ラベル付き有向グラフに対する Shape Expression Schema の抽出
Extracting Shape Expression Schemas from Graphs

学籍番号：201821622
氏名：坪井　悠冬里
Tsuboi Yutori

グラフとは，モノとモノとの関係をノード（点）とエッジ（線）で表現する表現方法である．現代社会において，様々な種類のデータがグラフで表現されるようになった．具体的には，人と人との関係を表現したソーシャルグラフや出版物と出版物の引用関係を表現した引用グラフが挙げられる．

多くの場合，グラフは非常に大規模であり数多くのノードやエッジを有する．そのようなグラフの特徴を知るために，グラフの概形（構造）が得られれば有用である．スキーマはグラフの簡潔な表現であり，グラフから「適切な」スキーマを抽出できれば，抽出されたスキーマを利用して効果的なグラフデータ管理を行うことができる．そこで本研究では，スキーマとして Shape Expression Schema (ShEx) に焦点を当て，RDF/グラフデータから ShEx スキーマを抽出することを考える．

グラフデータからスキーマを抽出する既存のアルゴリズムとしては，大規模なグラフから小さなグラフ（スキーマグラフ）を抽出するものが多い．主な先行研究として，以下の 2 つの種類がある．1 つ目は，ノードが持つラベルの類似度を基にした研究である．エッジラベルなしの無向グラフに対してグラフの要約を行っている．2 つ目は，グラフデータのパスを基にした研究である．同じラベルパスで到達できるノード同士を 1 つにまとめることで，スキーマ抽出を行う．しかし，サイクルを含むグラフデータに対してスキーマを抽出する際，スキーマが元のグラフデータより大きくなる可能性がある．これらのアルゴリズムは，大規模なグラフからスキーマグラフを抽出するアルゴリズムを提案している．一方，本研究のアルゴリズムはサイクルを含むグラフデータに対しても効率を悪化させることなく，より表現力の高い ShEx スキーマを抽出する．

提案アルゴリズムは，抽出されたスキーマの効率性と精度の両方を管理するために，2 つのスキーマ抽出ステップで構成される：(i) ノードの入出力近傍を考慮してノードの類似度を測り，類似度が高いノードに対して同じ型を割り当てる．(ii) そこで得られた型の割り当てを基に，入力グラフに対して妥当な ShEx スキーマを生成する．提案アルゴリズムを実装し評価実験を行った結果，概ね適切に ShEx スキーマを抽出できることが示唆された．また，アルゴリズムの実行時間はデータサイズに対して概ね線形であることが分かった．

今後の課題として，アルゴリズムの実行時間をさらに短縮する方法を考案すること，主記憶に収まらない大規模なグラフデータに対しても対応できるようにすることが挙げられる．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こと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．

研究指導教員：鈴木　伸崇
副研究指導教員：永森　光晴