Formal Language Foundations and Schema Languages

Stefan Tittel
University of Dortmund

Seminar: Theoretical Foundations of XML Data Processing, February 2006

XML Languages and Grammars

Introduction and Basics

Definition of

- XML grammars,
- Dyck primes,
- $L_G(X)$,
- contexts,
- $F_a(L)$.

Characterization

Definition of

- traces,
- surfaces.

Theorem 1 *Languages over $A \cup \overline{A}$ are XML languages iff*

- $L \subset D_\alpha$ *for some* $\alpha \in A$,
- $C_L(w) = C_L(w')$ *for all* $a \in A$ *and* $w, w' \in F_a(L)$,
- $S_a(L)$ *is a regular set for all* $a \in A$.

Theorem 2 *XML languages are closed neither under union nor difference.*

More results:

- XML languages are closed under intersection.
- For each XML language L there is exactly one reduced XML grammar generating L if variable names and entities are ignored.
- It is decidable if an XML language L is included in or equal to another XML language M.
- It is decidable if a regular language $L \subset D_A$ is an XML language.
- It is undecidable if a context-free language is an XML language.

One-Unambiguous Regular Languages

Introduction and Basics

Definition of

- unambiguity,
- one-unambiguity,
- the marking of regular expressions,
- first, last, and follow.

Theorem 3 *A regular expression E is one-unambiguous iff*

1. $\forall x, y \in \mathit{first}(E') : x \neq y \Rightarrow x^\natural \neq y^\natural$,
2. $\forall z \in \mathit{sym}(E') \wedge x, y \in \mathit{follow}(E', z) : x \neq y \Rightarrow x^\natural \neq y^\natural$,

where sym(E') is the set of symbols occurring in E'.

Definition of

- Glushkov automata.

Theorem 4 *A regular expression E is one-unambiguous iff G_E is a DFA.*

Recognition

Definition of

- orbits,
- gates,
- the orbit property,
- orbit automata and orbit languages.

Theorem 5 *Let M be a minimal DFA. Iff*

- *M has the orbit property,*
- *all of the orbit languages of M are one-unambiguous,*

then $L(M)$ is one-unambiguous.

Definition of

- trivial orbits,
- M-consistency,
- the S-Cut.

Theorem 6 *Let*

- *M be a minimal DFA,*
- *S be an M-consistent set of symbols,*

now iff

- *M_S satisfies the orbit property,*
- *all of the orbit languages of M_S are one-unambiguous,*

then $L(M)$ is one-unambiguous.

One-unambiguous regular languages are

- closed under derivatives,
- not closed under union, concatenation, star.

Analysis of XML Schema Languages

Introduction and Basics

Definition of

- XML schemas and XML schema languages,
- model groups,
- regular tree grammars,
- the normal form 1 (NF1),
- $contentModel(A)$.

Language Classes

Definition of

- the tree locality constraint and local tree grammars,
- single-type constraint languages.

These two language classes are

- closed under intersection,
- not closed under union and difference.

Theorem 7 *Local tree languages are proper subclasses of single-type constraint languages.*

Evaluating XML Schema Languages

DTD

- TDLL(1),
- local tree grammar.

DSD

- No constraints on the production rules,
- theoretically any regular tree grammar,
- practically not due to parser construction,
- TDLL(1) is suspected.

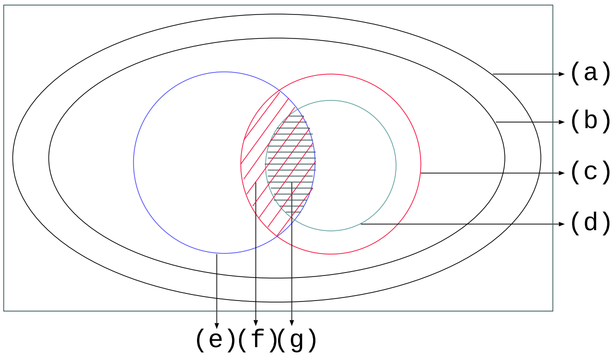
XML Schema

- TDLL(1) with single-type constraint,
- context-free content models are possible.

RELAX

- Any regular tree grammar.

This figure is from [3].



(a)	regular tree grammars (*RELAX*, *XDuce*)
(b)	TD(1) grammars
(c)	single-type constraint grammars
(d)	local tree grammars
(e)	TDLL(1) grammars (*DSD?*)
(f)	TDLL(1) w/ single-type constraint (*XML Schema*)
(g)	TDLL(1) w/ tree-locality constraint (*DTD*)

References

[1] Jean Berstel, Luc Boasson. *Formal Properties of XML Grammars and Languages.* Acta Informatica, 38:649–671, 2002.

[2] Anne Brüggemann-Klein, Derick Wood. *One-Unambiguous Regular Languages.* Information and Computation, 140:229–253, 1998.

[3] Dongwon Lee, Murali Mani, Makoto Murata. *Reasoning about XML Schema Languages using Formal Language Theory.* Technical Report, IBM Almaden Research Center, 2000. Log #95071.

[4] Thomas Schwentick. *Formal Methods for XML: Algorithms & Complexity.* Internet: ⟨`http://lrb.cs.uni-dortmund.de/~tick/Talks/edbtp.pdf`⟩, 2004 (cited 2006–01–26).

[5] Anders Møller, Michael I. Schwartzbach. *The XML Revolution: Technologies for the future Web.* Internet: ⟨`http://www.brics.dk/~amoeller/XML/`⟩, 2003 (cited 2006–01–26).

[6] Dongwon Lee, Murali Mani, Makoto Murata. *Taxonomy of XML Schema Languages Using Formal Language Theory.* Proceedings of the 2001 Conference on Extreme Markup Languages, 2001.