Data Integration: XML

Jan Chomicki

University at Buffalo and Warsaw University

April 5, 2007

XML documents (simplified)

XML tree [KSS03]

- finite, ordered, unranked tree
- element, attribute and text nodes
- element and attribute node labels from a finite label alphabet Σ
- attribute and text (PCDATA) values from an infinite domain D
- only element nodes have children
- document order (left-to-right prefix order)

XML trees represent well-formed documents:

- matching, properly nested opening and closing tags
- single root element

Regular expressions over Σ

$$E := \varepsilon \mid a \mid E \cup E \mid E\ E \mid E^*$$

where $a \in \Sigma$.

Defining valid XML documents

XML schema definitions

- Document Type Definitions (DTDs)
- specialized DTDs
- XML Schema
- ...

DTD (over Σ)

- element-only content: a function mapping node labels from Σ to a regular expression to which the concatenated children of the node must conform
- also text-only, mixed, empty, and unrestricted content
- attributes: text-valued (CDATA), enumerations, ID, IDREF

Specialized DTDs

A pair:

- a DTD over a finite set of types
- a function mapping types to elements of Σ

XML Schema

Simple types

- base types (many)
- derived types (by constraining facets)
- list/union types

Complex types

- content model: `sequence`, `all`, `choice`
- attribute declarations
- types can be recursive or anonymous
- element types can be locally declared

Integrity constraints

- keys
- foreign keys

Tree automata [Nev02]

Nondeterministic tree automaton (NTA)

A tuple $B = (Q, \Sigma, \delta, F)$:

- Q: a finite set of states
- $F \subseteq Q$: the set of final states
- $\delta : Q \times \Sigma \to 2^{Q^*}$, where Q^* is the set of finite words over Q and $\delta(q, a)$ is a regular string language over Q for every $q \in Q$ and $a \in \Sigma$.

Run of B over a tree t

Labeling $\lambda : Dom(t) \to Q$ such that for every $v \in Dom(t)$ with n children $v_1, \ldots v_n$, $\lambda(v_1) \cdots \lambda(v_n) \in \delta(\lambda(v), label^t(v))$. A run is accepting if $\lambda(root) \in F$.

Expressive power

DTDs (w/o attributes and text) and specialized DTDs describe tree languages recognized by NTAs.

Logic

Relational vocabulary $\tau_\Sigma = (E, <, (O_a)_{a \in \Sigma})$:

- E: the parent relation
- $<$: ordering of node's children
- (O_a) unary relations

Monadic second-order logic (MSO)

First-order logic extended with quantification over sets.

Theorem (Doner, Thatcher, Wright)

A set of trees L is recognized by an NTA iff there is an MSO formula φ such that $L = \{t \mid t \models \varphi\}$.


N. Klarlund, T. Schwentick, and D. Suciu.
XML: Model, Schemas, Types, Logics, and Queries.
In J. Chomicki, R. van der Meyden, and G. Saake, editors, *Logics for Emerging Applications of Databases*, pages 1–41. Springer-Verlag, 2003.


F. Neven.
Automata, Logic, and XML.
In *Workshop on Computer Science Logic (CSL)*, pages 2–26. Springer, LNCS 2471, 2002.
Invited talk.