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
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Clustering*, *Search process*; I.7.1 [**Document and Text Processing**]: Document and Text Editing – *Document management*; I.7.2 [**Document Preparation**]: Document Preparation – *Markup languages*.

General Terms
Algorithms, Measurement, Performance, Design, Experimentation.

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 and meta-data representations. Due to the ever-increasing web availability of XML-based multimedia content, methods for comparing XML data become crucial in the areas of multimedia databases and information retrieval.

XML similarity is central in version control, change management and data warehousing (finding, scoring and browsing changes between different versions of a document, as well as index maintenance) [1] [7], XML query systems (finding and ranking results according to their similarity in order to retrieve the best results possible) [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 the provision of efficient storage, retrieval and indexing facilities, 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 a given grammar to revalidate corresponding XML documents [3][4]).

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.
ACM 978-1-60558-303-7/08/10.

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 form 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 test and 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, subsequently 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 document/grammar comparison, and [16] for grammar/grammar matching. It is extensible to other XML-based comparison approaches (a combined structural and semantic similarity measure has been recently implemented in XS³ [10], integrating the traditional IR vector space model).

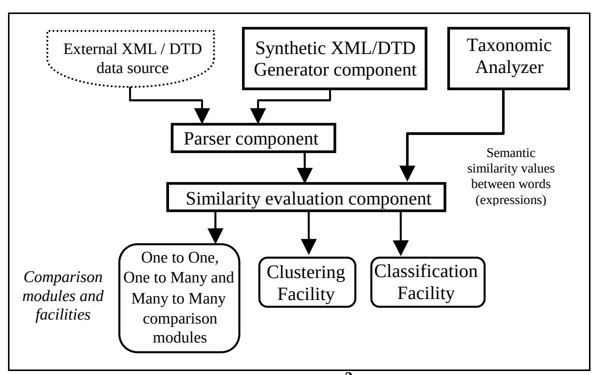

Figure 1. Overall XS³ architecture.

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, 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 XS^3 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 X_1 to a set of XML documents/grammars and vice-versa, ranking the documents/definitions according to their similarity to X_1) 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 XS^3, 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.

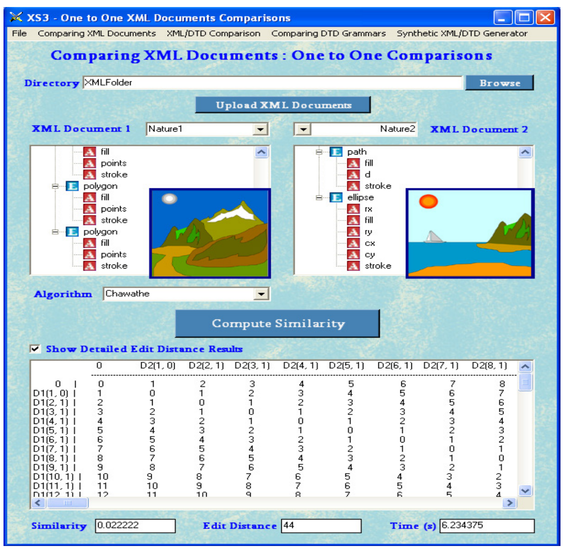

Figure 2. XS^3's *One to One* document comparison interface.

We will focus on XML-based multimedia data (mainly SVG and MPEG-7) and will show how XS^3 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 the WordNet[1] taxonomy [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 in an original manner (i.e., specially devised XML document-related *precision* and *recall* metrics coupled with hierarchical classification/clustering algorithms) to be utilized as means for comparing the efficiency and accuracy of different XML similarity approaches in various application scenarios.

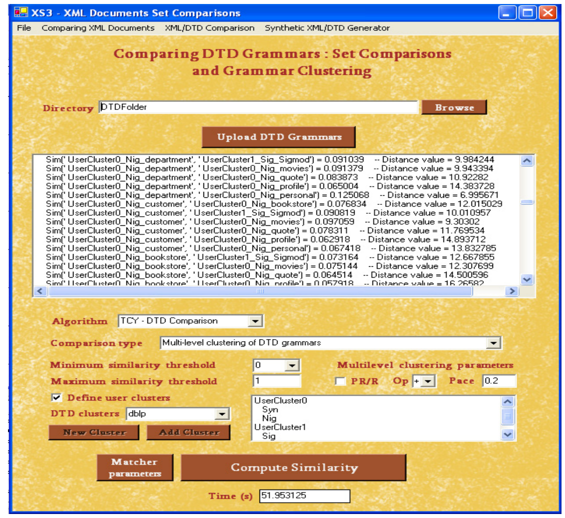

Figure 3. Snapshot of XS3's grammar clustering interface.

3. REFERENCES

[1] Chawathe S., Comparing Hierarchical Data in External Memory. In *Proc. of the VLDB Conference*, pp. 90-101, 1999.

[2] Dalamagas, T., Cheng, T., Winkel, K., and Sellis, T. 2006. A methodology for clustering XML documents by structure. *IS Journal, 31, 3*, pp. 187-228, 2006.

[3] Doan A., Domingos P. and Halevy A.Y., Reconciling Schemas of Disparate Data Sources: A Machine Learning Approach. *ACM SIGMOD*, 2001. Jeong B., Lee D., Cho H. and Lee J., A Novel Method for Measuring Semantic Similarity for XML Schema Matching. *Expert Systems with Applications*, 34-3:1651-1658, 2008.

[4] Leonardi E., Hoai T.T., Bhowmick S.S. and Madria S., DTD-Diff: A Change Detection Algorithm for DTDs. In *Proc. of DASFAA*, 2006.

[5] Lin D., An Information-Theoretic Definition of Similarity. *In Proc. of the 15th Int. Conf. on Machine Learning*, 296-304, 1998.

[6] Miller G. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4), 1990.

[7] Nierman A. and Jagadish H. V., Evaluating structural similarity in XML documents. In *ACM SIGMOD WebDB*, pp. 61-66, 2002.

[8] Pereira F., Technologies for Digital Multimedia Communications: An Evolution Analysis of MPEG Standards. *China Communications J.*, 2006.

[9] Rahm E. and Bernstein P.A., A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*, 10:334-350, 2001.

[10] Sanz I. et al., *ArHeX* : An Approximate Retrieval System for Highly Heterogeneous XML Document Collections, In *EDBT*,192-206, 2006.

[11] Schenkel R., Theobald A. and Weikum G., Semantic Similarity Search on Semistructured Data with the XXL Engine, *IR Journal* 521-545, 2005.

[12] Schlieder T., Similarity Search in XML Data Using Cost-based Query Transformations. In *ACM SIGMOD WebDB*, pp. 19-24, 2001.

[13] Tekli J., Chbeir R. and Yetongnon K., A Fine-grained XML Structural Comparison Approach. In *Proc. ER*, pp. 582–598, 2007.

[14] Tekli J., Chbeir R. and Yetongnon K., Efficient XML Structural Similarity Detection using Sub-tree Commonalities. In *SBBD* and *SIGMOD DiSC*, (Best paper), 2007.

[15] Tekli J., Chbeir R. and Yetongnon K., Structural Similarity Evaluation between XML Documents and DTDs. In *Proc. WISE.*, pp. 186-211, 2007.

[16] Tekli J., Chbeir R. and Yetongnon K., XML Grammar Matching and Comparison: A User-based Framework. Submitted to ICDE 2008.

[17] Tekli J., Chbeir R. and Yetongnon K., An XML Document Comparison Framework. Submitted to *IS Journal*, 2008.

[18] Wu Z. and Palmer M. 1994. Verb Semantics and Lexical Selection. In Proceedings of the 32nd Annual Meeting of the Associations of Computational Linguistics, pp. 133-138.

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