Available online at www.sciencedirect.com
SCIENCE DIRECT®
Data & Knowledge Engineering 52 (2005) 231–247
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

On the consistency of XML DTDs

Shiyong Lu *, Yezhou Sun, Mustafa Atay, Farshad Fotouhi

Department of Computer Science, Wayne State University, 430 State Hall, 5143 Cass Avenue, Detroit, MI 48202, USA

Received 21 May 2004; received in revised form 21 May 2004; accepted 21 May 2004
Available online 2 July 2004

Abstract

DTD has been widely used as the schema language for XML documents. A DTD describes the structure of a collection of similar XML documents. The consistency problem of XML DTDs concerns the question that given a DTD *D*, if there exists any finite XML document that conforms to *D*. This issue is important because one wants to know whether a DTD specification is meaningful. In this paper, we formalize the notion of the consistency of DTDs, identify a sufficient and necessary condition for a DTD to be consistent, and propose a linear algorithm, *DTDCon*, for the consistency checking problem.

Keywords: DTD; XML; Consistency; Algorithm

1. Introduction

XML (eXtensible Markup Language [3]) has become the standard for data representation and exchange on the Web. In contrast to HTML, which describes how data should be displayed to humans, XML describes the meanings and structures of data elements themselves, and therefore makes data self-describing and interpretable to software agents. Currently, XML has been used in a wide range of applications as this is facilitated by standard interfaces such as SAX [15] and DOM [1], and by the development of techniques and tools for XML such as XSL (Extensible

* Corresponding author.
E-mail addresses: shiyong@cs.wayne.edu (S. Lu), sunny@cs.wayne.edu (Y. Sun), matay@cs.wayne.edu (M. Atay), fotouh@cs.wayne.edu (F. Fotouhi).

0169-023X/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.datak.2004.05.007

Stylesheet Language) [6], XSLT (XSL Transformation) [4], XPath [5], XLink [13], XPointer [14] and XML parsers. It is well recognized that XML will continue to play an essential role in the development of the Semantic Web [10], the next generation web.

XML data usually comes with a DTD (Document Type Definition [3]) which describes the structure of a collection of similar XML documents. Therefore, the DTD definition language can be considered as a schema language for XML documents. With a DTD, independent groups of people can agree to use a common DTD for interchanging data. In addition, an application can use a standard DTD to verify if the data that the application receives from the outside world is valid or not. The World Wide Web Consortium has defined the grammar for DTDs [3]. Essentially, a DTD defines the constraints on the logical structure of XML documents, and an XML document is valid if it has an associated DTD and if the document complies with the constraints expressed in the DTD. Unfortunately, a syntactically correct DTD might be inconsistent in the sense that there exist no XML documents conforming to the structure imposed by the DTD. Fig. 4 shows some of such inconsistent DTDs. Inconsistent DTDs are of course meaningless, and they should be avoided. In practice, the consistency of small DTDs can be ensured by careful observation based on common sense; tools of checking consistency of DTDs might be desirable for large DTDs or for DTDs that are generated automatically from other data models such as the ER model and the relational model.

This paper is an extended version of [12], where a sufficient and necessary condition for the consistency of DTDs is identified. The condition implies a consistency checking algorithm that is linear for DTDs without *or* operators, but is exponential for DTDs with *or* operators. This paper extends the result presented in [12] with a linear algorithm *DTDCon* for the DTD consistency checking problem (Section 4) and with an experimental study of the performance of *DTDCon* (Section 5).

Organization. The rest of the paper is organized as follows: Section 2 formalizes the notion of the *consistency* of an XML DTD: an XML DTD is consistent if and only if there exists at least one XML document that is valid w.r.t. it. Section 3 identifies a sufficient and necessary condition for an XML DTD to be consistent. Section 4 presents a linear algorithm *DTDCon* for the consistency-checking problem. Section 5 gives the experimental results of evaluating the performance of *DTDCon*. Finally, Section 6 discusses related work, and Section 7 concludes the paper.

2. The consistency of DTDs

XML Document Type Definitions (DTDs) [3] describe the structure of XML documents and are considered as the schemas for XML documents. A DTD example is shown in Fig. 1 for memorandum XML documents.

In this paper, without loss of generality, we model both XML elements and XML attributes as XML elements since XML attributes can be considered as XML elements without further nesting structure. A DTD D is modeled as a set of *XML element definitions* $\{d_1, d_2, \ldots, d_k\}$. Each XML element definition d_i $(i = 1, \ldots, k)$ is in the form of $n_i = e_i$, where n_i is the name of an XML element, and e_i is a *DTD expression*. Each DTD expression is composed from XML element names (called *primitive DTD expressions*) and other DTD subexpressions using the following operators: