

Optimising XPath Queries



Project Leader

Peter Wood

ptw@dcs.bbk.ac.uk

Project staff

Manizheh Montazerian

gmont05@dcs.bbk.ac.uk

Project Details

PhD Research

Oct. 2002 -

Project Web Site

<http://www.dcs.bbk.ac.uk/~ptw/optimize.html>

Keywords

XPath, Query Optimisation, Document Type Definition, Query Containment, Query Satisfiability

Project Aims

XML queries and transformations usually make use of XPath expressions, which provide a way of navigating an XML tree (corresponding to a document) and return the set of nodes which are reachable from one or more starting nodes through the paths specified by the expressions. The efficiency of expression evaluation depends on the size of the expression, so it is important to have queries of minimum size. This project is investigating:

- the constraints which can be inferred from a Document Type Definition (DTD) and enable us to simplify queries,
- the containment and satisfiability problems for common fragments of XPath, and
- those real-world situations in which XPath expressions can be minimised efficiently.

Query Containment and Satisfiability

It has been shown that the problem of minimising a given XPath expression is closely related to the problem of checking whether there are two sub-expressions p and q such that p is *contained* in q . This means that every node returned by expression p is also returned by expression q . Thus query containment for XPath has been an active area of research. The query containment problem is made more complicated when the documents being queried are known to be valid with respect to a particular DTD or schema. On the other hand, containments or redundant sub-expressions are more likely to be found when the constraints imposed by a DTD or schema are taken into account.

Another related problem is to determine the *satisfiability* of an XPath expression, that is to determine whether or not the

expression returns the empty answer for all input documents. It can be shown that the satisfiability problem is (the negation of) a special case of the containment problem. The satisfiability problem can be used in query optimization to avoid the evaluation of unsatisfiable queries. Thus, applying the satisfiability test before executing a query can save processing time. Studying the satisfiability problem separately from the containment problem is worthwhile because it turns out that the containment problem has high complexity (and is sometime undecidable) for certain fragments of XPath.

Useful Properties of Real-World Data

Although recent results have shown that problems of containment and satisfiability under DTDs are NP-hard or worse for most fragments of XPath, we have investigated whether real-world DTDs in fact exhibit properties that make these problems easier to solve. We have identified two properties, called covering and duplicate-free, and shown that, in 100 real-world DTDs comprising over 5500 rules, fewer than 1% of the rules satisfy neither of the properties. In addition, we have identified a number of XPath fragments for which the complexity of the satisfiability problem reduces to PTIME when duplicate-free or covering DTDs are used. We are also studying the containment problem for different XPath fragments under real-world DTDs.

Key publication

M. Montazerian, P.T. Wood and S.R. Mousavi, "XPath Query Satisfiability is in PTIME for Real-World DTDs," in *Proc. Fifth Int. XML Database Symposium*, (Vienna, Austria, Sept. 23-24), LNCS 4704, Springer-Verlag, 2007, pp. 17-30.