4N-9

DBpediaを対象にしたリソースのランキング手法における一考察

一瀬詩織 † 小林一郎 † 岩爪道昭 ‡ 田中康司 ‡
† 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
‡ 独立行政法人 情報通信研究機構

1 はじめに

近年，Semantic Web の枠組みにおいて，Linked Open Data（LOD）の構築とその利用への取組みが盛んになってきている．LOD クラウドから必要なリソースを抽出するには，直接 URI を参照する，Semantic Web 検索エンジンを利用する，データセットで提供されている SPARQL エンドポイントを用いる，などの方法がある．このうち SPARQL エンドポイントを用いた手法では，ユーザは SPARQL クエリを用いて検索を行い，クエリの構造と一致した複数のリソースを取得することができる．取得したリソースが大量にあった場合，さらに情報の絞り込みを行うため，ユーザの要望に応じたリソースのランキングを提供することは有用である．本研究では代表的な LOD データセットの１つである DBpedia[1] において SPARQL によるリソース取得を行った場合に，ユーザの要求に合わせ，リソースを適切にランキングする手法について検討する．なお，SPARQL では条件にマッチした属性 (以下 property) も取得することが可能であるが，今回はリソースを取得する場合についてのみ考察する．

2 リソースの評価指標

SPARQL 検索結果のランキングとしては他に Mulay ら [2] による，LOD のグラフ構造を利用したランキング手法がある．本研究ではユーザの要求に応じたリソースのランキングを行うため，このようなデータセットのグラフ構造に加え，取得したリソースに関する情報を利用する必要があると考えた．PageRank アルゴリズム [3] は Web ページのグラフ構造からページの重要度を評価するアルゴリズムで現在 Web 検索のランキングに広く用いられており，LOD のような RDF トリプルによるグラフ構造にも適用可能である．今回は以下の 2 つの調査実験から，SPARQL クエリ検索結果を適切にランキングする手法についての考察を行った．

(1) DBpedia に PageRank アルゴリズムを適用した場合のランキング結果の特性と問題点

(2) SPARQL クエリの違いによる，取得リソースの持つ情報の変化

3 実験

3.1 実験環境

実験は [1] から提供されている DBpedia データセットを用いて行った．実験で使用したデータセットと計算環境を表 1 に示す．

表 1: 実験環境

CPU	Intel Core i7-3770K
メモリ	32GB
OS	Ubuntu 12.10
データセット	DBpedia 3.8

3.2 DBpedia データセットの PageRank 計算

DBpedia データセット内の RDF トリプルの Subject, property，Object の関係は，Subject と Object をノード，property を Subject から Object へのエッジとした有向グラフとして表すことができる．DBpedia 上のすべてのリソース数を n，あるリソース s へ $x \rightarrow s$ のエッジを持つリソース x の集合を B_s，リソース x から出るエッジの本数を c_x とし，以下の計算式に基づいた PageRank 値の計算を行った．

$$r_s = \frac{1-d}{n} + d \sum_{x \in B_s} \frac{r_x}{c_x}$$

Page ら [3] に従い，Dumping Factor は d=0.85 に設定した．また計算はべき乗法により行い，収束条件は $|r_s^k - r_s^{k-1}| < 1E-10$ とした．結果，PageRank 値が上位 5 件の DBpedia リソースを表 2 に示す．

A Study on a Ranking Method of Resources in DBpedia
†Shiori ICHINOSE(ichinose.shiori@is.ocha.ac.jp),
†Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
‡Michiaki IWAZUME(iwazume@nict.go.jp)
‡Kouji TANAKA(tanaka@nict.go.jp)
†Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University,2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610
‡National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289

表 2: PageRank 上位 5 件の DBpedia リソース (評価対象 9427434 件)

順位	URI(http://dbpedia.org/resource/〜)	PageRank
1	United_States	0.002770
2	France	0.001065
3	United_Kingdom	0.001031
4	Germany	9.41E-04
5	Race_and_ethnicity_in_the_United_States_Census	8.91E-04

3.3 SPARQL クエリ検索結果リソースの property 情報調査

クエリによる取得リソースの持つ情報の違いを調査するため，以下の SPARQL クエリを用いてリソースの取得を行った．

```
SELECT ?res WHERE {
{ ?res <?http://www.w3.org/1999/02/22-rdf-
syntax-ns#type > R .}
```

クエリ内の Object 要素 R に以下の 4 つの要素を用い，それぞれのクエリに対する結果リソースを取得した．また取得したリソースに対し，そのリソースを主語に持つトリプルの property 情報を取得した．

- http://dbpedia.org/ontology/Writer
- http://dbpedia.org/ontology/Actor
- http://dbpedia.org/ontology/Country
- ?x (すべてのリソース)

表 3 は R に http://dbpedia.org/ontology/Writer を用いた際の，出現頻度上位 10 件の property である．それぞれのクエリに対する結果リソース数と property 数を纏めたものを表 4 に示す．また，それぞれのクエリで出現頻度の高かった上位 100 件の property の共通出現率をまとめたものを表 5 に示す．

表 3: R:Writer の場合の出現 property 上位 10 件

順位	URI	出現回数	出現率
1	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	13743	1.0000
2	http://dbpedia.org/ontology/wikiPageRevisionID	13730	0.9991
3	http://www.w3.org/2000/01/rdf-schema#label	13730	0.9991
4	http://dbpedia.org/ontology/abstract	13730	0.9991
5	http://www.w3.org/2000/01/rdf-schema#comment	13730	0.9991
6	http://xmlns.com/foaf/0.1/isPrimaryTopicOf	13730	0.9991
7	http://dbpedia.org/ontology/wikiPageWikiLink	13730	0.9991
8	http://www.w3.org/ns/prov#wasDerivedFrom	13730	0.9991
9	http://dbpedia.org/ontology/wikiPageID	13730	0.9991
10	http://dbpedia.org/property/wikiPageUsesTemplate	13725	0.9987

表 4: 各問い合わせ結果の取得リソース数と property 数

R	取得リソース	property
All	1048576	41478738
Writer	13743	505540
Actor	2431	93593
Country	2710	143315

表 5: 出現頻度上位 100 件の property の共通出現率

R	All	Writer	Actor	Country
All	100	40	44	32
Writer	40	100	55	19
Actor	44	55	100	20
Country	32	19	20	100

3.4 考察

実験課題 (1) と (2) について，以下に考察を述べる．

(1) PageRank は参照数が多いページが上位にランキングされやすい傾向にある．表 2 のように，DBpedia における PageRank 計算では国などの土地のリソースが上位に多く出現した．土地のリソースは人やイベントなどの他のリソースから参照されることが多く，このような結果となったのは妥当であると考えられる．しかし，PageRank 値を SPARQL クエリ検索のランキングに適用した場合，どのようなクエリに対しても国のリソースが上位にランキングされてしまうような状況は望ましくない．クエリや検索結果の情報を利用し，よりユーザの求めている結果を上位にランキングさせるような仕組みが必要であると考えられる．

(2) property 情報の調査において，出現頻度が最も高い property はそれぞれのクエリによらず共通しており，表 3 の 10 件の property のうち 8 件は他のクエリにおいても出現頻度上位の 10 件に含まれていた．しかし，表 5 では property の共通出現率は比較的低くなっており，出現頻度が低い property にはクエリに共通でない，それぞれのリソース集合に特徴的な property が多く含まれていることが分かる．また Writer と Actor の property の共通出現率よりも Writer と Country の property の共通出現率の方が低い値となっていることから，同じ「人間」である要素を用いたクエリ検索では結果リソースの property も共通のものが多く出現することが分かった．このような property はそのリソース集合の特徴的な property であり，ユーザが絞り込みを行う際の有用な手掛かりとなることが考えられる．

4 おわりに

DBpedia での SPARQL 検索の結果リソースに対して適切なランキングを行うため，2 つの実験によってランキング手法の検討を行った．PageRank アルゴリズムはグラフ構造から重要度の評価を行う有効な手法の一つであるが，SPARQL 検索結果のランキングを行うためにはユーザの求める情報を考慮し，ランキング結果に反映させる必要がある．今後は，今回の実験で判明したクエリ結果のリソースに特徴的な property を利用し，リソースをランキングする手法について引き続き検討を行いたい．

参考文献

[1] http://dbpedia.org/

[2] Kunal Mulay and P Sreenivasa Kumar, SPRING: Ranking the results of SPARQL queries on Linked Data, Proceedings of the 17th International Conference on Management of Data (COMAD), Bangalore, India, December 2011.

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