XMLCorrector v1.0: A Tool for Correcting XML Documents With Respect to a Schema

Joshua Amavi[1] Béatrice Bouchou[2] Agata Savary[2]

LIFO - Université d'Orléans, B.P. 6759, 45067 Orléans cedex 2, France[1]
Université François Rabelais Tours, LI, Blois Campus, France[2]
E-mail: joshua.amavi@univ-orleans.fr,
{beatrice.bouchou+agata.savary}@univ-tours.fr

Abstract. XMLCorrector[1] is an implementation of an algorithm allowing to correct an XML document with respect to schema constraints expressed as a DTD. Namely, given a well-formed XML document t seen as a tree, a schema S and a non negative threshold th the algorithm finds every tree t' valid with respect to S such that the edit distance between t and t' is no bigger than th. The algorithm is based on a recursive exploration of the finite-state automata representing structural constraints imposed by the schema, as well as on the construction of an edit distance matrix storing edit sequences leading to correction candidate trees. The algorithm is an extension of ideas announced in [1, 2]. It has been made public under the GNU LGPL v3 license. To the best of our knowledge, this is the first full-fledged study of the XML tree-to-schema correction problem.

Keywords: XML Processing; Tree-to-Schema Correction; Tree Edit Distance.

Operations on an XML tree and distance between two trees. The correction of an XML document t *w.r.t* a set of schema constraints S consists in computing new documents that verify the set of structural specifications stated in S and that are close to t. For correcting an XML tree, we allow three kind of elementary operations on nodes: (*i*) insertion of a node at a position in the tree, (*ii*) deletion of a leaf of the tree, (*iii*) relabeling of a node in the tree. A sequence of node operations transforms a tree t into another tree t'. Figure 1 shows an example of an initial tree t and of a tree t' resulting from t by the application of an operation sequence. Each node operation has a cost. The cost of the node operation sequence is equal to the sum of the costs of each operation in the sequence. The distance between t and t' is defined as the minimal cost of all operation sequences allowing to transform t into t'. For instance, if we admit cost 1 for each node operation, the operation sequence in Fugure 1 has cost 3. Note that there exists non operation sequence transforming t into t' with a lower cost, thus the distance between t and t' is 3. The aim of XMLCorrector, for a given well-formed XML document, a DTD and a threshold, is to find all correction

[1] http://www.info.univ-tours.fr/ savary/English/xmlcorrector.html

candidates, i.e. all valid documents whose distance from the original document does not exceed the threshold. The program provides the list of all operation sequences leading to such correction candidates, as well as the resulting XML documents themeselves.

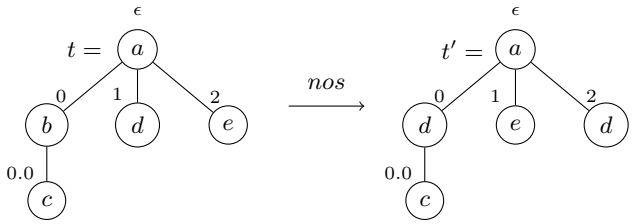

Fig. 1. Application of $nos = \langle (add, 1, e), (delete, 2, /), (relabel, 0, d) \rangle$ on the tree t, $cost(nos) = 3$

Applications. Applications of this problem are important and vary widely. We have: *XML data integration*, *web service searching and composition*, performing *consistent queries on XML databases*, *XML document classification*. But the most common scenario is inherently associated with the web: there is a constant *need of evolution*, for both XML documents and schemas.

Screenshots. (Figure 2)

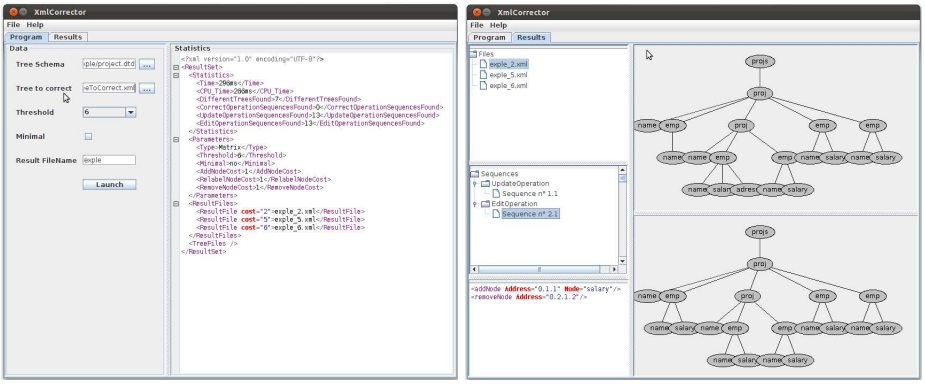

Fig. 2. Main Window and Content of the `Results` tab

References

1. Bouchou, B., Cheriat, A., Halfeld Ferrari Alves, M., Savary, A.: Integrating Correction into Incremental Validation. In: Proceeding of Bases de Données Avancées (BDA 2006). Lille (2006)
2. Bouchou, B., Cheriat, A., Halfeld Ferrari Alves, M., Savary, A.: XML Document Correction: Incremental Approach Activated by Schema Validation. In: Tenth International Database Engineering and Applications Symposium (IDEAS 2006). pp. 228–238 (2006)