Describing and Querying Semistructured Data: Some Expressiveness Results

Natasha Alechina[1] and Maarten de Rijke[2]

[1] School of Computer Science, University of Birmingham
Birmingham, B15 2TT, England
N.Alechina@cs.bham.ac.uk
WWW home page: http://www.cs.bham.ac.uk/~nxa
[2] ILLC, University of Amsterdam, Pl. Muidergracht 24
1018 TV Amsterdam, The Netherlands
mdr@wins.uva.nl
WWW home page: http://www.wins.uva.nl/~mdr

Data in traditional relational and object-oriented databases is highly structured and subject to explicit schemas. Lots of data, for example on the world-wide web is only *semistructured.* There may be some regularities, but not all data need adhere to it, and the format itself may be subject to frequent change.

The important issues in the area of semistructured data are: how to describe (or constrain) semistructured data, and how to query it. It is generally agreed that the appropriate data model for semistructured data is an edge-labeled graph, but beyond that there are many competing proposals. Various constraint languages and query languages have been proposed, but what is lacking so far are 'sound theoretical foundations, possibly a logic in the style of relational calculus. So, there is a need for more works on calculi for semistructured data and algebraizations of these calculi' [Abiteboul 1997].

One of the main methodological points of this paper is the following. There are many areas in computer science and beyond in which describing and reasoning about finite graphs is a key issue. There exists a large body of work in areas such as feature structures (see, for example, [Rounds 1996]) or process algebra [Baeten, Weijland 1990,Milner 1989] which can be usefully applied in database theory. In particular, many results from modal logic are relevant here. The aim of the present contribution is to map new languages for semistructured data to well-studied formal languages and by doing so characterise their complexity and expressive power.

Using the above strategy, we study several languages proposed to express information about the format of semistructured data, namely data guides [Goldman, Widom 1997], graph schemas [Buneman et al. 1997] and some classes of path constraints [Abiteboul, Vianu 1997]. Among the results we have obtained are the following:

Theorem 1. *Every set of (graph) databases defined by a data guide is definable by an existential first-order formula.*

Theorem 2. *Every set of (graph) databases conforming to an acyclic graph schema is definable by a universal formula.*

Every set of (graph) databases conforming to an arbitrary graph schema is definable by a countable set of universal formulas from the restricted fragment of first-order logic.

We also argue that first-order logic with transitive closure FO(TC) introduced in [Immerman 1987] is a logical formalism which is best suited to model navigational query languages such as Lorel [Abiteboul et al. 1997] and UnQL [Buneman et al. 1997]. We give a translation from a Lorel-like language into FO(TC) and show that the image under translation is strictly less expressive than full first-order logic with binary transitive closure.

Theorem 3. *Every static navigational query is expressible in FO(TC).*

Theorem 4. *The image under translation into FO(TC) of static navigational queries is strictly weaker than FO(TC) with a single ternary predicate Edge and binary transitive closure.*

In our ongoing work, we are investigating the use of FO(TC) for query optimisation, and we are determining complexity characterisations of the relevant fragments of FO(TC), building on recent results by [Immerman, Vardi 1997] on the use of FO(TC) for model checking.

References

[Abiteboul 1997] Abiteboul, S.: Querying semi-structured data. Proc. ICDT'97 (1997)

[Abiteboul et al. 1997] Abiteboul, S., Quass, D., McHugh, J., Widom, J., Wiener, J.L.: The Lorel query language for semistructured data. J. of Digital Libraries, **1** (1997) 68–88

[Abiteboul, Vianu 1997] Abiteboul, S., Vianu, V.: Regular path queries with constraints. Proc. PODS'97 (1997)

[Baeten, Weijland 1990] Baeten, J.C.M., Weijland, W.P.: Process Algebra. Tracts in Theoretical Computer Science, Vol. 18. Cambridge University Press (1990)

[Buneman et al. 1997] Buneman, P., Davidson, S., Fernandez, M., Suciu, D.: Adding structure to unstructured data. Proc. ICDT'97 (1997)

[Goldman, Widom 1997] Goldman, R., Widom, J.: DataGuides: enabling query formulation and optimization in semistructured databases. Proc. VLDB'97 (1997)

[Immerman 1987] Immerman, N.: Languages that capture complexity classes. SIAM J. of Comput. **16** (1987) 760 – 778

[Immerman, Vardi 1997] Immerman, N., Vardi, M.: Model Checking and Transitive Closure Logic. Proc. CAV'97 (1997) 291–302

[Milner 1989] Milner, R.: Communication and Concurrency. Prentic Hall (1989)

[Rounds 1996] Rounds, W.C.: Feature logics. In: van Benthem, J., ter Meulen, A. (eds.): Handbook of Logic and Language. Elsevier (1996)