XPath 式に対する K 最適修正候補発見アルゴリズム

池田　光雪

近年，様々なデータを柔軟に記述できるフォーマットとして XML(Extensible Markup Language)が普及している．大量の XML データを管理・蓄積する場合，DTD 等のスキーマ言語を用いて XML データの構造を予め定義しておき，それに関して妥当な XML データを扱うのが一般的である．また，XML データへの問い合わせ言語としては XPath(XML Path Language)がよく使われる．

DTD とそれに関して妥当な XML データが存在する状況において，XPath 式で問い合わせを行うことを考える．問い合わせを行う際は XML データもしくは DTD の構造を理解する必要があるが，これらが複雑であったりユーザが DTD や XPath 式の記述に不慣れな場合，DTD に関して妥当な XPath 式を記述することは必ずしも容易ではない．また，DTD やその下にある XML データが管理者によって更新され，検索対象のデータの構造が変化する場合もある．この場合，妥当であった XPath 式の妥当性が失われる可能性がある．

上記のような場合，DTD D に関して妥当でない XPath 式 p に対して，D に関して妥当かつ p との類似度が高い XPath 式をユーザに提示できれば，XPath 式記述の支援に有用であると考えられる．ただし，そのような XPath 式は一般に複数個存在するため，複数の候補からユーザが所望の式を選択できることが望ましい．そこで本論文では，XPath 式 p, DTD D, および正整数 K に対して，D に関して妥当な XPath 式を p に類似したものから K 個列挙するアルゴリズムを提案する．ただし本論文では，XPath 式に軸として child, descendant-or-self, following-sibling, preceding-sibling のみを許し，ノードテストは要素名のみを許すという制限を加えている．さらに，XPath 式には絶対表記と相対表記の 2 通りの記述があるが，本論文では絶対表記であることを前提としている．また，XPath 式間の類似度判定のため，本論文では XPath に関する編集距離を導入する．XPath 式 p_1 と p_2 間の編集距離を，p_1 を p_2 に更新するために必要な編集操作(要素名の置換，軸の置換など)のコストの総和と考える．本アルゴリズムの特徴は，DTD の全体構造を把握していなくても，目的の要素名がある程度特定できれば妥当な XPath 式が得られることである．

本アルゴリズムを Ruby で実装し，評価実験を行った．その結果，妥当でない XPath 式に対して，概ね適切な修正候補をユーザに提示できるとの見込みが得られた．

（指導教員　鈴木伸崇）