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　６ 年　 ５ 月 ２４ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１９００

研究課題名（和文）グラフデータにおける問合せ式充足可能性問題の計算複雑さおよび判定アルゴリズム

研究課題名（英文）Computational Complexity and Algorithm for Query Satisfiability Problem on Graph Data

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）　　2,400,000円

研究成果の概要（和文）：本研究ではグラフデータを対象に，スキーマ下での問合せ式充足可能性問題について考察した．問合せ言語としてConjunctive Property Pathおよびパターン問合せを対象とし，スキーマ言語としてShape Expression (ShEx)を対象とした．これらに基づいて，ShExスキーマ下での問合せ式充足可能性問題を解くためのアルゴリズムを開発した．次に，実装したアルゴリズムについて評価実験を行い，動作効率が概ね良好であること等の結果を得た．また，本テーマの発展として，ライフサイエンス分野のデータに対する適用について検討を行った．

研究成果の学術的意義や社会的意義
問合せ式の充足可能性問題や包含性問題は，問合せ式の検証・作成支援や最適化など，多くの応用に繋がる重要な問題である．例えば，もしqが充足不能な部分式を含む（または，q自体が充足不能である）場合，その部分の評価・実行は不要である．もし充足不能な部分式を効率よく検出できれば，多くの不要な問合せ処理を省くことができる．近年のグラフデータはサイズが非常に大きく，問合せ式の実行には相当の時間を要する．そのため，充足不能な問合せ式を効率よく検出し，その実行を回避することは極めて重要である．

研究成果の概要（英文）：In this study, we consider a query satisfiability problem under schema for graph data. We used Conjunctive Property Path and Pattern query as the query languages, and used Shape Expression (ShEx) as the schema language. Based on these languages, we developed algorithms for solving the query satisfiability problem under ShEx schema. Next, we conducted evaluation experiments on the algorithms and obtained results showing generally good performance of efficiency. In addition, as a further development of this theme, we studied its application to data in the life science field.

研究分野：データベース

キーワード：グラフデータ　スキーマ　充足可能性問題

１．研究開始当初の背景

関係データベースや XML 等では，標準的なスキーマ言語(XML の DTD や XML Schema 等)が確立しており，問合せ式の検証や最適化など多くの場面でスキーマが活用されている．一方，RDF/グラフデータ(以下，単にグラフデータ)においては，標準的なスキーマ言語が確立しているとは必ずしも言えない状況にあった．しかし近年，Shape Expression (ShEx)や Shapes Constraint Language (SHACL)など，記述力が高く充分な実用性をもつスキーマ言語が提案・策定され，グラフデータにおいてもスキーマの利用が着実に進みつつある．

本研究では，スキーマ下での問合せ式充足可能性問題という，最も基本的かつ重要な問題に着目する．ここで，スキーマ S と問合せ式 q に対して，もし S に妥当なグラフデータで q の検索結果が空でないものが存在するならば，q は S の下で充足可能であるといい，そうでなければ充足不能であるという．この問題は，問合せ式の検証・作成支援や最適化など，多くの応用に繋がる重要な問題である．例えば，もし q が充足不能な部分式を含む（または，q 自体が充足不能である）場合，その部分の評価・実行は不要である．もし充足不能な部分式を効率よく検出できれば，多くの不要な問合せ処理を省くことができる．近年のグラフデータはサイズが非常に大きく，問合せ式の実行には相当の時間を要する．そのため，充足不能な問合せ式を効率よく検出し，その実行を回避することは極めて重要である．

２．研究の目的

本研究では，グラフデータにおいて，スキーマ下での問合せ式充足可能性判定問題について考察し，(a)本問題の計算複雑さを明らかにした上で，本問題を解くための効率の良いアルゴリズムを設計すること，および，(b)得られたアルゴリズムを実装して評価実験を行い，提案アルゴリズムの有効性を明らかにすることを主な目的とする．

対象とする問合せ式として，グラフデータの問合せとして一般的な Conjunctive Property Path (CPP)およびパターン問合せを採り上げる．また，対象とするスキーマ言語は ShEx である(図 1)．ShEx は Shape Expressions W3C Community Group で仕様が検討されているスキーマ言語であり，基本部分の表現力は SHACL と同等であるため，本研究の成果は SHACL にも適用できる．ShEx は従来のグラフデータ用のスキーマ（グラフスキーマ等）より真に高い表現力をもち，既に様々な分野で利用されている．なお，RDF のスキーマ言語としては RDF Schema (RDFS)も存在する．しかし，RDFS はオントロジー記述言語としての性格が強く，データ構造を明確かつ厳密に記述・定義すべきスキーマ言語としては必ずしも適していないと考えられる．そのため，本研究では ShEx 対象とした．

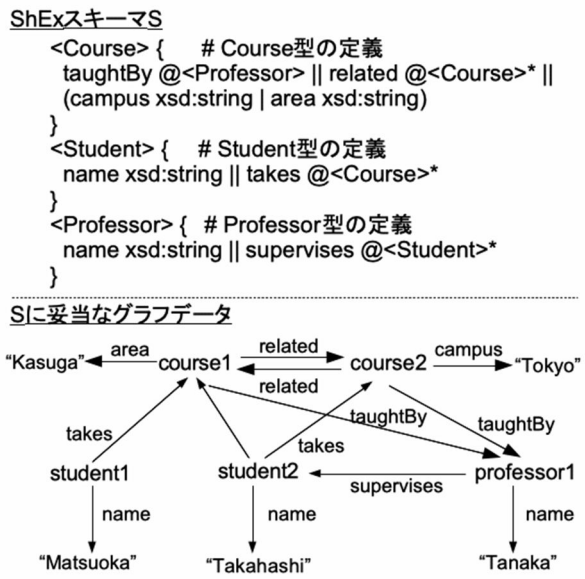

図 1：ShEx スキーマと妥当なグラフデータ

３．研究の方法

CPP およびパターン問合せの両者について，充足可能性問題を解くためのアルゴリズムを開発した．

CPP については，次の方法で充足可能性を判定する．まず CPP q を有向グラフ G_q に変換する．次に ShEx S と q を構成する各 Property Path p をそれぞれオートマトン M_S , M_p に変換する．各 Property Path p について，M_S の各ノードを M_p の開始状態と見立てて M_p の受理状態まで遷移できるかを調べる．一つでも条件を満たすパスが見つかれば充足可能，見つからなければ充足不能と判定することができる．G_q のエッジに沿って各 Property Path についてこの判定を行い，解として得られたノード候補を保持していく．判定を進めながらその都度ノード候補を更新するという手順を全ての Property Path の判定が終わるまで繰り返すことで q を満たす解が存在するかを判定する．

次に，パターン問合せについては，パターン問合せの包含性問題を解くためのアルゴリズムを開発することとした．ここで，ShEx S および問合せ p,q に対して，S に妥当などのデータに対しても p の解が q の解を含むとき，S の下で p は q を包含しているという．充足可能性問題は包含性問題の部分問題であるため，包含性問題を解くためのアルゴリズムが得られれば，充足可能性問題も解くことができる．

包含性問題を解くためのアルゴリズムを，次のように構成した．
A)　まず，与えられたパターン問合せ p,q に対して，p と q の間のノードの対応を求める．
B)　得られたノードの対応を基に，p と q の包含関係を判定する
まず A)について，パターン問合せ p,q はグラフであるから，隣接行列として表すことができる（各成分がエッジのラベルを表す）．p と q の隣接行列で共通の成分があれば，その部分は p と q で共通のエッジが存在することを表している．ノードの並び順に応じて異なる隣接行列が得られるため，p の隣接行列と q の隣接行列のうち，最も共通部分の大きいものを選び，共通部分を対応関係として抽出する．ただし，このような隣接行列の数はノード数に関して指数関数的に増加するため，全ての組み合わせを比較するのは非効率である．そこで，S の型を利用してノードの取り得る型を求め，探索空間を削減している．

次に，B)は次のような処理を行う．まず，q の各エッジ e に対して，それが S の下で p の解を減少させるものでない（追加的な制約を与えない）もので，かつ，e に対応するエッジが p に存在しない場合，e に対応するエッジを p に追加する．このような処理で得られた p,q に対して，もし q が p の部分グラフであれば，q は p を包含していると判定する．なお，この部分グラフか否かの判定は一般には計算困難であるが，本アルゴリズムでは p と q の間のノードの対応がわかっているため効率よく行うことが可能である．

４．研究成果

上記３で得られたアルゴリズムを計算機上に実装し，動作効率等を評価した．

まず，CPP に関する評価実験について述べる．グラフデータに対して充足不能な問合せの実行を防ぐためには，問合せ実行前に効率よく検出することが求められる．したがって，提案手法の実行時間とグラフデータに対する問合せ時間を比較し，前者が十分に短いことを確認する．

まず，提案アルゴリズムを Ruby を用いて実装した．次に，SP2Bench および BSBM という RDF のベンチマークツールを用いて RDF データを生成した．なお，これらのツールにはいずれも ShEx スキーマは定義されていないため，それぞれに対して ShEx スキーマを定義した．そして，それら ShEx スキーマに対して充足不能な CPP を作成した．ここで，充足不能な CPP は，報告者らが作成した Ruby プログラムによる自動生成であり，サイズが 5,10,15,20,25 のものをそれぞれ 10 個作成した．以上で得られた RDF データ，ShEx スキーマ，および，CPP を用いて，提案アルゴリズムの実行および問合せに要する時間を計測した．なお，CPP の検索を RDF データに対して行う際は，ApacheJena Fuseki を使用した．使用した計算機は Intel Core i5 CPU 2.3GHz デュアルコア，16.00GB RAM，macOS Monterey 12.0.1 を搭載した PC である．この結果，SP2Bench から生成されたデータにおいては，(a)提案アルゴリズムと(b)RDF 問合せの時間比率(a/b)が 0.00614 から 0.421，BSBM から生成されたデータにおいては 0.0324 から 0.510 であり，いずれの場合も提案アルゴリズムの実行時間が十分に小さいという結果が得られた．

次に，パターン問合せに関する評価実験について述べる．評価実験には 2 種類のデータを使用しており，1 つは SP2Bench，もう 1 つは Wikidata である．

評価に用いる問合せ式について，まず基本問合せを生成し，そこから導出問合せを生成した．ここで，基本問合せ q は，q の初期状態を空として，S からエッジをランダムに選んで q に追加していくことで生成する．本実験では，SP2Bench と Wikidata それぞれに対して 5 つの基本問合せを作成した．次に，q からの派生問合せ q’は次のようにして生成する．まず，q’の初期状態を空として，q からランダムにエッジを選んで q’に追加する．このとき，q と q’の類似度が一定以上となるように追加する．次に，q’に対して，q’のノードに隣接したノードを q’が一定のサイズとなるまで追加する．本実験では，各基本問合せに対して 15 の導出問合せを生成した．よって，SP2Bench と Wikidata それぞれに対して，15 の問合せからなる 5 つの問合せ集合が得られたことになる．

ベースラインを部分グラフ同型問題を解くためのアルゴリズムとし，提案アルゴリズムとベースラインアルゴリズムを比較した．アルゴリズムは Python で実装し，使用した計算機は Quad-Core Intel Core i5 CPU, 8.00GB RAM, Mac OS Monterey 12.2.1 を搭載したものである．まず，再現率では，提案アルゴリズムが 10 の問合せ集合全てで 1 であったのに対し，ベースラインアルゴリズムでは 0.178 から 0.688 という値であった．また，実行時間については，提案アルゴリズムが 0.000366 秒から 0.000965 秒であったのに対し，ベースラインアルゴリズムでは 0.128 秒から 3.83 秒であった．

以上から，CPP については SP2Bench と BSBM，パターン問合せについては SP2Bench と Wikidata に関して良好な結果が得られているが，他のデータについても同様の結果が得られるかどうか確認することが望ましい．これは今後の課題である．

また，今後の発展に向けて，ライフサイエンス分野のデータに対する適用について検討を行った．抗老化遺伝子情報に関する情報を ShEx および RDF を用いてモデル化し，データセットを作成して RDF ストアに格納した．また，比較対象として関係データベースを用い，関係データベースにもこのデータを格納した．両者を比較した結果，関係データベースと比較して自然なモデル化が実現できており，問合せも効率よく行えているという結果が得られた．これを基に，３で開発したアルゴリズムの適用可能性について検討を行っている．

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 H. Fujimoto and N. Suzuki	４．巻 -
２．論文標題 A Simple Algorithm for Checking Pattern Query Containment under Shape Expression Schema	５．発行年 2022年
３．雑誌名 Proceedintgs of the 18th International Conference on Web Information Systems and Technologies	６．最初と最後の頁 278-285
掲載論文のDOI（デジタルオブジェクト識別子） 10.5220/0011536800003318	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 前田祐紀，鈴木伸崇	４．巻 -
２．論文標題 Shape Expression Schemaの下でのConjunctive Property Path充足可能性判定手法	５．発行年 2022年
３．雑誌名 第14回データ工学と情報マネジメントに関するフォーラム(DEIM 2022)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. Maeda and N. Suzuki	４．巻 -
２．論文標題 Detecting Unsatisfiable Conjunctive Property Path under Shape Expression Schema	５．発行年 2022年
３．雑誌名 Automated Systems, Data, and Sustainable Computing	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） 10.55432/978-1-6692-0001-7_2	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 H. Fujimoto, N. Suzuki and Y. Kwon	４．巻 4
２．論文標題 Checking Pattern Query Containment Under Shape Expression	５．発行年 2023年
３．雑誌名 SN Computer Science	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） 10.1007/s42979-023-02142-z	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 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　2件）

１．発表者名
前田祐紀

２．発表標題
Shape Expression Schemaの下でのConjunctive Property Path充足可能性判定手法

３．学会等名
第14回データ工学と情報マネジメントに関するフォーラム(DEIM 2022)

４．発表年
2022年

１．発表者名
Y. Maeda

２．発表標題
Detecting Unsatisfiable Conjunctive Property Path under Shape Expression Schema

３．学会等名
International Conference on Data Technology and Engineering (CDTE)（国際学会）

４．発表年
2021年

１．発表者名
H. Fujimoto

２．発表標題
A Simple Algorithm for Checking Pattern Query Containment under Shape Expression Schema

３．学会等名
AP-iConcerence 2021（国際学会）

４．発表年
2021年

１．発表者名
權娟大

２．発表標題
RDFを用いた健康長寿や抗老化関連情報のモデル化

３．学会等名
日本薬学会第144年会

４．発表年
2024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
研究分担者	權　娟大 (Kwon Yeondae) (80597097)	国立研究開発法人農業・食品産業技術総合研究機構・農業情報研究センター・上級研究員 (82111)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関