ラベル付き有向グラフにおける node embedding を用いたスキーマ抽出

木下 一歩

グラフデータは情報や実態の相互関係を記述することに優れており，様々な用途で使用されているほか，近年大きく利用が広がっている Web や SNS におけるユーザや投稿等の関係を記述することにも利用可能である．これらの背景から高度に情報化が進む現代において，グラフデータの潜在的な利用可能性は高まっていると言える．さらに，グラフはスキーマを与えることで情報検索や構造理解が容易になる．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のが現状である．そこで，本研究では，グラフデータからスキーマの表現を抽出する方法を提案する．また，出力されるスキーマは高い表現力を持つ ShapeExpression (ShEx)スキーマを対象とした．ShEx はそれぞれの型について，プロパティと接続先ノードの型について記述する．

同様の課題に取り組む先行研究はすでにいくつか存在しており，なんらかの指標を定義または利用し，クラスタリングを行うものが多く挙げられる．その一方で，貪欲法等によって愚直に解析を行うようなアプローチもあるが，この方法では複雑さが NP 完全であることもあり，処理効率に課題がある．

このような背景から，本研究では，node2vec を拡張した手法を用いてノードの埋め込みを行い，類似度計算の効率化を図るという既存手法とは異なるアプローチを試みる．ノードの持つ型についての情報を極力損なうことなくかつスキーマ構造そのものの比較を避け，類似度計算を効率的かつ自然にすることを目指している．具体的には，node2vec のノード探索であるランダムウォークについて拡張を行った処理を実行し，skip-gram で埋め込み表現を学習する．さらにその後クラスタリングを行い，各クラスタについてすべてのノードのラベルを集計することで妥当なスキーマを得る．また，本問題ではクラスタ数を事前に知ることができないため，クラスタ数をなんらかの手法によって決定する必要があるが，これには Elbow 法を用いている．

SP2Bench や WatDiv によって生成した RDF を対象とした評価実験によって，提案手法が概ね適切にスキーマを抽出でき，実行時間はデータ量に対してほぼ線形に推移することがわかった．また，対象とするデータの特徴によって結果に差があることもわかり，実験を通して相性の良いデータの特徴を明らかにした．

（指導教員　鈴木 伸崇）