Information Sciences 295 (2015) 258–302

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Approximate XML structure validation based on document–grammar tree similarity

Joe Tekli [a], Richard Chbeir [b,*], Agma J.M. Traina [c], Caetano Traina Jr. [c], Renato Fileto [d]

[a] Dept. of Elec. and Compt. Eng., SOE, Lebanese American University (LAU), 36 Byblos, Lebanon
[b] LIUPPA Laboratory, University of Pau and Adour Countries (UPPA), 64200 Anglet, France
[c] ICMC, University of Sao Paulo (USP), 13566-590 São Carlos, SP, Brazil
[d] Federal University of Santa Catarina (UFSC), 88040-900 Florianopolis, SC, Brazil

ARTICLE INFO

Article history:
Received 21 April 2014
Received in revised form 15 September 2014
Accepted 25 September 2014
Available online 12 October 2014

Keywords:
XML
Semi-structured data
XML grammar
Structural similarity
Tree edit distance
Document classification

ABSTRACT

Comparing XML documents with XML grammars, also known as XML document and grammar validation, is useful in various applications such as: XML document classification, document transformation, grammar evolution, XML retrieval, and the selective dissemination of information. While exact (Boolean) XML validation has been extensively investigated in the literature, the more general problem of approximate (similarity-based) XML validation, i.e., document–grammar similarity evaluation, has not yet received strong attention. In this paper, we propose an original method for measuring the structural similarity between an XML document and an XML grammar (DTD or XSD), considering their most common operators that designate constraints on the existence, repeatability and alternativeness of XML elements/attributes (e.g., ?, *, *MinOccurs*, *MaxOccurs*, etc.). Our approach exploits the concept of tree edit distance, introducing a novel edit distance recurrence and dedicated algorithms to effectively compare XML documents and grammar structures, modeled as ordered labeled trees. Our method also inherently performs exact validation by imposing a maximum similarity threshold (minimum edit distance) on the returned results. We implemented a prototype and conducted several experiments on large sets of real and synthetic XML documents and grammars. Results underline our approach's effectiveness in classifying similar documents with respect to predefined grammars, accurately detecting document and/or grammar modifications, and performing document and grammar relevance ranking. Time and space analysis were also conducted.

1. Introduction

The structural and self-describing nature of XML promotes a number of emerging techniques ranging from XML version control, intelligent Web search, and data integration, to message translation and clustering/classification, requiring, in one way or another, some notion of XML structural similarity. In XML similarity-related research, most work has focused on estimating similarity at the XML data layer (comparing XML documents, e.g., [26,33,48]), while quite a few studies have targeted the XML type layer (comparing XML grammars, e.g., [5,28,61]). Nonetheless, few efforts have been dedicated to similarity evaluation in-between the XML data and type (document/grammar) layers.

* Corresponding author. Tel.: +33 559574337; fax: +33 559574308.
E-mail address: richard.chbeir@univ-pau.fr (R. Chbeir).

http://dx.doi.org/10.1016/j.ins.2014.09.044
0020-0255/

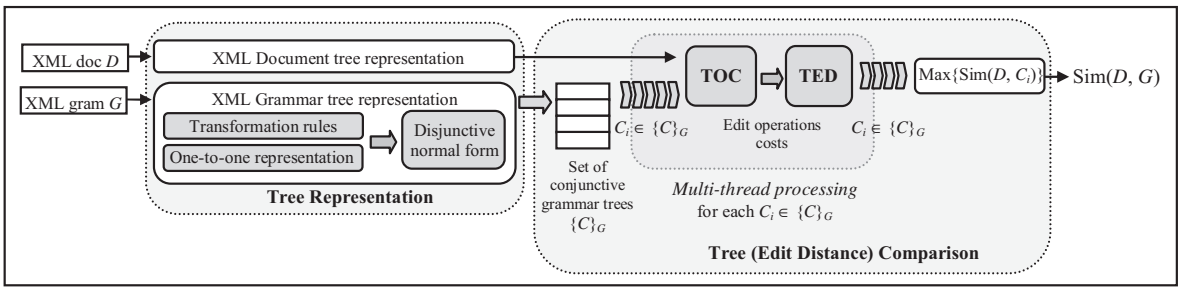

Fig. 1. Simplified activity diagram describing our XML document/grammar comparison framework.

Traditionally, most studies related to XML document/grammar comparison have targeted XML validation [7,8,49], i.e., a specific case of Boolean XML comparison, designed to verify whether an XML document is valid (or not) with respect to (w.r.t.) a given XML grammar (DTD [16] or XSD [31]). Yet with the proliferation of heterogeneous XML data on the Web (i.e., documents originating from different data-sources and not conforming to the same grammar, or documents lacking predefined grammars), there is an increasing need to perform ranked XML document/grammar comparison, which we refer to as 'approximate XML validation': identifying those documents which are not necessarily valid w.r.t. the user grammar, but which share a certain amount of similarity with the grammar, ranked following their similarity scores.

Evaluating the similarity between heterogeneous documents and grammars can be exploited in various application scenarios requiring accurate and ranked detection of XML structural similarities [10,62], ranging over: XML document classification against a set of grammars declared in an XML database [10,80], (just as DB schemas are necessary in traditional DBMS for the provision of efficient storage, retrieval and indexing facilities, the same is true for DTDs and/or XSDs in XML repositories), XML ranked document retrieval via structural queries [32,55] (a structural query being represented as a DTD/XSD in which additional constraints on content can be defined), the selective dissemination of XML documents [10] (user profiles being expressed as DTDs/XSDs against which the incoming XML data stream is matched), as well as Web service matching and SOAP processing (searching and ranking services which best match WSDL[1] service requests, and comparing outgoing SOAP messages to sender-side WSDLs, processing only those parts of the messages which differ from the WSDL descriptions in order to avoid unnecessary overhead, and thus reduce processing cost in SOAP parsing [74], serialization [2], and communications [72,78]).

In this study, we focus on the problem of evaluating the structural similarity between an XML document and an XML grammar, i.e., comparing the structural arrangement and ordering of XML elements/attributes in the XML document and the XML grammar. Different from previous approaches which are either generic (disregarding XML grammar constraints, e.g., the *Or* operator, ?, *, +, etc.) [32,50,75], developed for the DTD language and do not consider more complex and expressive XSD-based constraints (e.g., *MinOccurs* and *MaxOccurs*) [9,10], or restricted to Boolean results (i.e., traditional XML validation methods [7,8,49]), we aim at providing a method which is:

- Fine-grained in detecting and identifying the structural similarities and disparities between XML documents and grammars, in comparison with current generic [32,75] and alternative [9,10] approaches.
- Considering the more expressive XSD grammar constraints (namely *MinOccurs* and *MaxOccurs*), in comparison with less expressive DTD-based constraints (e.g., ?, *, +) handled in existing methods [9,10].
- Producing a ranked similarity result, in comparison with existing Boolean (validation) methods, e.g., [7,8,49].

To achieve these goals, we provide a new approach that extends well-known dynamic programming techniques for finding the edit distance between tree structures, XML documents and grammars being modeled as Rooted Ordered Labeled Trees. Our approach consists of two main phases: (i) XML document/grammar tree representation and (ii) XML document/grammar tree comparison (cf. overall architecture in Fig. 1). While XML documents can be naturally represented as labeled trees, XML grammars are usually more intricate, due to the various types of constraints on the existence, repeatability and alternativeness of XML nodes (e.g., ?, *, + operators in DTDs, *MinOccurs*, *MaxOccurs* cardinality operators in XSD, as well as the *And* sequence operator and *Or* alternativeness operator). These would have to be considered to obtain an accurate similarity measure. Hence, we address the problem of comparing an XML document with an XML grammar as that of: producing a tree representation for the XML grammar (comparable to the XML document tree representation) with additional components to describe cardinality constraints (namely the *MinOccurs* and *MaxOccurs* operators), and then applying a tree-to-tree edit distance function to compute document-to-grammar structural similarly, taking into account XML grammar constraints. We introduce dedicated grammar transformation rules to simplify grammar expressions (while preserving their

[1] Web Service Description Language (WSDL) is a special XML grammar structure that supports the machine-readable description of a Web service's interface and the operation it supports, including corresponding SOAP message formats.