Integrating Efficient XPath Evaluation Techniques into a Complete XQuery Engine

Philippe Michiels Jérôme Siméon Mary Fernández
George A. Mihăilă

Thanks to its self-describing and flexible nature, XML is naturally well suited for integrating information originating from multiple sources, such as databases, Web repositories or even messages. Depending on the source of the data, an XQuery engine may decide to use different access methods and processing techniques. Thus, for example, if an XML document is stored in a database it may be beneficial to make use of specific indices that have been pre-built. On the other hand, if the document is retrieved from the Web or has a transitory nature, such indices will not be available. Sometimes an index can be built on the fly, if that is anticipated to reduce the overall processing time. In other cases, it may be better to process the document in a streaming fashion, with little or no intermediate materialization.

Various storage schemes have been developed to support these different scenarios. Usually, these storage schemes are accompanied with special algorithms that enable the efficient evaluation of particular XPath fragments for their specific data model. Since these algorithms are specifically tuned towards the underlying storage scheme, they are usually applied in isolation.

In this work, we provide an algebraic framework that enables the integration of XPath evaluation algorithms and storage schemes inside a complete XQuery engine. The goal is to study how well these techniques perform on various physical representation of the XML data such as tree, shredded relation and streams. By providing a common infrastructure for physical XPath evaluation algorithms, our approach opens up possibilities for combining parts from different algorithms in new and useful ways.