ISS-P-043

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

坪井　悠冬里†　　鈴木　伸崇††
† 筑波大学　知識情報・図書館学類　　†† 筑波大学　図書館情報メディア系

1. はじめに

近年，グラフデータは，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．スキーマを利用することにより，ユーザーが問合せ式を作成する際の支援や問合せ式を実行する際に効率的に処理を行えるようになる．しかし，従来の手法([1,2]等)では，ラベル付き有向グラフに対してスキーマを効率よく抽出することが困難である．そこで本研究では，スキーマとして Shape Expression Schema (ShEx)を対象とし，グラフデータから ShEx を抽出するアルゴリズムを提案する．

2. 諸定義

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．

形式的には，ShEx は 3 次組 S = (Σ, Γ, δ)の事である．ここで，Σはエッジラベルの有限集合，Γは型の有限集合，δは型を定義する関数(Γの要素をΣ × Γに対する bag 言語に対応させる)である．ShEx には，1 つのノードに 2 つ以上の型を割り当てるか単一の型を割り当てるかによって，multi-type と single-type のセマンティックスがある．本研究では single-type を対象とする．

3. 提案手法

次の条件を満たす ShEx スキーマを最適と定義する．

A)　全ての型の非適合度が閾値以下

B)　A)の条件を満たすスキーマの中で，型の数が最小

本研究では，与えられたグラフに対して，最適な ShEx スキーマを抽出する問題が NP 困難であることを示した．そこで，貪欲法に基づいて ShEx スキーマを抽出する多項式時間アルゴリズムを提案する．

入力：グラフ G = (V,E)，閾値 s
出力：ShEx S = (Σ, Γ, δ)

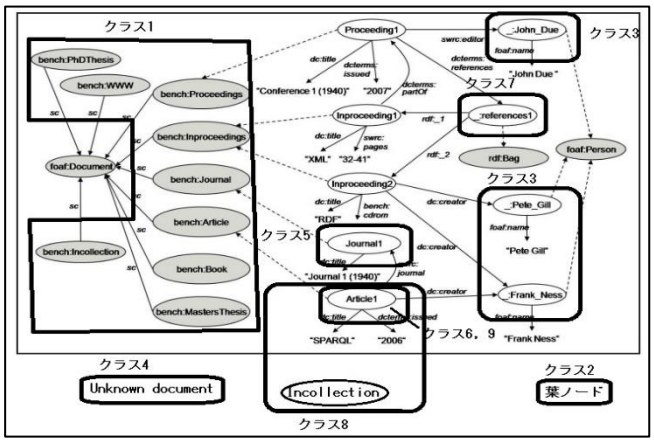
Loop
　/*類似した規則を持っている型どうしをマージする*/
　ForAll 任意の型の組(t_1,t_2)∈ 「×Γに対して
　　型(t_1,t_2)をマージした際，各型の非適合度の中で最大となる値 maxh(t_1,t_2)を計算する
　EndFor
　マージした際，maxh(t_1,t_2)が一番小さくなる型の組(t_1,t_2)を選ぶ
　If すべての型の非適合度が閾値 s より小さい **then**
　　型(t_1,t_2)をマージする
　Else
　　繰り返しの終了
　EndIf
EndLoop
Return ShEx S = (Σ, Γ, δ)

4. 評価実験

SP2Bench[3]により以下の2つの RDF データを生成し，評価実験を行った．

データサイズ(KB)	ノードの数	エッジの数
52.125	343	461
105.512	687	964

提案アルゴリズムでスキーマ抽出を行った結果を下図に示す．いずれのクラスも，同種の型をもつノードがまとめられており，概ね適切なスキーマが抽出できている．



5. まとめ

本研究では，ラベル付き有向グラフに対するスキーマ抽出方法を提案した．今後の課題は，アルゴリズムの効率化と対象データの大規模化を図ることである．

謝辞

本研究は JSPS 科研費(17K00150)の助成を受けたものである．

参考文献

[1] S. Navlakva, R. Rastogi, N. Shrivastava. ”Graph Summarization with Bounded Error”, SIGMOD 2008.

[2] R. Goldman and J. Widom. ”DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases”, VLDB 1997.

[3] M. Schmidt, T. Hornung, G. Lausen, and C. Pinkel. SP2Bench: a SPARQL Performance Benchmark. ICDE’09.