日本語DBpedia における SPARQL クエリログの分析

Log analysis of a SPARQL query in Japanese DBpedia

濱崎雅弘 [1*]　加藤文彦 [2]
Masahiro Hamasaki[1]　Fumihiko Kato[2]

[1] 産業技術総合研究所
[1] National Institute of Advanced Industrial Science and Technology (AIST)
[2] 情報・システム研究機構
[2] Research Organization of Information and Systems (ROIS)

Abstract: 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. 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，日本語 DBpedia の SPARQL クエリログの分析結果を報告する．日本語 DBpedia は，Wikipedia を Linked Open Data (LOD) 化した DBpedia [1] の日本語版であり，国内 LOD のハブ的存在である．本研究では日本語 DBpedia のアクセスログ 1300 万件の中から SPARQL クエリログを抽出・分析し，ユーザの SPARQL 利用状況について考察をする．

LOD はオープンデータのための情報共有の枠組みとして世界的にも多くの注目を集め，我が国においても政府機関や各種データプロバイダから LOD によるデータ公開が相次いでいる．LOD は大規模かつ分散したデータセットであり，その実体は複雑かつ多様なグラフデータである．LOD は一般の関係データベースや Web API のように応用のために設計されたデータ構造やインタフェースを持つわけではない．このため，LOD を利用したいユーザは，検索を通してデータセットの中身を知り，また，検索によってデータセットから必要な部分を取り出す必要がある．つまり LOD にとってクエリとは，データセットを理解するためのプローブ（探査針）であり，データセットを活用するためのスキーマ（データ構造）であるといえる．

このように LOD において検索と利活用（LOD アプリケーションの開発）は不可分な関係にある．一方で LOD 検索には，LOD が持つ複雑性と異種性のためにクエリ作成が困難であること，さらに LOD が持つ大規模性と分散性のためにクエリ実行に不安定性や性能劣化が伴うこと，という問題がある．これらは LOD 検索における根本的問題として数多くの研究が行われている．クエリ作成については様々なクエリ作成支援インタフェースが提案されている [Kiefer 07, Jarrar 08]．一方でクエリ実行に対しては，サーバ側の負荷を軽減しつつ鮮度の高いデータを効率よく検索するために，キャッシングやインデキシングを用いた様々な手法が提案されている [Hartig 13, Lorey 13, Verborgh 14]．

クエリの作成と実行を支援するにあたって重要となるのが，実際にユーザがどのように SPARQL クエリを用いて LOD にアクセスしているのか，ユーザとデータとのあいだにどのようなインタラクションが発生しているかを知ることである．これが不明瞭なままでは，適切なクエリ作成支援やクエリ実行支援の実現は容易ではない．国外では，実際の SPARQL クエリログを用いた研究用データセットの開発や，SPARQL クエリログの分析などが行われているが，国内においてはまだそのような取り組みが十分であるとはいえない．LOD はデータそのものが多種多様であるため，利用状況の分析についてはそれぞれ分析が必要であると考える．そこで本研究では，国内 LOD のハブ的存在である日本語 DBpedia に対する SPARQL クエリログの分析を行い，ユーザの LOD および SPARQL 利活用状況について考察をする．

本稿の構成は以下の通りである．2 章にて，分析対象となる日本語 DBpedia と，実際に用いたアクセスロ

*連絡先： 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
E-mail: masahiro.hamasaki@aist.go.jp

[1] http://dbpedia.org

グについて概説する．3 章にて，SPARQL クエリログの分析結果を示し，4 章にて，分析結果をふまえた考察を述べる．5 章にて，既存の LOD アクセスログおよび SPARQL クエリログを分析した研究について述べる．6 章にて本稿をまとめる．

2　日本語 DBpedia

日本語 DBpedia (DBpedia Japanese) とは，日本語 Wikipedia [2] から生成された DBpedia である．DBpedia [Bizer 09] とは Wikipedia から構造化データを抽出して様々な形式で公開するコミュニティプロジェクトであり，2015 年 7 月現在，本家の英語版 DBpedia をはじめ 18 言語の DBpedia が存在する [3]．日本語版は 2012 年 4 月に始まり，2015 年 6 月時点で 100,090,381 トリプルを有する．日本語特有のローカライゼーション [Kato 13] に加え，日本語 WordNet と紐づけているといった特徴もある [小出 13]．本家の英語版と比較すると，トリプル数においても語彙数においてもまだまだ少ないが，国内の LOD においてはハブ的役割を果たしている．

図 1 は，日本語 DBpedia へのアクセス数を示したものである．毎月のページビューを積算した値が示されており，2013 年 10 月から 2014 年 12 月までの 1 年 2 カ月の間に，約 1200 万件のアクセスがあった．なお，集計にあたっては後述する日本語 DBpedia のサーバへのアクセスログを用いている．図 2 は，日本語 DBpedia へアクセスしたマシンの IP の異なり数を示したものである．これは先ほどの積算グラフと異なり，月ごとの IP の異なり数（ユニーク IP 数）を示している．2013 年 11 月に大きく増えてから，2014 年 4 月まではゆるやかに減少傾向にあったが，それ以降は増加傾向にあり，最終的に 2014 年 12 月には月間ユニーク IP 数は 22,574 件となっている．

図 3 は，毎月のアクセス数をタイプ別に示したものである．sparql は SPARQL クエリを用いたアクセス，res はリソースへのアクセス，page はリソースのページへのアクセス，other はその他のページへのアクセスである．そのほかのページとは，日本語 DBpedia のトップページやヘルプページなが含まれる．リソースよりも SPARQL によるアクセスの方が多く，SPARQL が LOD へのアクセスの基本的な手段となっていることがわかる．

[2] http://ja.wikipedia.org
[3] http://wiki.dbpedia.org/about/language-chapters

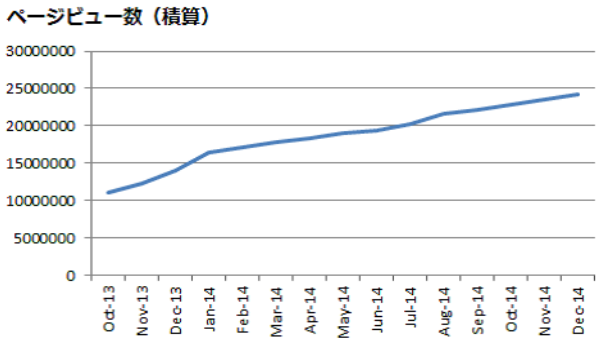

図 1: ページビューの積算（2013 年 10 月から 2014 年 12 月まで）

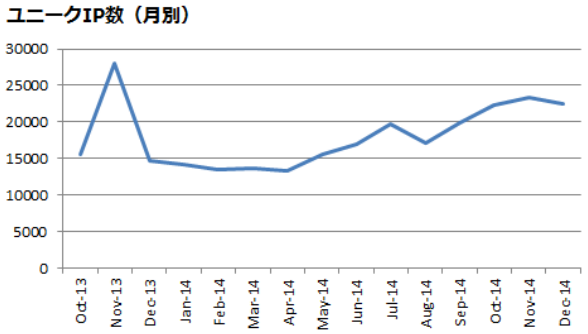

図 2: 月別のユニーク IP 数（2013 年 10 月から 2014 年 12 月まで）

3　SPARQL クエリのログ分析

3.1　データの概要

分析に用いるのは 2013 年 6 月から 2015 年 1 月までの DBpedia Japanese の Web サーバのアクセスログである．アクセスログに記録された GET リクエスト約 1300 万件を分析する．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．なお，文法チェックは行っていないため，文法エラーとなる SPARQL クエリや検索結果が 0 件の SPARQL クエリも含まれている．

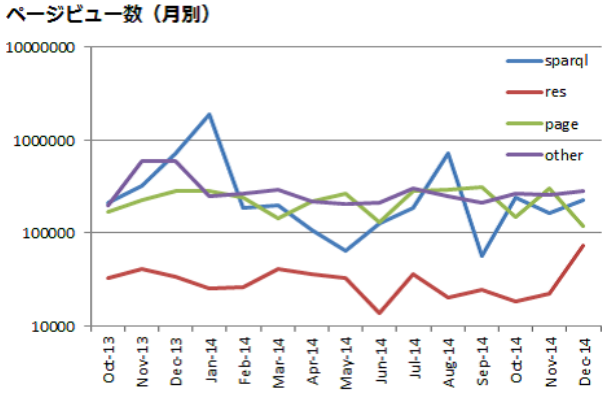

図 3: タイプ別の月間ページビュー数.

3.2 全体の利用傾向

図 4 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，表 1 はクエリパターン，ここでは SPARQL クエリ中に出現する SPARQL キーワードの組とする，のうち利用したユニーク IP 数が多いもの上位 10 件を示したものである．なお，ここでいう SPARQL キーワードとは，SPARQL 1.1 の仕様書にて公開されている 113 個のキーワード [4] を指す．

図 4 から，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ことがわかる．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 3.9 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 5 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 6 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 7 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

[4] http://www.w3.org/TR/2013/REC-sparql11-query-20130321/#sparqlGrammar

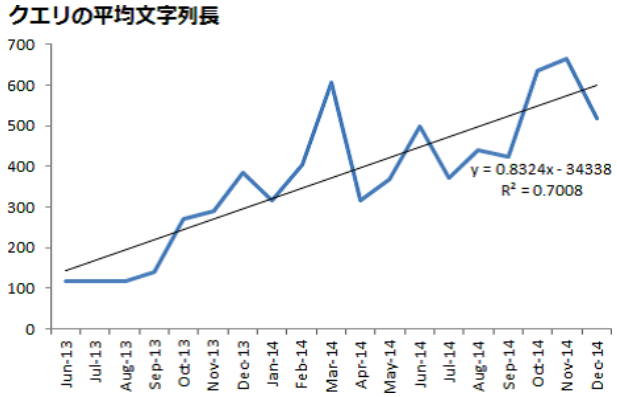

図 5: SPARQL クエリの平均長の時間変化．直線と数式は線形近似の結果.（縦軸：クエリの長さの平均値，横軸：年月）

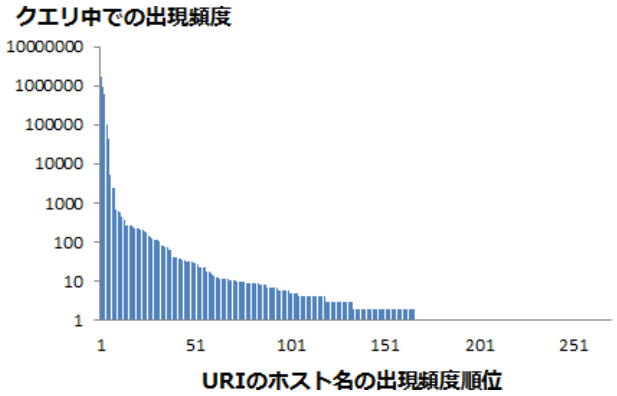

図 6: SPARQL クエリ中に現れる URI のホスト名の出現頻度（縦軸：出現頻度，横軸：出現頻度の順位）

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 7 よりわかる．

3.3 ユーザごとの利用傾向

ここで，1 つのユニーク IP は 1 人のユーザもしくは 1 つのユーザグループを表している仮定し，ユニーク IP ごとのログを分析する．図 8 は，ユーザ（ユニーク IP）ごとの利用回数と利用日数を示している．ユニーク IP 数はすべてで 5585 個であった．利用日数が 1 日のみは 4150 人（0.74），2 日以上は 1435 人（0.26）であった．また，利用回数が 1 回のみは 1835 人，10 回未満は 4183 人（0.75）であった．

2 日以上の利用ログがある 1435 人のログから，利用初日の利用状況と，全期間での利用状況を比較した．表 2 は，利用初日と全期間でのキーワード数・ホスト数

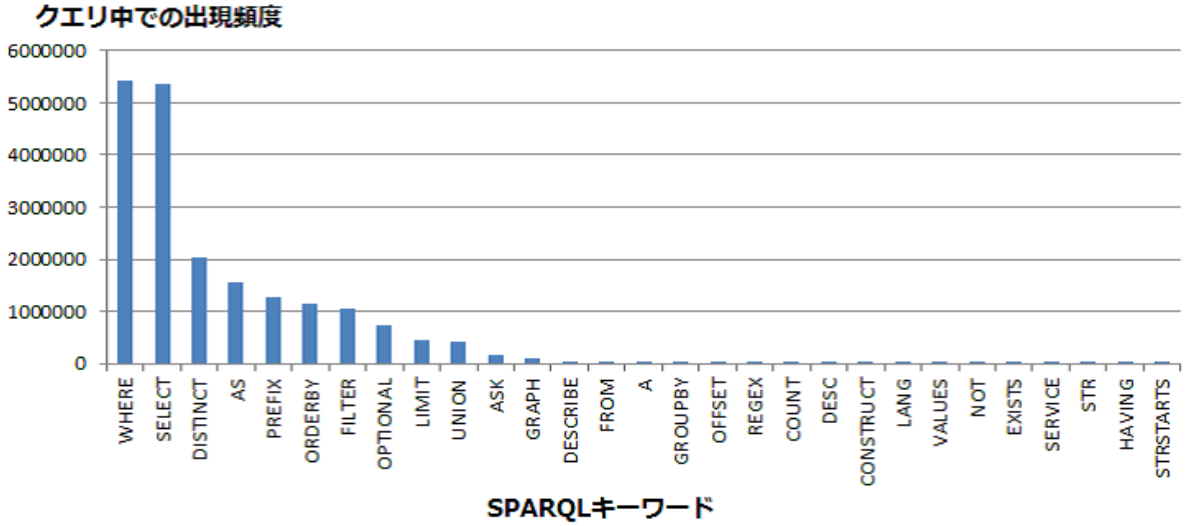

図 4: SPARQL クエリ中に現れる SPARQL キーワードの出現頻度（縦軸：出現頻度，横軸：出現頻度の順位）．出現頻度が 500 以上のもののみを表示している．

表 1: 利用したユニーク IP 数の多いクエリパターン（クエリ含まれる SPARQL キーワードセット）上位 10 件

クエリパターン	出現回数	ユニーク IP 数	クエリの異なり数
SELECT, DISTINCT, WHERE	586957	2339	162747
SELECT, DISTINCT, WHERE, OPTIONAL	6603	909	1955
DESCRIBE	55390	787	40150
SELECT, DISTINCT, WHERE, ORDER BY	2277	586	446
SELECT, WHERE	274752	379	64076
SELECT, DISTINCT, WHERE, LIMIT	96156	358	35533
SELECT, WHERE, LIMIT	58418	302	4792
SELECT, DISTINCT, FILTER, OPTIONAL, LIMIT	7952	278	6608
PREFIX, SELECT, WHERE	377630	254	28079
PREFIX, SELECT, DISTINCT, WHERE	154481	219	56158

の比較と，増加したユーザ数およびその割合を示している．

キーワード数が初日から増加したユーザは全体の半分弱，ホスト数は約 3 割であった．これらの値から複数日にわたって利用したユーザの平均像を描くと以下の様になる：初日にキーワード 3 個でクエリを作成し，最終的にはもう一つ新しい SPARQL キーワードを使うようになる．ホスト数は最初から最後まで変わらずたかだか 1 個である．

一方で最大値を見ると，キーワード数は初日で 17 個，全期間で 27 個であった．ホスト数は初日 14 個，全期間 30 個であった．これは極端な例であるが，SPARQL を十分に使いこなしているのはまだ少数であることがわかる．

4 考察

本章では，SPARQL クエリログを用いた既存研究について述べる．さらに，それらの知見と今回のログ分析結果の知見を踏まえて，LOD 利活用にはどのような支援が有効であるかについて議論する．

4.1 関連研究

UseWOD [5] は Linked Open Data の利用状況分析を行うワークショップで，研究用データセットとして DBpedia のログ（UseWOD 2014 SPARQL データセット）を公開している．Rietveld らはこのログと SPARQL エ

[5] http://usewod.org/

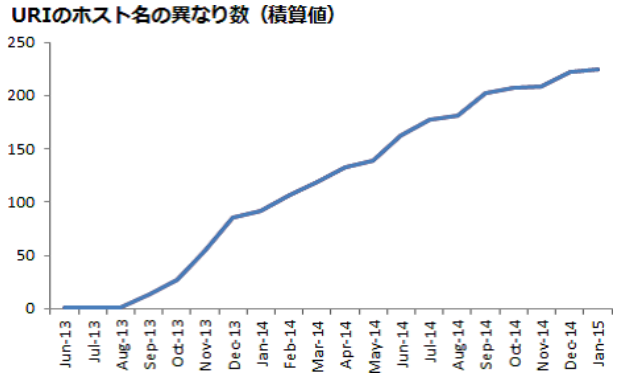

図 7: SPARQL クエリ中に現れる URI のホスト名が初めて出現した年月（縦軸：ホスト名の異なり数の累計，横軸：ホスト名の初出年月）

表 2: 利用初日と全期間でのキーワード数・ホスト数の比較と，増加したユーザ数およびその割合．

	利用初日	全期間	増加したユーザの数
キーワード数	平均 3.43 中央 3 最大 17	平均 3.92 中央 4 最大 27	669 (0.47)
ホスト数	平均 0.86 中央 1 最大 14	平均 1.04 中央 1 最大 30	434 (0.30)

ディタ YASGUI [6] のログを比較することで，ユーザが SPARQL エディタを用いた場合の SPARQL クエリの特徴を明らかにしようとしている [Rietveld 14]．分析の結果，LIMIT や FILTER，UNION などの利用回数において，YASGUI 方が多いという差が見られた．また，クエリに含まれるトリプル数も，YASGUI の方が多いという結果であった．ユーザインタフェースの充実によって，ユーザはもっと複雑なクエリを作れるようになることを示す結果といえる．

Huelss らも同じ UseWOD データセットを用いて，SPARQL クエリ中の共起関係からトリプル間の関連性を発見できないかを調査している [Huelss 15]．彼らの報告によると，データセットに含まれる 30 万クエリのうち半数近い 14.6 万クエリが bot よる同一クエリであった．我々の日本語 DBpedia のアクセスログも 1300 万件中 750 万件が bot のログであり，bot を取り除く処理を行わなければ結果が大きく変わってしまう．ログからユーザの利用分析をするにあたっては注意が必要である．なお，目的であったトリプル間の関連性発

[6] http://yasgui.org/

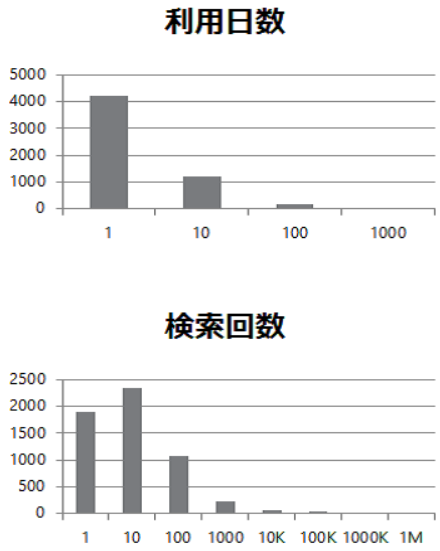

図 8: ユニーク IP ごとの SPARQL 検索の利用日数と利用回数の頻度分布．（縦軸：該当する利用日数または検索回数のユニーク IP の数，横軸：利用日数または検索回数）

見についてはまだ十分な成果が得られていないと述べている．

SEMLEX [Mazumdar 11] は SPARQL クエリログ分析のための可視化ツールである．コンセプトグラフは，SPARQL クエリ中の出現パターンから得られた特徴をもとに，クラスをネットワーク図として可視化する．述語遷移ツリーは，SPARQL クエリ中におけるトリプルの出現位置の順序関係をもとに，ある述語を使ったあとに，どの述語が使われやすいかをツリー形式で可視化したものである．

利用分析とは異なるが，SPARQL クエリログを分析して，SPARQL ベンチマーク用の適切なクエリセットを作成しようとする取り組みもある [Morsey 11]．クエリ間の類似度を定義してクラスタリングし，ベンチマーク用に用いる典型的なクエリを探そうとしている．頻出する SPARQL キーワードを取り除き，変数名を統一したうえで，文字列編集距離を用いて類似度を定義している．クラスタリングの結果，3.5 万個のクエリが 1.2 万クラスタ（うち 1/4 は属するクエリは一つ）に分類された．具体的な数字は示されていないが，クラスタに属するクエリの数の分布はロングテールとなり，ごく少数のクラスタにたくさんのクエリが集まった．ここから，典型的なクエリは存在し，類似度ベースのサジェストによってクエリ作成が支援可能なこと，一方でクエリは多様であり，それだけでは不十分である，ということが考えられる．典型的なクエリがあることを利活用するという点では，クエリのキャッシュ機

構が考えられるが，これについては様々な手法が提案されている [Hartig 13, Lorey 13, Verborgh 14].

4.2 LOD 利活用支援

我々のログ分析から，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．既存研究では時間変化については分析されていないため，同様の傾向があるかどうかは明らかではないが，クエリパターンの頻度分析からも，やはり大多数は単純なクエリしか利活用できていないと考えられる．

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．一方で，クエリの文字列長や利用される URI が増加していることから，利用されるクエリの種類は収束することなく増大し続けると考えられ，ユーザによる新たなクエリの発見（気づき）もまた重要であるといえる．その場合，LOD データセットおよびクエリの可視化や，クエリの推薦といった機能が重要になると考えられる．

5 おわりに

本稿では，日本語 LOD の SPARQL クエリログから，LOD の利用状況について考察を行った．分析にあたっては日本語 DBpedia の 1300 万件のアクセスログを利用した．分析結果からは，日本語 DBpedia は利用者数は増加傾向にあること，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．しかし，ごく少数のユーザ以外は LOD および SPARQL が持つ機能を十分に使いきっているとは言い難い状況にあることがわかった．

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．本稿では，日本語 LOD の SPARQL クエリログから，困難な状況を明らかにした．今後は，この知見を用いてユーザの LOD 利活用支援に取り組んでいきたい．

謝辞

本研究の一部は JSPS 科研費 15H02781 の助成を受けたものです．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

[Bizer 09] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S.: DBpedia - A Crystallization Point for the Web of Data, *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, pp. 154–165 (2009)

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

[Huelss 15] Huelss, J. and Paulheim, H.: What SPARQL Query Logs Tell and do not Tell about Semantic Relatedness in LOD, in *Proc. of NoISE 2015* (2015)

[Jarrar 08] Jarrar, M. and Dikaiakos, M. D.: MashQL: a query-by-diagram topping SPARQL, in *Proc. ONISW '08*, pp. 89–96 (2008)

[Kato 13] Kato, F., Takeda, H., Koide, S., and Ohmukai, I.: Building DBpedia Japanese and Linked Data Cloud in Japanese, in *Proc. of Linked Data in Practice Workshop* (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)

[小出 13] 小出 誠二, 武田 英明, 加藤 文彦, 大向 一輝 : 日本語 WordNet と IPAdic 辞書の RDF 化と DBpedia リンク, 第 27 回人工知能学会全国大会論文集 (2013)

[Lorey 13] Lorey, J.: Caching and Prefetching strategies for SPARQL queries, in *Proc. of USEWOD 2013* (2013)

[Mazumdar 11] Mazumdar, S., Elbedweihy, K., Cano, A. E., Wrigley, S. N., and Ciravegna, F.: SEMLEX - A Framework for Visually Exploring Semantic Query Log Analysis, in *Proc. of ISWC 2011* (2011)

[Morsey 11] Morsey, M., Lehmann, J., Auer, S., and Ngomo, A.-C. N.: DBpedia SPARQL Benchmark ? Performance Assessment with Real Queries on Real Data, in *Proc. of ISWC 2011* (2011)

[Rietveld 14] Rietveld, L. and Hoekstra, R.: Man vs. Machine Differences in SPARQL Queries, in *Proc. of USEWOD 2014* (2014)

[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)