外部記憶アルゴリズムを用いた大規模グラフデータからのスキーマ抽出手法

関根　吉紀

人間同士の関係や交通ネットワークのような，関係データベース（RDB）では扱いにくいデータを保存，管理する手段としてグラフデータが広く用いられている．グラフデータにおいては，スキーマを設計する必要は通常ないため，スキーマが付与されているグラフデータは稀である．しかし，そのようなグラフデータにおいても，スキーマを抽出できれば，ユーザが問合せ式を記述する手助けをすること，問合せ処理系が問合せを実行する上での効率を向上させるという 2 つの重要な恩恵を受けることができる．

近年，計算機の処理能力の向上や HDD，SSD 等の外部記憶装置の低価格化と大容量化によって，計算機上で処理・蓄積されるデータが大幅に増加している．このため，主記憶のサイズを大幅に超えるデータを処理することのできるアルゴリズムが必要となっている．しかし，これまで提案されてきたスキーマ抽出アルゴリズムは，データを主記憶に格納し処理することを前提としているため，データが非常に大きい場合には適用困難である．

そこで本研究では，サイズが大きく全体を主記憶上で処理できないグラフデータに対応したスキーマ抽出手法を提案する．スキーマ抽出は，各ノードが属するクラスの抽出と，それらクラス間のエッジを抽出することにより行われる．具体的には，増分クラスタリング法（incremental clustering method）を用いた作成法である先行研究のアルゴリズムを外部記憶アルゴリズムとして拡張する．先行研究では，グラフデータも含む，スキーマ抽出に必要なすべての情報を主記憶に格納し処理していたのに対し，本研究では，グラフデータをシーケンシャルに読み込み，必要なデータのみを主記憶上に置き逐次的にクラス抽出を行う．また，スキーマ抽出に必要だが主記憶に乗り切らない情報は外部ファイルに書き出し，それらも同様にシーケンシャルに読み込んで分析・処理することによってスキーマ間のエッジ抽出を実現している．

本研究はラベル付き有向グラフを対象として評価実験を行い，入力データサイズに対して線形に近い実行時間でスキーマ抽出できること，そして入力データサイズに対して約 2 割の主記憶使用量でスキーマ抽出を行うことができることを示した．

（指導教員　鈴木伸崇）