Report on the First "XQuery Implementation, Experience and Perspectives" Workshop (XIME-P)

Ioana Manolescu

INRIA Futurs, Gemo group, France
Ioana.Manolescu@inria.fr

Yannis Papakonstantinou

Computer Science and Engineering
UC San Diego, USA
yannis@cs.ucsd.edu

Summary

The *XQuery Implementation, Experience and Perspectives (XIME-P)* workshop was organized by Ioana Manolescu and Yannis Papakonstantinou in cooperation with the ACM SIGMOD Conference, and was held in *Maison de la Chimie*, in Paris, France, on June 17 and 18, 2004. This report summarizes the goals and topics of the workshop, presents the major workshop highlights and the main issues discussed during the workshop.

Workshop Goals and Topic

Work has been going on in the last four years at the W3C on the standardization of XQuery, a query language for XML. In parallel, several XML query processing engines have been developed, some of which leverage previous efforts on SGML, HTML, OEM, and XML querying.

XQuery is a complex language, whose specification relies on a functional programming model. Its expressive power covers SQL-like data transformations, document search and restructuring, ordered and unordered querying, function-based extensibility, and other features. This complexity caters to widely different requirements, and makes XQuery useful in many contexts, such as application integration, information integration, content management, stream processing, information retrieval , etc. At the same time, XQuery implementations draw from multiple paradigms and technologies, such as relational and object-oriented database systems, automata-based techniques, and native XML querying frameworks.

The great pluralism in applications and implementations creates a need for increased communication in the community. At this crucial point in XQuery standardization, we proposed the organization of XIME-P as a forum where researchers from the industry and academia can jointly assess the existing efforts on XQuery implementation, envision its future, and discuss the commonalities and differences among the various applications and implementations.

XQuery Standardization

The first XIME-P session was devoted to the W3C standardization process and resulting specifications. The keynote speaker, Mary Fernàndez (AT&T), presented a short history of the standardization process, outlining the roles and influences of existing technologies, formalisms, industrial efforts and standard committee members themselves, on the XQuery specifications. Mary emphasized the need for the XQuery standard to be compatible with pre-existing XPath and XSL specifications. This was an important requirement for the W3C and the main reason for the differences between XQuery and the database-oriented existing query languages.

Then XIME-P featured a panel on the standardization and usage of XQuery for various communities. The panel featured Daniela Florescu (BEA), Jérôme Siméon (IBM Watson), Dan Suciu (U. Washington) and Vassilis Vassalos (AUEB, Greece). Participants raised concerns that the complexity of the W3C specification makes it difficult to faithfully and efficiently support it in its entirety. It was suggested that isolating a clean-cut "XQuery subset" capable of performing most usual operations, but lacking some of the complex features (such as full support of the type system etc.) might help in making XQuery better understood and easier to adopt and address in research works.

State of Industrial XQuery Implementations

XIME-P featured a session on XQuery implementations produced in the industry, either as commercial products or research prototypes. This session featured three paper presentations, and a panel featuring Michael Carey (BEA), Fatma Özcan (IBM), Markos Zaharioudakis (Xyleme SA) and Torsten Fiebig (Software AG).

The speakers expressed different points of view on the language adoption, implementation, and perceived applications. For BEA, XQuery is at the center of an application integration suite, based on Web service messaging; BEA does provide a comprehensive industrial-strength XQuery implementation. IBM does not currently provide an XQuery implementation as part of a product; IBM research presented in XIME-P is focused on the perceived need of a *group by* construct, which is currently lacking from the language specification. Software AG's Tamino supports a subset of XQuery, and provides furthermore some text retrieval and update facilities which are not covered by the current language. Xyleme provides a suite for managing large corpora of XML documents in a cluster architecture; these documents are queried using a relatively restricted XQuery subset, which does not currently support updates.

The session did not produce a consensus view as to the adoption of the full XQuery specification in the industry. Some companies do support it, or intend to do so, some others orient their business plan around simpler XML management/query models and others plan on extensions (such as *group-by*) and updates. The complexity of the standard was perceived as a big obstacle in the adoption of XQuery for tasks routinely delegated to database processing languages such as SQL; this opinion was shared also by the audience.

XQuery Research and Applications

Seven refereed papers described implementation architectures, particular technical points on specific language feature, and envisioned applications.

Three processing models were presented: processing on top of a persistent storage (database-style), streaming XQuery processing, and compiling XQuery into some imperative program that handles XML documents in memory as custom program data structures. In the persistent store scenario, the aspects targeted by presentations were index design in ToX (University of Toronto) and join ordering in Natix (University of Mannheim). They both do not support the full language, rather, they approach XQuery implementation based on classical database principles, such as physical data independence and algebraic optimization.

The main question emerging from this session was whether there are genuinely new problems raised by XQuery processing (and if so, do they pertain to the database field ?) or is XQuery research re-discovering known (and solved) problems? The difficulty in raising a consensus on such a question seems to stem from the conflicting viewpoints on what the language actually is: a complex functional programming language or a high-level query language (thus prone to optimization using techniques from the database field). Adopting the first answer has lead to complete, yet (still) inefficient, implementations; adopting the second leads to efficient storage models and algebraic frameworks for XML, but at the price of abstracting away the language until it gets easy to manage in such a framework. The audience made clear that reaching a consensus will greatly increase the chances for efficient *and* comprehensive implementations.

XQuery Perspectives

The final panel featured Serge Abiteboul (INRIA), Stefano Ceri (Politecnico di Milano), Alin Deutsch (UCSD) and Ioana Manolescu (INRIA). All panelists teach XQuery in undergraduate or graduate university classes. The panelists argued that many XML query research topics are still open and present novel challenges: algebraic frameworks for query optimization, view-based query rewriting, extensions for updates, and distributed XML data management. At the same time, the lack of clear underlying paradigms currently makes it difficult to compare the multiple works on XQuery processing and assess the success that the community has had in meeting the challenges raised by XQuery.

Numbers, Feedback and Conclusions

XIME-P received 30 paper submissions, out of which 10 were selected by a 13-member program committee. Each paper was rated by 3 reviewers. The call for papers reached a sizeable audience, as witnessed by the multiple submissions from industry and academia. The workshop had 70 attendees, which, for a first issue, is a success, especially given the "competition" by WebDB, which happened at the same time.

Many of the attendees have provided enthusiastic feedback, especially regarding the lively discussions that addressed controversial aspects of XQuery. Interestingly, the organizers had received before the workshop some negative opinions on XQuery as a framework for database research. While most of the proponents of the negative viewpoints did not attend the workshop, a fair deal of criticism was voiced, mainly concerning the language complexity. The presence of W3C standardization committee members and of an interested audience enabled the discussion and clarification of some of these concerns, shed light on XQuery's current status and requirements and provided ideas on how the involvement of the database community in XQuery can continue in ever more productive ways. We consider this to be among the most sucessful components of the workshop.

The feedback has encouraged us to set up the succession of XIME-P. An application is currently under preparation for the second XIME-P workshop, organized in cooperation with SIGMOD 2005, and co-chaired by Hamid Pirahesh (IBM) and Daniela Florescu.

Workshop Web Site (On-line Proceedings)

http://www-rocq.inria.fr/gemo/Gemo/Projects/XIME-P
(thanks to our Web master Andrei Arion)
XIME-P 2004 contents also appear on the ACM Digital Library and DiSC 2005.