

制約つきグラフ探索を実現する異種データベース統合技術に関する研究

Heterogeneous Database Integration Method for Constrained Graph Search

鈴木 源吾*

Gengo Suzuki

本研究では、異種データベース統合技術をグラフデータ操作に拡張し、グラフ探索とノード・エッジのプロパティに対する制約条件とを組み合わせる、制約つきグラフ探索を可能とする手法を提案する。グラフデータベースを含む複数のデータベーススキーマを統合し、対応関係のメタデータを構築する。検索要求時にそのメタデータを探索し、情報源に対する問合せの組合せを決定する動的な側面と、情報源の能力に応じてグラフ探索を最適化できる点に特長がある。グラフデータベースへの探索の一部を、情報源側に適切にプッシュダウンすることで高速化することが可能である。時間が限定されたサービスが実施されているノードを経由する最短経路問題の一種である、時間制約つき寄り道経路探索に本手法を適用し有効性を示した。

Information services using graph data have widely been developed and used. In near future integrated search, which uses graph databases and exiting databases will get more importance. In this paper, we propose a new integration search method, we call it "conditional graph search", which is combination of graph search and constraint conditions on properties of nodes and edges in graphs. Our method has two merits. First merit is a dynamic determination of databases to retrieve when retrieves are requested. Second merit is graph search optimization adapting abilities of information sources. Graph search is separated to integration side operations and information source side operations, and maximal operations are pushed down to information source, and it improves system performance. We adapted this method to "time constrained trip planning search" and showed efficiency under distributed graph databases.

1. はじめに

交通検索・地図検索・SNS検索等でグラフ情報を活用するサービスが、近年多く開発され実用化されている。交通・地図検索としては駅探のような鉄道路線検索やバス路線の検

索サービス、Google Mapによる道路網のナビゲーションサービス等がWeb上で容易に利用可能となっている。また、最近公開されたFacebookのグラフ検索では、人の属性や関係性（例：サイクリングが好きで近くに住んでいる友達）を表す問合せによってSNSグラフを検索することができる。これらのサービスは、グラフ情報を扱う専用のデータベースであるグラフデータベースによって実現されるようになりつつある。グラフデータベースでは、情報はノードとそれをつなぐエッジ、それぞれに付与されるプロパティでモデル化され、最短経路探索等のグラフ探索、SPARQLによる知識検索、パターンマッチ等のグラフ独自処理を高速に実行することができる。

現在のグラフ情報利用サービスは、最短経路探索等の比較的単純なグラフ探索技術を利用することが多いが、今後はグラフに対して様々な条件を指定した経路探索の必要性が増えると考えられる。図1は、SNSに対する条件探索の例であり、「ある人と、2ホップ以内のつながりで趣味が音楽の人」までの経路を求める探索である。この場合、ノードのプロパティである「趣味」に対して、制約条件をつけて探索を行っている。図2は、交通網と店とのつながりに関するグラフの探索であり、ノードのプロパティに対する条件として、店のカテゴリを指定しており、店の開店時間に対する時間制約も指定している（この例は、時間制約つき寄り道探索と呼んでおり、本研究の主要な応用として後に詳述する）。

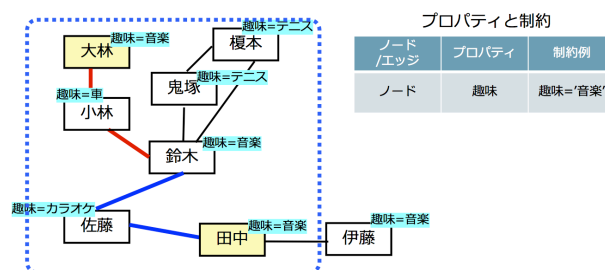


図1 制約つきグラフ探索: SNS の条件探索

Fig.1 Graph Search Example: SNS search

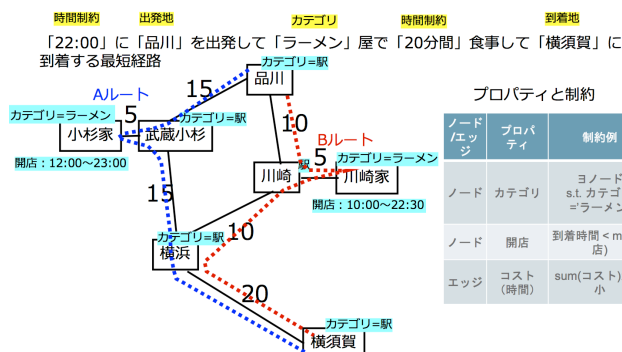


図2 制約つきグラフ探索: 時間制約つき寄り道探索

Fig.2 Graph Search Example: trip planning search

このような、グラフのノードとエッジのプロパティに対して条件を指定するグラフ経路探索を、「制約つきグラフ探索」

* 会員 NTT ソフトウェアイノベーションセンタ
suzuki.gengo@lab.ntt.co.jp

と呼ぶこととする。制約つきグラフ探索は、本論文で定義した用語であり、過去に一般的に用いられているものではない。その定義は次章で示しているが、その基本的特徴は以下の3つである。

- 対象データはグラフ：座標の集まりではなく、ノードとエッジであり、軌跡探索などの地理情報的なモデルではない。
- グラフ上のコストで評価：エッジのプロパティとしてコスト（時間・距離等）を与える。ユークリッド距離（座標間の距離）による評価ではない。
- 探索は経路探索：プロパティで指定した制約を満たす「経路」を求める。グラフのパターンによる探索（例：化学構造探索）は対象外である。

制約つきグラフ探索では、グラフのノード・エッジ情報と制約条件のプロパティは、複数の情報源にまたがるが多い。例えば、レストラン情報と交通情報、SNSのつながりとレビューサイトのカテゴリ情報、ネット上のRDF情報と社内DB等があげられる。これらの情報源を組み合わせることによって、制約つきグラフ探索を実現する必要があるが、既存

次世代のグラフデータベース活用サービスの確立に向けた、制約つきグラフ探索の異種分散データベース環境での実現

の情報源とグラフデータベースは、そもそも組み合わせて連携することを前提に設計されていない。よって、これらの情報源・データベースは分散して様々な異種性を持っていることがほとんどである。

そこで本研究では以下を研究目標とし、そのイメージを図3に示す。

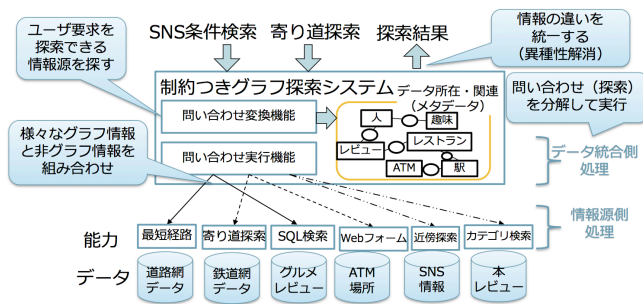


図3 本研究の目標

Fig.3 This research's goal

様々なグラフデータと既存情報源が分散して存在するとき、それらを統合検索できる制約つきグラフ探索システムが、ユーザからのSNS条件探索、寄り道探索等の様々な問い合わせ要求に対し、探索結果を求めることができる。制約つきグラフ探索システムは、ユーザ要求から探索できる情報源を探して特定し（問い合わせ変換機能）、問い合わせを分解して、様々なグラフ情報と非グラフ情報に対して探索・問い合わせ要求を発行し、その結果を組み合わせ、情報の違いを統一しユーザに返却する（問い合わせ実行機能）。

2. 研究の課題

この研究目標を実現するために、以下の2つの課題がある。
【課題1：スキーマ間の異種性解消と統合】

独立して設計された情報源には、情報の構造・表現形式・命名に違い（異種性）がある。その違いを解消しつつデータ統合検索を実現する必要がある。その異種性を解消するための、複数の情報源の突き合わせ（スキーマ統合）には多大な稼働がかかり、完全自動化が困難であることが知られている[1]。その稼働削減には、適切なモデル化・支援方式・ツールがあることが望ましい。しかし、グラフデータベースは新しい情報源であり、そのスキーマ統合手法は未確立である。

【課題2：情報源能力のばらつきに対応した問い合わせ処理】

グラフ利用サービスが発展途上であるため、グラフデータベースの機能はまだ確立しておらず流動的である。よって、グラフデータベースには、RDBにおけるSQLのような標準言語が存在しない、また、Web経由でサービスが提供されるケースが多いため、そのポリシーによってすべてのグラフ探索APIが公開されるとも限らない。これらの理由から、グラフデータベースを利用する情報源には、グラフ探索できる能力にばらつきがある。よって、制約つきグラフ探索システムは、このようなばらつきを考慮しつつユーザからの要求を適切に、情報源による探索処理（情報源側処理）と制約つきグラフ探索システム内部の処理（データ統合側処理）に分解して、処理する必要がある。グラフ探索処理は、頻繁にノード・エッジにアクセスすることが一般的であるために、できるだけ情報源側へ処理を移譲するような最適化が望ましい。今後、グラフ探索処理の分担を説明するときに、この「情報源側」「データ統合側」という用語を用いることとする。また、グラフ利用サービスは、探索サービスの内容や利用する項目の頻繁な変更が多いため、その能力のばらつきも変化が大きい。よって、スキーマ変更への追従が容易な方式が望ましい。

3. 提案手法

これらの課題の解決のために、本論文では、2つの手法を提案する。

まず、データ統合に必要な、データ項目と概念の対応関係・概念同士のつながり・概念とデータ項目の表現形式対応のメタデータを効率的に構築するために、概念グラフを用いたスキーマ統合技術[2][3]を確立した（手法1）。このスキーマ統合技術は、グラフデータベースを含む様々な情報源スキーマを「概念グラフ」と呼ばれるモデルに変換して異種性を解消することが特徴である。概念グラフは、ERモデル等と比較して、データモデル構成要素が少ないため、同じ意味を異なる構造で表現する異種性である構造異種性を回避しやすいことが特長である。また、データ項目が一定のルールで標準的に構造化されている場合に、データ項目を複数の概念に分解し、マッチングを増すことができる。概念間の類似度計算法を組み込んだスキーマ統合支援ツールを実現することにより、スキーマ統合稼働の削減の可能性を示し、課題1を解決した。

次に、動的に異種性解消するデータ統合検索技術（手法2）[4][5]を確立した。利用側の要求とデータベースの構造を事前にSQL等の言語によって固定的に結びつけるのではなく、検索要求時に情報源の所在や表現形式等のメタデータを探索して、検索要求に答えることのできる情報源への問合せの組合せ（問合せ候補と呼ぶ）を生成し、動的に命名・構造・表現の異種性を解消する。そのイメージを図4に示す。メタ

導出法	基本導出法			寄り道の push down	動的導出法		
	観点1→(能力)	L1	L2		L3	L1	L2
集中	基本	寄り道	時間寄り道		基本	基本	時間寄り道
分散1	基本	寄り道	寄り道		基本	基本	基本
分散2	基本	基本	基本		基本	基本	基本

図 8 最適化観点の組み合わせと情報源側の処理

Fig.8 Combination of viewpoints and information source side processing

4. おわりに

本手法の適用範囲について考察する。本手法は、グラフ探索処理を Web インターフェースとして仮想化されている場合に、一般的に利用できると言える。本論文では、寄り道探索の POI は一箇所という仮定であったが、それが複数箇所になった場合でも適用は可能である。グラフ探索は、そのパラメータ指定はノードやそのプロパティであることが多いため、適用範囲は広い。また、処理のプッシュダウン制御によって、仮に情報源側で高度な能力を持っていない場合でも、基本能力が公開されていれば、基本能力を利用して、データ統合側で高度な能力を使うことによって、論理的にはすべてのグラフ処理を行うことが可能になる(グラフ探索は基本能力の組み合わせで実現できる)。ただし、実験結果に示したように実用的な性能を得ることは難しい。

ただし、本手法は情報源の仮想化は関係モデルによっている。寄り道探索の返却結果は、経路情報を JSON 等の文字列で返却されることが想定されている。この文字列を分解し、必要な情報を取り出す処理は利用者(データ統合検索の開発者)側で実施する必要がある。経路情報の表現形式変換関数も利用者責任で作成する必要がある。また、グラフ類似検索・軌跡検索のようにグラフをパラメータとして渡す場合、それらを文字列等に変換して情報源側に渡せば適用は可能であるが、項目の対応を基礎としている本手法のメリットは必ずしも生かせない。

また、本手法は、情報源を跨るグラフ探索には対応できていない。情報源を跨るデータ統合処理は、関係演算である結合と和に限られている。例えば、鉄道網の最短経路探索が各鉄道会社毎に構築されているとき、それぞれの路線に閉じた最短経路を求め和として返却することは可能であるが、両社の路線を辿って検索する結果を求めることができない。このような分散的なグラフ探索は問題としては難しく、個別の探索毎に有効な手法も異なるため、一般的な異種分散データベースのフレームワーク内で整理するのは、今後の課題である。本手法は、例えば、情報源にない階層カテゴリを付加して寄り道探索を行うなど、単一情報源では実現できないデータ統合検索を容易に実現することができる有用な手法であって、最適化制御の考え方も導入したグラフデータベースを統合検索するための基本的ではあるが重要な一歩である。

[文献]

[1] Batini, C. , Lenzarini, M. , Navathe, S.B. : A Comparative analysis of methodologies for database

schema integration, ACM Computing Survey, Vol.18, No. 4, pp.323-364 (1986).

[2] 山室雅司, 鈴木源吾: データ標準化と概念グラフへの変換を利用したスキーマ統合支援法, 電子情報通信学会論文誌 D-1 Vol.J79-D-I No.11 pp.966-974 (1996).

[3] Suzuki, G. , Yamamuro, M. : Schema Integration Methodology Including Structural Conflict Resolution and Checking Conceptual Similarity - Conceptual Graphs Approach - , International Workshop on Database Reengineering and Interoperability, pp.229-242 (1995).

[4] Honishi, T. , Suzuki, G. , Kobayashi, N. , Konishi, K. : A Mediation System Based on Universal Relation Modeling, 20th International Conference on Conceptual Modeling Proceedings (ER2001), SE3, pp.1-4 (2001).

[5] Suzuki, G. , Iizuka, Y. , Kasuga, S. : Integration of Keyword Bases Source Search and Structure Bases Information Retrieval, 7th International CODATA Conference, pp.149-158 (2000).

[6] 鈴木源吾, 榎本俊文, 小林伸幸, 山室雅司, 鬼塚真: 時間制約をもつ寄り道経路探索システムの実現と評価, 情報処理学会論文誌 Vol.53 No.2 pp.857-867(2012).

[7] 大沢裕, 藤野和久: 前処理を必要としない道路ネットワーク上での最短寄り道経路探索アルゴリズム, 電子情報通信学会論文誌 D, Vol.J93-D, No.3, pp. 203--210 (2010).

[8] 鈴木源吾, 鬼塚真, 榎本俊文, 小林伸幸: 制約つきグラフ探索を実現する異種データベース統合技術, 情報処理学会研究報告 2013-DPS-154 (2013).

鈴木 源吾 Suzuki GENGO

1990 年東北大学大学院理学研究科修士課程修了。同年日本電信電話株式会社入社。データベース設計, マルチデータベースシステム, XML データ管理, グラフデータ管理等の研究開発に従事。電子情報通信学会, 日本データベース学会各会員。