1G3-OS-08b-2

Linked Open Data のための SPARQL クエリ共有システムの提案

Proposal of SPARQL Query Sharing System for Linked Open Data

濱崎 雅弘 *1
Masahiro Hamasaki

*1産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

We propose a SPARQL query sharing system for Linked Open Data. Recently, many and various datasets are published as a Linked Open Data (LOD). SPARQL is an RDF query language and it provides a powerful way to access LOD. However, it is not easy to utilize them because it requires not only techniques of SPARQL but also knowledge of datasets and vocabularies that they used for users. Therefore, we propose sharing SPARQL query as a solution of this problem. Sharing SPARQL query is a simple solution but it has an important role for LOD. In this paper, we analyze access logs of a SPARQL endpoint to show a difficulty of utilizing LOD and introduce our prototype system for sharing SPARQL query.

1. はじめに

本稿では，特にデータセットに対して知識が不十分なユーザを支援対象とした，Linked Open Data (LOD) のための SPARQL クエリ共有システムを提案する．LOD は様々な応用が期待されるが，膨大かつ多様なデータセットであるため，適切なクエリを作成するのは容易ではない．これは SPARQL や LOD の初学者だけの話ではなく，熟練者にとっても起こりうる問題である．なぜなら LOD は既存のデータセットと異なりオープンな系であるため，常に自分から見て未知のデータセットが存在しうるためである．つまり LOD の熟練者であっても，あるデータセットに対して初心者であるという状況が常に起こる．そこで提案システムでは，SPARQL クエリを共有し検索・推薦可能にすることで，一部の熟練ユーザが作成した SPARQL クエリの多くのユーザが利活用できるようにする．

Linked Open Data (LOD) はオープンライセンスの元で公開された Linked Data である．LOD は機械可読なデータであるが，データが作られた出自を考えると，元のデータを Web データ化することが目的であり，ある特定のサービスで便利になるように作られたものではない．一般にデータベースは，あるサービスのために構造が定義（スキーマ設計）され，構造化データが蓄積される．そのため，サービスからのデータ利用は容易である．しかし LOD はデータ利用者ではなくデータ所有者（厳密にはデータそのもの）の都合に合わせて構造が定義され，構造化データが蓄積される．もちろん，様々な利用可能性が開かれているという点で，特定の利用者のためにデータ構造が設計されるのではなく，元データをより適切に共有するためにデータ構造が設計されることは正しい．しかしながらこの特徴ゆえに，Linked Open Data はアクセス可能・検索可能になっただけでは利活用は容易であるとはいえない．サービスにとって適した構造化がなされているとは限らないため，ユーザはデータセットの構造をよく理解しクエリを工夫して作成しなくてはならない．利用者によるデータセットの理解を支援する目的で，データセットに関するメタデータの提案や公開も進められている．そしてデータセットを理解するための手段の一つとして，SPARQL クエリによる検索が挙げられる．

検索，とくに検索クエリの作成が困難な場合に，支援を行う技術は数多く提案されている．多くは検索対象となるデータをシステム側が把握することで，ユーザが目的とするクエリを予測し補完やサジェストを行うが，LOD はその性質上，検索対象となるデータをシステム側が把握することは困難である．また，クエリの補完やサジェストはユーザの目的を予測することで成立するが，検索ニーズは必ずしも予想可能なものだけではない．探索的検索 [Marchionini 06] のような，LOD の新しい利活用方法を発見するような検索行為を支援するには，予測というアプローチでは難しい．

そこで本研究では，SPARQL クエリをユーザ間で共有するアプローチを提案する．LOD のデータ構造が，データ所有者によって創発的に作られるように，それを利活用する SPARQL クエリもまた創発的に作ることを可能にする．SPARQL クエリはエンドユーザが自由に作成できるものであるが，ここでいう創発的に作るとは，ユーザ間で SPARQL クエリを共有することで，SPARQL クエリの再編集や派生クエリの作成によるクエリの洗練，拡張，発展を指す．

本稿の構成は以下の通りである．まず 2 章にて，DBpedia Japanese の SPARQL クエリログ分析から，ユーザによる SPAQRL 検索の現状とクエリ共有の可能性について調べる．次に 3 章にて，提案システムの概要とプロトタイピングした Web アプリケーションを紹介する．4 章にて，SPARQL クエリ推薦に必要な SPARQL クエリ間の類似度について議論する．5 章にて関連研究を述べ，6 章にて本稿をまとめる．

2. 日本語 DBpedia のクエリログ分析

2.1 データの概要

本章では，クエリ共有のための予備調査として，DBpedia Japanese *1 のログを分析する．DBpedia Japanese は日本版 DBpedia であり [Kato 13]，国内 LOD のハブ的存在である [加藤 14]．今回対象とするのは 2013 年 6 月から 2015 年 1 月までの Web サーバへのアクセスログであり，その中でも GET リクエスト約 1300 万を分析する．POST リクエストは除外するが，これはリクエスト全体に占める割合は大きくない

連絡先: 濱崎雅弘，産業技術総合研究所，〒 305-8568 茨城県つくば市梅園 1-1-1 中央第二事業所，Tel: 029-861-3885，masahiro.hamasaki@aist.go.jp

∗1 http://ja.dbpedia.org

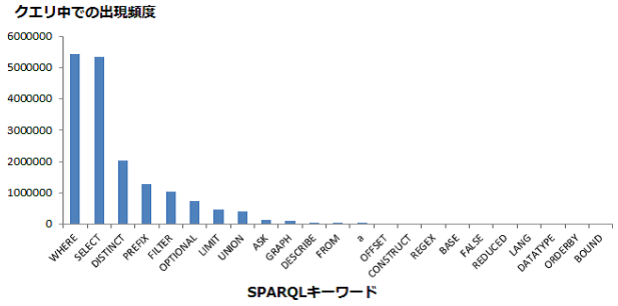

図 1: SPARQL クエリ中に現れる SPARQL キーワードの出現頻度（縦軸：出現頻度，横軸：出現頻度の順位）

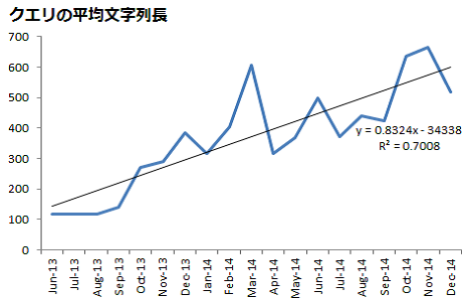

図 2: SPARQL クエリの平均長の時間変化．直線と数式は線形近似の結果．（縦軸：クエリの長さの平均値，横軸：年月）

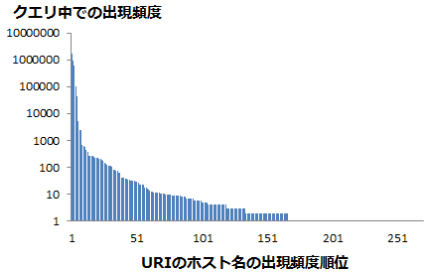

図 3: SPARQL クエリ中に現れる URI のホスト名の出現頻度（縦軸：出現頻度，横軸：出現頻度の順位）

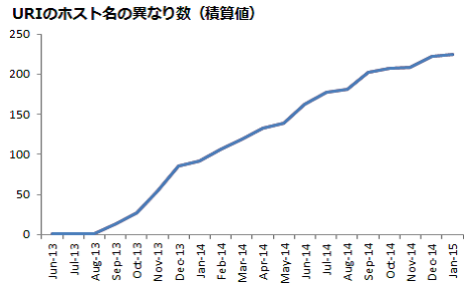

図 4: SPARQL クエリ中に現れる URI のホスト名が初めて出現した年月（縦軸：ホスト名の異なり数の累計，横軸：ホスト名の初出年月）

ため（2014 年 1 月以降は常に全体の 1〜3%程度），分析に影響は与えないものと考える．

GET リクエストのログには，アクセス IP，アクセス日時，ブラウザエージェント，リクエストパラメータが含まれている．分析にあたっては，まずブラウザエージェントの名前をもとに Bot によるアクセスログを削除し，次にリクエストパラメータに埋め込まれた SPARQL クエリを抽出した．以上の手順により GET リクエスト 13,213,311 件のうち，SPARQL リクエスト 5,486,190 件を取得した．

2.2 クエリログ分析

図 1 は SPARQL キーワードの全 SPARQL クエリにおける出現頻度を示したものである．WHERE と SELECT がほとんどのクエリにおいて出現するが，これら以外は大幅に出現頻度が減る．ユニークなアクセス IP 一つを 1 ユーザと見立てた場合，ユーザが 1 回でも利用したことがある SPARQL キーワードの数は平均 3.3 語であった．多くのユーザは，まだ SPARQL クエリの文法における学習の余地があると考えられる．

図 2 はクエリの文字列長の月ごとの平均値をプロットしたものである．なお，2015 年 1 月は月の途中までしかログがなかったため除外している．クエリの文字列長は増加傾向にあることがわかる．クエリの長さとクエリの複雑さに相関があると仮定すると，時間経過とともにデータセットが理解され，より複雑なクエリが投げられるようになっていると解釈できる．

図 3 は SPARQL クエリ中に書かれた URI のホスト名の出現頻度を片対数グラフで示したものである．縦軸は出現頻度，横軸は出現頻度順に並べたときの順位である．図 4 はこれらのホスト名が時間経過とともに増えていく様子を示したものである．縦軸がホスト名の異なり数，横軸がそのホスト名を含む URI が初めて SPARQL クエリ中に出現した年月を示している．さまざまな外部データセットとのリンクを生み出す点で重要な情報であるが，これも SPARQL キーワード同様，一部のホスト（にあるリソース）が頻繁に参照され，そうでないものが多数存在する．また，そういった外部データセットとのつながりがユーザによって徐々に発見され利用されていくことが図 4 よりわかる．

2.3 クエリ共有に関する考察

以上のログ分析から，SPARQL キーワードについても参照 URI についても，人気のものとそうでないものが存在し，人気のものはかなり高い頻度で利用されていることがわかった．また，クエリの文字列長やクエリで利用される URI が時間経過とともに増えていることから，ユーザコミュニティは時間をかけて LOD データセットを利活用できるようになっていくことがわかった．

人気が偏ることについては，これはユーザ間で似たようなクエリを作成する可能性が高いことを示唆しており，過去に他のユーザが入力した SPARQL クエリを再利用することで，ユーザが入力しようとしているクエリを予測し補完できると考えられる．しかし本研究では多くのユーザが利用するヘッド部分ではなく，むしろそれ以外の部分に着目したいと考える．なぜなら，LOD の最大の特色は分野を越えた様々なデータが統一フォーマットでオープン化されている点にあり，これを横断利用することが LOD の特色を最大限に活かす利用法であると考えるためである．そのためには，必ずしも一般的ではないデータセットやデータの使い方に気づく必要があるが，システムから一方的に目新しい SPARQL クエリを推薦されても，ユーザは受け入れるのは難しい．

そこで本研究では，SPARQL クエリを一つの Web ページとしてコンテンツ化し，SPARQL クエリをブラウジングするという新しい検索インタラクションによって，新しい SPARQL

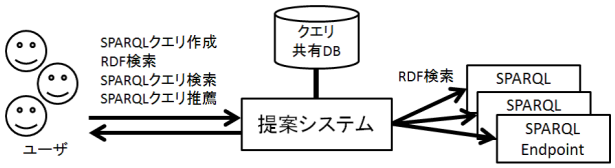

図 5: システム構成図

クエリとの出会いを可能にするシステムを提案する．ユーザ間でクエリに関する知識を共有することにより，ユーザコミュニティが LOD データセットを利用する速度を増すことができると考える．

3. 提案システム

3.1 システム概要

提案システムでは，ユーザ間で SPARQL クエリの共有を行う．システムは RDF データを持たず，他の SPARQL Endpoint への検索インタフェースとなる．図 5 はシステム構成図である。ユーザがシステムに SPARQL クエリを入力すると，システムは SPARQL Endpoint に問い合わせ，得られた検索結果を表示する．ただしクエリの入力にあたっては，ユーザは SPARQL クエリ（とリクエストを投げる SPARQL Endpoint）だけでなく，クエリを説明するタイトル，説明文，タグ，コメントも入力する．システムは，*(a) SPARQL* **クエリ**に加えて，これらのユーザが入力した *(b)* **メタデータ**，さらに Endpoint から返ってきた *(c)* **検索結果**の合計三種類のデータをひとまとまりにしてデータベースに登録する．

ユーザによって登録された SPARQL クエリは，**クエリページ**として他のユーザからも閲覧可能である．プロトタイプシステムのクエリページの例を図 6 に示す．クエリページにはタイトルや説明文，タグなどのメタデータに加え，SPARQL クエリ本文とその検索結果が表示される．検索結果は毎回 SPARQL Endpoint に問い合わせているのではなく，クエリの更新や更新リクエストがない限りはキャッシュされたものが表示される．クエリによっては検索結果が膨大になるため，プロトタイプシステムでは SPARQL クエリに対して強制的に LIMIT 文を追加している．

3.2 SPARQL クエリの検索・推薦

ユーザは通常のキーワード検索によってクエリページを見つけることができる．クエリページ検索は，SPARQL クエリ，メタデータ，検索結果すべてに対する全文検索となる．よってタイトルや説明文，タグなどをキーワードで検索することもできるし，リソースの URI を入力して，それを用いたクエリやそれがヒットするクエリを検索することもできる．

ユーザが見つけたクエリは，ユーザ自身の目的に完全に一致しているとは限らない．提案システムではクエリの編集および派生クエリの作成が可能になっている．クエリの編集とは，現在閲覧中のクエリを直接編集することであり，派生クエリの作成とは，現在閲覧中のクエリをドラフトとして新しいクエリを作成することである．ユーザであれば誰でもクエリの作成，編集，派生クエリの作成が行える．

SPARQL クエリが誰でもアクセス可能な場所に置かれ，また，検索可能になったとしても，ユーザが目的の SPARQL クエリに出会えるかどうかは確かではない．ユーザはクエリを検索するために必要なキーワードを知らないかもしれないし，そ

図 6: クエリページの画面例

もそも探索的検索のように目的が不明瞭であるかもしれない．そこで提案システムでは SPARQL クエリ推薦機能によってこれを支援する．

SPARQL クエリ推薦機能とは，現在閲覧中のクエリページに対して関連するクエリページを推薦する機能である．プロトタイプシステムではメタデータによるつながりがあるクエリページと，派生関係によるつながりがあるクエリページを推薦している．現在のプロトタイプでは，クエリ中のキーワードおよび URI の一致度合いに基づくものと，検索結果中に含まれる URI の一致度合いに基づくものの，二種類の推薦を行っている．どのような推薦が効果的であるかは今後の検討課題である．この点については次章に考察する．

4. 考察

提案システムは，ユーザが SPARQL クエリを次々にブラウジングしていくという新しい検索インタラクションを実現するために，SPARQL クエリ推薦機能を持つ．本章では，どのような SPARQL クエリが推薦されるとユーザにとって有用であるかについて議論する．

4.1 人気度に基づく推薦

ユーザ間で SPARQL クエリを共有することで得られるメリットの一つとして，クエリの人気度（利用頻度）が計算可能になることが挙げられる．より多くのユーザが閲覧したクエリ（クエリページ）は，典型性の高いお手本となるクエリであると考えられる．より多くの派生クエリが創られたクエリは，汎用性の高いテンプレート的なクエリであると考えられる．これらのクエリを推薦することは，特に SPARQL クエリの作成に不慣れなユーザに有用であると考えられる．

4.2 クエリの類似度に基づく推薦

ユーザが現在閲覧中のクエリに対して，SPARQL クエリ文そのものが類似するクエリを推薦する．情報推薦における基本的なアプローチであり，様々な手法が提案されている．これを SPARQL クエリ推薦に適用する場合，二種類の類似性が考えられる．

一つはクエリの言語レベルの類似性である．これは利用している関数や予約語の一致に基づく類似性である．この類似性を用いたクエリ推薦は，現在閲覧中のクエリが用いている関数や予約語の利用例を示すものといえ，特に初学者にとって有用であると考えられる．

もう一つはクエリで参照しているリソースの類似性である．日本語 DBpedia の検索フォーム [*2] で入力例として出ている「select distinct * where <http://ja.dbpedia. org/resource/東京都> ?p ?o . 」を例とすると，「<http://〜/東京都>」を利用しているクエリを推薦する．この類似性を用いたクエリ推薦は，現在閲覧中のクエリの拡張例を示すものといえる．例えば元クエリが東京都を出身地とするに人物名を集めていた場合，「東京都出身だが現在は東京都以外に住んでいる人」といった絞り込み条件を追加したクエリや，「東京都出身の人と結婚した人」といった RDF グラフを辿って異なるリソースを抽出してくるクエリなどが推薦される．これはユーザが興味のあるデータの周辺情報を提供していることになり，探索的検索にとって有用であると考えられる．

4.3 検索結果の類似度に基づく推薦

ユーザが現在閲覧中のクエリに対して，検索結果が類似するクエリを推薦する．これはつまり，違う言い方で同じものを得られているケースを推薦することになる．

クエリを作成するユーザからすると，クエリの改修案を示すものといえる．似たような結果を得られるが，クエリ本文がより簡潔であれば，より効率がよいクエリかもしれない．似ているがより多くの検索結果が得られるなら，網羅性が高いクエリとして参考になるかもしれない．類似のしかたによって何を推薦するかを使い分けることで，初学者から熟練者まで有用なアプローチと考える．

5. 関連研究

セマンティックウェブおよび Linked Open Data (LOD) は複雑な構造を持つ膨大なデータの集まりである．事前にデータセットについての理解がなければ，適切なクエリを構築することができない．そこで，可視化 [Kiefer 07][Russell 08] やユーザインタフェース [Morbidoni 07][Jarrar 12][Goto 12] によって RDF データの検索を支援する研究が多く存在する．

これらの研究はいずれも複雑な SPARQL クエリをどう作成するか，もしくは，SPARQL クエリを直接作成せずに Smantic Web/LOD 検索を行うか，という点にフォーカスしている．これはつまり SPARQL クエリの作成が本質的に難しいことを示している．見方を変えると，これらの手法を用いて作成した複雑なクエリは，それ自体が有用なデータであるといえる．本稿で提案するアプローチは有用なデータであるクエリを共有するものであり，検索クエリ作成支援技術とは補完関係にある．

蓄積された SPARQL クエリを再利用することで，SPARQL クエリの実行を効率化する研究もある [Hartig 13][Verborgh 14]．再利用を促すという点で，過去に使われた SPARQL クエリをサジェストし，クエリ実行だけでなく作成そのものも支援可能である．データベース分野では SQL クエリを対象として同様の研究がある [Chatzopoulou 09][Akbarnejad 10]．本研究と共通する点は多いが，上記の研究ではより多くの人が同じクエリを利用して効率化されることを狙うが，本研究ではむしろ新しいクエリの発見を促す点で，目的が異なる．

*2 http://ja.dbpedia.org/sparql

6. おわりに

本稿では，Linked Open Data (LOD) のための SPARQL クエリ共有システムを提案した．日本語 LOD の SPARQL クエリログから，クエリ共有の可能背について検討した．提案システムは，SPARQL クエリを共有し検索・推薦可能にすることで，ユーザが作成した SPARQL クエリの他の多くのユーザが利活用できるようにする．

LOD は様々な応用が期待されるが，膨大かつ多様なデータセットであるため，適切なクエリを作成するのは容易ではない．しかも利用可能なデータセットは日々増えていくため，データセット全体を熟知するということは不可能である．よってユーザ全体でデータセット利用のための知識としての SPARQL クエリを共有することは LOD の利活用を推進するうえで強力な支援になると考える．

謝辞

本研究を行うにあたり議論していただいた国立情報学研究所武田英明教授，大向一輝准教授，加藤文彦研究員および LODAC Project のメンバーの皆様に感謝いたします．

参考文献

[Akbarnejad 10] Akbarnejad, J., Chatzopoulou, G., Eirinaki, M., Koshy, S., Mittal, S., On, D., Polyzotis, N., and Varman, J. S. V.: SQL QueRIE Recommendations, *Proc. VLDB Endow.*, Vol. 3, No. 1-2, pp. 1597–1600 (2010)

[Chatzopoulou 09] Chatzopoulou, G., Eirinaki, M., and Polyzotis, N.: Query Recommendations for Interactive Database Exploration, in *Proc. of SSDBM '09* (2009)

[Goto 12] Goto, T., Hamasaki, M., and Takeda, H.: DashSearch LD: Exploratory Search for Linked Data, in *Proc. JIST 2012* (2012)

[Hartig 13] Hartig, O.: An Overview on Execution Strategies for Linked Data Queries, *Datenbank-Spektrum*, Vol. 13, No. 2, pp. 89–99 (2013)

[Jarrar 12] Jarrar, M. and Dikaiakos, M. D.: A Query Formulation Language for the Data Web, *IEEE Trans. on Knowledge and Data Engineering*, Vol. 24, No. 5, pp. 783–798 (2012)

[Kato 13] Kato, F., Takeda, H., Koide, S., and Ohmukai, I.: Building DBpedia Japanese and Linked Data Cloud in Japanese, in *Proc. of LDPW '13*, pp. 1–11 (2013)

[Kiefer 07] Kiefer, C., Bernstein, A., and Stocker, M.: The Fundamentals of iSPARQL: A Virtual Triple Approach For Similarity-Based Semantic Web Tasks, in *Proc. ISWC 2007* (2007)

[Marchionini 06] Marchionini, G.: *Exploratory search: from finding to understanding*, Vol. 49 of *Communications of the ACM*, pp. 41–46, ACM (2006)

[Morbidoni 07] Morbidoni, C., Polleres, A., and Tummarello, G.: Who the FOAF knows Alice? A needed step toward Semantic Web Pipes, in *New Forms of Reasoning for the Semantic Web'07* (2007)

[Russell 08] Russell, A., Smart, P. R., Braines, D., and Shadbolt, N. R.: NITELIGHT: A Graphical Tool for Semantic Query Construction, in *Proc. SWUI 2008* (2008)

[Verborgh 14] Verborgh, R., Sande, M. V., Colpaert, P., Coppens, S., Mannens, E., and Walle, de R. V.: Web-Scale Querying through Linked Data Fragments, in *Proc. LODW 2014* (2014)

[加藤 14] 加藤 文彦, 武田 英明, 小出 誠二, 大向 一輝：日本語 Linked Data Cloud の現状, 第 28 回人工知能学会全国大会論文集 (2014)