On Database Theory and XML

Dan Suciu
University of Washington

Abstract

Over the years, the connection between database theory and database practice has weakened. We argue here that the new challenges posed by XML and its applications are strengthening this connection today. We illustrate three examples of theoretical problems arising from XML applications, based on our own research.

1 On Database Theory

The field of relational databases is the product of a theoretician, E.F. Codd, from the early 70s. Relational databases had to struggle for a while against the industry proposal CODASYL [58, but then became universally adopted and today we have both a strong industry and a flourishing research field. The end of 70's and early 80's were golden years for database theoreticians. Theory had two principal threads [61], relational database theory (dependency theory, universal relation theory, acyclic hypergraph theory), and transaction processing. Many contributions of this period influenced database systems and the industry. Papers from that time analyzing the state of theoretical database research and its relationship to database systems concluded that the field was in a healthy state [61, 52]. Over time, however, theoretical research became less connected to database systems. New fields, like deductive databases (recursive queries) [60, 7, 8], nested relations [1, 32, 51] theory of object-oriented query languages [35, 36, 2, 3, 4], ended up having little or no influence on the industry, despite their practical motivation and excellent results. Other areas, especially in the 90's, where directly motivated by theoretical questions and were totally ignored by practitioners: query languages and complexity classes [63, 30, 31, 39], topics in finite model theory [57, 10], topological and constraint databases [34, 47, 45].

In 1995 Christos Papadimitriou wrote a thought provoking essay entitled Database Metatheory: Asking the Big Queries [46], analyzing the state of database theory at that time. Two of its ideas are especially relevant to our discussion. The first is that theory in Computer Science is inevitable. Unlike traditional natural sciences which study an objective world, Computer Science is a science of the artificial, studying artifacts that only exists as a result of scientific activity. The second is to adapt Kuhn's paradigm principle for natural sciences to Computer Science. Under this principle, a science evolves in a cycle consisting of three states: from normal science to crisis, to revolution, then back to normal science. In Kuhn's theory the crisis occurs when new observations and measurements about the objective reality fail to agree with the science's accepted paradigm; during a revolution a new paradigm is developed, consistent with the new observations. Clearly, this definition of a crisis does not apply to a field of Computer Science, since there is no real world to observe.

Instead, Papadimitriou suggest another definition. Consider the graph of all research units1 in a given field. An edge from x to y represents the fact that x influenced y. Normal science is defined by a graph which is multiply connected (i.e. remains connected even if several nodes are removed); a science in crisis is defined by one which is almost disconnected, with two major connected components corresponding to theoretical and applied projects respectively. This definition is interesting for two reasons. First, it should be possible to automatically monitor the healthiness of our research field, e.g. by using the DBLP bibliography database [41]. Second, it reinforces our general belief that our science is healthy if theory and practice are strongly connected: the general feeling in the mid 90's was that this connection was weakening.

What caused this rift in a field which started with theoreticians and practitioners working so closely ? While there are probably a large number of reasons that we don’t claim to know, there is one in particular that is of importance to our discussion: the tension between the theoreticians’ quest to explore new avenues, branching out of Codd’s pure relational framework, and the practitioners’ need to improve systems in the existing framework (relational/SQL) and for existing applications (client/server). Theoreticians explore new query languages (e.g. with recursion [60]), new data models [51], new usages of data (e.g. incomplete information [29]), or entirely new frameworks (e.g. constraint databases [34]). Practitioners improve the execution of relational operators [24] or improve relational optimizers [33, 25, 26, 1]

1i.e. projects, papers, group – Papadimitriou leaves purposely the term undefined.
in academic and industrial research labs before being adopted by the industry or standards committees. While more academics are now getting involved in the W3C working groups, standards and concepts are sometimes created faster than research communities can validate them.

3 DB Theory and XML

We see a double role of theory in the Web age. One is a long term of conceptualization and rationalization which can lead to improvements in existing standards. This is a traditional role, and it will probably take years; perhaps an early example in this category is the recent work on keys for XML by Buneman et al. [11]. The second is a more short term: to answer critical technical questions that may either help the working groups or the implementors. We discuss here three particular problems: XML publishing, XML typechecking, and XML storage. Their particular choice is influenced by our own previous work in [21, 20, 43, 19], and is not intended to be an exhaustive, or even representative list of XML research problems.

XML Publishing XML publishing is the problem of transforming existing, relational data into XML. Conceptually, this is the same as defining an XML view over the relational data, but this view is considerably more complex than relational views. The relational data is normalized: flat and fragmented into many relations. Often the relational schema is also proprietary, since it represents the company’s internal organization (relation names may correspond to business units) or to the company’s policy (whether price is an attribute of Product or of the Product - Customer relationship). By contrast, XML data is unnormalized: nested and monolithic. Moreover, its raison d'être is to have a public schema, shared across a community, or, at least between a few partners; for example www.biztalk.org lists about 458 public XML schemas.

Consider the following relational schema:

| Product(pid, pname, price) |
| Customer(cid, cname, address) |
| Orders(oid, pid, cid, date) |
| Complaints(oid, text) |

Suppose we want to export it into an XML document with the following DTD:

```xml
<!ELEMENT products (product)>
<!ELEMENT product (productID, name, price, order*, complaint*)>
<!ELEMENT order (customerId, date, 
    customerName, customerAddress)>
<!ELEMENT complaint (customerId, text)>
```
CREATE VIEW product AS /* <product> nodes */
SELECT DISTINCT product1.pid, product1.name, product1.price
FROM Product product1

CREATE VIEW order AS /* <order> nodes */
SELECT DISTINCT product2.pid, orders2.date,
customer2.cid, customer2.name, customer2.address
FROM Product product2, Orders orders2,
Customer customer2
WHERE product2.pid = orders2.pid AND
orders2.cid = customer2.cid

CREATE VIEW complaint AS /* <complaint> nodes */
SELECT DISTINCT product3.pid,
customer3.cid, product3.pid
FROM Product product3, Orders orders3,
Customer customer3, Complaint complaint3
WHERE product3.pid = orders3.pid AND
orders3.cid = customer3.cid AND
order3.oid = complaint3.oid AND
order2.date = "1999"

SELECT DISTINCT product1.pid, complaint3.text
FROM Product product1, Product product2,
Orders orders2, Customer customer2
WHERE product1.pid = product2.pid AND
product2.pid = orders2.pid AND
orders2.cid = customer2.cid AND
order3.oid = complaint3.oid AND
order2.date = "1999"

SELECT DISTINCT product1.pid, complaint3.text
FROM Product product1, Orders orders2,
Customer customer2, Complaint complaint3
WHERE product1.pid = product2.pid AND
product2.pid = orders2.pid AND
orders2.cid = customer2.cid AND
order2.oid = complaint3.oid AND
order2.date = "1999"

The problem becomes clear now: this query performs more joins than necessary. In other words, this query is not minimal. The reason is that we generated this query automatically, by combining join conditions from three different relational views, and they happened to share some common subexpressions. Some redundant joins can be easily eliminated, for example it follows from the viewtree structure that the tuple variables product2 and product3 are redundant, but others may require a general-purpose minimization algorithm. We may also need to take into account constraints in the databases. In our example, if we assume that pid, cid is a key in Orders, the query can simplified to:

SELECT DISTINCT product1.pid, complaint3.text
FROM Product product1, Orders orders2,
Customer customer2, Complaint complaint3
WHERE product1.pid = product2.pid AND
orders2.cid = customer2.cid AND
order2.oid = complaint3.oid AND
order2.date = "1999"

This phenomenon is not new: when views are unfolded in SQL queries they can also introduce repeated subexpressions (hence, non-minimal queries). However, joins across multiple views are rare in the relational world, so query minimization is not critical.

3Materialized XML views are discussed in [54, 20].
4SilkRoute uses XML-QL rather than XQuery.
By contrast, in XML publishing every node type is defined by a different relational view, and we should expect queries to frequently join many of them: here minimization, either in the engine or in some middleware, is a necessity.

While query containment (and, hence, minimization) is NP complete for conjunctive queries, recent advances in theoretical database research have created powerful tools that could lead to practical minimization algorithms. Chekuri and Rajaraman [16] have shown that containment can be checked efficiently if the queries have a small query width. Kolaitis and Vardi [38] established additional relationships to tree-widths, first order logic with bounded variables, and the constraint satisfaction problem (see also [64]).

XML Typechecking Although XML can be schema-less, most XML instances will probably be associated to some form of schema, either a DTD or an XML-Schema [9]. Given a program that generates an XML document, typechecking is the problem of deciding whether the generated XML document always conforms to a given output DTD (or XML-Schema); this should not be confused with validation, which checks conformance of a given XML document to a DTD (or XML-Schema). It is possible to do typechecking dynamically, by validating the generated XML document at runtime, but this creates the possibility of runtime errors, and slows down an application. Static typechecking is more desirable, but also more difficult, since it requires a thorough analysis of the program generating the XML document.

Static XML typechecking is undecidable if an arbitrary C++ or Java program is generating the XML document, due to Rice’s theorem. The more interesting question is whether typechecking is possible for more restricted languages, like XSLT [17] or XQuery [13], since in all likelihood, most dynamically generated XML documents will be created by scripts/queries written in these languages. In the following discussion we assume the XML types to be given by DTDs: everything carries over to XML-Schemas too. Given two DTDs $\tau_1$, $\tau_2$, and given an XML transformation $f: \text{XML} \rightarrow \text{XML}$ expressed in some query language, the typechecking problem asks whether for every document $D$ conforming to $\tau_1$ (in notation: $D \in \tau_1$), $f(D)$ conforms to $\tau_2$:

$$\forall D \in \tau_1, f(D) \in \tau_2$$

All systems that offer static typechecking today, including the XQuery Algebra [13], base their typechecking algorithm on type inference, which we formalize here as follows: given an input DTD $\tau_1$ and transformation $f$, compute the “output DTD”

$$f(\tau_1) \overset{\text{def}}{=} \{ f(D) \mid D \in \tau_1 \}$$

If we knew how to do type inference, then type checking is easy: given $\tau_1$, $\tau_2$, and $f$, first infer the output type (i.e. compute $f(\tau_1)$), then check whether $f(\tau_1) \subseteq \tau_2$, which is possible based on the fact that containment of two regular expressions is decidable\(^2\).

This is the point where we need some input from theory. It turns out that type inference is not possible, not even for the simplest subsets of query languages. Systems that do type inference compute an approximate output DTD that will lead to false negatives in the typechecking procedure. To see such an example, consider the input DTD $\tau_1$:

$$\tau_1 = < \text{!ELEMENT root (elm*)} >$$

The query below iterates over the input document three times, producing first an $<$el$>$ element for each <elm>, then a $<$b$>$ element, then a $<$c$>$ element. We give $f$ below, expressed in XQuery:

$$\begin{align*}
    f &= <
    \text{result}>
\quad \text{FOR } x \text{ IN } "\text{doc.xml"/@root/elm RETURN }<a> x/\text{text}() </a>
\quad \text{FOR } x \text{ IN } "\text{doc.xml"/@root/elm RETURN }<b> x/\text{text}() </b>
\quad \text{FOR } x \text{ IN } "\text{doc.xml"/@root/elm RETURN }<c> x/\text{text}() </c>
    \</\text{result}>
\end{align*}$$

Thus, for the input XML document $<$root$<$elm$/> <elm/> ...$<elm/> $</root$>, in short elf\(^a\), the query will return the output document $a^*b^*c^*$. Today’s systems will infer the following output DTD for this query:

$$\tau_2 = < \text{!ELEMENT result (a*,b*,c*)} >$$

While this seems reasonable, it is obviously not the correct one, which, according to the definition should be:

$$f(\tau_1) = < \text{!ELEMENT result ((a*,b*,c*) n \geq 0)} >$$

Obviously, this is not a real DTD, and perhaps not very interesting in practice, and one may wonder if one should accept, for all practical purposes, the inferred output DTD $a^*b^*c^*$. The problem, however, is that with this inferred DTD one may fail to correctly typecheck. To see an example, consider the following output DTD:

$$\tau_2 = < \text{!ELEMENT result (((a,a)*(b,b)*(c,c)*)) | ((a,a)*,a,b,b*,(c,c)*,c*)} >$$

It says that there are either an even number of $a$’s, b’s, and c’s, or an odd number of $a$’s, b’s, and c’s.

\(^2\)In addition, both DTDs and XML-Schema restrict the regular expressions such that their associated automaton is deterministic. If $\tau_2$ has this property, then one can check $f(\tau_1) \subseteq \tau_2$ in PTIME.
Clearly $f$ typechecks with respect to $\tau_1$ and $\tau_2$, but, using the “inferred” type $\tau$, we have $\tau \not\subseteq \tau_2$, hence the typechecking procedure based on type inference will fail.

This is a serious limitation of the type inference approach. If users want to generate XML documents that conform to the type $\tau_2$ and correctly write the program $f$ for that purpose, the system will reject it claiming (incorrectly) that it does not typecheck. This limitation needs to be investigated by theoretical research.

One may argue that $\tau_2$ is not a “practical” output type. But this begs for a definition of “practical” DTDs. Even if we had one, the question remains whether type inference is complete for that restricted set of DTDs.

Another question is whether typechecking is possible without using type inference. One answer, given in [43], is a surprising “yes”, for a large class of XML transformations that include recursive traversal of the XML tree, e.g. like in XSLT. However, these transformations do not allow us to do joins: this may be acceptable for XSLT, but certainly such a result is useless for query languages like XQuery, or extensions of SQL with XML publishing features. When one adds joins, typechecking becomes undecidable, even if one attempts to restrict DTD’s in various ways to “practical” DTDs [6, 5].

Today the most promising approach to typechecking remains that based on type inference. The XDuce language [27, 28] defines a type inference system for a functional language with recursion; the XQuery algebra defines a type inference system using XML Schema as its type system. Since we know that this approach cannot be as robust as typechecking in general-purpose programming languages, a study of its applicability and limitations is needed.

**XML Storage** XML data is a labeled tree; a relation is a table. The problem of storing XML data in one or several tables, suggested in Fig. 2, is a challenging one, both for theoreticians and practitioners. Since the tree is meant to describe some irregular structure while tables are by definition regular, we are attempting to store some irregular data into a regular data type. In addition to the pure combinatorial aspect, there is a logical aspect to the storage problem: given a storage mapping, one needs to be able to translate queries formulated over the XML data into relational queries formulated over the relational storage. The combination of combinatorics and logic make the problem particularly appealing.

Several approaches have been tried so far. The simplest is to store XML as a graph, in a ternary relation (two columns for the edges, the third for the labels and/or data values). This approach is explored by Florescu and Kossman in [22]. The price one pays for its simplicity is that many self-joins of the edge table are required in order to reconstruct a given XML element: one join for each subelement. Shanmugasundaram et al. [55] use the DTD (or XML-Schema) to derive a relational schema. One table is created for each element type that can occur in a collection position. This technique works well in practice whenever one has a schema for the XML document. A subtle problem is that the resulting storage is very sensitive to that schema. For example if the content of <person> changes from (name, phone) to (name, phone*) then we need to move all phone numbers to a separate table, although perhaps the XML document has changed very little.

The case when the XML document has no schema, or when the schema changes frequently is harder, and has a more dramatic impact on performance. An approach is proposed in [19], which uses data mining on the XML instance to infer a relational schema. The idea is to find regularities that may exists in a given XML data instance, and to organize the storage based on those regularities. Another approach is proposed in [18]. Here all data values of the XML document are stored together and a smart index allows efficient access based on the path expression and, possibly, based on the data value.

The challenge in any storage schema is that it has to be flexible enough to accommodate any XML data, yet it has to be as efficient as regular data storage when the XML data happens to be regular. Finding the largest regular subset in an irregular data instance is a problem which can be formulated and addressed theoretically. A related problem is that of quantifying the degree of irregularity in a semistructured data instance. For instance, in the simple case of flat XML data, elements can be viewed as records with variable fields; a boolean matrix $S$ describes the structure completely, with $S_{ij}$ equal 1 when element $i$ contains field $j$, and 0 otherwise. It is easy to see that if the data instance can be stored perfectly in $k$ separate tables then the rank of the matrix is $k$. Here $k$ is a measure of semistructuredness, since the larger it is, the more irregular the data. Finding a general definition of semistructuredness remains a topic for future research.

## 4 Conclusions

We have described three XML research problems, inspired from our own work. XML’s semistructured data model represents paradigm shift for theoretical
database research. It is not the first one: for example the object-oriented data model can also be considered a paradigm shift, which generated a vast amount of theoretical and applied research. This time, however, the shift comes from outside the community (XML) and this, at least, settles easily the question of applicability. It offers us both a chance both to apply research on old topics (query containment) and to conduct research on new topics (typechecking).

Acknowledgment I would like to thank Gerome Miklau for his comments.

References


