Processing XML Documents with Overlapping Hierarchies

Ionut E. Iacob*
Department of Computer Science
University of Kentucky
Lexington, KY, USA
eiaco0@cs.uky.edu

Alex Dekhtyar†
Department of Computer Science
University of Kentucky
Lexington, KY, USA
dekhtyar@cs.uky.edu

ABSTRACT

The problem of overlapping markup hierarchies, first mentioned in the context of SGML, often occurs in XML text encoding applications for humanities. Previous solutions to the problem rely on manual maintenance of the markup and address only the problem of representing overlapping features in XML, leaving the issues of automated maintenance and querying open. As a consequence, traditional XML tools are of little practical use when dealing with overlapping markup. In this work we demonstrate the implementation of our framework for management of concurrent XML hierarchies from a computer science perspective. We propose an underlying model, data structures, APIs, and algorithms so that the most of the burden of managing concurrent XML hierarchies would be born by the software.

Categories and Subject Descriptors

E.1 [**Data structures**]: Trees, Graphs; H.3.3 [**Information Search and Retrieval**]: Query formulation, Search proces

General Terms

Algorithms, Management

Keywords

XML, overlapping hierarchies, querying

1. DEMONSTRATION OVERVIEW

This work attempts to bridge the gap between the apparent necessity for concurrent markup and the lack of software support for it by proposing a framework for the creation, storage, maintenance, transforming, and querying the concurrent XML markup.

The following problems are addressed in our framework:

• *Data model for concurrent document-centric XML.* We use the GODDAG data structure [5], a generalization of DOM trees for XML, to represent multihierarchical XML documents. Our demo includes a DOM-like API for the GODDAG data structure and a GODDAG parser from a variety of XML representations of multihierarchical documents [3].

• *Querying concurrent XML.* XPath and XQuery are inefficient in expressing certain important information needs over concurrent XML documents (e.g., requests for overlapping content given two tags). In addition, XPath is defined on the DOM Tree structure, whereas concurrent XML documents are modelled using GODDAG graphs. We redefine the XPath semantics on GODDAG (we call it the Extended XPath), and extend it with features that are specific to processing of concurrent XML [4], such as the overlapping axis. We show an efficient implementation of the Extended XPath and how it can be used to answer meaninful queries in the context of multihierarchical XML.

• *Authoring tools for document-centric XML.* Our software suite includes xTagger, a specialized editor for multihierarchical document-centric XML. xTagger allows users to select a document fragment and choose the approriate markup for it (from any of the XML hierarchies associated with the document). It implements prevalidation checking, which detects encodings that cannot be extended to valid XML with further markup insertions[2].

• *Document manipulation.* One of the advantages of the proposed framework is its flexibility: concurrent XML can be imported into/exported from our software suite from/to a wide range of representations ([1]). We demo the filtering feature for partially viewing and/or exporting a subset of document encodings.

*Work supported, in part, by the NEH grant RZ-20887-02.

†Work supported, in part, by the NSF grant ITR-0219924 and the NSF grant ITR-0325063.

JCDL'05, June 7–11, 2005, Denver, Colorado, USA.

2. REFERENCES

[1] A. Dekhtyar and I. E. Iacob. A Framework for Management of Concurrent XML Markup. *Data and Knowledge Engineering*, 52 (2005), 185 – 208.

[2] I. E. Iacob, A. Dekhtyar, and M. I. Dekhtyar. Checking Potential Validity of XML Documents. In *Proc. of WebDB'04*, 91–96, 2004.

[3] I. E. Iacob, A. Dekhtyar, and K. Kaneko. Parsing Concurrent XML. In *Proc. of WIDM'04*, November 2004.

[4] I. E. Iacob, A. Dekhtyar, and W. Zhao. XPath Extension for Querying Concurrent XML Markup. TR 394-04, U. of Kentucky, Dept. of Computer Science, Feb. 2004.

[5] C. M. Sperberg-McQueen and C. Huitfeldt. GODDAG: A Data Structure for Overlapping Hierarchies. In *Proc. of DDEP/PODDP 2000*, 139–160, Sept. 2000.