NeT & CoT: Inferring XML Schemas from Relational World

Dongwon Lee UCLA / CSD dongwon@cs.ucla.edu

Murali Mani * UCLA / CSD mani@cs.ucla.edu

Frank Chiu UCLA / CSD fchiu@ucla.edu

Wesley W. Chu UCLA / CSD wwc@cs.ucla.edu

With XML emerging as *the* data format of the Internet era, there is a substantial increase in the amount of data encoded in XML. However, the majority of everyday data is still stored and maintained in relational databases. Therefore, we expect the needs to convert such relational data into XML documents will grow substantially as well.

Two conversion algorithms, called NeT and CoT, to translate relational schemas to XML schemas using various semantic constraints are presented. We first present a language-independent formalism named *XSchema* so that our algorithms are able to generate output schema in various XML schema language proposals. The benefits of such formalism are that it is both precise and concise. Based on *XSchema* formalism, our proposed algorithms have the following characteristics: (1) NeT derives a nested structure from a flat relational model by repeatedly applying the *nest* operator so that the resulting XML schema becomes hierarchical, and (2) CoT considers not only the structure of relational schemas, but also inclusion dependencies during the translation so that relational schemas where multiple tables are interconnected through inclusion dependencies can be handled as well. Overview of our approach is illustrated in Figure 1.

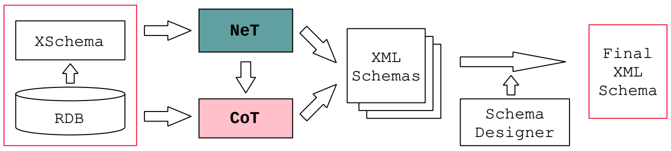

Figure 1. Overview of our approach.

NeT: In [1], we proposed the NeT, an improvement over a straightforward relational to XML *flat* translation (FT). Since FT maps the flat relational model to the flat XML model in a one-to-one manner, it does not utilize the hierarchical nature of XML model. Our idea was to find a more intuitive element content model that uses the "*" or "+" using the *nest* operator. The details are omitted for brevity.

CoT: Although NeT infers hidden characteristics of data by nesting, it is only applicable to a single table at a time. Therefore, it is unable to capture a correct "big picture" of relational schema where many tables are interconnected. To remedy this problem, we present the second proposal called *Constraints-based Translation* (CoT); CoT considers inclusion dependencies during the translation. Such constraints can be acquired from database through ODBC/JDBC interface or provided by human experts who are familiar with the semantics of the relational schema being translated. CoT is capable of generating a more intuitive XML schema than what NeT or other tools compared in [1] generate.

CoT is essentially trying to merge multiple interconnected tables into a coherent and hierarchical parent-child structure. The intuitive idea to merge two tables with a foreign key is to (1) identify a *parent* table that captures the core meaning of the schema and a *child* table that captures the auxiliary information, and (2) embed the child table into the parent table as a *sub-element* in the content model. Furthermore, when a table s is related with more than one tables, say t_1 and t_2, via inclusion dependencies, naive mapping would exhibit the same phenomenon known as *update anomaly* in database theory. To avoid this anomaly, the redundant foreign key is captured through a *reference* (e.g., `IDREF`) instead of parent-child relationship. For the general case where a set of tables are interconnected through a set of complicated inclusion dependencies, we capture the relationships by the notion of *IND-Graph* and use a BFS-based scan algorithm for accurate schema conversion. For details of algorithms, a proof of concept, and extensive experimentations, refer to [2].

*Partially supported by the NSF grants - 0086116, 0085773, 9817773.

References

[1] D. Lee, M. Mani, F. Chiu, and W. W. Chu. "Nesting-based Relational-to-XML Schema Translation". In *Int'l Workshop on the Web and Databases (WebDB)*, Santa Barbara, CA, May 2001.

[2] D. Lee, M. Mani, F. Chiu, and W. W. Chu. "Translating Relational Schemas to XML Schemas using Semantic Constraints". 2001. Subitted for publication.