正規木文法の差分抽出問題に関する研究

堀江　和磨

XML データをデータベース等で継続的に蓄積・管理する場合，格納すべきデータの構造をスキーマで定義しておき，それに沿った構造のデータを作成・格納することが一般的である．また，時間の経過と共に格納すべきデータの構造や種類が変化し，それに応じてスキーマ定義が更新されることも多い．このような状況では，スキーマの更新履歴の管理，スキーマの更新に応じた XML データの修正等が必要となるため，スキーマの更新内容を適切に把握しておく必要がある．特に，管理者が複数で更新内容の共有が必要な場合や，スキーマが複雑で更新内容が多岐にわたる場合等は，スキーマの更新内容を把握することがより重要となる．スキーマの更新内容を把握するには更新前後のスキーマ間で差分抽出を行う必要があるが，これを適切に行える手法はこれまでほとんど提案されていない．本研究は，スキーマ定義言語としては最も表現力の高い正規木文法を対象とし，正規木文法のための差分抽出問題について考察する．

これまで，文字列間の差分(編集操作列) に関しては基本的なアルゴリズムが確立しており，順序木や XML データ間の差分抽出に関しても，順序木の編集操作列を求めるアルゴリズムがいくつか提案されている．しかし，これら既存のアルゴリズムは木文法の意味を解することができないため，正規木文法の適切な差分抽出を行うことは困難である．

正規木文法は終端記号(要素名)の集合，非終端記号(要素の型)の集合，開始記号，および生成規則の集合から構成される．正規木文法の差分抽出は，2 つの正規木文法 G と G'が与えられた時に，G を G'へ更新するために必要なコスト最小の編集操作列を求めることをいう．ここで，編集操作列とは「生成規則の追加・削除」等の編集操作の系列である．本研究では，まず，正規木文法の差分抽出問題が計算困難であることを示す．次に，差分抽出が効率よく行えるための十分条件を求め，その十分条件の下で正規木文法の差分抽出を行う多項式時間アルゴリズムを構成する．最後に，そのアルゴリズムに関する評価実験を行う．

本アルゴリズムを Ruby で実装し，評価を行った．結果，抽出される差分はごく少量であるにもかかわらず，相当の手間を省くことができるという結果が得られた．より差分の量が多い場合，さらに差分抽出の効率化を図ることができると予想される．したがって，本アルゴリズムは更新されたスキーマの変更内容を把握するのに有用であるという見通しが得られた．

（指導教員　鈴木伸崇）