intended meaning of the specification and ensures that all aspects are taken into account.
3. The results may be used to query optimization. We can use the formal semantics to prove that some expressions in a query can be replaced by equivalent expressions, which are less expensive to evaluate. However, this aspect was not addressed directly in the thesis (I think that, for example, some problems of query containment might be discussed using the developed formalization).

Overall evaluation

The candidate worked on a difficult and very relevant issue – finding appropriate means for defining formal semantics for XPath, XQuery and XSLT in a unified way, and applying this formalization for identifying the cores of these languages. In this way the candidate proves his ability for creative scientific work. I find that this thesis describes original work, with some challenges remaining as to be expected, and I recommend that this Ph.D. thesis be accepted.