1N3-3

SPARQL 生成支援のための RDF グラフ構造解析技術の開発

Structural analysis of RDF graph for assisting in writing a SPARQL query

山口 敦子[*1] 小林 紀郎[*2] 戀津 魁[*2] 山本 泰智[*1] 古崎 晃司[*3]
Atsuko Yamaguchi Norio Kobayashi Kai Lenz Yasunori Yamamoto Kouji Kozaki

[*1] 情報・システム研究機構 ライフサイエンス統合データベースセンター
Database Center for Life Science, Research Organization of Information and Systems

[*2] 理化学研究所 情報基盤センター
Advanced Center for Computing and Communication, RIKEN

[*3] 大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

Many databases in life science are provided in Resource Description Framework (RDF) model. However, it may be difficult for users who are not familiar with Semantic Web technologies to understand RDF datasets and write a semantic query for them. In this study, for such users, we propose a human-understandable representation of RDF datasets based on class-class relationships appearing in the datasets. The two technologies for the representation are introduced: a labeled multi graph named class graph to display class-class relationships and an RDF specification named SPARQL Builder metadata to obtain and store required metadata for construction of a class graph. In addition, as a practical application, we introduce the SPARQL Builder system, which assists users in writing semantic queries for RDF datasets.

1. 背景

多種多様かつ膨大なデータを統合的に扱うための基盤技術として，生命科学分野ではセマンティックウェブ技術の利用がすすめられつつある．特に，Resource Description Framework (RDF)と呼ばれる，セマンティックウェブ技術における標準データモデルを採用するデータベースは年々増加の一途を辿っている[Redaschi2009, Belleau2008, Jupp2014, Fu2015]. RDF 化された生命科学データベースを一般の生物学研究者が有効活用できるようにするためには，研究者の要求に沿ったデータを柔軟に取得できることが必要である．しかしながら，RDF のグラフ構造はデータベース毎に異なるため，セマンティックウェブ技術に習熟していないユーザが RDF データの構造や仕様を理解し，自分の欲しいデータを取得することは難しい．

そこで，著者らは RDF データの構造を把握するための表現として，プロパティを介したクラス間関係に着目した．本発表では，クラス間関係を計算し提示するためのデータ構造であるクラスグラフ，および，クラスグラフを構築するため事前に RDF データから SPARQL エンドポイントを通して取得蓄積することを想定したメタデータの設計について述べる．さらに，クラスグラフおよびメタデータを基盤としたクラス間関係提示の応用例として，RDF に対する標準クエリ言語 SPARQL のクエリ生成支援システム SPARQL Builder を紹介する．

2. クラス間関係提示

RDF はセマンティックウェブ技術における標準データモデルである．RDF モデルを用いることによって，各データベースはデータとその関係からなるグラフをなし，さらに，多数のデータベースは互いにつながって大きなグラフをなす．ユーザが欲しいデータを RDF 化されたデータベースから適切に抽出するためには，なんらかの形でそのデータベースに内在するつながりをユーザに提示して理解させ，そのつながりを用いたデータ検索を行うことが望ましい．しかしながら，すべてのデータ間のグラフ内でのつながりを提示することは，一定以上の大きさのデータベースでは現実的ではない．そこで，本研究では，データ間関係を，そのデータが属するクラスでまとめ，クラス間関係を利用したデータベース構造の提示と，クラス間関係に基づいたデータ取得を提案する．

連絡先：山口敦子，atsuko@dbcls.rois.ac.jp

2.1 クラスグラフ

クラス間関係を効率的に取り扱うために，特にクラス間パスの計算や提示のため，クラスグラフという構造を利用する．クラスグラフとは，クラスを頂点，プロパティを辺とするラベル付き有向グラフであり，厳密には以下のように定義される: R を RDF データセットとし，C を R に含まれるクラスの集合，P を R に含まれるプロパティの集合とする．このとき，R に対するクラスグラフはラベル付き有向グラフ $G_R = (V, E, c, p)$である．ただし，Vは大きさ$|C|$の頂点集合，c は Vから C への一対一関数である．E は $V \times V$ 上の多重集合であり，p は Eから P への関数であり，Rから次のように構成される: 2 つのクラス $class_d$, $class_r$, およびプロパティ $prop$ について，(条件 1) 2 つのトリプル($prop$, rdfs:domain, $class_d$), ($prop$, rdfs:range, $class_r$)が Rに含まれる，(条件 2) 3 つのトリプル(s, $prop$, o), (s rdf:type $class_d$), (o rdf:type $class_r$)が Rに含まれる，のいずれかが成り立つとき，またそのときに限って，頂点 $v = c^{-1}(class_d)$から $u = c^{-1}(class_d)$の辺 e_{prop}を E に加え，$p(e_{prop}) = prop$ とする．

任意のクラス間関係はクラスグラフ上のパスによって表現することができる．ユーザが望むクラス間関係をクラスパスによって提示できる可能性を高くするためには，クラスパスをできるだけ多く計算することが望ましい．しかしながら，通常のパス探索問題と違い，クラスグラフは多重辺グラフであり，さらにクラスパスも一般に単純パスではないため，パスの長さを増やすにつれてパス数が爆発的に増大する．そのため，クラスグラフを無向でシン

プルなグラフへ，辺の方向とラベルを取り払い，多重辺は一つの辺へと束ねられたものとすることで変換し，その後，変換後のグラフ上でパスを探索し，探索によって得られたパスに対し，多重辺のラベルを組み合わせで付加することで，ラベル付き多重辺へ戻すという手法をとった．この手法を使うことにより，パス探索の計算時間は大きく改良され，その結果，より多くのクラスパスを提示することが可能となった．

3. メタデータ設計と利用

クラス間関係の提示のためには，クラスグラフ構築に必要な情報を SPARQL エンドポイントから取り出す必要がある．当初は必要な情報を必要なだけ動的に SPARQL エンドポイントから取り出すことも検討したが，現実的な時間でクラスグラフを構築するためには事前に抽出し蓄積する方法が妥当であるという結論に至った．そのため，クラスグラフ構築に必要なデータを事前に SPARQL エンドポイントに過剰な負担をかけずに取得できることが望ましい．その目的のもと，取得すべきメタデータを洗い上げてスキーマを設計し，さらにそれらのメタデータを取得するための SPARQL 文を定義した．

設計したメタデータのスキーマ仕様(SPARQL Builder Metadata)について，大まかには，SPARQL エンドポイント→エンドポイントに含まれるデータセット→データセットのメタデータの階層構造になっており，各メタデータ部分にはクラスリスト，プロパティリスト，クラス-プロパティ-クラス関係，さらにそれらに関連するインスタンス数やトリプル数等の統計情報が含まれる．詳しくは，SPARQL Builder Metadata Version Sep. 2015 (http://www.sparqlbuilder.org/doc/sbm_2015sep/)を参照されたい．

4. SPARQL Builder

SPARQL Builder (http://sparqlbuilder.org/)とは，SPARQL 言語の知識がなくとも，また，対象データセットの構造を知らなくても，クラス間関係提示を用いた対話的な GUI を介して SPARQL クエリを生成することができることを目指して開発されたウェブ上のサービスである[Yamaguchi2014]．ユーザは，まず，入力クラスと出力クラスをそれぞれクラスのリストから選ぶ．たとえば，ユーザがタンパク質のリストを持っており，それらと代謝経路の関係に興味がある場合は，入力クラスとして Protein，出力クラスとして Pathway を選ぶことになる．入出力の二つのクラスが確定すると，クラスグラフにおけるクラスパスが計算され，その結果を利用して，データ内に含まれるクラス間関係のリストがユーザに提示される．ユーザがクラス間関係を一つ選ぶと，そのパスのクラス間関係に対応する SPARQL クエリが自動生成される．

図 1 は SPARQL Builder のシステム構成の概要である．事前に対象となる SPARQL エンドポイントからクラスのリストやクラス間関係，さらに，インスタンス数やトリプル数などの統計情報など，必要なメタデータを取得し格納しておく(1)．ユーザが GUI からシステムにアクセスすると，メタデータからクラスのリストが取り出され(2)，ウェブ API を通じて(3)，GUI 上に提示される．ユーザがクラスのリストから入出力クラスを選択すると，メタデータから作られたクラス間関係を表すグラフであるクラスグラフを用いてクラス間パスが計算され(4)，クラス間パスのリストが GUI 上に提示される．ユーザがパスをひとつ選ぶと，パスから生成した SPARQL クエリが GUI 上に提示される．システム概要からわかるように，システムの鍵となるのは，先述したクラスグラフの構築およびクラスグラフ構築に必要なメタデータの取得となる．本システムでは，2016 年 3 月現在，38 のデータセットに対するメタデータを取得蓄積し，これらのデータセットに対するクエリ生成支援サービスとして運用している．

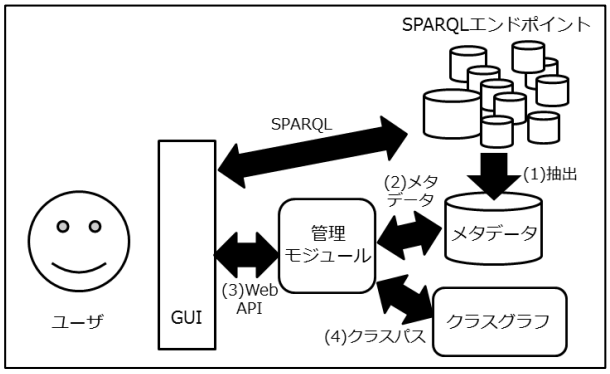

図 1 SPARQL Builder システム概要

5. まとめ

RDF データセットのデータ間関係を，そのデータが属するクラスでまとめることで，クラス間関係を利用した RDF データ構造の提示を提案した．さらにクラスグラフ構築の技術を開発し，メタデータの設計を行った．さらに，これらの技術に基づいて，SPARQL クエリ生成支援システム SPARQL Builder を開発し，サービスを運用している．

今後の課題として，クラスパスのランキングやメタデータ取得方法の改良を行いたい．そして，複数の SPARQL エンドポイントを利用したフェデレート検索についても対応していきたい．

謝辞

本研究は独立行政法人科学技術振興機構(JST)、バイオサイエンスデータベースセンター (NBDC) の助成および科学研究費補助金基盤研究(B) 25280081 の助成を受けて行った．

参考文献

[Redaschi2009] Redaschi, N. and Consortium UniProt: UniProt in RDF: Tackling Data Integration and Distributed Annotation with the Semantic Web. Nature Precedings, <http://dx.doi.org/10.1038/npre.2009.3193.1> (2009).

[Belleau2008] Belleau, F., Nolin, M. A., Tourigny, N., Rigault, P., Morissette J.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J. Biomed. Inform. 41(5), 706-716 (2008).

[Jupp2014] Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S. M., Martin, M., Le Novére, N., Parkinson, H., Birney, E., Jenkinson, A. M.: The EBI RDF platform: linked open data for the life sciences. Bioinformatics 30(9), 1338-1339 (2014).

[Fu2015] Fu, G., Batchelor, C., Dumontier, M., Hastings, J., Willighagen E., and Bolton, E.: PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. Journal of Cheminformatics, 7(34), doi:10.1186/s13321-015-0084-4 (2015).

[Yamaguchi2014] Yamaguchi, A., Kozaki, K., Lenz, K., Wu, H., Kobayashi, N.: An Intelligent SPARQL Query Builder for Exploration of Various Life-science Databases, CEUR Workshop Proceedings 1279, The 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014), Riva del Garda, Italy