1N4-OS-19a-4

SPARQL取得結果に対するレイティング手法の検討

A Study on a Rating Method for the Results Retrieved by 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

‡お茶の水女子大学 基幹研究院 自然科学系
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

SPARQL queries are used for retrieving data from Linked Open Data (LOD) on the Web. In large-scale data sets, it is often difficult to grasp the data structure of a huge amount of retrieval results. In this study, we discuss appropriate rating scores for the retrieved results by SPARQL query which helps us to acquire the knowledge of data structure.

1. はじめに

Linked Open Data(LOD) の利用において，巨大なデータセットから一部のデータを取得するための手段として，SPARQL によるクエリ問い合わせが用いられている．DBpedia[*1] や data.gov[*2] といった大規模なデータセットでは SPARQL のためのエンドポイントを公開し，利用者が自由にデータセット内の探索を行える環境を提供している．しかし，容易にデータが利用できる環境が整備されていても，利用者が SPARQL を用いて目的のデータ集合を取得するには 2 つの課題がある．

ひとつは，利用者が SPARQL クエリを作成するためには，事前にデータセット内部のデータ構造の知識を必要とする点である．この課題に対しては様々な試みがなされている．利用者の構造理解を支援するためのメタデータの作成，構造の可視化，あるいは利用者のクエリ作成を支援するためのクエリ推薦，クエリの自動生成などである．

もうひとつの課題は，得られる検索結果の件数が多く，利用者が検索結果の全ての内容を把握できない点である．大規模なデータセットではひとつのクエリに対して 100 件以上の検索結果が得られることが多い．例えば DBpedia japanese で「http://ja.dbpedia.org/resource/東京都」を主語とするトリプルを検索すると 779 件の検索結果が返ってくる [*3]．これが問題になるのは特に，利用者が目的のデータを定めず探索的に検索を行っている場合である．検索結果が多く内容を把握しきれないために，利用者が潜在的に興味を持っているデータや，そのようなデータに繋がるプロパティを見落とす可能性がある．

本研究では，SPARQL 検索結果を特定のスコア順に並び変えることによって，ふたつめの課題である「検索結果からのデータ発見」を支援するためのアプローチを行う．あらかじめ目的データとなる可能性の高い有用なリソースやプロパティに高いスコアを付与しておき，検索結果を表示する際にスコア順で並び替えを行うことで，通常の検索結果よりも目的データの発見が容易になると考えられる．本稿では次数中心性などの指標をレイティングスコアに適用し，実際に DBpedia japanese のダンプデータに適用してリソースとプロパティのレイティングを行った結果から，SPARQL 探索における適切なレイティングスコアについて検討を行う．

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

∗1 http://dbpedia.org/

∗2 http://data.gov/

∗3 2016 年 3 月現在

2. 関連研究

問い合わせ検索結果のランキング手法として，著者らはこれまで検索結果のグラフから結果の重要度を計算する手法を提案し，被験者実験によりプロパティを指定したリソースの問い合わせにおいては有効な手法であるという結果を得た [Ichinose 15]．

Thalhammer ら [Thalhammer 14] は一つのリソースに対する複数の結果を集約するための指標として，DBpedia のリソース間関係に PageRank アルゴリズムを適用した “DBpedia PageRank” データセットを作成している．このスコアは実際に DBpedia のエンドポイントで利用することができ，SPARQL の ORDER BY 句を利用して検索結果をスコア順に表示することが可能となっている [*4]．DBpedia PageRank はリソースに対する指標であり，プロパティの検索結果に対しては並び替えが行えないため，著者らの目的であるクエリ検索結果のレイティングは部分的にしか達成しない．しかし既存の SPARQL の構文を用いているため，検索結果取得後のスコア計算と再配列のコストがかからないという点で，前述の著者らの手法より優れている．

本稿では Thalhammer らの指標と同様，グラフパターンを利用した指標を用いてリソースとプロパティのレイティングを行う．得られたスコアはクエリ内容に依存しないため，データセットへの組み込みと指標による並び替えが可能である．

3. クエリ検索結果からの知識発見

検索結果の内容を理解するための方法として，検索結果に含まれている，データセットでよく利用されている代表的なデータを把握することは有効であると考えられる．また，データセット全体では利用頻度が少ないが，その検索結果には含まれているデータはそうでないデータよりも情報量が大きいと考えられる．

LOD データセットのトリプルの集合はリソース (リテラル)，プロパティからなる有向グラフとして表せ，データセット中の

∗4 http://people.aifb.kit.edu/ath/

中心的なデータはグラフの中心性として求めることができる．本稿ではデータの中心性を表す指標として，次数中心性を用いる．

すべてのトリプルの集合を T，そのうち主語 s の出現するトリプルの集合を $S(s)$, 目的語 o の出現するトリプルの集合を $O(o)$，プロパティ p の出現するトリプルの集合を $P(p)$ とし，リソース r，プロパティ p の次数中心性 DC を以下のように定義する．

$$DC(r) = \frac{|S(r)| + |O(r)|}{2|T(r)|}$$

$$DC(p) = \frac{|P(p)|}{|T(p)|}$$

またデータセット全体では出現頻度が少ないが，その検索結果にのみ含まれているデータのスコアとして選択情報量を用いる．先に定義した次数中心性をそのデータの出現頻度とみなし，選択情報量 I を以下のように定義する．

$$I(d) = -\log_2 DC(d)$$

3.1 実験

先に定義したリソース，プロパティのスコアを計算し，その後 SPARQL クエリに対してスコアを用いた並べ替えを行う．データセットには DBpedia japanese で提供されている 2014 年 12 月 30 日時点のデータダンプを用いる．リソースに対する指標としては先に定義した次数中心性，選択情報量，さらに比較として Thalhammer らも利用している PageRank[page 98] のスコアを用いる．またプロパティに対する指標として次数中心性，選択情報量を用いる．SPARQL 問い合わせのパターンは「主語を指定したプロパティと目的語の問い合わせ」「目的語を指定した主語とプロパティの問い合わせ」の 2 つの基本的なパターンとする．実際の問い合わせに用いるクエリを表 1 に示す．

表 1: 実験に使用した SPARQL クエリ

1	SELECT ?p ?o WHERE{ <http://ja.dbpedia.org/resource/東京都> ?p ?o.}
2	SELECT ?s ?p WHERE{ ?s ?p <http://ja.dbpedia.org/resource/株式会社>.}

3.2 結果・考察

リソース，プロパティの選択情報量のスコア分布を図 1 に示す．情報量のスコア分布は冪乗則に従っており，DBpedia japanese では一部のリソースやプロパティは非常に多くのトリプルに出現するが，大多数は出現数が少ない (スケールフリー性を持つ) ことが分かる．

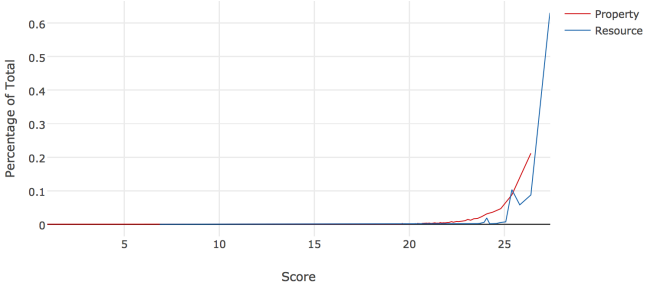

図 1: 選択情報量のスコア分布

クエリへの問い合わせ結果について，リソースの並び替え結果を表 2，表 4 に示す．表 2 では次数中心性と PageRank で wiki ページのテンプレートが上位に多く見られる．DBpedia japanese は実際に wiki ページのテンプレートから生成されたデータセットであり，データセットにもその情報が多く含まれていることがこの結果から分かる．情報量スコアは外部ページへのリンクが上位を占めているが，これは外部ページの URL はデータセット内における参照数がほとんど 1 となっているためである．これら外部ページの URL 同士をレイティングする場合はデータセット内部の構造だけでなく，ドメインの信頼性など外部の情報も利用する必要がある．

またプロパティの並び替え結果を表 3，表 5 に示す．プロパティの並び替えでは次数中心性において同一のプロパティが上位となった (表 3，表 5)．次数中心性は利用数の多いプロパティの方がスコアが高くなる．そのためプロパティの性質として同一主語に対して複数利用されるものと 1 つのみ利用されるものでは前者の方がスコアが高くなる傾向にある．データセット内の中心的なプロパティをこの性質によらずレイティングするためには，プロパティがひとつの主語に平均的に用いられている回数により，次数中心性を補正する必要があると考えられる．情報量スコアにおいては，クエリ 1 については土地に関する，クエリ 2 については，会社に関するプロパティが高いスコアを得た．これらのプロパティは特定の主語に対して利用され，データセット全体で中心的に利用されているプロパティと比較して利用数が少ないため，高いスコアを得たと考えられる．

4. おわりに

本稿では次数中心性，選択情報量をレイティングスコアとして DBpedia japanese の SPARQL 検索結果の並び替えを行い，SPARQL 探索における適切なレイティングスコアについて検討を行った．その結果，中心性のスコアではどの結果にも含まれるような共通のリソース，プロパティを上位に並べ，情報量のスコアではデータセット外部の URL や特定のリソースでしか利用されないプロパティを上位に並べる傾向にあることが分かった．またプロパティの性質により，主語に対して使用される数に違いがあることが分かり，中心性のスコアはこの性質を考慮して調整を行う必要がある．今後は wiki ベースでないデータセットに対しても同様のスコア傾向となるか検証を行い，被験者実験によるスコアの支援効果の評価を行いたいと考えている．

謝辞

本研究は、平成２７年度お茶の水女子大学大学院生研究補助金の補助を受けたものである．

参考文献

[Ichinose 15] 一瀬詩織，小林一郎，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，第 29 回人工知能学会全国大会論文集 (2015)

[Thalhammer 14] Thalhammer, A., & Rettinger, A.: Browsing dbpedia entities with summaries, In The Semantic Web: ESWC 2014 Satellite Events, pp. 511-515, Springer International Publishing (2014)

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

表 2: [クエリ 1] 検索結果?o の上位 5 件 (779 件中)

標準 (http://)	DC (http://)	I (http://)	PageRank (http://)
schema.org/AdministrativeArea	ja.dbpedia.org/resource/Template:Reflist	www.kanko.metro.tokyo.jp/	ja.dbpedia.org/resource/Template:Main
schema.org/Place	schema.org/Place	www.tcvb.or.jp/	ja.dbpedia.org/resource/Template:Commons&cat
www.ontologydesignpatterns.org/ont/d0.owl#Location	www.ontologydesignpatterns.org/ont/d0.owl#Location	commons.wikimedia.org/wiki/Special:FilePath/Tokyo_Montage_2012.png	ja.dbpedia.org/resource/Template:See_also
dbpedia.org/ontology/AdministrativeRegion	http://dbpedia.org/ontology/Place	www.data.jma.go.jp/obd/stats/etrn/view/nml_sfc_ym.php?...	ja.dbpedia.org/resource/Template:YouTube_channel
dbpedia.org/ontology/Place	ja.dbpedia.org/resource/Template:Flagicon	commons.wikimedia.org/wiki/Special:FilePath/Tokyo_Montage_2012.png?width=300	ja.dbpedia.org/resource/AFN

表 3: [クエリ 1] 検索結果?p の上位 5 件 (779 件中)

標準 (http://)	DC (http://)	I (http://)
www.w3.org/1999/02/22-rdf-syntax-ns#type	dbpedia.org/ontology/wikiPageWikiLink	dbpedia.org/ontology/leaderName
www.w3.org/1999/02/22-rdf-syntax-ns#type	dbpedia.org/ontology/wikiPageWikiLink	dbpedia.org/ontology/areaCode
www.w3.org/1999/02/22-rdf-syntax-ns#type	dbpedia.org/ontology/wikiPageWikiLink	dbpedia.org/ontology/areaCode
www.w3.org/1999/02/22-rdf-syntax-ns#type	dbpedia.org/ontology/wikiPageWikiLink	ja.dbpedia.org/property/code
www.w3.org/1999/02/22-rdf-syntax-ns#type	dbpedia.org/ontology/wikiPageWikiLink	ja.dbpedia.org/property/iso

表 4: [クエリ 2] 検索結果?s の上位 5 件 (10000 件中)

標準 (http://)	DC (http://)	I (http://)	PageRank (http://)
ja.dbpedia.org/resource/五戸美樹	ja.dbpedia.org/resource/ZIP-FM	ja.dbpedia.org/resource/Holux	ja.dbpedia.org/resource/4-Legs
ja.dbpedia.org/resource/新保友映	ja.dbpedia.org/resource/IMAGICA	ja.dbpedia.org/resource/GET-Film	ja.dbpedia.org/resource/Holux
ja.dbpedia.org/resource/ミランカ	ja.dbpedia.org/resource/C.A.L	ja.dbpedia.org/resource/Maneo	ja.dbpedia.org/resource/Amazon.co.jp
ja.dbpedia.org/resource/キャタピラー_(企業)	ja.dbpedia.org/resource/SUPER_GT	ja.dbpedia.org/resource/PRISM_VIDEO	ja.dbpedia.org/resource/Mmbi
ja.dbpedia.org/resource/桐生かえみ	ja.dbpedia.org/resource/SNK	ja.dbpedia.org/resource/QOOV	ja.dbpedia.org/resource/PFU

表 5: [クエリ 2] 検索結果?p の上位 5 件 (10000 件中)

標準 (http://)	DC (http://)	I (http://)
dbpedia.org/ontology/affiliation	dbpedia.org/ontology/legalForm	dbpedia.org/ontology/author
dbpedia.org/ontology/affiliation	dbpedia.org/ontology/legalForm	dbpedia.org/ontology/category
dbpedia.org/ontology/author	dbpedia.org/ontology/legalForm	dbpedia.org/ontology/company
dbpedia.org/ontology/category	dbpedia.org/ontology/legalForm	dbpedia.org/ontology/affiliation
dbpedia.org/ontology/company	dbpedia.org/ontology/legalForm	dbpedia.org/ontology/affiliation