XS³: A System for Similarity Evaluation in Multimedia-based Heterogeneous XML Repositories

Joe Tekli  
LE2I Laboratory UMR-CNRS  
University of Bourgogne  
21078 Dijon Cedex  
France  
joe.tekli@u-bourgogne.fr

Richard Chbeir  
LE2I Laboratory UMR-CNRS  
University of Bourgogne  
21078 Dijon Cedex  
France  
richard.chbeir@u-bourgogne.fr

Kokou Yetongnon  
LE2I Laboratory UMR-CNRS  
University of Bourgogne  
21078 Dijon Cedex  
France  
kokou.yetongnon@u-bourgogne.fr

Categories and Subject Descriptors  

General Terms  

Keywords  
XML, Semi-structured data, Multimedia data and metadata, Structural similarity, Tree edit distance, Semantic similarity.

1. INTRODUCTION

For the last two decades, multimedia data have become increasingly available, especially on the web considered as the largest multimedia database to date. Its applications include video-on-demand systems, video conferencing, medical imaging, on-line encyclopedia, cartography, etc. Since the value of (multimedia) content depends on how easy it is to search and manage [8], the need to efficiently index, store, and retrieve multimedia data is becoming very high. This is why, W3C’s XML (eXtensible Mark-up Language) has been accepted as a major means for complex (multimedia) data management and exchange. Making use of XML to index, represent, retrieve and compare complex objects has been proven successful, particularly in multimedia applications. SVG, SMIL, X3D and MPEG-7 are only some examples of XML-based multimedia data/meta-data representations. Due to the increasing availability of XML-based multimedia content, comparing XML data becomes crucial in the areas of multimedia databases and information retrieval (IR).

XML similarity is central in version control, change management and data warehousing (identifying and browsing changes between different versions of a document) [1] [7], XML query systems (finding and ranking results according to their similarity) [10][11][12], classification and clustering of XML documents gathered from the web against a set of DTDs declared in an XML database (just as schemas are necessary in traditional DBMS for efficient storage, retrieval and indexing, the same is true for DTDs and XML repositories) [7][2], data and schema integration [3][9] message translation (central in B2B applications) [9], as well as XML data maintenance and schema evolution (detecting differences between different versions of an XML grammar to revalidate corresponding documents [3][4]).

In this demonstration, we aim to present XS³, a system for XML Structural and Semantic Similarity assessment. It allows the comparison of heterogeneous XML documents (originating from different data sources), the comparison and matching of XML grammars (DTDs/XML Schemas), as well as the relatively novel trend of comparing XML documents and grammars, based on their structural and semantic features.

In comparison with existing DB and IR-related systems involving XML similarity assessment, our prototype is not tied to a specific application nor to a specific context (it does not extend or propose a new XML querying language as in [11][12], nor does it focus on one single application such as document clustering [2] or structural pattern matching [10]). In fact, it implements low-level algorithms and similarity evaluation methods that could be exploited in various application scenarios, enabling the user to evaluate their efficiency in each application domain, and thus choose the one that is most adapted to her needs.

2. SYSTEM ARCHITECTURE

The XS³ prototype, implemented using C# .Net, is made of four independent and interactive components, as well as various comparison modules and facilities (cf. Figure 1).

The parser component starts by verifying the integrity of XML documents and DTDs, transforming them into ordered labeled trees to be treated by the similarity evaluation component.

The similarity evaluation component consists of several autonomous algorithms (mostly based on the concept of tree edit distance), among which [1][2][7][13][14] dedicated to XML document/document comparison, [15] for documentgrammar comparison, and [16] for grammar/grammar matching. It is extensible to other XML comparison approaches (a combined structural/semantic similarity measure has been recently added [10], integrating the traditional IR vector space model).

![Figure 1. Overall XS³ architecture.](image-url)
The **Synthetic XML/DTD generator** produces sets of XML documents and DTD definitions, based on specific user input requirements (e.g., a variability parameter for document generation, a controlled vocabulary for generating synthetic DTDs, number of 'And/Or' operators, operator disposition ...).

Furthermore, a **taxonomic analyzer component** was introduced to compute semantic similarity values between words (expressions) in a given reference knowledge base (e.g., WordNet), to be subsequently exploited in evaluating XML element/attribute label similarity [10]. It currently encompasses measures in [5][18] and is extensible to others.

Built upon the main system components of XS³ are different modules and facilities for assessing XML similarity. These range over **One to One comparisons** (comparing one XML document/grammar to another document/grammar), **One to Many comparisons** (comparing one XML document/grammar to a set of XML documents/grammars and vice-versa, ranking the documents/definitions according to their similarity to Xₐ) and the **Many to Many comparison module** (comparing sets of XML documents/grammars, consequently enabling XML documents/grammars clustering and classification). In the demonstration of XS³, we will provide an overview of the various components and functionalities of the system (cf. Figure 2 and 3) and how it enables XML similarity evaluation.

![Figure 2](Image 70x257 to 278x458)

**Figure 2. XS³’s One to One document comparison interface.**

We will focus on XML-based multimedia data (mainly SVG and MPEG-7) and will show how XS³ can be exploited in XML multimedia ranked **search-by-document** and **search-by-grammar** applications, as well as classic data warehousing and version control ones (edit script and mappings generation). We aim to stress on our system’s efficiency in a multimedia framework (using multimedia specific knowledge bases, particularly in the MPEG-7 domain) as well as in a generic IR context (using fragments of WordNet[6]). We will show that adding semantic assessment to the comparison process yields more accurate results - having an accurate, domain specific and complete knowledge base - while demonstrating its impact on time complexity.

We will also focus on the clustering and classification facilities which integrate information retrieval concepts and metrics (i.e., specially devised XML document-related precision and recall) to be utilized for comparing the accuracy and efficiency of different XML similarity methods in various application scenarios.

![Figure 3](Image 334x490 to 542x661)

**Figure 3. Snapshot of XS3’s grammar clustering interface.**

3. REFERENCES


---

1. http://www.cogsi.princeton.edu/cgi-bin/webwn