Applications of Automata in XML Processing

Christoph Koch

Cornell University
Ithaca, NY, USA
koch@cs.cornell.edu

XML is at once a document format and a semistructed data model, and has become a de-facto standard for exchanging data on the Internet. XML documents can alternatively be viewed as labeled trees, and tree automata are natural mechanisms for a wide range of processing tasks on XML documents. In this talk, I survey applications of automata in XML processing with an emphasis on those directions of work that so far have had the greatest practical impact. The talk will consist of three parts. In the first, I will discuss XML validation. The standard schema formalisms for XML, Document Type Definitions and XML Schema, are regular tree grammars at their core. These official standards of the World Wide Web Consortium are well-founded in automata theory and formal language theory, and are designed to incorporate special restrictions to facilitate the creation of automata for document validation. The second part will cover XML stream processing techniques and XML publish-subscribe systems, an area in which a number of exciting automata-based systems have been built. The third and final part covers XML query processing using automata, and applications in Web information extraction.

S. Maneth (Ed.): CIAA 2009, LNCS 5642, p. 2, 2009.