様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　９　月　　５　日現在

機関番号：　８２６２６
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７８１

研究課題名（和文）Ｌｉｎｋｅｄ　Ｏｐｅｎ　Ｄａｔａ 利活用のためのクエリ共有手法に関する研究

研究課題名（英文）Study on query sharing for utilization of Linked Open Data

研究代表者
濱崎　雅弘（Hamasaki, Masahiro）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究グループ長

研究者番号：５０４１９０１６

交付決定額（研究期間全体）：（直接経費）　　12,400,000円

研究成果の概要（和文）：本研究では，Linked Open Data(LOD)の利活用を阻害する最大の要因であるLOD検索の困難さに対して，クエリ共有が有効であることを明らかにするため，(課題1)クエリ生成の支援に有効な共有クエリ推薦技術の研究開発，(課題2)クエリ実行の高速化に有効なクエリキャッシュ技術の研究開発，の二つの研究を実施した．既存のSPARQLエンドポイントのアクセスログ解析，プロトタイプを用いたユーザ評価，さらにはRDF 問合せ最適化のための基礎技術の開発を行った．

研究成果の概要（英文）：In this research, we investigate query sharing among SPARQL users based on query recommendation and query caching. 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. We developed query recommendation methods for sharing queries among users, and query caching methods to improveme searching with SPARQL queries.

研究分野：　セマンティックウェブ

キーワード：　情報システム　Linked Open Data　クエリ共有　情報推薦　キャッシング　セマンティックウェブ　RDF　SPARQL

１．研究開始当初の背景

本研究は，Linked Open Data (LOD) の利活用を阻害する LOD 検索の困難さに対して，クエリ共有という新しいアプローチでの解決を図るものである．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．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．クエリ生成については可視化[1] や対話型インタフェース[2] によるアプローチが多く取られてきている．最近では，既存の Web 検索のように検索ログを利用してユーザが入力したいクエリを予測補完するための研究[3]も行われている．クエリ実行については，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[4] が提案されている．

既存研究では，主に個々のユーザが目的とするクエリを容易に生成し正確かつ高速に実行することを支援している．これに対して，本研究では LOD 利活用のための検索という考えから，検索はアプリケーションの設計行為の一部であり，試行錯誤による探索的な検索の支援が肝要とする．そのため，ユーザにとって未知である LOD データセットに対しても興味深いクエリを生成できるか，かつ，鮮度・正確性を犠牲にしてでも，高速性・安定性のあるクエリが実行できるか，という点が重要となる．この前提のもと，本研究では，LOD の試行錯誤的な検索を可能にするために，クエリを共有し再利用・再編集可能にする新しい検索基盤を提案する．蓄積された共有クエリとその検索結果を利用することで，過去のクエリから検索のための知識を得ることができ，また，LOD に対する安定したクエリの実行を実現することができる．そのために本研究課題では，(1) 共有クエリ間の異なる種類の関係を用いたクエリ推薦と，(2) 共有クエリの包含関係に基づくクエリキャッシングの，二つの要素技術の研究開発を行う．

２．研究の目的

本研究では，Linked Open Data(LOD)の利活用を阻害する最大の要因である LOD 検索の困難さに対して，クエリ共有が有効であることを明らかにする．具体的には，(課題 1)クエリ生成の支援に有効な共有クエリ推薦技術の研究開発，(課題 2)クエリ実行の高速化に有効なクエリキャッシュ技術の研究開発，の二つの研究を実施する．

筆者らは LOD の調査や構築，利活用に関する様々な研究を行ってきた．これらの経験から，SPARQL クエリが LOD におけるプローブとスキーマの役割を果たしており，データを検索しながら試行錯誤して SPARQL クエリ作成する行為こそが LOD アプリケーションのラピッドプロトタイピングであるという着想を得た．さらに，相互互助関係が成立する利用者コミュニティが生まれ，クエリ共有というアプローチが実現可能な土壌ができていることから，セマンティックウェブ技術に加え，応募者らが培ってきたユーザ参加型システム技術，ソーシャルメディア分析技術，関係抽出技術およびデータベース技術を組わせることで，LOD 検索における根本的問題であるクエリ生成と実行の問題に対して，クエリ共有による新しい解法が提案できるという考えに至った．

３．研究の方法

本研究では，クエリ共有に基づくクエリ推薦とクエリキャッシュが，クエリの生成と実行における問題に対して有効であることを示す．まず，既存の SPARQL エンドポイントの利用ログを解析し，ユーザが生成したクエリのパターンや特徴を明らかにする．次に，クエリ推薦およびクエリキャッシュの基盤技術となるクエリ間類似度を定義し，その計算方法を提案する．これに基づいてクエリ推薦アルゴリズムとクエリキャッシュアルゴリズムを構築し，評価実験により有効性を確かめる．

本研究の学術的な特色として以下の 3 点が挙げられる．(1) クエリ共有によりクエリの生成と実行の問題を一挙に解決する．(2) LOD 検索を社会的な行為として捉え，知識（クエリ）の共有・再利用により問題を解決する．(3) 正確性や単純な実行速度よりも，試行錯誤の繰り返しを快適に行えることを目指す．これは LOD において検索行為は，データセット理解およびアプリケーション開発のための探索・発想行為でもあると認識しているためである．

４．研究成果

（課題 1）
クエリ生成の支援に有効な共有クエリ推薦技術の研究開発として，SPARQL クエリユーザの振る舞いについての分析を行った．具体的には既存の SPARQL エンドポイント（日本語 DBpedia）のアクセスログ解析を行った．そこから，持ちられる SPARQL クエリの長さが年々増えている傾向にあること（図 1）や，クエリには頻出パターンがあること（表 1）がわかり，クエリ推薦の必要性等について知見が得られた．

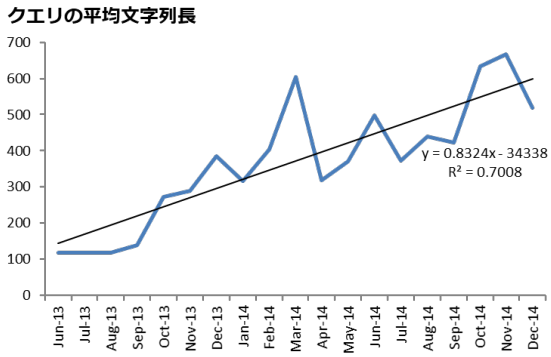

図 1: SPARQL クエリの平均長の時間変化．直線と数式は線形近似の結果．（縦軸：クエリの長さの平均値，横軸：年月）

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

さらに，実際に Web 上で公開されている LOD データセットおよび SPARQL エンドポイントについて調査を行った．これは統合クエリ（Federated Query）を活用するには複数の LOD データセットおよび SPARQL エンドポイントが必要となるが，それらについての情報の整備が特に国内においては不十分だからである．可用性や処理性能，他の LOD との接続性といった点について，国内 50 件以上の LOD データセットおよび SPARQL エンドポイントに対して調査を行った．

また一方でクエリ共有インタフェースに関する研究を進め，クエリ共有を音楽コンテンツに適用したプロトタイプシステムを構築しユーザ評価を行った．ここではクエリ推薦を用いたクエリ共有を新しい検索インタラクションとして提案し（図 2），このプロトタイプを用いてユーザ評価を行った．これらの成果はオープンコラボレーションに関する国際会議にて発表した．

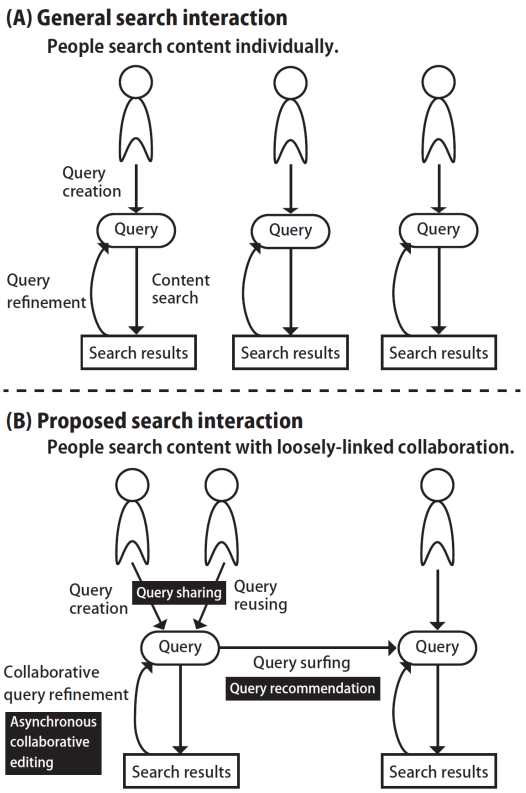

図 2：クエリ共有インタラクション

（課題 2）
クエリ実行の高速化に有効なクエリキャッシュ技術の研究開発については，結合選択率の間接的見積り手法を新たに提案し，RDF 問合せ最適化のための基礎技術の開発を行った．本手法は結合選択率の間接的見積もりを用いることにより精度改善を達成したものである．本研究の特徴は，不完全な統計情報下でも利用可能な，条件付き確率を用いた数理モデルに基づく算出法を提案している点にある．これによりキャッシュサーバに残された部分的なデータでも問合せコストを見積もることが可能となる．実データに適用して提案手法による見積値がある程度妥当に得られることも示した．本成果は高く評価され国内学会にて表彰された（情報処理学会 2016 年度山下記念研究賞）．

統合クエリのキャッシュ技術についても研究開発を進めた．提案手法では完全な検索結果を返すのではなく，類似した検索結果を返すというアプローチを取る．Coverage（より多くの検索結果が含まれる），Freshness（より新しい情報が含まれる），Diversity（より多くの情報源から取得する）の三つの観点から評価した結果，既存手法と比較して最大 30%の向上を達成した．

研究クエリキャッシュ技術のより実践的な改良を行うために，研究用データセットの構築に取り組んだ．これまで SPRQL クエリログは研究用データセットとしてあったが SPRQL クエリ検索結果を含めた研究用データセットはこれまで存在しなかった．そこで SPARQL クエリキャッシュ研究をさらに発展

させる基盤技術として，匿名化処理による SPARQL クエリログの研究用データセット化システムを研究開発した．また，RDF トリプル等に含まれる URI が参照するデータのタイプを自動推定する手法の研究開発に取り組み，この成果を Web インテリジェンスに関する国際会議にて発表した．

これら課題１および 2 に関する研究の推進に加え，本プロジェクトは SPARQL の利活用により LOD アプリケーション開発（ラピッドプロトタイピング）が容易になるという考えに立脚しているが，これを体現するデモシステムとして D'ownLOD を Web 公開した．D’ownLOD を用いれば，Web サーバ上の D’ownLOD プラットフォームに Web ブラウザでアクセスし SPARQL クエリを登録するだけで LOD アプリケーションを開発できる(図 3)．この LOD アプリケーションは D’ownLOD プラットフォーム上だけでなく自分の手元でも動かすことができる（図 4)．

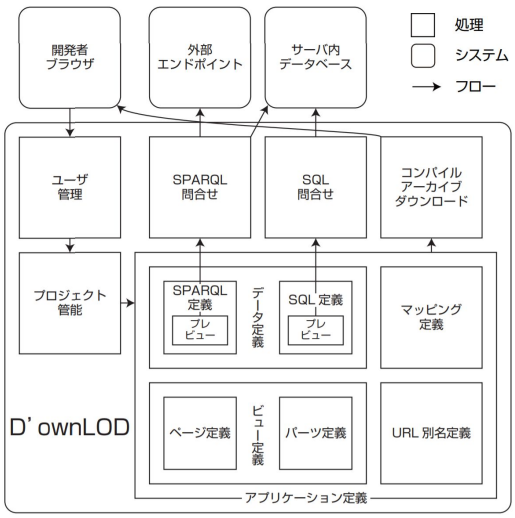

図 3：D'ownLOD プラットフォーム

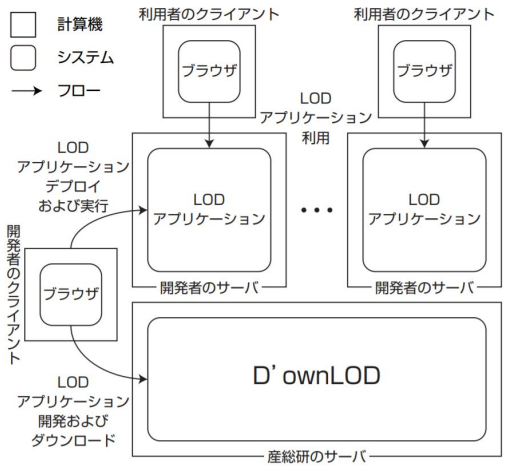

図 4：D'ownLOD アプリケーション

D’ownLOD の特徴を以下に示す．SPARQL クエリが１つあれば LOD アプリケーションを１つ構築できる，SPARQL による LOD アプリケーションはのラピッドプロトタイピングを実現するシステムである．

- 動的ページ：GET パラメタに応じてページが動的に変更
- SPARQL クエリ発行：任意の EP に任意の SPARQL を発行
- DB キャッシュ：SPARQL とその結果にキャッシュ．高速化．
- SQL 結果の JSON 変換：SQL 結果を自動で JSON として出版
- Web API：JSON を返す URL を提供
- モビリティ：すべて１つのファイルにパッケージ．ダウンロード可能．
- 即時実行可能：アプリ実行条件は Java のみ
- 自動テーブル定義：DB のテーブルは SPARQL から動的に生成
- データとビューの分離：データとビューの再利用性を向上
- 別名 URL：URL の短縮やトップページなどの特殊ページの作成
- プレビュー：SPARQL や SQL を試行錯誤的に定義できる

図 5：D’ownLOD の特徴

本研究では，クエリ共有の基盤技術となるクエリ生成の支援に有効な共有クエリ推薦技術の研究開発，クエリ実行の高速化に有効なクエリキャッシュ技術の研究開発に取り組んできた．クエリ推薦に関してはユーザの SPARQL クエリ利用分析やクエリ共有インタフェースのプロトタイピングとユーザ評価を行い知見が得られた．クエリキャッシュに関しては結合選択率の間接的見積もりを用いた手法を提案し性能向上を達成した．

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．しかし LOD を横断的に利用しようとすればするほど，クエリ生成の困難さとクエリ実行の性能劣化の問題が深刻となる．本研究は，国内外におけるオープンデータ化の流れを，データ公開でとどめず，新たな価値創造（データイノベーション）へと展開していくために資するものと考える．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　３件）

①Marie Katsurai, Ikki Ohmukai, Hideaki Takeda, Topic Representation of Researchers Interests in a Large-Scale Academic Database and Its Application to

Author Disambiguation, IEICE Transactions on Information and Systems, 査読有, Vol. E99.D, 2016, 1010-1018
②大向 一輝、図書館とデジタルアーカイブ：相互運用性に関する課題と展望、図書館雑誌、111 巻、2017、369-372
③嘉村 哲郎、大向 一輝、人文科学における Linked Open Data の活用、人工知能、32 巻、2017、401-407

〔学会発表〕（計　7 件）
①濱崎 雅弘、加藤 文彦、日本語 DBpedia における SPARQL クエリログの分析、第 36 回セマンティックウェブとオントロジー研究会、2015
②的野 晃整、小川 宏高、統計情報制限下における RDF 問合せ最適化のための結合選択率の間接的見積り手法、第 8 回 Web とデータベースに関するフォーラム（WebDB Forum 2015）、査読有、2015
③荒木 将貴、桂井 麻里衣、大向 一輝、武田英明、研究成果データベースを用いた異分野の共同研究者の推薦、第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM 2016）、2016
④ Steven Lynden, Makoto Yui, Akiyoshi Matono, Akihito Nakamura, Hirotaka Ogawa, Isao Kojima, Optimising Coverage, Freshness and Diversity in Live Exploration-based Linked Data Queries, the 6th International Conference on Web Intelligence, Mining and Semantics (WI 2016), 査読有, 2016.
⑤的野 晃整、D'ownLOD: LOD のためのラピッドプロトタイピングプラットフォーム、第 41 回セマンティックウェブとオントロジー研究会、2017
⑥ Masahiro Hamasaki, Masataka Goto, QueryShare: Working Together to Facilitate Exploratory Multimedia Searches without Skill in Creating (OpenSym 2017), 査読有, 2017.
⑦Steven J. Lynden, Analysis of semantic URLs to support automated linking of structured data on the web, International Conference on Web Intelligence, Mining and Semantics (WIMS 2017), 査読有, 2017.

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

D'ownLOD
http://downlod.linkedopendata.net/

６．研究組織
(1)研究代表者
濱崎 雅弘（HAMASAKI, Masahiro）
産業技術総合研究所・情報技術研究部門・研究グループ長
研究者番号：５０４１９０１６

(2)研究分担者
的野 晃整（MATONO, Akiyoshi）
産業技術総合研究所・人工知能研究センター・主任研究員
研究者番号：１０４４３２２７

大向 一輝（OHMUKAI, Ikki）
国立情報学研究所・コンテンツ科学研究系・准教授
研究者番号：３０４１３９２５

リンデン スティーブン（LYNDEN, Steven）
産業技術総合研究所・人工知能研究センター・主任研究員
研究者番号：３０５２８２７９