XML Navigation and Tarski's Relation Algebras

Maarten Marx

Informatics Institute, Universiteit van Amsterdam
The Netherlands

Navigation is at the core of most XML processing tasks. The W3C endorsed navigation language XPath is part of XPointer (for creating links between elements in (different) XML documents), XSLT (for transforming XML documents) and XQuery (for, indeed, querying XML documents). Navigation in an XML document tree is the task of moving from a given node to another node by following a path specified by a certain formula. Hence formulas in navigation languages denote paths, or stated otherwise binary relations between nodes. Binary relations can be expressed in XPath or with first or second order formulas in two free variables. The problem with all of these formalisms is that they are not compositional in the sense that each subexpression also specifies a binary relation. This makes a mathematical study of these languages complicated because one has to deal with objects of different sorts. Fortunately there exists an algebraic formalism which is created solely to study binary relations. This formalism goes back to logic pioneers as de Morgan, Peirce and Schröder and has been formalized by Tarski as *relation algebras* [7]. (Cf., [5] for a monograph on this topic, and [8] for a database oriented introduction). A relation algebra is a boolean algebra with three additional operations. In its natural representation each element in the domain of the algebra denotes a binary relation. The three extra operations are a constant denoting the identity relation, a unary conversion operation, and a binary operation denoting the composition of two relations. The elements in the algebra denote *first order definable* relations. Later Tarski and Ng added the Kleene star as an additional operator, denoting the transitive reflexive closure of a relation [6].

We will show that the formalism of relation algebras is very well suited for defining navigation paths in XML documents. One of its attractive features is that it does not contain variables, a feature shared by XPath 1.0 and the regular path expressions of [1]. The connection between relation algebras and XPath was first made in [4].

The aim of this talk is to show that relation algebras (possibly expanded with the Kleene star) can serve as a unifying framework in which many of the proposed navigation languages can be embedded. Examples of these embeddings are

1. Every Core XPath definable path is definable using composition, union and the counterdomain operator $\sim$ with semantics $\sim R = \{(x, x) \mid \mathsf{not}\ \exists y : xRy\}$.
2. Every first order definable path is definable by a relation algebraic expression.
3. Every first order definable path is definable by a positive relation algebraic expression which may use the Kleene star.

L. Ong (Ed.): CSL 2005, LNCS 3634, pp. 1–2, 2005.

4. The paths definable by tree walk automata and certain tree walk automata with pebbles can be characterized by natural fragments of relation algebras with the Kleene star.

All these results hold restricted to the class of finite unranked sibling ordered trees. The main open problem is the expressive power of relation algebras expanded with the Kleene star, interpreted on this class of models. Is this formalism equally expressive as binary first order logic with transitive closure of binary formulas? Whether the latter is equivalent to binary monadic second order logic is also open [2, 3]. So in particular we do not know whether each regular tree language can be defined in relation algebras with the Kleene star.

References

1. S. Abiteboul, P. Buneman, and D. Suciu. *Data on the web*. Morgan Kaufman, 2000.
2. J. Engelfriet and H. Hoogeboom. Tree-walking pebble automata. In *Jewels are Forever, Contributions on Theoretical Computer Science in Honor of Arto Salomaa*, pages 72–83. Springer-Verlag, 1999.
3. J. Engelfriet and H. Hoogeboom. Automata with nested pebbles capture first-order logic with transitive closure. Technical Report 05-02, LIACS, 2005.
4. J. Hidders. Satisfiability of XPath expressions. In *Proceedings DBPL*, number 2921 in LNCS, pages 21–36, 2003.
5. R. Hirsch and I. Hodkinson. *Relation algebras by games*. Number 147 in Studies in Logic and the Foundations of Mathematics. North-Holland,, 2002.
6. K. Ng. *Relation Algebras with Transitive Closure*. PhD thesis, University of California, Berkeley, 1984.
7. A. Tarski. On the calculus of relations. *Journal of Symbolic Logic*, 6:73–89, 1941.
8. J. Van den Bussche. Applications of Alfred Tarski's ideas in database theory. *Lecture Notes in Computer Science*, 2142:20–37, 2001.