様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ２ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００１５０

研究課題名（和文）大規模グラフデータからスキーマを抽出するための外部記憶アルゴリズムの開発

研究課題名（英文）Developing an Algorithm for Extracting Schema from External Graph

研究代表者
鈴木　伸崇（Suzuki, Nobutaka）

筑波大学・図書館情報メディア系・教授

研究者番号：６０３０５７７９

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究課題では，グラフデータからスキーマを抽出するためのアルゴリズムを開発した．対象とするスキーマは，グラフスキーマとShape Expressionである．まず，グラフスキーマに対して，新たな効用関数と処理の並列化を併用することなどにより，大規模なグラフデータに対しても効率よくスキーマを抽出することの可能なアルゴリズムを開発した．さらに，より表現力のShape Expressionへの対応についても検討を行った．

研究成果の学術的意義や社会的意義
グラフデータは，ソーシャルグラフ，生物データ，Linked Open Dataなどのオープンデータなど，様々な場面で用いられているが，データ量が増大しその構造も複雑化している．一方，グラフデータ対してShape Expressionなど新たなスキーマ言語が提案されており，今後はその活用が進むものと期待されている．グラフデータからスキーマを抽出することができれば，得られたスキーマを用いて，大規模なグラフデータの管理をより効率よく行うことが可能となる．

研究成果の概要（英文）：In this study, we constructed an algorithm for extracting schema from graph data. The target schema languages are graph schema and Shape Expression. First, we constructed an algorithm for extracting schema efficiently by introducing new utility function and parallelization. Then we considered extending our algorithm to extract Shape Expression schema from graph data.

研究分野：情報学

キーワード：グラフデータ　スキーマ抽出

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

関係データベースや XML などの文書データベースなど，既存のデータベースでは多くの場合スキーマが用いられている．その場合，ユーザが格納すべきデータの構造をあらかじめスキーマとして定義しておき，そのスキーマに適合したデータを作成・格納する，という使われ方をされることが一般的である．スキーマには格納すべきデータの構造が明確に定義されており，かつ，その記述量が格納されるデータのサイズと比較して極めて小さいという特徴がある．このような特徴から，スキーマは問合せ式の記述支援や問合せ最適化等に広く用いられている．一方，近年では，ソーシャルグラフ，生物データ，Linked Open Data などのオープンデータなど，グラフとして表現されるデータ（グラフデータ）が急速に増加している．しかし，グラフデータにはこれまで標準的なスキーマ言語がほとんど存在していなかったという背景もあり，多くの場合グラフデータにはスキーマが与えられていないのが現状である．しかし，Shape Expression など，高い表現力をもったスキーマ言語制定の動きがあり，今後はグラフデータに対してもスキーマの利活用が進展するものと期待される．このような状況から，スキーマをもたないグラフデータからスキーマを効率良く抽出することができれば，グラフデータにおける問合せ式の記述支援や問合せ最適化等に活用でき極めて有用であると考えられる．

２．研究の目的

本研究の目的は，グラフデータからスキーマを効率良く抽出する効率の良いアルゴリズムを開発することである．本研究のアルゴリズムはラベル付き有向グラフを対象とするが，無向グラフなどにも適用させることは容易である．なお，抽出対象となるスキーマについては，まず一般的なグラフスキーマを主な対象とし，さらにより表現力の高い Shape Expression についても対応方法を考察する．スキーマ抽出にあたっては，下記２点の要件を満たすことが望ましいため，これらを満たすようアルゴリズムを開発する．

- 近年のグラフデータはサイズが大きいものが多いため，そのようなグラフデータに対しても効率よくスキーマが抽出できるようにアルゴリズムを構成する必要がある．
- グラフデータからのスキーマの抽出にあたっては，できるだけ「適切」なスキーマを抽出することが望ましい．ここで，「適切」なスキーマとは，「できるだけ類似した構造をもつノードが同じクラス（型）にまとまっており，かつ，クラス（型）の数を低減できている」という性質を満たすものであるとするのが自然である．そこで，そのような性質を満たすスキーマを効グラフデータから効率よく抽出するアルゴリズムを開発する必要がある．

３．研究の方法

本研究で扱うスキーマは，グラフスキーマと Shape Expression である．まず前者においては，スキーマを「グラフデータ G において類似した構造(ノードのもつ入出力辺のラベルや数)をもつノードを同じクラスにまとめることで得られる，G の概形を表すグラフ」と定め，これを G から抽出するアルゴリズムを開発する．適切なスキーマを抽出するため，「できるだけ類似した構造のノード同士を同じクラスにまとめ，クラス数を可能な限り低減できているか」というスキーマの品質を算出する関数（効用関数）を定義した．そして，効用関数の値ができるだけ大きくなるようなスキーマを抽出するアルゴリズムについて検討を行った．ここで，効用関数が最大となるスキーマを抽出する問題は計算困難と考えられるため，以下のようにしてアルゴリズムを開発した．

1. まず，本問題が計算困難であることを形式的に証明した．
2. 1 の結果を踏まえて，スキーマ抽出において計算の手間を要する部分について検討を行った．
3. 本問題を解くための効率の良いアルゴリズムを開発する．ただし，厳密な最適解(スキーマ)を効率よく求めるのは困難と考えられるので，2 の考察の結果を踏まえつつ，ヒューリスティクス的手法に基づき，効用関数の効率化，並列化などの手法を採り入れ，効率の良いアルゴリズムを開発した．
4. 得られたアルゴリズムを計算機上に実装して評価実験を行い，動作効率や抽出されたスキーマの品質等の観点から，本アルゴリズムの有効性について検証を行った．

次に，後者の Shape Expression について，Shape Expression はグラフスキーマとは異なり，出力先の型が明示的に記述される．そのため，グラフスキーマを前提とした手法に基づいて類似したノードを同じ型にまとめると，それらが属する型が変化することになり，それらを参照している他のノード（型）にも影響が生じる．そこで，出力先の型の変化を親ノードに伝搬させながらスキーマ内の型の修正を繰り返すことでスキーマ抽出を行うアルゴリズムを構成した．ここで得られたアルゴリズムについても，計算機上に実装して評価実験を行い，動作効率や抽出されたスキーマの品質等の観点から，本アルゴリズムの有効性を評価した．

４．研究成果

得られたアルゴリズムを実装し，評価実験を行なった．まず，グラフスキーマの評価実験について述べる．実行環境は Intel Xeon E5-2623 v3 3.0GHz CPU，16GB RAM，2TB SATA HDD，and Linux CentOS 7 64bit である．アルゴリズムの実装は Ruby を用いて行なった．使用したデータは SP2Bench と DBPedia である．ここで，前者は RDF データのベンチマークツールであり，DBPL

に基づいた任意のサイズの RDF データを生成することができる．後者は，Wikipedia から情報を抽出して Linked Open Data としたものである．以下では，SP2Bench の結果を示す．SP2Bench で生成したデータの詳細を表 1 に示す．

表 1：SP2Bench による生成データ

エッジ数	ノード数（リテラルを除く）	ラベルの異なり数	データサイズ(GB)
1,000,009	187,066	24	0.10
10,000,457	1,730,250	26	1.04
100,000,380	17,823,525	26	10.35

表 1 のデータに対するクラス抽出の実行時間(秒)を表 2 に示す．２行目が従来の効用関数を用いた手法[Q. Y. Wang et al., 2000]を用いてスキーマ抽出を行った場合の実行時間であり，３行目が提案アルゴリズムで用いてスキーマ抽出を行った場合の実行時間である．効用関数の改良，および，アルゴリズムの並列化などの改善により，従来手法と比べて実行時間が大きく短縮されていることがわかる．

表 2：クラス抽出の実行時間

	1,000,009	10,000,457	100,000,380
従来手法	13.26	11.43	1065.99
提案アルゴリズム	6.70	64.23	651.07

次に，抽出されたスキーマの品質について評価を行なった．その結果を表 3 に示す．ここで，スコア 1 とスコア 2 は，入力 RDF データの各ノードに付与されている RDF Schama の型を正解とみなして，抽出されたクラスとの比較により算出したスコアである．スコア 1 は抽出されたクラスのノードがもつ RDF Schema の型が少ないほど大きな値となり，スコア 2 は RDF Schema のそれぞれの型に対して，その型が属するクラスの数が小さいほど大きな値となる．表 3 に示されるように，実行時間は従来のアルゴリズムよりも大幅に短縮されているにも関わらず，スコアは概ね従来のアルゴリズムと同等であることがわかる．

表 3：抽出されたクラスのスコア

	スコア 1	スコア 2	平均
手法	97.02	95.32	96.17
提案アルゴリズム（α=1）	72.53	100.00	86.26
提案アルゴリズム（α=10）	99.45	88.83	94.14

次に，Shape Expression の評価について概要を述べる．こちらも Ruby を用いて実装し，環境も上記のグラフスキーマにおける評価実験と同じである．使用したデータは SP2Bench と LodPaddle の RDF データである．後者は，ナント・メトロポールにある文化施設の場所，住所，URL，電話番号等の情報をラベル付き有向グラフとして表現したデータとなっている．以下では，後者の結果について述べる．LodPaddle のデータサイズは 127,647KB，ノード数は 2178，エッジ数は 3252 であった．抽出したスキーマについて，10 個の型が抽出され，精度の指標となる Rand 尺度は 0.93 と高い値を示した．ただし，評価に用いたデータのサイズが小さいため，より大きなデータを用いて更なる評価を行うことが今後の課題である．

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Y. Sekine and N. Suzuki	４．巻 -
２．論文標題 Extracting Schemas from Large Graphs with Utility Function and Parallelization	５．発行年 2018年
３．雑誌名 Proceedings of the Second International Workshop on Graph Data Management and Analysis (GDMA 2018), LNCS 10829	６．最初と最後の頁 125-140
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-319-91455-8_13	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Y. Sekine and N. Suzuki	４．巻 -
２．論文標題 A Schema Extraction Algorithm for External Memory Graphs Based on Novel Utility Function	５．発行年 2018年
３．雑誌名 第10回データ工学と情報マネジメントに関するフォーラム講演論文集	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Y. Tsuboi and N. Suzuki	４．巻 -
２．論文標題 An Algorithm for Extracting Shape Expression Schemas from Graphs	５．発行年 2019年
３．雑誌名 Proceedings of the ACM Symposium on Document Engineering 2019	６．最初と最後の頁 1-4
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3342558.3345417	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 赤澤豪樹，松原尚利，鈴木伸崇	４．巻 -
２．論文標題 Shape Expression Schemaのスキーマ進化に対するProperty Path式修正アルゴリズム	５．発行年 2020年
３．雑誌名 第12回データ工学と情報マネジメントに関するフォーラム講演論文集	６．最初と最後の頁 1-6
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　2件）

１．発表者名
N. Suzuki

２．発表標題
Extracting Schemas from Large Graphs with Utility Function and Parallelization

３．学会等名
Proceedings of the Second International Workshop on Graph Data Management and Analysis (GDMA 2018)（国際学会）

４．発表年
2018年

１．発表者名
坪井悠冬里

２．発表標題
ラベル付き有向グラフに対する Shape Expression Schema の抽出

３．学会等名
2018年電子情報通信学会総合大会学生ポスターセッション

４．発表年
2018年

１．発表者名
赤澤豪樹

２．発表標題
Shape Expression Schemaのスキーマ進化に対するProperty Path式修正アルゴリズム

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム

４．発表年
2020年

１．発表者名
Y. Tsuboi

２．発表標題
An Algorithm for Extracting Shape Expression Schemas from Graphs

３．学会等名
ACM Symposium on Document Engineering 2019（国際学会）

４．発表年
2019年

１．発表者名
Y. Sekine

２．発表標題
A Schema Extraction Algorithm for External Memory Graphs Based on Novel Utility Function

３．学会等名
第10回データ工学と情報マネジメントに関するフォーラム

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考