3O1-1in

SPARQL取得結果に対するプロパティに基づいた評価手法の検証

Verification of a Property-based Evaluation Method for the Retrieval Results with SPARQL Queries

一瀬詩織　　小林一郎
Shiori Ichinose　Ichiro Kobayashi

お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学領域
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu　University

In recent years, Linked Open Data (LOD) has been attracting attention as the technology of integrating the utilization of online resources has been developed. LOD datasets have published in the RDF framework, and can be read automatically by computers. SPARQL is an RDF query language and can retrieve data which satisfy specific requirements. Because the retrieved data are equally treated, this often causes a problem to find out useful resources from many retrieval results. Considering this, in this study, we examine the ability of our proposed methods of ranking the retrieval results with SPARQL queries, and compare the method with other methods.

1. はじめに

近年，Web 上に公開された情報を統合利用するための技術として，Linked Open Data(LOD) が注目されている．LOD データセットは RDF フレームワークで公開されており，機械による自動処理がしやすいという利点がある．LOD データセットから条件を指定してデータを取得する場合，RDF クエリ言語の 1 つである SPARQL が用いられる．取得された結果はすべて同列に扱われるため，整理されていない多数の取得結果から有用なデータを探すことはしばしば困難である．本研究では，先に提案した SPARQL クエリ取得結果の重要度計算手法 [Ichinose 14] を用いて取得結果のランキングを行い，結果を他手法と比較することで，手法の有効性についての検証を行った．

2. 関連研究

SPARQL クエリによる問い合わせ結果の重要度計算手法として，Mulay らによって定義された，リソース評価のフレームワーク [Mulay 11] がある．この手法では LOD データセット間の rdf:sameAs 等のリンクを考慮し，データセット，リソース，トリプルの 3 つのレイヤーでそれぞれスコアリングを行っている．LOD 全体を俯瞰したスコアリングを行える一方，この手法ではリソースの評価に利用している情報はデータセット間のリンク情報のみであり，データセット内の関係性については考慮していない．Bamba らの研究 [Bamba 04] では，Semantic Web データセットへ RDF クエリを用いた問い合わせを行った場合，リソースの重要度とプロパティの頻度情報，グラフの大きさを用いた検索結果のランキングを行う手法を提案している．先に提案した重要度計算手法 [Ichinose 14] のアルゴリズムは Bamba らの手法に基づいており，リソース，プロパティの重要度をデータセット内部におけるリソース間関係に基づいて算出している．本研究では手法の検証のため，Bamba らのアルゴリズムを一部変更したものをベースラインとして設定し，手法間のランキング比較を行った．

連絡先: 一瀬詩織，お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学領域，〒 112-8610 東京都文京区大塚 2-1-1，03-5978-5708，ichinose.shiori@is.ocha.ac.jp

3. SPARQL 検索結果のランキング手法

SPARQL クエリによって得られた検索結果をクエリに代入し，クエリ構造を持った結果グラフとして定義する．それぞれのグラフについて提案手法のアルゴリズムを用いたスコア付けを行う．グラフには複数のリソースとプロパティが含まれており，アルゴリズムによってそれぞれの重要度を重み付けすることでグラフ全体のスコア付けをする．すべてのグラフのスコア付けが行われた後，スコアに基づいて決定された順位を検索結果のランキングとして出力する．

3.1 リソースとプロパティの重要度

リソースの重要度には情報検索の分野で利用される PageRank アルゴリズム [Page 98] を用いる．またプロパティの重要度計算にはプロパティの主語が属するクラスに着目し，あるクラスにおけるプロパティの出現頻度とクラスに対する希少性を考慮した指標 PF・ICF を定義する．プロパティ頻度 (PF) はあるクラスにおいて，そのプロパティが使われる頻度を表す．逆クラス頻度 (ICF) はすべてのクラスにおいてそのプロパティが使われる頻度の逆数を表す．この指標は文書処理の分野で用いられる TF・IDF を参考にしたもので，作家や大学などの特定のクラスにおいて多く出現するプロパティに高い値を与える．

3.2 PFICF を用いたリソース評価手法（提案手法 1）

リソースをノード，プロパティをエッジとし，検索結果から作成したグラフに [Bamba 04] のクエリ評価アルゴリズムを適用してグラフの評価を行う．ノードとエッジの重要度には 3.1 で定義した指標を用いる．

3.3 RDF トリプルを単位とした評価手法（提案手法 2）

RDF データの情報はプロパティとリソースの 3 つ組 (トリプル) により記述される．[Bamba 04] における結果グラフの評価アルゴリズムではグラフ中のリソース (ノード) とプロパティ (エッジ) を別々に評価しているが，提案手法 2 ではノードとエッジを同時に評価するトリプルの評価式をアルゴリズムに導入し，トリプル単位でのスコア付けを行う．

RDF トリプルの主語，述語，目的語を n_s，e，n_o で表したとき，評価式 $TripleScore$ を以下のように定義する．

$$TripleScore = \frac{Imp(n_s) \times PFICF(n_s, e) \times Imp(n_o)}{linkNum(n_s) + linkNum(n_o) - 1} \quad (1)$$

$Imp(n)$，$PFICF(n, e)$ はそれぞれリソースの重要度，プロパティの重要度を表す．また，$linkNum(n)$ はノード n から出るエッジの本数を表す．ここで $linkNum(n)$ によるスコアの分割は，グラフ評価アルゴリズムにおいて同じリソースを主語として，目的語としてなど複数回評価する場合に行われる．

手法 2 におけるグラフ評価のアルゴリズムを以下に示す．ここで $decayFactor$ はユーザの興味の強さの減退を表す定数である．1.0 以下の値に設定することで SELECT 節で選ばれたリソースやプロパティとグラフ上の距離が遠いトリプルの重要度を低減させる．

アルゴリズム:

1. $decay = 1.0$, $score = 0.0$ に初期化する．
2. Adj を SELECT 節で選択されたノードを含んだトリプルの集合で初期化する．
3. Adj が空になるまで以下を繰り返す:
   (a) $ClassedEdges$ を Adj のノードのクラスとノードから伸びたエッジの集合とし，(c, e) で表す．
   (b) $score(r) += \sum_{t \in Adj} TripleScore[t] * decay$
   (c) $decay* = decayFactor$ $(decayFactor < 1.0)$
   (d) Adj と隣接した，まだ訪れていないトリプルで Adj を初期化する．

4. 手法の検証

提案手法は SPARQL クエリの問い合わせ構文，SELECT 節と WHERE 節を用いた問い合わせについて想定した手法である．本研究では具体的な問い合わせ例として「大学について問い合わせたい場合」を想定し，2 つの検証を行った．まず「大学という属性を持つリソース」を問い合わせる場合について取得結果をランキングした場合，有用な大学が上位に含まれているかどうかの検証を行った．有用な大学であるかどうかの指標には，上海交通大学による「世界大学学術ランキング (ARWU)」とイギリスの TIMES Higher Education による「THE 世界大学ランキング (THE-TR)」の 2 つのランキングデータを利用した．また「特定の大学についての情報」を問い合わせる場合については，問い合わせ結果に対して複数手法でランキングを行い，手法ごとの順位の付き方について考察を行った．比較手法には Bamba らのアルゴリズムを利用した．ただしリソースの重要度については提案手法と同じ PageRank スコアを用いている．すべての実験において，データセットには DBpedia3.8 を用いた．

表 1: 手法の検証に用いた SPARQL クエリ

ID	クエリ
1	SELECT ?property ?object WHERE{ ?subject rdf:type dbo:University.}
2	SELECT ?subject WHERE{ :University_of_Tokyo ?property ?object.}

4.1 属性を指定したリソース問い合わせ

ある地域に住んでいる人，あるジャンルに属する映画など，属性を指定して主語となるリソースを問い合わせるパターンはしばしば用いられる．今回は「大学という属性を持つリソース」に対し外部の大学ランキングとの比較を行うことで，この問い合わせに対するランキングの検証を行った．提案手法は大学ランキングに特化した手法ではないが，外部のランキングとの比較を行うことで一般的に評価が高い，つまり有用な大学がどれくらい上位に含まれているかという検証ができると考える．比較データには 2014 年度の ARWU[*1] と THE-TR[*2] のデータを用いた．手順として，まず表 1-1 のクエリにより問い合わせを行い，得られた 16,338 件の取得結果について提案手法 1 を用いたランキングを行った．次にランキング結果と比較データのそれぞれ上位 10 件，20 件，30 件，40 件，50 件の大学集合に対し，Jaccard 係数による類似度を求めた．有用な大学が上位に含まれているかどうかの基準は，提案手法の大学集合に含まれたリソースの 70%が比較データの大学集合に含まれていた場合，すなわち類似度が 0.54 以上である場合とした．提案手法 1 により得られたランキング結果の上位 10 件を表**??**に示す．また提案手法と比較データとの類似度を表 2 に示す．なお使用したクエリの構造から提案手法 1，2 のランキング結果は同一となるため，類似度の表は 2 つの提案手法で共通である．

表 2: ランキング間の類似度

比較対象	10	20	30	40	50
提案手法 - ARWU	0.54	0.54	0.54	0.48	0.41
提案手法 - THE	0.54	0.60	0.50	0.48	0.45

表 2 より，上位 20 件までのランキング結果において有効なランキングが得られていることが確認できる．上位 50 件までの比較においても 50%以上のリソースの一致が見られ，提案手法がこの問い合わせ結果のランキングについて一定の有効性を持っていると言える．リソースの一致率が下がっている一因として，提案手法では英語圏外の地域の大学についてリソースのスコアが低く計算される傾向が見られた．使用したデータセットは英語版の DBpedia であり，英語圏の大学のデータがその他の大学よりも多くなりがちであることがリソーススコアの偏りを招いていると考えられる．また，提案手法では大学のみをランキング対象としているが，比較データでは研究所も対象に入れている，といった対象範囲の不一致も見られた．

4.2 リソースの関連情報問い合わせ

特定の地域に関する情報，特定の人物に関する情報など，あるリソースを指定し関連するプロパティを求める問い合わせもよく利用されるパターンの 1 つである．今回は「東京大学のリソースを主語としたプロパティ，目的語」を問い合わせた場合についてベースラインの手法と提案手法によるランキングを行い，各手法での順位のつけ方について比較を行った．表 1-2 のクエリを用いて問い合わせを行い，290 件の取得結果が得られた．scoreEdge=0.7, scoreNode=0.3, decayFactor=0.5 に設定し，それぞれベースライン，提案手法 1，提案手法 2 を用いてランキングを行った．結果を表 5，表 6，表 7 に示す．また，大学に関する有用なプロパティとして，属性 dbo:University

*1 http://www.shanghairanking.com/ARWU2014.html
*2 http://www.timeshighereducation.co.uk/world-university-rankings/2013-14/world-ranking

表 3: 大学に属するプロパティ

順位	URI	スコア
1	http://dbpedia.org/ontology/city	1.89
2	http://dbpedia.org/property/campus	1.50
3	http://dbpedia.org/ontology/campus	1.28
4	http://dbpedia.org/property/established	1.22
5	http://dbpedia.org/ontology/state	1.19
6	http://dbpedia.org/property/city	1.078
7	http://dbpedia.org/ontology/affiliation	0.97
8	http://dbpedia.org/property/students	0.95
9	http://dbpedia.org/ontology/type	0.93
10	http://dbpedia.org/ontology/country	0.93

に属するプロパティスコア (PFICF 値) を調査した．結果を表 3 に示す．

表 3 は dbo:University の属性を持つリソースにのみ多く見られるプロパティであり，これらのプロパティを含むデータは大学特有の情報を持つデータであると考えられる．提案手法 1 はプロパティ“dbo:country”，“dbo:city”，“dbo:campus”，“dbo:state”，“dbo:established” を含んだデータが高スコアを付けられており，他手法よりもリソースの特性を重視したランキング結果となっている．提案手法 2 もプロパティ“dbo:country”，“dbo:state” などが高スコアを付けられているが，こちらは目的語に “:Tokyo” など土地のリソースを含むデータがより高くスコア付けされていた．土地のリソースは他のリソースよりも一般にリソーススコアが高いため，手法 2 は手法 1 よりもリソーススコアの影響を受けやすいと考えられる．

どのランキングにも共通して，複数の「似た」データがランキングに入るという問題が見られた．‘ontology’ と ‘property’ とで同じ性質を持つプロパティが二重に定義されており，結果として同じようなデータが複数上位に含まれてしまっている．これらは本手法の目的である有用なリソースの発見を阻害している．この問題を解決するためにはデータの類似度を定義し，しきい値以上のデータを統合するなどの対策が必要であると考えられる．

5. おわりに

本研究では先に提案した SPARQL クエリ検索結果の重要度計算手法について，具体的に大学情報の問い合わせを想定したランキングを行い，結果の検証を行った．大学リソースの問い合わせにおいては外部の専門的な大学ランキングとの比較を行い，20 位までのランキング結果に対して有用なリソースを上位にランキングできていることを示した．また特定の大学データの問い合わせにおいては東京大学のリソースに関するデータのランキングを行い，提案手法 1，2 が，ベースラインの手法よりも大学特有のデータを高くスコア付けできていることを示した．これらの検証結果から，提案手法によるランキングは属性を指定したリソースの問い合わせ，あるリソースに関するデータの問い合わせにおいて，有用なデータを見つけやすい検索結果を提供できると考えられる．一方で問題点として，プロパティを含んだ検索結果において，複数の意味の等しいデータが含まれてしまう点が挙げられた．今後はプロパティの類似度を考慮し，意味の等しいデータをランキング結果から省くことで，ランキング結果を改善する手法について検討していきたい．

参考文献

[Ichinose 14] Ichinose, S., Kobayashi, I., Iwazume, M., and Tanaka, K.: Ranking the Results of DBpedia Retrieval with SPARQL Query. Semantic Technology. Springer International Publishing, pp.306-319 (2014)

[Bamba 04] Bamba, B. and Mukherjea, S.: Utilizing Resource Importance for Ranking Semantic Web Query Results. Proc. Semantic Web and databases, 2nd Int. Workshop (SWDB 2004), Toronto, Canada, Revised selected Papers, pp.185-198 (2004)

[Mulay 11] Mulay, K. and Kumar, P. S.: SPRING: Ranking the results of SPARQL queries on Linked Data, Proc. 17th International Conference on Management of Data (COMAD), Bangalore, India (2011)

[Page 98] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank citation ranking: bringing order to the web (1998)

表 4: 提案手法によるリソースランキング上位 10 件

順位	URI
1	http://dbpedia.org/resource/Harvard_University
2	http://dbpedia.org/resource/University_of_Oxford
3	http://dbpedia.org/resource/University_of_Cambridge
4	http://dbpedia.org/resource/Columbia_University
5	http://dbpedia.org/resource/Yale_University
6	http://dbpedia.org/resource/Stanford_University
7	http://dbpedia.org/resource/University_of_California,_Berkeley
8	http://dbpedia.org/resource/Massachusetts_Institute_of_Technology
9	http://dbpedia.org/resource/University_of_Chicago
10	http://dbpedia.org/resource/Princeton_University

表 5: 東京大学のデータ上位 10 件 (ベースライン)

順位	?property	?object	スコア
1	http://dbpedia.org/property/country	http://dbpedia.org/resource/Japan	3.06
2	http://dbpedia.org/ontology/country	http://dbpedia.org/resource/Japan	2.81
3	http://dbpedia.org/property/latinName	Universitas Tociensis@en	2.43
4	http://dbpedia.org/ontology/wikiPageExternalLink	http://www.u-tokyo.ac.jp/	2.37
5	http://dbpedia.org/ontology/wikiPageExternalLink	http://search.japantimes.co.jp/cgi-bin/nn20090811i1.html	2.37
6	http://dbpedia.org/ontology/wikiPageExternalLink	http://www.u-tokyo.ac.jp/index_e.html	2.37
7	http://dbpedia.org/ontology/wikiPageWikiLink	http://dbpedia.org/resource/World_War_II	2.36
8	http://dbpedia.org/property/doctoral	6022(ttp://www.w3.org/2001/XMLSchema#int	2.15
9	http://dbpedia.org/property/imageSize	220(ttp://www.w3.org/2001/XMLSchema#int	1.98
10	http://dbpedia.org/ontology/wikiPageWikiLink	http://dbpedia.org/resource/Japan	1.95

表 6: 東京大学のデータ上位 10 件 (提案手法 1)

順位	?property	?object	スコア
1	http://dbpedia.org/ontology/country	http://dbpedia.org/resource/Japan	4.12
2	http://dbpedia.org/ontology/city	http://dbpedia.org/resource/Bunkyo,_Tokyo	3.99
3	http://dbpedia.org/ontology/city	http://dbpedia.org/resource/Bunky%C5%8D,_Tokyo	3.978
4	http://dbpedia.org/property/campus	http://dbpedia.org/resource/Urban_area	3.34
5	http://dbpedia.org/property/country	http://dbpedia.org/resource/Japan	3.11
6	http://dbpedia.org/ontology/state	http://dbpedia.org/resource/Tokyo	2.86
7	http://dbpedia.org/ontology/campus	http://dbpedia.org/resource/Urban_area	2.83
8	http://dbpedia.org/ontology/wikiPageWikiLink	http://dbpedia.org/resource/World_War_II	2.73
9	http://dbpedia.org/property/established	1877(ttp://www.w3.org/2001/XMLSchema#int	2.39
10	http://dbpedia.org/ontology/wikiPageWikiLink	http://dbpedia.org/resource/Japan	2.33

表 7: 東京大学のデータ上位 10 件 (提案手法 2)

順位	?property	?object	スコア（ ^ E-12）
1	http://dbpedia.org/ontology/country	http://dbpedia.org/resource/Japan	11.0
2	http://dbpedia.org/property/country	http://dbpedia.org/resource/Japan	5.89
3	http://dbpedia.org/ontology/state	http://dbpedia.org/resource/Tokyo	3.11
4	http://dbpedia.org/ontology/wikiPageWikiLink	http://dbpedia.org/resource/World_War_II	2.31
5	http://dbpedia.org/property/campus	http://dbpedia.org/resource/Urban_area	2.17
6	http://dbpedia.org/ontology/wikiPageWikiLink	http://dbpedia.org/resource/Japan	1.97
7	http://dbpedia.org/ontology/campus	http://dbpedia.org/resource/Urban_area	1.85
8	http://dbpedia.org/property/state	http://dbpedia.org/resource/Tokyo	1.21
9	http://dbpedia.org/ontology/wikiPageWikiLink	http://dbpedia.org/resource/Tokyo	0.43
10	http://dbpedia.org/ontology/wikiPageWikiLink	http://dbpedia.org/resource/Urban_area	0.24