RDF データに対する Shape Expression Schema の妥当性検証アルゴリズム

藤永　健

RDF データはグラフとして表現され，複数の情報のつながりを表現することに適している．また，スキーマを定義することで，問合せ式を記述する際の参照やその実行処理効率の向上といった恩恵を受けられる利点がある．一方，スキーマが定義されている場合，そのスキーマがグラフデータに対して本当に妥当であるか検証を行う必要がある．

従来の RDF のスキーマ言語として，RDF スキーマが提案されている．しかし, RDF スキーマには，スキーマ言語としての形式的なセマンティクスが仕様上定義されていないなどの問題がある．このため，データの構造を厳密に定義して妥当性検証を行うための記述言語としては必ずしも適さないことが指摘されている．そこで提案されたのが Shape Expression Schema (ShEx)である．ShEx は型を Regular Bag Expression という規則に基づいて表し，グラフのノードには型が割り当てられるというスキーマである．ShEx では各ノードが複数の型を持つことを許しており，より表現の幅が広いことも利点として挙げられる．

先行研究において，ShEx の型の定義に一定の制限を置いた場合，妥当性検証が効率よく行えることが Boneva らによって示されている．同時に彼らは，複数の妥当性検証アルゴリズムを提案している．特に，妥当性検証において核となっているのは Refine という処理である．この処理によって型がそのノードに妥当か否かを判別している．この手法では，データ全体を主記憶に収めて処理可能であることを前提としている．しかし，近年では大規模な RDF データが増加しており，メモリ消費の観点からも効率よく妥当性検証を行える手法が必要になると考えられる．

そこで本研究では，より少ないメモリで RDF データの妥当性検証が可能なアルゴリズムを提案する．具体的には，まず各ノードが，リテラルのテキストノードでない葉ノードであるかを判別する．葉ノードとは，出力辺を持たないノードのことであり，型の候補はただ一つに決まるため，主記憶に型を保持しておく必要がない．次に，葉ノード以外のノード（中間ノード）に対して型の初期値を付与する．グラフデータは，各ノードに対して出力ノードとラベルのペアを列挙した形で表し，ノード毎に読み込んでいく．読み込まれたノードに対して Refine を行い，妥当でない型を型の候補から除く．この処理を型の候補が変化しなくなるまで続ける．本論文の提案する手法では，主記憶に格納するノードの数が従来手法よりも少なく，より大きな RDF データの検証を行うことができる．

提案アルゴリムを実装し，RDF データを対象とした評価実験を行った．その結果，提案アルゴリズムが妥当性検証を行えること，および処理時間とメモリ消費量に関して効率よく妥当性検証を行えることを確認した．

（指導教員　鈴木伸崇）