

Linked Data から潜在的な関係を探すための クエリグラフパターン最適化

酒 井 理 江^{†1} 飯 塚 京 士^{†1} 佐 藤 宏 之^{†1}
村 山 隆 彦^{†1} 小 林 透^{†1}
服 部 宏 充^{†2} 石 田 亨^{†2}

セマンティック Web や Linked Data の広がりにより様々なデータをつなげることが可能となり、直接つながりがなかったため従来は見つからなかった情報が探索可能となった。異なるクエリグラフパターンを用いることにより様々な関係を発掘することが可能だが、巨大な Linked Data には膨大なクエリの候補が存在するため、全関係をたどることは困難である。また、類似した意味のつながりが多数存在することがある。今まで発見できなかった潜在的な関係に基づく情報が見つかるようになった一方、得られた情報すべてに信頼性があるわけではない。このように膨大な数のクエリグラフパターンから、信頼性が確保でき、かつ効率的に探索するために数を絞り込むことが課題と考える。本論文では信頼性の高い情報にたどり着けるクエリグラフパターンを、ユーザごとに異なるニーズに対応するためバリエーションを残しつつ絞り込むための方法を提案する。信頼性を確保するためには複数のパスからなるクエリグラフパターンを選択した。数を絞り込むためにはクエリのサブグラフパターンを分類し、各分類から選出した代表パターンの組合せを用いることにより、バリエーションが減らないようにした。実データで検証した結果、代表からなるクエリグラフパターンのみを抽出することにより、全体の 9.1% に絞り込むことができた。さらに、定性的な評価により信頼性が確保できる情報にたどり着けることが確認できた。

Query Graph Pattern Optimization for Efficient Discovery of Implicit Information within Linked Data

RIE SAKAI,^{†1} KYOJI IIDUKA,^{†1} HIROYUKI SATO,^{†1}
TAKAHIKO MURAYAMA,^{†1} TORU KOBAYASHI,^{†1}
HIROMITSU HATTORI^{†2} and TORU ISHIDA^{†2}

Spreading the use of Semantic Web and Linked Data has made it possible to

explore new connections between the data which were not linked before. The use of various query graph patterns allows us to explore such connections and lead us to find potential information; however, this information may not be reliable due to its linkages, for example, the linkage between the data is weak. Also in a huge Linked Data, there are too many patterns to follow and many of these may contain similar semantic connections. Our goal is to select query graph patterns which allow us to find reliable information efficiently from a huge number of patterns. We propose a method to reduce the number of query graph patterns while maintaining its variations to meet different users' needs. To achieve high reliability of information reached, we select the patterns made of two paths. To ensure wide variations, we classify the existing patterns into some representative categories and keep all the categories while we reduce the number. By verifying with real datasets, we confirmed that we could reduce the number up to 9.1% of its original, and also underpinned the reliability of the resulting information.

1. はじめに

近年、セマンティック Web の基礎となる RDF (Resource Description Framework) の仕様に従った Linked Data が大量に提供されはじめ、複数のサイトで保有されていた異種のデータを統合し、新しい価値を提供するサービスの可能性が見えはじめている¹⁾。現在 Linked Data に対応している主なデータソースとして、英語版 Wikipedia の中の約 2.9 億件 (2010/3 時点) を RDF 化した DBpedia²⁾、地理データの Wikipedia と呼ばれる GeoNames の情報中、約 9,300 万件を RDF 化した Geo Semantic Web³⁾、2003 年 7 月より開始された大規模な SNS である MySpace のユーザを RDF で表現するサービス⁴⁾ 等があげられる。

このようにセマンティック Web を用いたデータが増えていくなかで、性質の異なるデータ、たとえば Wikipedia のようにコミュニティの目を通してある程度正しさが保証されている情報と陳腐化する可能性のある情報の両方が含まれているもの、また SNS のように新鮮ではあるが正しさが保障されない可能性のある情報とが互いにつながり、リンクをたどって一方のデータから他方のデータの内容を探索することが可能となる。異種データをつなげた利点、つまり単一データでは見えなかった発見的な情報の探索が可能となることを活かし

^{†1} 日本電信電話株式会社 NTT 情報流通プラットフォーム研究所
NTT Information Sharing Platform Laboratories, NTT Corporation

^{†2} 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

つつ、異種データをつなげたことの問題点、つまり性質が異なるデータを混ぜ合わせて探索する際の信頼性や、データが巨大化するなかでの探索の効率性を確保するには、どのようにしたら良いのかを検討する。

本論文では、2章で RDF の統合データを探索する際の課題、3章で信頼性や効率性を確保した RDF データを探索するためのクエリグラフパターン生成手法の提案、4章で提案手法を検証するための実験と結果、5章で実験結果の考察を述べる。

2. RDF グラフデータ探索の現状と課題

2.1 RDF データ探索手法の現状

グラフ構造を持つ RDF データを探索するには、大きく 2 つの方法がある。RDF データ用のブラウザを用いる方法と、クエリをグラフパターンで表現できる RDF 用のクエリ言語を用いる方法である。

2.1.1 RDF ブラウザ

ここでは RDF トリプルをユーザがたどって情報を閲覧してゆくアプリケーションを RDF ブラウザと呼んでいる。RDF ブラウザでは、アプリケーション側でクエリ生成は行わず、ユーザが自ら必要なリンクをたどりながら探すという方法をとっている。World Wide Web やセマンティック Web を提唱した Tim Berners-Lee が考案した Tabulator⁵⁾ や Freie Universität Berlin で研究されている Marbles⁶⁾ 等がこれにあたる。これらのアプリケーションは、ユーザが RDF データのノードに相当する URI を指定すると、その URI の属性のほか、“owl:sameAs” と “rdfs:seeAlso” といった関係を用いて複数のデータソースから関連する情報を抽出して表示し、ユーザは表示された情報をたどってさらなる情報を閲覧できる。データの参照に有効なツールであるが、膨大なデータから特定のデータ間の関係を探し出したりするには、RDF データ用のクエリ言語を解釈する仕組みが別に必要となる。

2.1.2 RDF クエリ言語

W3C では 2004 年から、SPARQL⁷⁾ と呼ばれる RDF データに対する汎用クエリ言語の標準化を進めている。クエリグラフパターンを作成すれば、RDF データからグラフパターンに一致する対象を探索することができる。しかし、あらかじめデータ構造を知らなければクエリに用いるグラフパターンを作成することはできない。また、データ構造が把握できたとしても、たどりたいデータ間の関係はクエリとして手で表現する必要がある。特に Linked Data の場合、新しいデータが次々と追加されグラフ構造が変化していくため、それに対応してクエリも次々と新しいものを用意していかなければならないという問題がある。

2.2 RDF データ探索における課題

グラフ構造の RDF データが巨大になれば、あるデータからある情報にたどり着くためのクエリグラフパターンも膨大となる。その中からユーザが、そのユーザにとって意味があったり、興味深かったりするつながりを持つ情報にたどり着くのは手間がかかる作業である。あるデータに関連する情報を探すときに出発点となったデータからたどり着くルートが 1 つしかない場合、その情報は出発点のデータと関連が薄いため、信頼性が低いとユーザは判断するかもしれない。情報探索において、出発点となるデータとたどり着いた情報との関係についての信頼性を確保できるグラフパターンのルートを絞り込めるかが課題である。また、巨大なデータの中には似たような構造、似たような意味を持つつながりが多数存在する。個々のユーザによって意味を持つつながりや興味深いつながりは異なるので、つながりに関するバリエーションを確保しつつ、効率良く探索するクエリグラフパターンの数を絞り込めるかが課題である。

3. RDF グラフ探索用クエリ抽出手法の提案

2.2 節であげた RDF データの探索における課題であるクエリグラフパターンの数を絞り込む際の仮説と、そのアルゴリズムを提案する。

3.1 仮説

- (1) 探索の始点となるデータから終点となる情報にたどり着くパスが 2 本存在するクエリグラフパターンであれば、関係の信頼性が確保できる。
- (2) クエリの性質として、同じデータセットのみからなり、クエリの両端が同じクラスのもの、お互いのクエリの解の網羅率が高いと考える。これを利用して、すべてのクエリグラフパターンをこの単位に分割、同データセット内のクエリの両端のクラスが同じものでグループ分けし、それぞれのグループから代表を抽出、それらの代表からのみなるクエリグラフパターンに絞り込めばバリエーションを確保しつつ数を絞り込め、エンドユーザは効率良く関係を参照できる。なお、ここでいう解の定義は 3.2 節の用語の説明に、解の網羅率の算出方法は 3.2 節のアルゴリズムの Step4-1 に示す。

3.2 アルゴリズム

3.1 節の仮説を実現するためのアルゴリズムを提案する。

RDF のノードを構成するデータは、RDF スキーマやオントロジ記述言語 OWL で定義されたクラスのインスタンスとして存在している。たとえば、ある技術に關係の深い人を探る際に「技術」と「人」の關係を抽出するには、「技術」を表すクラスのインスタンスと、

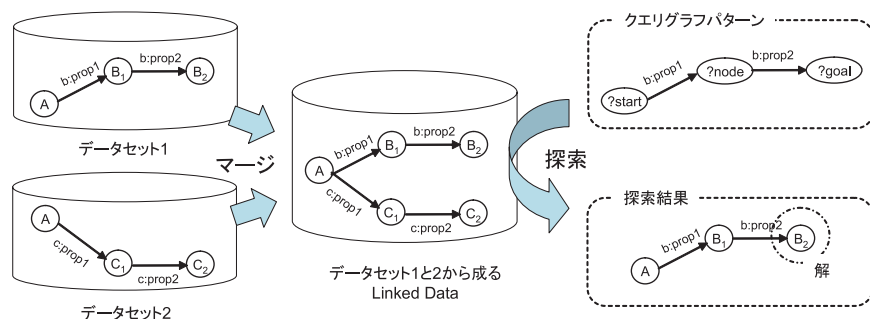


図 1 用語の定義 (クエリグラフパターンと解)

Fig. 1 Terms of query graph patterns and results.

「人」を表すクラスのインスタンスとの間にある関係を探ることになる。以後、本論文では、あるクラスの任意のインスタンスを表現する際に、変数として“?”の記号を用いて表記する。たとえば、「技術」クラスのインスタンスとして存在するデータは「?技術」と表記する。

図 1 と図 2 を利用して本アルゴリズムで用いる用語を説明する。

- データセット：図 1 に示すように、データセットとは別々の目的で作成された RDF のデータベースを指す。また、それを複数マージしたものを Linked Data と呼ぶ。
- 解：図 1 に示すように、クエリグラフパターンを用いて Linked Data からパターンにマッチした情報を探索する。その際、探索の始点となるインスタンスのクラスを Start、終点となるインスタンスのクラスを Goal とし、それに対応する任意のインスタンスをそれぞれ ?start, ?goal とする。探索結果の“?goal”に対応するインスタンスを解と呼ぶ。
- パス：図 2 に示すように、1 本のルートで“?start”から“?goal”にたどり着くグラフパターンをパスと呼ぶ。

提案するアルゴリズムの適用対象となるクエリグラフパターンの全候補は、図 3 に示す方法を用いてデータセットから抽出する^{8)–10)}。あらかじめ定めた 2 つのクラスのそれぞれの任意のインスタンス (“?start”と“?goal”) 間に存在するパスを、特定のホップ数 (?start と ?goal の間のノードの個数) の範囲で探索し、インスタンス間に存在するつながりの解析からクエリグラフパターンの全候補を同定している。以下のアルゴリズムで、これらのクエリグラフパターンを絞り込む方法を示す。

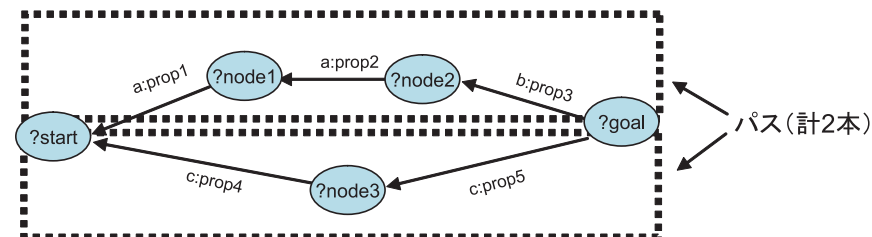


図 2 用語の定義 (パス)

Fig. 2 Term of paths.

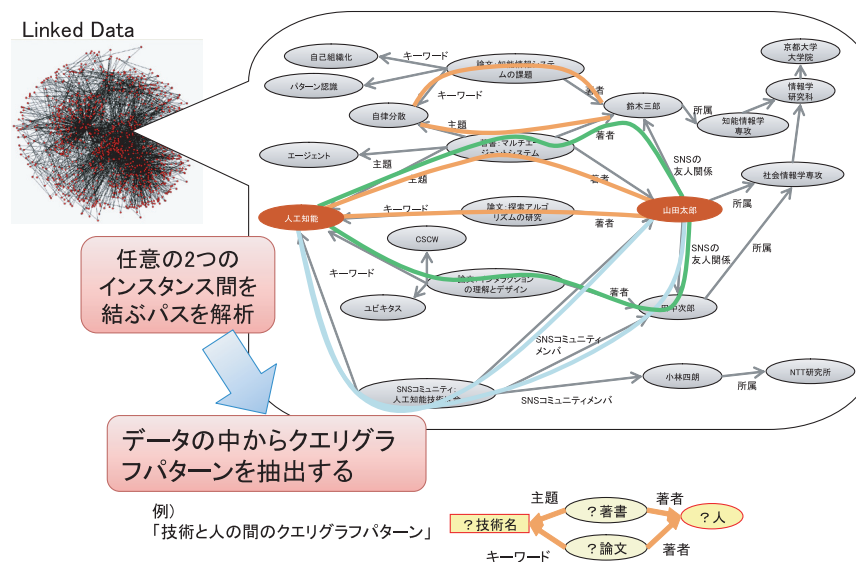
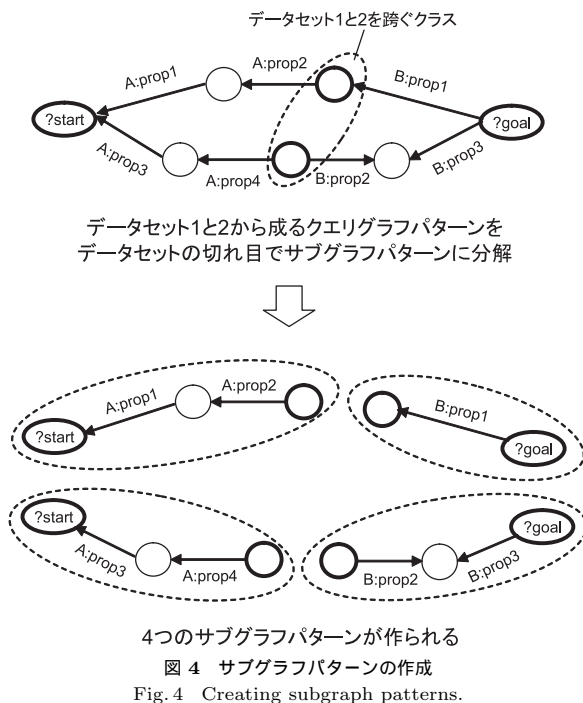


図 3 クエリグラフパターンの抽出

Fig. 3 Extracting query graph patterns.

Step1 ?start と ?goal の間に 2 本のパスが存在するクエリグラフパターンに絞り込む。

Step2 Step1 で抽出されたクエリグラフパターンに関して、各パスごとにデータセットの切れ目のクラス、つまり 2 つのデータセットをまたぐクラスですべてのクエリグラフパターンを分解する。この分解されたクエリグラフパターンを、サブグラフパターンと



呼ぶ(図4参照)。

Step3 サブグラフパターンのそれぞれの端点におけるクラスの同一性とデータセットの種類に基づき、サブグラフパターンをグループに分ける。

Step4 Step3で分類したそれぞれのグループから代表を抽出し、代表サブグラフパターンのみの組合せから構成されるクエリグラフパターンに絞り込む。代表の抽出方法は以下のとおりである：

Step4-1 各グループの代表サブグラフパターンは、代表以外のサブグラフパターンで探索できる情報ある程度網羅できるものとする。Step3の各グループ内の任意のサブグラフパターン“クエリ1”が、同じグループ内の他のサブグラフパターン“クエリ2”の解をどの程度網羅できているかを以下の計算式を用いて計算する。

K : ?start のインスタンス, 検索キーワードの集合

$k_i \in K$

$Q_1(k_i)$: 検索キーワード k_i を ?start とした際のクエリ1の解の集合

$Q_2(k_i)$: 検索キーワード k_i を ?start とした際のクエリ2の解の集合

検索キーワード k_i において, Q_1 が Q_2 の解をどの程度網羅しているかの平均値 m を以下の計算式で表す：

$$m = \frac{1}{|K|} \sum_{i=1}^k \frac{|Q_1(k_i)|}{|Q_1(k_i) \cup Q_2(k_i)|} \quad (1)$$

同じグループ内の他のすべてのサブグラフパターンの解の網羅率が閾値を超える場合は、そのサブグラフパターンを代表にする。閾値に満たない場合は、代表以外のサブグラフパターンの解が複数の代表で閾値以上網羅できるまで、代表を追加してゆく。

Step4-2 Step4-1で代表となったサブグラフパターンの組合せからのみなるクエリグラフパターンのバリエーションを確認する。存在するすべてのバリエーションが保てていない場合、同等になるまで Step4-1で解の網羅率が高い順に代表サブグラフを追加する。

4. 実験

4.1 実験概要

企業に所属するユーザが産学連携先としてある技術に詳しい大学教職員を探す際に用いることのできる「KnowWho サービス¹¹⁾」に本手法を適用した。この KnowWho サービスでは、大学等の機関が提供している教職員のデータベースのみではなく、性質の異なるデータをまたいで技術と教職員の関係を見つけることができる。キーワードとして技術名を入力すると、その技術に関連する教職員が、どのように技術とつながっているかも含めて表示される(図5と図6参照)。

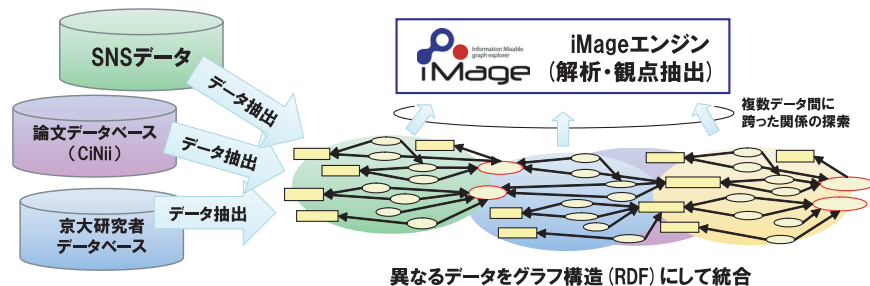
4.1.1 利用データセット

性質の異なるデータを用いるため、情報は正確であるが情報更新が頻繁に実施されず、また以前に登録された古い情報が残り続ける可能性のある“京都大学総覧DB”と“CiNii論文データ”(以降、これらをまとめて“レガシーデータ”と呼ぶ)、情報は最新であるが必ずしも正しさが担保されない“産学連携推進SNSデータ”の大きめに分けて2種類のデータセットを利用した。以下に各データセットの詳細を示す。なお、データの合計数は69,019



図 5 KnowWho サービス

Fig. 5 KnowWho service.



異なるデータをグラフ構造 (RDF) にして統合

図 6 サービス構成図

Fig. 6 Service diagram.

トリプルである。

(1) 京都大学総覧 DB

正式名称「京都大学研究者総覧データベース」¹²⁾とは、研究者の専門分野、研究業績等教育研究活動に関する情報を広く社会に公開することを目的として京都大学が運営する検索システムである。

京都大学総覧 DB から抽出し、登録したデータの内訳は、教職員データ 196 件、研究課題データ 225 件、専門分野データ 179 件、論文データ 5,930 件、特許データ 66 件である。なお、今回登録したのは「産学連携推進 SNS」に参加している情報学研究所の教職員のためのデータである。

本データは教職員名や所属等の基本データ以外は教職員本人が登録するため、それ以外のデータが登録されていない。また登録はされているがアップデートされていないものが含まれる。

(2) CiNii 論文データ

CiNii (NII 論文情報ナビゲータ)¹³⁾とは、学協会刊行物・大学研究紀要・国立国会図書館の雑誌記事索引データベース等、学術論文情報を検索の対象とする論文データベース・サービスで、国立情報学研究所 (NII) が運営している。更新頻度はかなり高い。

情報学研究所の教職員が執筆した論文情報をさらに補うことを目的に、該当教職員の論文 5,173 件を追加している。

(3) 産学連携推進 SNS データ

正式名称「京都大学 ICT 連携推進 SNS」¹⁴⁾とは、産学間の技術ニーズ・シーズの効率的なマッチングのためのコーディネーション基盤を目的として始められた SNS である。参加者は京都大学情報学研究所の教職員、学生、学外 (企業等) のメンバを含めて約 630 名である。

SNS データに関しては京都大学の学生の協力を得て、統制実験を実施して取得した。統制実験では 12 名の学生を 4 グループに分け、グループごとに「組み込み」と「エコロジ」のシンポジウム開催について SNS 上で議論してもらった。実験用に新たな SNS コミュニティを作成し、学生にそのコミュニティのメンバとなってもらい、そこで記事やコメントを書いてもらった。

統制実験から抽出したデータの内訳は、今回の実験で書き込まれた記事 307 件、その中で話題となった教職員名 194 件、プロフィールに登録されたキーワード (興味

のある技術名) 156 件である。

4.1.2 RDF のスキーマ構造

レガシーデータと SNS データをつなぐために設計し、実験に利用した RDF のスキーマ構造を図 7 に示す。

レガシーデータである「論文」「研究課題データ」クラスと SNS データである「記事」クラスと人クラスである「学生」は、人クラスである「教職員」とキーワードクラスである「キーワード」を介してつながっている。なお、人クラスの「学外」と SNS データである「イベント」「コミュニティ」クラスに対応するデータは、今回の実験データには含まれていない。

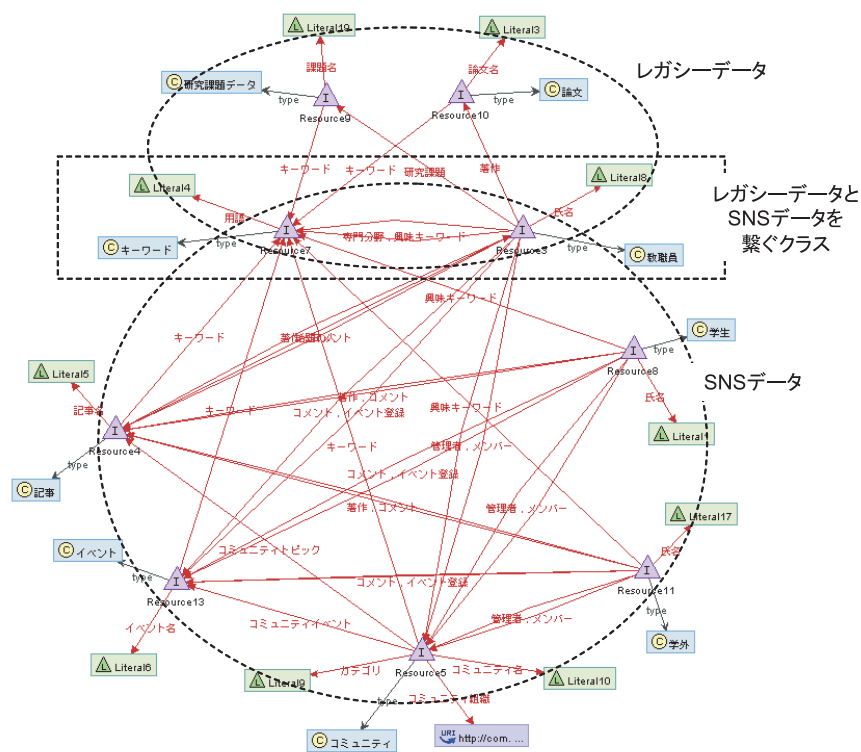


図 7 RDF のスキーマ構造

Fig. 7 Structure of RDF used in the experiment.

4.2 クエリグラフパターン抽出

4.2.1 全クエリグラフパターンの抽出

KnowWho サービスのためのクエリグラフパターンを抽出するということは、「キーワード」と「教職員」を結ぶグラフパターンを抽出すればよい。2 点間を結ぶクエリグラフパターンを抽出するエンジンとして、iMage (旧 CSM)⁸⁾⁻¹⁰⁾ を利用する。今回の実験では、ホップ数は最大 4 個、クエリグラフパターンを構成するパス数は最小 1 本、最大 2 本でクエリグラフパターンを抽出した。なお、抽出にあたって「キーワード」と「教職員」クラスのそれぞれのインスタンス間に存在するつながりを解析をする際に、?start として「組み込み」と「エコロジ」、?goal は任意の「教職員」クラスのインスタンス?教職員を指定して行った。クエリグラフパターンは 583 個抽出された。以下のクエリグラフパターンはすべて?キーワードと?教職員間を結ぶものとする。

4.2.2 クエリグラフパターンの絞り込み

以下の項で、3.2 節で提案したアルゴリズムの実施方法と実施結果を示す。

- (1) 2 本パスのクエリグラフパターン抽出
アルゴリズムの Step1 を実行し、構成パス数が 2 本のクエリグラフパターンを抽出、その結果 583 個から 543 個 (93.1%) に絞り込まれた。
- (2) サブグラフパターンの分類
アルゴリズムの Step2 より、543 個のクエリグラフパターンをサブグラフパターンに分解し、Step3 より 6 つのグループに分類した結果を図 8 に示す。レガシーデータと SNS データをまたぐクラスである?キーワードと?教職員がサブグラフパターンの終端ノードとなった。なお、たとえば (SB-2) は 2 つの (SC-1) からなる等、サブグラフパターンは必ずしも最小単位ではない。
図 8 中の「/」は「or」の意味である。たとえば (SA-3) のプロパティ「sns:コメント/著作」は「sns:コメント」または「sns:著作」であるため、(SA-1) は 2 種類存在する。
- (3) クエリグラフパターンのバリエーション
543 個のクエリグラフパターンはすべて 2 本のパスからなるが、それぞれのパスが図 8 のどのグループのサブグラフパターンの組合せからなっているかを示したのが表 1 である。表 1 では、たとえばパス「SA-SC」が 2 本からなるクエリグラフパターンの場合は「SA-SC」を 1、パス「SA-SC」と「SA-LC」の 2 本からなるクエリグラフパターンの場合は「SA-SC」と「SA-LC」のそれぞれを 1 とカウントしている。
- (4) 解の網羅率を用いた代表サブグラフパターンの選択

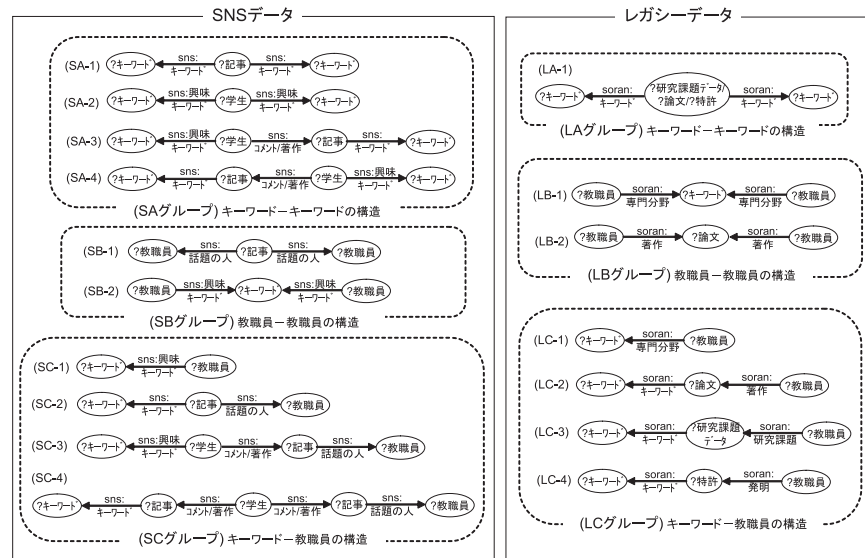


図 8 構造によるサブグラフパターンの分類

Fig. 8 Classification of subgraph patterns according to their structures.

表 1 パスの分布
Table 1 Paths variations.

	SA	SB	SC	LA	LB	LC
SA	x					
SB	x	x				
SC	129	41	x			
LA	x	x	15	x		
LB	x	x	60	x	x	
LC	311	36	x	91	52	x
-	x	x	238	x	x	26

アルゴリズムの Step4-1 より, 式 (1) を用いて計算した結果を図 9 に示す. 図 9 は縦軸を式 (1) のクエリ 1, 横軸を式 (1) のクエリ 2 とした結果で, 解の網羅率がある一定以上, ここでは 7 割以上確保できるようなサブグラフパターンを各グループの代表第 1 候補とした. なお, SA グループと SC グループのいくつかのサブグラフパ

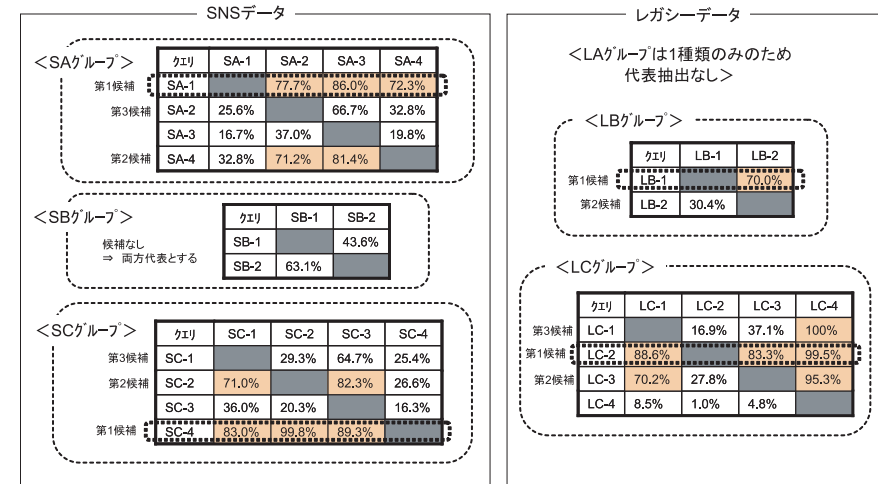


図 9 代表サブグラフパターン抽出結果

Fig. 9 Representative subgraph patterns of each category.

ターンには「コメント/著作」という 2 種類のプロパティが存在するが, 実験で用いた RDF データでは「著作」に比べ「コメント」の存在が圧倒的に多く, 「コメント」は「著作」の解を約 99%網羅しているため, 以降はすべて「コメント」に集約することとする.

- (5) クエリグラフパターンのバリエーションを用いた代表サブグラフパターンの選択
- Step4-1 で選択した代表からなるクエリグラフパターンのバリエーションを確認した結果, SC グループの代表を SC-4 のみとしてしまうと, パスの長さの制約 (中間ノードの数) より他のサブグラフパターンとの結合ができないため, クエリグラフパターンの一部のバリエーションが欠けてしまう. アルゴリズムの Step4-2 より, すべてのバリエーションを確保するため, 解の網羅率の高い順に SC グループの代表を SC-4, SC-2, SC-1 の 3 つにするとすべてのバリエーションが確保できる (表 2 参照).

4.3 実験結果

「キーワード」と「教職員」を結ぶクエリ 583 個を, 3.2 節にあげたアルゴリズムの Step1 より 543 個 (93.1%) に絞り込めた. また, 代表サブグラフパターンからのみなるものに絞り込むことにより, 最終的には 53 個 (9.1%) に減らすことができた. 絞り込んだクエリグ

表 2 パスの分布 (SC グループの SC-1, SC-2, SC-4 を代表とした場合)

Table 2 Paths variations (in the case of choosing SC-1, SC-2 and SC-4 as the representatives of the SC group).

	SA-1	SB	SC-1, 2, 4	LA	LB-1	LC-2
SA-1	x					
SB	x	x				
SC-1, 2, 4	8	8	x			
LA	x	x	5	x		
LB-1	x	x	8	x	x	
LC-2	10	10	x	9	7	x
-	x	x	22	x	x	8

ラフパターンは、削除されたクエリの解の 7 割以上が網羅され、バリエーションも削除前と同等であることも確認できた。

5. 考 察

実験による仮説の検証結果と産学連携への適応に関する考察を以下に示す。

5.1 仮説 (1) の検証

3.1 節であげた仮説 (1) より、1 本パスと比較して 2 本パスのクエリグラフパターンでは解の信頼性が確保できているのかを検証する。

5.1.1 クエリグラフパターン例 1

図 10 に示すクエリグラフパターンは「LC-2-LB-1」と「SC-2-LB-1」(図 8 参照)の異なるパス 2 本からなるものである。以下、図 10 で上のパスを「パス A」、下のパスを「パス B」と呼ぶ。表 3 に示したように、検索キーワードが「組み込み」の解の数と「エコロジ」の解の数は、パスが 2 本になることにより絞られている。

検索キーワードと解の関係をいくつか詳しく見てみると、信頼性の低い例として、S の?キーワードが「組み込み」の場合、パス A では A2 の?論文で「内分泌系モデルを組み込んだロボットハードウェア」より A3 の?教職員で「組み込み」に関係のある教授が抽出されたが、この教授の専門分野である「人工神経回路モデル」から A4 に「神経」が入り、結果としてたどり着いた G の?教職員は「組み込み」と関係が薄いと思われる生体学の教授であった。また、S の?キーワードが「エコロジ」の場合、パス B では環境教育等を教える教授が B3 の?教職員で抽出され、この教授の専門分野である「環境教育」から B4 に「環境」と「教育」が入り、結果として「エコロジ」と関係が薄いと思われる「音環境理解」を専門

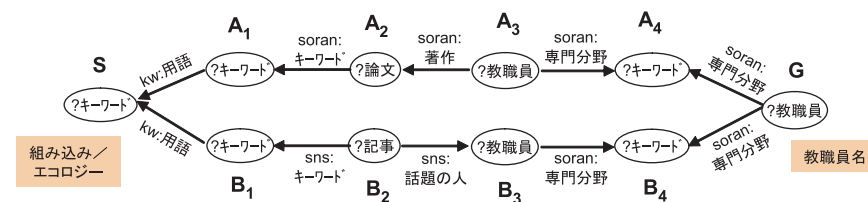


図 10 クエリグラフパターン例 1

Fig. 10 A sample result of the selected query graph patterns.

表 3 図 10 のパス A、パス B とクエリグラフパターンの解の数

Table 3 Number of the results of path A, path B and the query graph pattern of the Fig. 10.

	「組み込み」の解	「エコロジ」の解
パス A	22	14
パス B	33	17
パス A ∪ B	12	9

とする教授や、「言語教育」を専門とする教授が G の?教職員としてあがった。しかし、これら 2 本のパスを組み合わせた結果として、たとえば S の?キーワードが「組み込み」の場合、A4 と B4 の?キーワードには A3 と B3 の教授の専門分野である「集積回路」と「アーキテクチャ」がそれぞれ入ることにより G の?教職員には低消費電力設計やモバイル端末で利用可能なリコンフィギャラブルデバイスを研究されており「組み込み」と関係の深い教授があがった。

ここでは「教育」等の一般的な用語から 1 本パスでは解が発散してしまっているが、パスを 2 本にすることにより S の?キーワードに関係の深い教授を探索できている。

5.1.2 クエリグラフパターン例 2

図 11 に示すクエリグラフパターンは「LA-1-LC-2」と「SC-2-LB-1」(図 8 参照)の異なるパス 2 本からなるものである。以下、図 11 で上のパスを「パス C」、下のパスを「パス D」と呼ぶ。表 4 に示したように、5.1.1 項の例と同じくパスが 2 本になることにより解の数が絞られている。

検索キーワードと解の関係をいくつか詳しく見てみると、信頼性の低い例として、S の?キーワードが「組み込み」の場合、パス C では C2 の?論文で「組み込みシステムにおける複数のフリーリストに割り振るメモリ量の最適化」より、C3 の?キーワードで「メモリ」が抽出されたため、C4 の?論文で「形態と位置のワーキングメモリモデル」を執筆した「組み

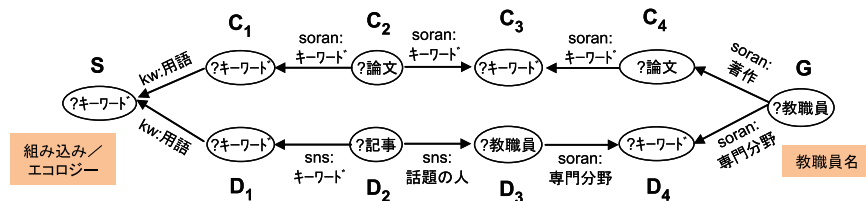


図 11 クエリグラフパターン例 2

Fig. 11 Another sample result of the selected query graph patterns.

表 4 図 11 のパス C, パス D とクエリグラフパターンの解の数

Table 4 Number of the results of path C, path D and the query graph pattern of the Fig. 11.

	「組み込み」の解	「エコロジ」の解
パス C	85	43
パス D	33	17
パス $C \cup D$	24	10

込み」と関係が薄いと思われる生体学の教授が G の?教職員としてあがった。

また、S の?キーワードが「エコロジ」の場合、パス D では環境デザイン等を教える教授が D3 の?教職員で抽出され、この教授の専門分野である「知能情報学」から D4 の?キーワードに「情報学」が入り、結果として「エコロジ」と関係が薄いと思われる「社会情報学」を専門とする教授が G の?教職員としてあがった。しかし、これら 2 本のパスを組み合わせた結果として、たとえば S の?キーワードが「エコロジ」の場合、C3 と D4 の?キーワードがそれぞれ「学習」と「環境教育」が入ることにより G の?教職員には環境教育支援や放牧管理の研究をされており「エコロジ」と関係の深い教授があがった。

ここでも「情報学」等の一般的な用語から発散した解を 2 本パスで絞り込めている。

これらの結果より、1 本パスでは発散し信頼性が低くなってしまった解を 2 本パスでより関係の深い解に絞り込む効果があると考えられる。

5.1.3 解の信頼性

本手法では再現率ではなく精度を重視した。産学連携先を探す場合を例にとると、たとえ正解が含まれていても膨大な数の連携先候補が抽出されることは、個別に調査やコンタクトすることを考えると望ましくないためである。

たとえば考察 5.1.1 項に示したクエリグラフパターンで再現率と精度を求めた結果、パス A では精度が 27.3%，再現率が 54.5%，パス B では精度が 21.2%，再現率が 72.7%であり、

この 2 つのパスを組み合わせたクエリグラフパターンでは精度が 50.0%，再現率が 54.5%であった。1 本パスと比較して 2 本パスの方が再現率は低いものの、たとえばパス B で 33 名出てきた解のうち、7 名が正解であるよりも、2 本パスで 12 名出てきた解のうち、6 名が正解である方が良いと考える。なお、本値での正解の定義は Google に「教職員名 京都大学 組み込み」と入力し（「教職員名」の部分は実際の教職員名に置き換え）、上位 10 件以内に人が確認してその教職員と「組み込み」に関連が見られるものとした。人手を介した理由は、検索結果上位に研究会の発表リストが多く、その中にはたとえば「組み込み」に関する発表リストとともに別分野で発表する該当教職員が同ページに記載されているものがあり、このようなケースを除外するためである。

5.2 仮説 (2) の検証

3.1 節であげた仮説 (2) より、絞り込んだクエリグラフパターンはバリエーションが確保できているのか、図 8 に示したサブグラフパターンの分類を用いて検証する。

様々な側面のバリエーションが存在するが、本手法のアルゴリズムより解のバリエーション（網羅率）とパスのレベルで見た際の構造のバリエーションに関しては確保できているため、ここでは意味、つまりサブグラフパターンの両端のクラスの間に存在する関係のバリエーションが確保できているかについて考察する。

図 8 に示した SA グループは SNS の記事や人を介した、LA グループは研究データを介した（検索）キーワードの言い換えであるといえる。同様に SB, LB グループは記事、人、研究データを介した“人の置き換え”，つまり“検索解の言い換え”，SC, LC グループは直接または記事、人、研究データを介した“キーワードと人のつながり”であるといえる。このことより、それぞれのグループ内のサブグラフパターンは意味的に類似しており、それぞれのグループから最低 1 つを代表とすれば、すべての意味的バリエーションも残せたと考える。

本手法では 5.1 節であげたように中間に存在するインスタンスを見てたどり着いた解までの関係を確認することができるため、解を採用するかどうかの最終的な判断はユーザが下すことが可能である。また上記であげたように、バリエーションを残しつつクエリグラフパターンの数は絞り込めていることから、ユーザは信頼性の高いクエリグラフパターンを効率的に探すことが可能になったと考える。

5.3 計算量

本手法と同様にクエリグラフパターンの絞り込みを行う別の手法との計算量の比較を試みる。比較手法は各クエリグラフパターンの解を全探索したうえで、たとえば頻出等の解の

分布特性を用いて上位や一部のクエリグラフパターンに絞り込むものとする。これらの手法を、今回の実験でも多くみられた両端のクラスとその中間に存在するクラスのうち1つ、合計3つのクラスがハブとなるようなクエリグラフパターンを対象にして計算量を比較する。ハブとなるようなクラス（以下、ハブクラス）とはインスタンスの次数がべき乗分布を示すクラスのこと、クエリグラフパターン内でのハブクラスの個数と位置が計算量を決めることとなる。

比較手法ではハブクラスがクエリグラフパターンの両端に1つずつと中間に1つ存在し、中間に存在するハブクラスに関してはそのハブクラスへのinのリンクとoutのリンクの2回計算が必要となるため、Nをインスタンスから接続するリンク数の最大値とすると、計算量の理論的オーダは $O(N^4)$ である。それに対し、本手法は中間に存在するハブクラスで分割し、分割後のそれぞれのサブグラフパターンには両端にハブクラスが存在するため計算量の理論的オーダは $O(N^2)$ の足し算となり、全体として $O(N^2)$ となる。なお、Nが大きい場合は、特定の期間に存在したデータのみを対象にする等のサンプリングに基づいてNを減らすことにより、計算量を減らすことも可能と考える。

5.4 関連研究との比較

ソーシャルネットワーク分析がさかんになるなか、グラフ探索が問題となっている。グラフ探索では組合せ爆発が発生するため、クエリグラフパターンを効率的に抽出することが主な課題である。多くのソーシャルネットワークはRDFデータとは異なりアークのラベルがない無向グラフというシンプルなグラフであるため、RDFデータのクエリグラフパターン探索とは質が異なるが、その中でクエリグラフパターン抽出法の代表的なものとしてランダムウォークと呼ばれる、各アークにスコアを振り検索範囲を制御する方法^{15)–17)}がある。しかし、この方法ではスコアを誰がどのようなルールに基づいて振るのが課題となる。RDFデータを対象としたクエリグラフパターンの絞り込みを行っているものとしてはクエリグラフパターンの長さに制限を設けることで絞り込む方法¹⁸⁾や、希少なパスに価値があると考え、類似するパス数の逆数でランキングを行う方法¹⁹⁾や、ユーザが関心のあるドメインに重み付けを行い、重みが高いクエリグラフパターンを抽出する方法^{20),21)}がある。しかし、長さで絞り込む方法では、長さの情報からは“?start”と“?goal”の関係が近いとはいえるが、それだけで信頼性が保障できるかは不確実であり、希少なパスを用いる方法では、目新しい関係を炙り出すことができそうだが、同じく信頼性が保障できるかは不確実である。また、重みを用いる方法では、ドメインごとに異なるであろう重みをどのようなルールで誰が付与するかが課題となる。本研究では信頼性を確保しつつ、ドメインにかかわ

らず同じ方法でクエリグラフパターンを絞り込むことが可能である。

また、実験で用いた数万トリプルレベルのRDFデータよりさらに巨大なものに本手法を用いようとした際、本手法自体のスケールビリティに関しては5.3節で触れたサンプリングを用いることにより確保ができると考えるが、本手法で絞り込んだクエリグラフパターンを用いた検索を実行するためには、処理の高速化をクラウド上の分散処理で実現する方式²²⁾や複雑なSPARQLクエリを最適化する方式²³⁾と組み合わせるとよいと考える。

5.5 産学連携への適用

産学連携先を探す際の問題となっているのが、産業界と大学での用語の違いである。たとえば産業界では技術の適応法である「組み込み」は、大学ではその要素技術としての「集積回路」であるように、産業界では実用化寄り、大学では基礎研究寄りのキーワードを利用するため、通常の異表記語や同義語の辞書ではカバーできる範囲の揺れではなく、マッチングが非常に難しい。今回の実験例でもあるように、「組み込み」に関する要素技術は京都大学でも実際に研究が行われているものの、京都大学が産学連携先検索用に用意した「京都大学総覧DB」で検索しても、たとえば「組み込み」で情報学研究科の教職員が2名、「エコロジ」で1名しか検索できず、産業界から問合せがあっても該当教職員にたどり着くのが困難な状況である。今回の実験のように、たとえばSNSのようなカジュアルで言い換えが多そうな情報と論文データのような正確な情報をLinked Dataとしてつなげるによりブリッジが形成されて産業界とつながり、本実験で絞り込んだクエリグラフパターンを用いることにより大学内での組み込みの要素技術に関する情報が産業界からも参照できるようになれば、産学連携先を探す際の問題も解決できると考える。

5.6 他のRDFデータへの本手法の適応に関して

本手法を適用するための条件は、個々のインスタンスに対してクラスが定義されたRDFデータの集合から異なるデータセットを対象とすることと、両データセットの共通のクラスのインスタンスを通して各インスタンス（グラフ構造におけるノード）がつながっていることである。今回の実験データで抽出されたクエリグラフパターンは、SNSデータセットとレガシーデータセットをつないでいるキーワードクラスと教職員クラスのインスタンスを仲介したものであった。本実験のデータのスキーマ構造を図7で示しているが、今回実験したキーワードクラスと教職員クラス以外のクラス、たとえば論文クラスや学生クラスを両端とするクエリグラフパターンにおいても、その中間にハブでありデータセットをつないでいるインスタンスが含まれている構造であるため、すべてのクラスの組合せにおいて本手法を適用することが可能である。

しかし、さらに巨大で複雑な Linked Data で同姓同名や表記揺れ等の名寄せに起因するリンク間違いの問題、また、本実験では信頼性の高い情報が信頼性の低い情報より多く含まれていたため、2本のパスにすることにより解の信頼性が確保できたが、より信頼性の低いデータが多く混在しているケースへの本手法の適応に関しては、今後の課題である。

6. ま と め

巨大な RDF データから 2 点間の関係に着目して探索する場合、膨大なクエリグラフパターンが生成されてしまうという問題がある。これらのクエリからデータ間の関係の信頼性を確保しつつ数を絞り込むことにより効率的な探索を行うという課題を解く方法を提案した。

KnowWho システムを想定した実データでの実験により、全クエリグラフパターン数の 9.1%まで絞り込むことができた。また絞り込んだクエリでは、データにたどり着くパスを 2 本としたことにより、1 本では発散してしまった解をより関係の深い解に絞り込めることが検証できた。抽出できたクエリグラフパターンでは論文や記事等を介した語句の言い換えも可能であるため、産学間での用語にギャップがある産学連携先検索に活用できることを示した。

謝辞 本研究の一部は京都大学グローバル COE プログラム「知識循環社会のための情報学教育研究拠点」による。

参 考 文 献

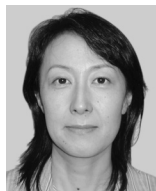
- 1) 武田英明：セマンティック Web と Linked Data , 技術研究報告 , SWIM2008-22, pp.25-28, 電子情報通信学会 (2008).
- 2) wiki.dbpedia.org: About. <http://wiki.dbpedia.org/About>
- 3) GeoNames Ontology – Geo Semantic Web. <http://www.geonames.org/ontology/>
- 4) DBTune – Myspace RDF Service. <http://dbtune.org/myspace/>
- 5) Tabulator: Generic data browser. <http://www.w3.org/2005/ajar/tab>
- 6) Marbles Linked Data Engine. <http://marbles.sourceforge.net/>
- 7) SPARQL Query Language for RDF. <http://www.w3.org/TR/rdf-sparql-query/>
- 8) Sato, H., Iiduka, K., Mukaigaito, T. and Murayama, T.: Finding Similarity and Comparability from Merged Hetero Data of the Semantic Web by Using Graph Pattern Matching, *WWW2005 Workshop, Activities on Semantic Web Technologies in Japan* (2005).
- 9) 飯塚京士, 佐藤宏之, イコ・プラムディオノ, 村山隆彦: RDF データを対象としたグ

ラフ検索におけるクエリ生成方式の検討, 人工知能学会 SIG-SWO-A502-08 (2005).

- 10) Sato, H., Pramudiono, I., Iiduka, K. and Murayama, T.: Automatic RDF Query Generation from Person Related Heterogeneous Data, *IWeb2006 Workshop on Information Integration on the Web in Conjunction with WWW2006* (2006).
- 11) 酒井理江, 飯塚京士, 山本具英, 大友健治, 坂本 啓, 佐藤宏之, 村山隆彦, 服部宏充, 石田 亨: SNS と大学成果情報の統合による産学連携マッチング支援の有効性, セマンティック Web コンファレンス 2009 (2009). <http://s-web.sfc.keio.ac.jp/conference2009/0405-sakai.pdf>
- 12) 京都大学研究者総覧データベース. <http://www.adm.kyoto-u.ac.jp/infolib/meta/help.html>
- 13) CiNii-NII 論文情報ナビゲータ. <http://ci.nii.ac.jp/>
- 14) 連携推進ソーシャルネットワークサービス (SNS), 京都大学 ICT 連携推進ネットワーク. <http://ict-nw.i.kyoto-u.ac.jp/sns/index.php>
- 15) Faloutsos, C., McCurley, K.S. and Tomkins, A.: Fast discovery of connection subgraphs, *KDD*, pp.118-127 (2004).
- 16) Tong, H. and Faloutsos, C.: Center-Piece Subgraphs: Problem Definition and Fast Solutions, *Proc. 12th ACM SIGKDD Conference*, pp.404-413 (2006).
- 17) Koren, Y., North, S.C. and Volinsky, C.: Measuring and extracting proximity in networks, *Proc. 12th ACM SIGKDD Conference*, pp.245-255 (2006).
- 18) Auer, S. and Lehmann, J.: What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content, *Proc. 4th ESWC*, pp.503-517 (2007).
- 19) Lin, S. and Chalupsky, H.: Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis, *Proc. 3rd IEEE International Conference on Data Mining*, pp.171-178 (2003).
- 20) Aleman-Meza, B., Halaschek-Wiener, C., Arpinar, I.B. and Sheth, A.P.: Context-Aware Semantic Association Ranking, *Proc. 1st SWDB*, pp.33-50 (2003).
- 21) Anyanwu, K., Maduko, A. and Sheth, A.P.: SemRank: ranking complex relationship search results on the semantic web, *Proc. 14th WWW2005*, pp.117-127 (2005).
- 22) Weaver, J. and Williams, G.T.: Scalable RDF query processing on clusters and supercomputers, *Proc. 5th International Workshop on Scalable Semantic Web Knowledge Base Systems*, pp.81-93 (2009).
- 23) Lampo, T., Ruckhaus, E., Sierra, J., Vidal, M.-E. and Martinez, A.: OneQL: An Ontology-based Architecture to Efficiently Query Resources on the Semantic Web, *Proc. 5th International Workshop on Scalable Semantic Web Knowledge Base Systems*, pp.65-80 (2009).

(平成 22 年 4 月 16 日受付)

(平成 22 年 9 月 17 日採録)



酒井 理江

1994 年パデュー大学電気工学科卒業。1996 年同大学大学院電気工学科修士課程修了。同年日本電信電話株式会社入社。NTT 情報流通プラットフォーム研究所研究員。以来、Web サービス、セマンティック Web 等の研究開発に従事。



飯塚 京士 (正会員)

1993 年名古屋大学大学院理学研究科数学専攻博士前期課程修了。同年日本電信電話株式会社入社。NTT 情報流通プラットフォーム研究所研究員。セマンティック Web を用いた情報共有基盤の研究開発に従事。人工知能学会員。2005 年度人工知能学会研究会優秀賞受賞。



佐藤 宏之 (正会員)

1996 年慶應義塾大学大学院政策・メディア研究科修士課程修了。同年日本電信電話株式会社入社。以来、知識共有支援、セマンティック Web を利用した情報統合、ログマイニング等の研究開発に従事。ICCE98 Outstanding Paper Award 受賞、2005 年度人工知能学会研究会優秀賞受賞、i-Society2006 Best Paper Award 受賞。セマンティック Web 委員会委員。筑波大学より博士 (情報学)。2010 年度より慶應義塾大学環境情報学部講師 (非常勤)。



村山 隆彦

1984 年東北大学工学部通信工学科卒業。1986 年同大学大学院工学研究科情報工学専攻修士課程修了。同年 NTT 入社。以来、知識工学、エージェント、Web サービス、セマンティック Web 等の研究開発に従事。2004 ~ 2010 年電気通信大学大学院情報システム学研究科客員准教授。電子情報通信学会、日本データベース学会各会員。



小林 透 (正会員)

1985 年東北大学工学部精密機械工学科卒業。1987 年同大学大学院工学研究科修士課程修了。同年 NTT 入社。以来、ソフトウェア生産技術、ユビキタスコンピューティング、情報セキュリティ、セマンティック Web、データマイニング等の研究開発に従事。現在、NTT サイバーソリューション研究所主幹研究員。電子情報通信学会会員。



服部 宏充 (正会員)

2004 年名古屋工業大学大学院博士課程修了。2004 年学術振興会特別研究員 (PD)。リバプール大学、マサチューセッツ工科大学客員研究員を経て、現在、京都大学大学院情報学研究科助教。工学博士。電子情報通信学会、人工知能学会各会員。マルチエージェントシミュレーション、行動モデリングに興味を持つ。



石田 亨 (フェロー)

京都大学情報学研究科社会情報学専攻教授。ミュンヘン工科大学、パリ第六大学、メリーランド大学、上海交通大学、清華大学客員教授等を経験。工学博士。情報処理学会、電子情報通信学会、IEEE フェロー。マルチエージェントシステム、フィールド情報学等の研究に取り組む。デジタルシティ、言語グリッド、異文化コラボレーション等のプロジェクトを推進。