Semantic Integration of Schema Conforming XML Data Sources

Dimitri Theodoratos[1], Theodore Dalamagas[2], and I-Ting Liu[1]

[1] Dept. of CS, NJIT
{dth,il2}@njit.edu
[2] School of EE and CE, NTUA
dalamag@dblab.ece.ntua.gr

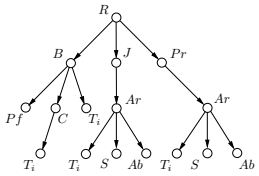
A challenging problem in Web engineering is the integration of XML data sources. Even if these data sources conform to schemas, they may have their schemas and the correspongind XML documents structured differently.

In this paper, we address the problem of integrating XML data sources (a) by adding semantic information to document schemas, and (b) by using a query language that allows a *partial specification* of tree patterns. The semantic information allows the grouping of elements into the so called *schema dimensions*.

Our approach allows querying data sources with different schemas in an integrated way. Users posing queries have the flexibility to specify structural constraints fully, partially or not at all. Our approach was initially developed for arbitrarily structured data sources [1]. Here, we show how this approach can be applied to tree-structured data sources that comply to schemas.

We consider XML data sources that conform to DTDs. The DTD of an XML data source can be represented as a *schema tree* whose nodes are labeled by the elements of the DTD and the edges denote element-subelement relationships. An example of a schema tree S_1 for a DTD is shown in Figure 2. Abbreviations for the names of the elements are shown in Figure 1.

Book	B
JournalIssue	J
Proceedings	Pr
Chapter	C
Article	Ar
Title	Ti
Preface	Pf
Abstract	Ab
Section	S
Publication	Pu
Medium	Me
Unit	U
Journal Abstr.	JAb
Proc. Abstr.	PAb



$\{R\} \rightarrow R$
$\{B,\ J,\ Pr\} \rightarrow Me$
$\{Pf,\ C,\ A\} \rightarrow U$
$\{Ti\} \rightarrow Ti$
$\{S\} \rightarrow S$
$\{R/J/A/Ab\} \rightarrow JAb$
$\{R/Pr/A/Ab\} \rightarrow PAb$

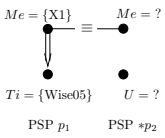


Fig. 1. Abbrv. **Fig. 2.** Schema tree $\mathcal{S}_1$ **Fig. 3.** Dim. set $\mathcal{D}_1$ **Fig. 4.** Query

The nodes of a schema tree may share common features. This semantic information allows the partitioning of the nodes into sets which are called *schema dimensions*. The semantic interpretation of the nodes is provided by the user. Figure 3 shows a partition of the nodes of $\mathcal{S}_1$ of Figure 2 into a dimension set $\mathcal{D}_1$. The name of a set of nodes follows the $\rightarrow$.

Queries are formed on dimension sets. A query on a dimension set provides a (possibly partial) specification of a tree pattern (PSTP). This query pattern is formed by a set of graphs of annotated schema dimensions called *partially specified paths (PSPs)*. The edges of PSPs are ancestor ($\Rightarrow$) or child ($\rightarrow$) relationships. A query also involves *node sharing expressions* denoted by an edge labeled by $\equiv$. This tree pattern is to be matched against an XML document of a data source. A node sharing expression forces nodes from different PSPs to be shared. Figure 4 shows an example of a query on $\mathcal{D}_1$.

We introduce *schema dimension graphs* (a) to guide the user in formulating queries, (b) to check queries for satisfiability, and (c) to support the evaluation of queries. Schema dimension graphs abstract the structural information of a schema tree based on the semantic equivalences of nodes defined by the dimension set.

A PSTP query can be computed by identifying a set of (completely specified) tree-pattern queries. Tree-pattern queries are constructed using schema dimension graphs. They are used to generate XPath expressions to be evaluated on the XML documents in the data sources.

In our data integration approach, we assume that a global set of schema dimension names is fixed and used among the different data sources to be integrated. Each data source creates its own *local* schema dimension graph. In addition, a *global* schema dimension graph is created that merges the dimension graphs of the different data sources. The global dimension graph is used to guide user query formulation and to identify queries that are globally unsatisfiable (that is, unsatisfiable with respect to the global schema dimension graph). A query issued against an integrated system is specified on the global set of schema dimensions. Its answer is the collection of its answers on every local data source. Figure 5 shows the different steps of the evaluation of a query in an integration system.

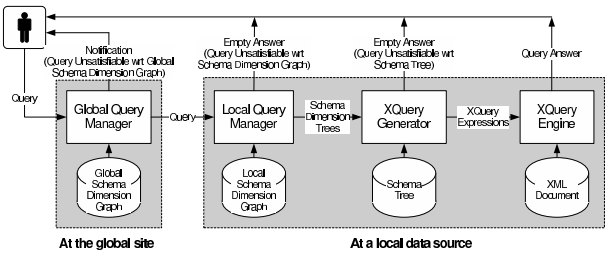


Fig. 5. Evaluation of queries in an integration system

References

1. D. Theodoratos, T. Dalamagas, A. Koufopoulos and N. Gehani. Semantic Querying of Tree-Structured Data Sources Using Partially Specified Tree Patterns. In *Proc. of the 14th Conference on Information and Knowledge Management (CIKM'05)*, 31st Oct - 5th Nov, Bremen, Germany, 2005.