Erratum to “Propagating XML constraints to relations”

Susan Davidson a, Wenfei Fan b,c, Carmem Hara d,*

a University of Pennsylvania, CIS Department, Philadelphia, PA 19104, USA
b University of Edinburgh, LFCS, School of Informatics, EH9 3JZ, United Kingdom
c Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974, USA
d Universidade Federal do Parana, Departamento de Informatica, Brazil

Received 5 September 2005
Available online 18 September 2007

We proposed a method for shredding XML data into normalized relations. Given a set \( \Sigma \) of XML keys and a mapping \( \sigma \) from XML to relations, we derive a relational schema \( R \) in BCNF or 3NF for storing XML documents \( T \) satisfying \( \Sigma \). The method is based on an algorithm for computing relational functional dependencies (FDs) propagated from XML keys. Given \( \Sigma \) and \( \sigma \), it computes a set \( \Gamma \) of FDs such that if \( T \) satisfies \( \Sigma \), then the relations \( \sigma(T) \) satisfy \( \Gamma \). Based on \( \Gamma \) we normalize \( R \).

To compute \( \Gamma \) we used an inference system \( I \) for XML key implication: for any set \( \Sigma \) of XML keys and a single key \( \varphi \), we conclude that \( \Sigma \) entails \( \varphi \) if \( \varphi \) can be proved from \( \Sigma \) using \( I \). The XML keys considered are a special case of XML keys studied in [1], of the form \((Q,(Q',S))\), where \( Q, Q' \) are XPATH expressions and \( S \) is a set of XML attributes. Here \( I \) is a variation of the inference system given in [1].

Mistake. Unfortunately, \( I \) is not complete in a special case, i.e., it may happen that \( \Sigma \) entails \( \varphi \) but \( \varphi \) cannot be proved from \( \Sigma \) using \( I \) when (a) \( \varphi = (\epsilon, (Q',\emptyset)) \), and (b) for any XML tree \( T \), if \( T \) satisfies \( \Sigma \) then there exist no node reachable from the root of \( T \) via \( Q' \). This is also a problem with the more general inference system of [1], for which the incompleteness was recently pointed out in [3].

Correction. Fortunately, the bug has little impact on our main results. Indeed, all FDs that the propagation algorithm fails to find, due to the incompleteness of \( I \), are of the form \( R : \emptyset \rightarrow l \) (and others that can be derived from them using Armstrong Axioms), where \( R \) is a relation schema of \( \mathcal{R} \), and \( l \) is a column of \( R \) which is guaranteed to have the constant value \textit{null} throughout the entire relation. Such FDs are not needed in deriving a normal form for \( \mathcal{R} \). Nevertheless, the incompleteness of \( I \) can be fixed with the addition of three rules for deducing the special case. Due to the space constraint we defer these rules to [2], which shows the following. (1) The revised inference system is sound and complete for the XML keys we considered. (2) Employing the revised system instead of \( I \), the propagation algorithm can be extended with a preprocessing phase that simply detects and removes those relation columns that are guaranteed to have the \textit{null} value, since such columns are not needed in XML shredding. (3) The complexity bounds remain precisely the same for the revised algorithms. More specifically, the implication problem for XML keys, i.e.,
for deciding whether or not $\Sigma$ entails $\varphi$, can still be decided in $O(|\Sigma|^2|\varphi|^2)$ time, and the propagation algorithm is in $O(|\Sigma|^2|\sigma|^4)$ time.

References