Blended Browsing and Querying of XML in a Lazy Mediator System

Kevin D. Munroe, Bertram Ludäscher, Yannis Papakonstantinou

{kmunroe,yannis}@cs.ucsd.edu , ludaesch@sdsc.edu

1. Background and Overview

Research in semistructured data and XML has focused on query languages, data extraction, and mediator systems for integrating heterogeneous sources. However, the issue of user interfaces for browsing and querying semistructured data has largely been ignored despite the fact that it is of primary importance as XML becomes more widely used as a data exchange format and information model on the Web.

To facilitate XML query construction and browsing of results by the non-expert user, the DTD-based graphical user interface *BBQ* (*B*lended *B*rowsing and *Q*uerying) has been developed as a front end to XML sources. It is currently used as a front-end to the virtual sources exported by the MIX mediator system [MIX,LPV00]. A virtual source may be an actual XML source or an XML view constructed by the mediator. BBQ supports intuitive *querying and browsing*, *seamless iterative query refinement*, and *structure discovery*. Data sources are displayed in a multi-document interface layout, where each source is assigned a window, with its data and schema shown side-by-side. Both the data and the schema are displayed as directory-like tree structures, which users can navigate and place conditions on.

BBQ combines the query paradigm used in relational databases with Web searching. As with relational database searches, querying in BBQ is *schema-driven* (using XML DTDs); but like Web searches the user may not know in advance into which query results to "drill down". Hence the BBQ system emphasizes navigation of schemas and data, and the iterative refinement of queries. The latter is achieved by having query results become sources for subsequent queries. Users can construct a query result document (a *virtual view* in MIX's case) and that document becomes a first-class data source within BBQ that can be browsed, queried, or used to construct further query result documents.[1] We facilitate the perception of a query result as a source by automatically inferring its DTD from the query statement and the source DTDs [PV99]. Furthermore, once the data for the query result are obtained, a *structure discovery module* can suggest an approximate structure for those elements in the source DTDs that were loosely specified (i.e., using the keyword ANY).

BBQ employs the MIX mediator's *lazy* implementation of the DOM client API, which provides an efficient platform for virtual, "on-demand" XML views. Since this architecture for *V*irtual *X*ML *D*ocuments (DOM-VXD) [LPV00] does not actually retrieve source objects until they are required by the user's navigation into the virtual view, browsing of potentially very large query results and subsequent query refinement become feasible.

2. BBQ Features and Demo Outline

BBQ provides a uniform directory tree-based user interface to both XML data (documents) and XML schema (DTDs). A BBQ *query session* can involve one or more *query cycles*. A query cycle involves the following steps: First, an XML query is composed[2] by attaching constraints to the relevant source DTD(s)

[1] The idea of combined query and browsing is borrowed from the "Query-in-Place" paradigm of PESTO [CHM+96].

[2] The XML query language of the MIX Project is called *XMAS* (for *X*ML *Ma*tching *A*nd *S*tructuring).

and combining information across several sources (e.g., by joining attributes from different source DTDs). From these user interactions, BBQ can construct the XMAS *query body*. The XMAS *query head* defines the structure of the XML answer document, based on the pieces extracted by the body. To this end, the user can specify the structure of the XML answer tree by (i) dragging and dropping XML elements from the source DTDs, (ii) introducing new elements, and/or (iii) *grouping* of elements according to values of other elements. Using a DTD inference component, the DTD of the result document is inferred. When execution of the query starts at the mediator, the incoming (partial) results are browsed by the user. This concludes a query cycle and the new (virtual) result document can be incorporated in the query session as a regular source. Further features of BBQ include the following:

- The representation of DTDs is faithful in the sense, that BBQ's tree-notation exactly mirrors the structure of the corresponding DTD (whereas, e.g., the first implementation of BBQ [BGL+99] preserved only the parent-child relation, but did not include precise representations of sequences (",") or alternatives ("|")).
- Whether or not order among siblings is relevant for the query can be specified.
- Multiple DTDs and XML documents can be incorporated simultaneously, and multiple query cycles can be handled in a single query session.

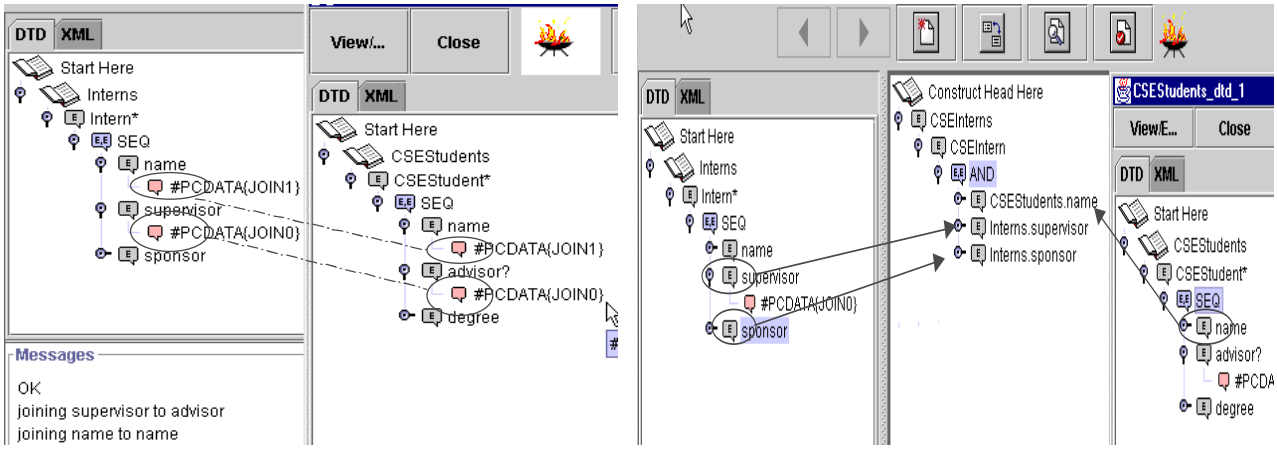

Fig. 1: Defining a join across data sources (left), and construction of the query head using drag-n-drop (right).

References

[BGL+99] C. Baru, A. Gupta, B. Ludäscher, R. Marciano, Y. Papakonstantinou, P. Velikhov, V. Chu. XML-Based Information Mediation with MIX. *ACM SIGMOD 1999, exhibition program*, pp. 597-599.

[CHM+96] M. Carey, L. Haas, V. Maganty, J. Williams. PESTO: An Integrated Query/Browser for Object Databases. *VLDB 1996*, pp. 203-214.

[LPV00] B. Ludäscher, Y. Papakonstantinou, P. Velikhov. Navigation-Driven Evaluation of Virtual Mediated Views, *EDBT 2000*.

[MIX] Mediation of Information Using XML, http://www.npaci.edu/DICE/MIX, http://www.db.ucsd.edu/Projects/MIX

[PV99] Y. Papakonstantinou, P. Velikhov, Enhancing Semistructured Data Mediators with Document Type Definitions, *ICDE 1999*.