意味情報に基づくオープンデータへのアクセス・利用効率化のための移動アプリケーション技術の応用的開発

研究代表者　　福　田　直　樹　　静岡大学 学術院情報学領域

1 研究の概観

本研究では，計算を継続しながら複数のコンピュータ上を移動可能な特性を持ったソフトウェア（移動アプリケーション）上から，効果的にオープンデータにアクセスし，意味情報に基づいて適切な情報を検索・提示するためのソフトウェア基盤を開発する．本ソフトウェア基盤の特長は，ソフトウェアが他のコンピュータ上へ移動することに伴って，必要な情報や支援が変化した場合に，必要となるデータの提供源を種々のオープンデータと，そのデータ格納基盤とオントロジーに基づいて動的に検索可能とする点である．ソフトウェアの構成要素となるコンポーネントおよびサービスについて，明示的な意味表現であるオントロジーに基づいた記述を扱うための枠組みを与え，資源配分方法をメカニズムデザインの知見に基づき自動化することにより，オープンデータを効果的に扱う移動アプリケーションの実現を可能とする．

本研究によって開発された技術により，アプリケーションソフトウェアの利用者は，利用者の身近にあるコンピュータを用いて，いつものアプリケーションソフトウェアを，安全に，継続して利用することが単に可能となるだけでなく，ユーザの周囲の情報や公共データなどのオープンデータを意味情報や文脈に基づいて動的に検索し，その特長を最大限活かして，アプリケーションをリッチに実行させるようなアプリケーション開発基盤が実現される．

なお，本学では，寄付金として受け入れたものに対しての，非公開の内容を含めた報告書の作成を含めた，寄付の対価として判断される可能性のある行為を禁じているが，本報告書（要約）にあるものは，基本的に末尾に記載した公表済み（近日学会等での公表予定のものを含む）の内容および研究テーマ申請書に記載した内容のみから構成されるため，この行為にはあたらない．

2 研究の背景と具体的なねらい

これまでに本研究代表者およびそのグループは，オントロジーマッピングの効果的な利用技術についての検討を行ってきている．たとえば，SPARQLoid[Fujino 12a][Fujino 12b][Fujino 14]では，オープンデータへのアクセスで用いられる SPARQL エンドポイントへのアクセスを，そのオープンデータそのものに対してのオントロジーに必ずしも熟知しない場合であっても，他のオントロジーからのマッピングを用いてアクセス可能とすると同時に，その際のマッピング精度に基づく取得データの順序付けを，クエリ書き換え技術により実現している[Fujino 12a]．

また，SPARQLoid ではさらにそのアクセスを１つのクエリから複数のエンドポイントにまたがった検索を行う FederatedQuery に拡張[Fujino 12b] しており，その有効性についての検証[Fujino 14] を行ってきている．また，オープンデータのエンドポイントそのものの探索問題[Ladwig 10] に対しては，たとえば，クエリ作成時にそのクエリの実行に適したエンドポイントを，その検索対象に対する文字列や既存オントロジーとのマッピングに基づいてエンドポイントの適合性を探索しながらそのクエリの実行を可能とする機構[Noguchi 13]の開発を進めてきている．

こうしたオープンデータへのアクセス方法の拡張やエンドポイント探索手法の実装では，それらのエンドポイントへのアクセス手法の効率化を行わないと，エンドポイントやその間の通信路に大きな負荷がかかってしまう点が課題となる[Kadono 14b]．この課題に対する１つのアプローチが，こうした探索問題をオンライン学習手法問題の１つである BLMAB(Budget-Limited Multi-Armed Bandit) 問題として定式化・拡張[Kadono 14b] し，BLMAB アルゴリズムやその派生アルゴリズム等[Kadono 14a] を用いて効率化を試みる方

法である．

分散データソースへのアクセス効率化手法としては，これらのアクセス時のクエリやアクセス先選択の効率以外にも，データソースとの間のネットワーク構成上の非対称性などを利用したアクセスの効率化手法があり，たとえば，P2P 型・モバイルエージェント[White 94] 型のアクセスの効率化手法が提案されてきている[Fukuta 12]．本研究の最初のステップでの目的は，これらの複数のデータアクセス手法の効率化手法を，オープンデータへのアクセス方法の容易化・拡張手法と組み合わせて利用可能とすることで，効率的で容易なオープンデータへのアクセスを実現するためのソフトウェア基盤の実現とした．本研究では，その実現のコアとなる機構の試作を中心として，その要素技術の開発もあわせて行う形で，その応用の可能性を検討した．

3 試作基盤フレームワークの概観

本基盤フレームワークは，モビリティを持たせたソフトウェアの開発実行プラットフォームである MiLog[Fukuta 01] を用いてその試作を進めた．MiLog は，文献[Fukuta 12] 等での実績があり，本試作における実装・検証の効率化には有効であると考えられる．

図 1 は，本システムの動作の概観を示すものである．図 1 では，SPARQL 拡張アクセス処理エンジンを，データソース(右下)，クライアントサイド(左上)，およびその中間となる proxy サイト(右上) となるホスト上にそれぞれ本実行環境を起動している．この例では，各ホストはそれぞれ OS X 10.10，UbuntuLinux 14.04(仮想環境上)，および Ubuntu Linux 10.04(仮想環境上) で動作しており，それぞれの動作環境上で MiLog を動作させ，本試作システム上で試作した機構を動作させている．ここでの仮想環境の実行には，Parallels 8.0 を用いている．各ホスト上での SPARQL 処理エンジンは，この例では仮に MiLog 上で簡易的に準備したものを用いている．

このように，必要に応じて処理エンジン等を異なるホスト上に配置可能とし，その処理性能および負荷等の観測を行えるようにしている．図 1 では，テストクエリを実行した際のクエリの動作の流れをモニタする状況を示している．

これら以外には，MiLog の持つ Web サーバ機能を用いて，クライアントサイドでの Web ブラウザ経由での簡易ユーザインタフェース等との連携機能の試作を進めてきており，これらの成果は近日中に公開の予定である．

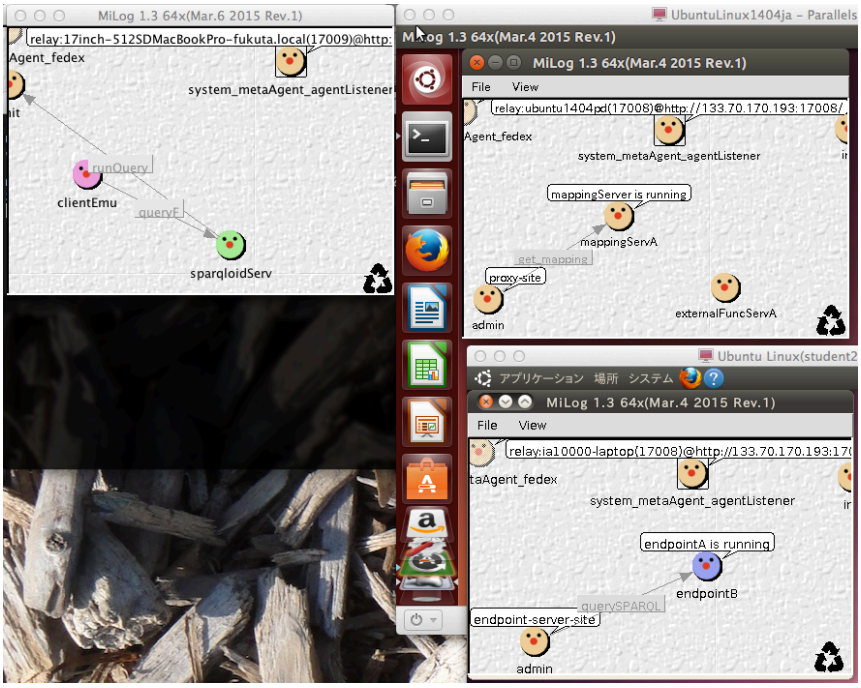

図 1．試作フレームワークに基づく動作例の概観（動作モニタ画面）

4 関連開発技術の概観

4-1 SPARQL クエリの実行時間予測技術の開発

オープンデータを相互に関係づけ可能にしたものは別名 LOD(Linked Open Data)とも呼ばれ，LOD 検索はその検索を受け付けるエンジンであるエンドポイントに対して RDF (Resource Description Framework)検索のクエリ標準言語である SPARQ に基づくクエリを発行して行う．オントロジーに基づく推論をエンドポイントで利用する場合，計算の複雑性の高さなどの課題があり，エンドポイントにおける推論器による推論機能の提供は重要な課題の 1 つとなっている．

LOD 検索において，問い合わせを受け取る側は，検索結果を得るまでにかかる時間をどの程度まで許容可能か，推論の厳密さをどの程度まで妥協することができるかなどの，問い合わせを送る側の意図を把握することが難しい．この課題に対処するために，LOD 検索クエリ作成者が，検索意図に応じた適切なクエリを作成する場面を支援するというアプローチも考えられる．しかし，クエリ作成者が，エンドポイントにおけるクエリの実行にどの程度の時間を要するかを常に予想しながら，適切なクエリを作成することは，容易でない．そうした洗練されたクエリを作成できない検索者から送られたクエリの実行のためには，エンドポイント側での多くの計算資源を消費してしまう.クエリの実行に伴うオントロジーに基づく推論時間を推定して，推論に関して極端に処理負荷の高いクエリの実行を抑制したり，あるいはより低い負荷で実行可能な近似したクエリの実行に置き換えたりすることで，そうしたエンドポイント側における推論の負荷を低減できる可能性がある．

本研究では，このような目的のために，LOD 検索クエリの問い合わせ文の実行における低負荷化と最適化をエンドポイント側で自動的に行うための手法の第一歩として，推論付き LOD 検索を対象とした，クエリとオントロジーに基づく検索実行時間の高負荷発生予測を行う手法を検討し，その機構を試作した．

本機構を用いてエンドポイントレベルで時間のかかるクエリ実行を回避する仕組みを提供することで，これらの問題に対処する仕組みは次のように実現可能となる．まず，検索者によって送られたクエリを受け取った際に，受け取ったクエリの実行に時間がかかるかどうかを判定する．時間のかかるクエリでないと判定された場合は，そのままクエリを実行する．時間のかかるクエリであると判定された場合には，クエリの実行を拒絶すること，または時間のかからないクエリに書き換えた後に実行し，実行結果とともにクエリを書き換えたことの通知を行うことが考えられる．

これらの機能の実現に向けて，時間のかかるクエリ，かからないクエリの判定機構，クエリ実行の拒絶を検索者に伝える機構などを備えた仲介システムを試作している.仲介システム（フロントエンド EP と呼ぶ）は，検索者とエンドポイント（バックンド EP と呼ぶ）の間に位置する．フロントエンド EP を介したクエリ実行の詳細は，本稿では割愛する．

クエリ実行時間を推測するための方法として，機械学習器を用いて判定を行う方法と，過去に行われたクエリの中から類似するクエリを探し出し，その実行時間を調べる方法が考えられる．本研究では，まずは機械学習器を用いた実行時間のかかるクエリ判定について検討し，判定機構を試作した．

本研究の範囲では，本手法の実現効果を検証する予備実験および手法の基本的なアイデアまでを検討したが，その後の継続的な発展的検討の過程で，アンサンブル学習手法である Bagged C4.5 および Boosted C4.5 により高い精度を得ることができることが，文献[Yamagata14b]で示された．

4-2 SPARQL エンドポイントの探索効率化技術の開発

Linked Open Data(LOD) が急速の普及が進み，2014 年 3 月現在，400 を超える情報提供源のエンドポイントが公開されている．エンドポイントに対してのアクセスは，すでに規格化されたクエリ言語である SPARQL によるクエリを実行することによって行うことができる．一方で，エンドポイントがどのような問い合わせに対して有効なデータを持っているかを知ることは，スキーマを事前に決めてデータをテーブル形式で格納する関係データベースとは異なり，LOD への検索では必ずしも容易ではない[Kadono 14b]．エンドポイントの持つ外形的な情報のみからでは，無数にあるエンドポイント中から適切なものを選ぶことは困難であり，なおかつ，LOD の相互接続性という特性を考えれば，ユーザが得たい情報は必ずしも一つのエンドポイント内で完結するとは限らないため，複数のエンドポイントにまたがる検索（横断的検索）を効果的にかつ少ない負担で行えるようにすることが，重要な課題の１つとなる[Noguchi 13]．

情報源としてのエンドポイントが検索対象として特定の問い合わせにとって有益であるかどうかを判断す

るには，なんらかの別の事前の問い合わせによりその有益さを推定する必要がある．その予備的な事前の問い合わせを，対象が無数にある場合に行おうとすると，それぞれのエンドポイントやその通信路となるネットワークに大きな負荷がかかってしまうため，その効率化の必要がある．本研究では，これらの問題に対して，MAB(Multi-Armed Bandit) モデルに基づく手法を用いてエンドポイント探索を支援するシステムを試作している．

標準 MAB モデル[Robbins 52] では，プレイヤーが K 個のアームが付いたスロットマシンに対し，アームを実行するとそれぞれのアームに設定された独立かつ未知の分布に従った報酬が得られる中で，どのアームを選択するかを考える問題を扱う．この問題の目的は，得られる報酬の合計を最大化することであり，焦点は，最も良いアームを探しつつも可能な限り得られる報酬を増やさなければならない点にある．

MAB 問題は，情報収集と情報活用のトレードオフがある中で報酬を最大化することを考える．エンドポイント探索においても，評価クエリによる情報収集と情報活用のバランスを考慮することで効率化を図りたいため，MAB モデルの利用を考えるが，エンドポイント探索においては，評価クエリの実行時間や，ネットワークへの負荷といったコストを考慮しなければならない．そこで本研究では，MAB モデルに予算とコストの概念を追加した Budget-Limited MAB(BLMAB) モデル[Tran-Thanh 10][Tran-Thanh 12a] を利用する．BLMAB モデルでは，各アームにコストが設定されており，アームを引くためにはそのコストを予算から支払わなければならず，プレイヤーは予算に収まるようにアームを引かなければならない．

エンドポイント探索における BLMAB のモデルは，探索する際に生じるネットワークの負荷と，評価クエリの実行にかかる時間という二重のコストが存在するため，従来の BLMAB モデルに対し，時間の制約という概念を追加したモデルを導入する．

拡張 BLMAB のモデルでは，プレイヤーが K 個のアームを持つスロットマシンにおいて，1 ステップごとに任意の個数のアームを選択し，実行する．それぞれのアーム i にはコスト c_i が設定されており，プレイヤーはアーム i を実行するごとにコスト c_i を支払い，アーム i に設定されたばらつきのある報酬 μ_i を得る．このときプレイヤーは報酬 μ_i を事前には知らない．また，プレイヤーは予算 B を持ち，支払うコストの合計は予算 B に収まっていなければならないと同時に，ステップ数は T($\in$N) に制限される．プレイヤーの目的は，T ステップ以内に予算 B を駆使し得られる報酬を最大化するようにアーム i を選択・実行することである．拡張 BLMAB モデルでは，これらを考慮して従来の BLMAB モデルにおける制約式を再定義する．

KDE アルゴリズム[Tran-Thanh 12b] は，BLMAB 問題に対する強力なアルゴリズムの一つである．本研究では，エンドポイント探索における拡張 BLMAB モデルを適用した拡張 KDE アルゴリズムを準備し，拡張 ε-greedy アルゴリズムによる初期性能の評価をシミュレーションに基づいて評価するとともに，拡張 KDE アルゴリズムの適用について検討している．予備的な検討の結果としては，ε 値が，通常の MAB 問題ではおよそ ε=0.1 程度であるとされるが，本拡張 BLMAB モデルでは，その ε の値が 0.4 から 0.6 程度となるときが最適な結果を得られることが解析されている．

5 成果と今後の課題

本研究では，効果的にオープンデータにアクセスし，意味情報に基づいて適切な情報を検索・提示するためのソフトウェア基盤におけるアクセス効率化機構の試作を行ってきており，前節までにその現状を述べた．本ソフトウェア基盤および試作した機構の特長には，オントロジーマッピングに基づく異種データ検索技術にクエリ変換技術およびオンライン学習技術などを適用することで，必要となるデータの提供源そのものをも，種々のオープンデータから動的にかつ効率的に検索可能とする技術の実現を行おうとしている点も加えることができる．

本稿ではそのアクセス効率化機構の試作の概要のみを述べたが，その性能の詳細な解析，およびこれらを応用したシステムの実装事例における利点の検証とその知見の考察は，今後の課題であり，現在もその検討を進めている．また，SPARQL クエリ実行の効率化手段としては，RDFS 等の構造を利用したクエリの書き換え[Bischof 13] など，様々なアプローチが試みられている．本研究グループでも，エンドポイント上での OWL 推論を可能とした場合における実行時間の増加に対処する具体的な手法の開発[Yamagata 14b][Yamagata 14a] も別途行ってきており，これらと本機構の統合的な利用を行えるようにすることも，今後の課題である．

【参考文献】

[Bischof 13] Bischof, S. and Pollers, A.: RDFS with Attribute Equations via SPARQL Rewriting, in Proc. the 10th Extended Semantic Web Conference (ESWC2013), pp. 335–350 (2013)

[Fujino 12a] Fujino, T. and Fukuta, N.: A SPARQL Query Rewriting Approach on Heterogeneous Ontologies with Mapping Reliability, in Proc. IIAI International Conference on Advanced Applied Informatics(IIAI-AAI 2012), pp. 230–235, Fukuoka, Japan (2012)

[Fujino 12b] Fujino, T. and Fukuta, N.: SPARQLoid – a Querying System using Own Ontology and Ontology Mappings with Reliability, in Posters and Demonstrations Track, The 11th International Semantic Web Conference(ISWC2012) (2012), (demonstration)

[Fujino 14] Fujino, T. and Fukuta, N.: Utilizing Weighted Ontology Mappings on Federated SPARQL Querying, in Kim, W., Ding, Y., and Kim, H.-G. eds., Lecture Notes in Computer Science, Vol. 8388, pp. 331–347, Springer-Verlag (2014)

[Fukuta 01] Fukuta, N., Ito, T., and Shintani, T.: A Logic-based Framework for Mobile Intelligent Information Agents, in Poster Proc. of the Tenth International World Wide Web Conference(WWW10), pp. 58–59 (2001)

[Fukuta 12] Fukuta, N.: A Mobile Agent Approach for P2P-based Semantic File Retrieval, Journal of Information Processing, Vol. 20, No. 3, pp. 607–613 (2012)

[Kadono 14a] Kadono, Y. and Fukuta, N.: LAKUBE: An Improved Multi-armed Bandit Algorithm for Strongly Budget-Constrained Conditions on Collecting Large-Scale Sensor Network Data, in Proc. 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI2014), pp. 1089–1095 (2014)

[Kadono 14b] Kadono, Y. and Fukuta, N.: An Online Learning-based Efficient Search System for Sufficient SPARQL Endpoints using Extended Multi-armed Bandit Algorithm, in Poster and Demo Proc. of the 4th Joint International Semantic Technology Conference (JIST2014) (2014), (poster with demonstration)

[Ladwig 10] Ladwig, G. and Tran, T.: Linked Data Query Processing Strategies, in Proc. International Semantic Web Conference (ISWC2010) PART I, pp. 453–469 (2010)

[Noguchi 13] Noguchi, H., Fujino, T., and Fukuta, N.: On Implementing SPARQLoid and its Query Coding Support Framework – Querying with Weighted Ontology Mappings, in Proc. The 3rd Joint International Semantic Technology Conference (JIST2013) (2013), (demonstration)

[White 94] White, J. E.: Mobile Agents Make a Network an Open Platform for Third-Party Developers, IEEE Computer, Vol. 27, No. 11, pp. 89–90 (1994)

[Yamagata 14a] Yamagata, Y. and Fukuta, N.: Approximating Inference-enabled Federated SPARQL Queries on Multiple Endpoints, in Proc. ISWC2014 Posters and Demonstrations Track, a track within the 13th International Semantic Web Conference(ISWC2014), pp. 441–444 (2014)

[Yamagata 14b] Yamagata, Y. and Fukuta, N.: A Dynamic Query Optimization on a SPARQL Endpoint by Approximate Inference Processing, in Proc. 3rd IIAI International Conference on Advanced Applied Informatics (IIAI AAI2014), pp. 161–166 (2014)

[Robbins 52] Robbins, H. Some aspects of the sequential design of experiments. Bulletin of the AMS 55:527-535, 1952.

[Tran-Thanh 10] Tran-Thanh, L. Chapman, A. Munoz De Cote Flores Luna, J. Rogers, A. and Jennings, N. Epsilon-First Policies for Budget-Limited Multi-Armed Bandits. In, Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010. , 1211-1216, 2010.

[Tran-Thanh 12a] Tran-Thanh, L. Chapman, A. Rogers, A. and Jennings, N. Knapsack based optimal policies for budget-limited multi-armed bandits. In, Twenty-Sixth AAAI Conference on Artificial Intelligence(AAAI-12), 1134-1140, 2012

[Tran-Thanh 12b] Tran-Thanh, L. Budget-limited multiarmed bandits. University of Southampton, Faculty ofPhysical and Applied Sciences, Doctoral Thesis, 2012

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
クエリ変換手法に基づく LOD 検索の高速化のための機構を備えた SPARQL エンドポイントの試作	第 28 回人工知能学会全国大会講演論文集, 1G4-OS-19a-5in	2014.5
Multi Armed Bandit モデルに基づくエンドポイント探索支援システムの試作	第 28 回人工知能学会全国大会講演論文集, 1G4-OS-19a-4in	2014.5
LOD 検索の高速化のための機構を備えた SPARQL エンドポイントにおけるクエリ実行性能の解析	人工知能と知識処理研究会, 電子情報通信学会技術研究報告,pp25--30	2014.8
An Online Learning-based Efficient Search System for Sufficient SPARQL Endpoints using Extended Multi-armed Bandit Algorithm	Poster and Demo Proc. of the 4th Joint International Semantic Technology Conference (JIST2014), CEUR Vol.1312, pp.136--139	2014.12
オントロジーマッピングを用いた意味情報に基づくオープンデータへのアクセス効率化機構の試作	第 29 回人工知能学会全国大会講演論文集, 1G3-OS-08b-3	2015.5
Toward an Agent-based Framework for Better Access to Open Data by using Ontology Mapping and their Underlying Semantics	Proc. 3rd IIAI-AAI International Conference on Smart Computing and Artificial Intelligence (ICSCAI2015)	2015.7