Processing XML in Database Systems

Albrecht Schmidt
CWI Amsterdam, The Netherlands
Albrecht.Schmidt@cwi.nl

Supervisor: Martin Kersten
Overview of Thesis

- Storage of XML Document in the Main-Memory DBMS Monet
- Algebraic Querying of XML Documents
- Nearest-Concept Queries for Ad-Hoc Users
- Query Optimiser Architecture
- XMark Benchmark for XML Processing
Storage of XML Documents

- Storage backend is the home-grown main-memory DBMS Monet.

- Binary storage schema helps to cope with potentially irregular structure of many documents.

- **Structural summary** created and maintained during bulkload; no DTD or schema information is required.

- Summary information used during query processing and provided to users for query formulation
Algebraic Querying of XML Documents

- Idea: extend Monet’s algebra with structural summaries and path expressions (and other helpers).

- Stages of query processing:
  1. Queries are translated to an extended relational algebra.
  2. Query processor rewrites queries using summary information.
  3. Monet’s kernel executes the query.
Nearest-Concept Queries for *Ad-Hoc* Users

- Extension of query algebra with the *meet* operator

- Idea: combine results of, for example, a fulltext search with *lowest common ancestor search* in XML syntax trees.

- Novice users can explore, browse and query a database *without being familiar* with the structure.

- Operator integrates with additional heuristics and can *re-use* existing query engine functionality.
Query Optimiser Architecture

- **CHOOSE operator** to define query equivalences

- Helps to exploit availability of different (equivalent) data sources and query expressions by letting the optimiser make **cost-based decisions**.

- Integrates seamlessly with **existing optimiser architecture**.

- Useful also in other application areas like GIS, data warehousing as well as for **semantic query optimisation** in general.
XMark Benchmark for XML Processing

- Database modelled after an Internet auction site with items, customers, auctions, annotations, emails, etc.

- Tries to identify, abstract and challenge query primitives in 20 queries.

- Provides help to assess existing technology, to find bottlenecks and to evaluate new ideas in an XML context.

- Tools are made available to the public on the project Web site at http://www.xml-benchmark.org.