ラベル付き有向グラフにおける Shape Expression Schema の抽出

坪井悠冬里

グラフデータは，情報同士の関係を表現するのに適した構造を持っており，Web グラフや Linked Open Data 等の様々な場面で幅広く利用されている．スキーマが利用できれば，問合せ式の作成支援や問い合わせ式の最適化などに有用である．しかし，グラフデータに対してスキーマはほとんど与えられていない．そこで本研究では，グラフデータが与えられた際に，その概形を求めることによってスキーマを抽出するアルゴリズムを提案する．

このアルゴリズムは，ラベル付き有向グラフを対象にスキーマ抽出を行うものである．ただし，ラベル付き無向グラフに対しても容易に応用可能である．また，スキーマとして Shape Expression Schema (ShEx)を考える．ShEx は，RDF グラフの構造を記述するためのスキーマであり，W3C により仕様策定中である．ShEx は，ノードとその近傍に構造的制約を課す型の集合であり，それぞれのノードに型が割り当てられる．型はノードが持つ出力辺と接続先のノードの型を規定する．このように，ShEx は表現能力が高いスキーマであるため，様々な状況を網羅するスキーマを抽出する事ができる．

スキーマ抽出においては，類似した構造を持つノードを 1 つの型にまとめることでグラフの概形を求めている．基本的な方針として，MDL 原理を基に，全ての型の非適合度が閾値以下であり，かつクラス数が最小であるスキーマを求めることとする．ただし，本研究では，厳密な最適解（スキーマ）を求めることは計算困難であることを示す．そこで，貪欲法に基づく多項式時間アルゴリズムを構成し，スキーマ抽出を行う．更に，求めたスキーマの妥当性を確保するために，どのノードに対しても 1 つの型を割り当て，かつどのノードも割り当てられた型を満たすようにスキーマを抽出する．

ラベル付き有向グラフを対象として，提案アルゴリズムの評価実験を行った．使用したデータは, RDF データのベンチマークツール SP2Bench を用いて生成した RDF データである．評価実験の結果，概ね適切にスキーマが抽出可能であることが分かった．提案アルゴリズムの実行時間は，グラフデータのサイズに対して線形より大きい割合で増加していることが分かった．

グラフデータは年々増加傾向にありかつ大規模化している．このため，大規模なグラフデータを効率的に処理する事が求められるようになっている．一方，本研究のスキーマ抽出アルゴリズムは，主記憶に収まらないような大規模グラフデータには対応していない．今後は，主記憶に収まらない大規模なグラフデータからでも効率よくスキーマを抽出可能なアルゴリズムを考案する予定である．

（指導教員　鈴木伸崇）