

XPath 式に対する K 最適修正候補発見アルゴリズム

An Algorithm for Finding K Correct XPath Expressions

学籍番号：201121707

氏名：池田 光雪

Kosetsu IKEDA

Web 上において様々なデータを柔軟に表現可能なフォーマットとして、XML が広く普及している。一般に、XML データの管理・蓄積には DTD 等のスキーマ言語が併用され、それらにより XML データの構造を定義し、データの品質を保証している。また、XML データへの問合せ言語としては XPath がよく使われている。

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 式で問合せを行う場合を考える。このとき、ユーザは対象となる XML データの構造を理解している必要がある。しかし近年、XML データのサイズは巨大化しているため、その構造を正しく把握し妥当な 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し、XPath 式の妥当性が失われる場合もある。これらのような状況の際、ユーザが記述した（妥当でない）XPath 式に近く、かつ DTD に対して妥当な式をいくつか提示することができれば XPath 式の記述・修正の支援に有用である。

本論文では、DTD D 、XPath 式 p 、正整数 K を入力とし、 D に関して妥当かつ p との類似度が高い K 個の XPath 式をユーザに提示するアルゴリズムを提案する。本論文では入力 XPath 式に近い XPath 式を提示するため、XPath に関する編集距離を導入している。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、コストの最も小さいものから K 個の妥当な XPath 式を出力する。提案アルゴリズムの特徴は、DTD の全体構造を把握していなくと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。これまでもユーザの記述した XPath 式を入力とし、その修正候補を提示するアルゴリズムは多く提案されてきたが、再帰を扱えない、XML データの入力を必要とするなどの問題があり、実用上の大きな制限となっていた。提案アルゴリズムでは、再帰も許した DTD を入力とすることで高速な修正候補の提示を可能としている。

本アルゴリズムを Ruby で実装し、実行速度と生成される式の両方に対する評価実験を行った。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実用上十分な速度でユーザに提示できるとの見込みが得られた。

研究指導教員：鈴木 伸崇

副研究指導教員：森嶋 厚行