2N5-OS-21b-4

DBpediaにおけるSPARQL検索結果のランキング手法

A Ranking Method for DBpedia Resources based on Retrieval Results with SPARQL Queries

一瀬詩織[*1] 小林一郎[*1] 岩爪道昭[*2] 田中康司[*2]
Shiori Ichinose Ichiro Kobayashi Michiaki Iwazume Kouji Tanaka

[*1]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 情報科学コース
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

[*2]独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

With the expansion of the Linked Open Data (LOD) cloud, methods retrieving data from their datasets have become more and more important. SPARQL is an RDF query language and is used for extracting data from LOD datasets. A SPARQL query are formed using triple patterns and can return elements matching with their patterns. However, SPARQL does not support the function of ranking query results and it is hard for us to find which resource is important. In this paper, we research on a ranking method for the results of SPARQL queries to DBpedia, a typical Semantic Web dataset in LOD cloud.

1. はじめに

近年，Linked Data の普及に伴い，RDF 形式などで構造化されたデータが Web 上に大量に公開されるようになった．Linked Data は構造化されたデータを相互にリンク付けして公開する手法であり，これによって様々な公開データセットがリンクで結びついた，グローバルなデータ共有の実現が可能となる．Linked Open Data (LOD) クラウドのデータセットは主に RDF 形式で公開されており，データの問い合わせにはしばしば SPARQL[*1] と呼ばれる RDF クエリ言語が用いられる．SPARQL クエリを用いることで構造化データの一部分を簡単に抽出することができるが，一方で SPARQL は抽出したリソースをランキングする手段を持たず，結果のうちどのリソースが重要であるかということは分からない．本研究では LOD クラウドのデータセットでも中心的な存在であり多様なリソースを持つ DBpedia[*2]を対象とし，SPARQL クエリによる検索を行った場合の結果のランキング手法について，以下の 2 点から考察を行った．

- PageRank アルゴリズムによるリソースの重要度評価
- SPARQL クエリによって取得したリソースの特性調査

2. 関連研究

Semantic Web のランキング手法については既に多くの研究がなされている．代表的な Semantic Web 文書の検索エンジン Swoogle[Ding 04] では PageRank アルゴリズムの考え方を Semantic Web へと適用した Ontology Rank を定義し，文章のスコアリングに用いている．一方で，RDF クエリ言語の問い合わせ結果に対するランキング手法についての研究はまだあまり行われていない．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．例として実際に生物関係の特許データベースに対し，提案手法を適用したランキングを行っているが，手法の正当性についての検証は十分にされていない．また，SPARQL クエリによる問い合わせ結果のランキング手法としては LOD クラウドにおけるリソース評価のフレームワークを定義した，Mulay らの SPRING[Mulay 11] がある．この手法ではデータセット間の 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．本研究では DBpedia データセットを対象とし，データセット内部におけるリソース間関係に基づいたリソースの重要度のスコアリング，およびスコアを利用したランキングについての考察を行う．また，リソースの持つプロパティの頻度情報についても調査し，結果リソースのスコアリングに対するプロパティ利用の可能性についても考察を行う．

3. PageRank アルゴリズムによるリソースの重要度評価

SPARQL 検索において，取得した複数のリソースをその重要度の順に並べることは有用であると考えられる．複数の DBpedia リソースとその間のプロパティ情報はリソースを点，プロパティ情報をリソース間のエッジとした有向グラフとして表すことができる．これは WWW におけるページとその間のリンクの関係に類似している．本研究ではこれを DBpedia のリソースグラフとして定義し，WWW において Web ページの重要度を決定するのに用いられている PageRank アルゴリズム [Page 98](3.2 に詳説) を用いて，リソースの重要度の評価を行った．

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コース 小林研究室，〒 112-8610 東京都文京区大塚 2-1-1，03-5978-5708，ichinose.shiori@is.ocha.ac.jp

∗1 http://www.w3.org/TR/rdf-sparql-query
∗2 http://dbpedia.org

3.1 リソースグラフの定義

DBpedia 内のトリプル t を (s,p,o) で表す．ここで s は主語，p は述語 (プロパティ)，o は目的語である．DBpedia データセットで定義されているすべての URI の集合を U，トリプルの集合を T としたとき，頂点集合を R，辺集合を E とするリソースグラフ $G=(R,E)$ を定義する．$r \in R$ は $s,o \in U$ であるようなトリプル $t \in T$ の主語 s または目的語 o であり，R は U の部分集合となる．また $t \in T$ について，$s,o \in R$ であった場合のリソース間関係 $s \to o$ を $e\{s,o\} \in E$ とする．

リソースグラフ G は DBpedia データセット内に含まれるリソースとリソース間の関係のみによって構成されたグラフであり，DBpedia データセット外部のリソースとの関係は含まれていない．

3.2 PageRank アルゴリズム

定義したリソースグラフについて，PageRank アルゴリズムを用いたリソースの重要度評価を行う．DBpedia 上のすべてのリソース数を $|R|$，あるリソース $r \in R$ へ $\{x,r\} \in E$ のエッジを持つリソース x の集合を B_r，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．

$$PR_r = \frac{1-d}{|R|} + d\sum_{x \in B_r} \frac{PR_x}{c_x}$$

d は Dumping Factor を表す．Page ら [Page 98] は経験的にこの値を 0.85 に定めており，本研究でも同様の値を用いて計算を行った．

3.3 PageRank 値の計算

リソースグラフに対し，べき乗法を用いた PageRank 値の計算を行った．収束条件は PageRank ベクトルの差分により判定した．安定した PageRank 値を得るため，収束条件 $|PR_r^k - PR_r^{k-1}| < 1E-X$ をそれぞれ $X=5,6,...,14$ に設定して計算を行い，$X=n$ と $X=n+1$ の場合における，PageRank 値によるリソースの順位変動を調べた．データセットは DBpedia で提供されている最新のデータセットである DBpedia3.8 を用いた．実験環境および使用したデータを表 1 に示す．また，べき乗法の繰り返し回数と計算時間，X の値によるリソースの順位変動をそれぞれ図 1，図 2 に示す．

図 1 のように，べき乗法の繰り返し回数と計算時間は X の値と比例して線形に増加した．その一方で図 2 より，順位変動するリソース数は X の値が増加するにつれて大きく減少し，$X=10$ 近辺では殆ど順位が安定していることが分かる．$X=10$ よりも値が大きい場合においては順位変動するリソース数は殆ど変化せず，むしろ 1000 位以上の大きな順位変動をするリソースの増加が見られたことから，X=10 の場合の PageRank 値をもっとも安定した値としてランキングに用いることとした．

表 1: 実験環境・使用データ

CPU	Intel Core i7-3770K
メモリ	32GB
OS	Ubuntu 12.10
データセット	DBpedia 3.8
総リソース数	9440897
総トリプル数	158373972

PageRank 値上位には，国や都市などの土地に関するリソースが多く見られた．表 2 は DBpedia リソースを PageRank 値

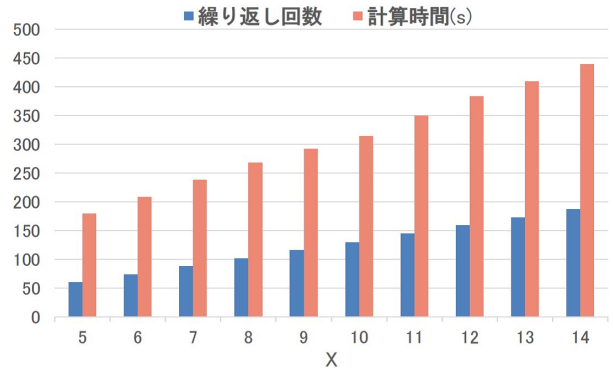

図 1: べき乗法の繰り返し回数と計算時間

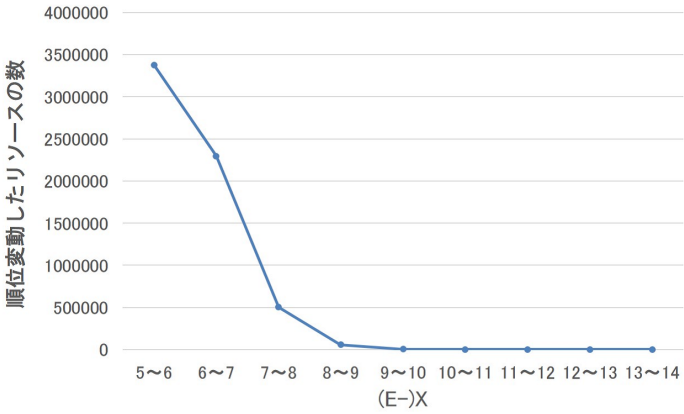

図 2: X の値によるリソースの順位変動

の順に並べた際の上位 10 件のリソースであるが，このうち 8 件が土地関係のリソースとなっている．PageRank アルゴリズムはその計算式により，多くのページからリンクされているほどページのスコアが高くなるという性質がある．土地情報はイベントの起こった場所，人物の生まれた場所，といったように様々なリソースとの関係を持ちやすく，参照される機会が多いために，このような高いスコアを持つ結果となったとことが考えられる．10 位以下のリソースについても国や，ロンドン，ニューヨークといった大都市のリソースが上位に見られた．PageRank 値を単純に SPARQL クエリ検索のランキングに適用した場合，このようなスコアの偏りがそのままランキングに現れ，常に土地のリソースが上位に来るような結果となることが考えられる．これを避けるためには，SPARQL クエリや検索結果の情報を取り入れ，ユーザの求めている結果をより上位にランキングさせるスコア付けを行うことが必要であると考えられる．

3.4 クエリ検索結果のランキング

以下の単純な SPARQL クエリに対して PageRank 値を用いた検索結果のランキングを行い，ランキング結果に対する考察を行った．SPARQL では，SELECT 節は取得したい要素を表す変数，WHERE 節は要素の条件を規定する．このクエリでは取得したいリソース `?res` に対し，そのタイプ情報 `A` を規定している．

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

表 2: PageRank 上位 10 件の DBpedia リソース
($X = 10$, 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
6	England	8.04E-04
7	World_War_II	7.56E-04
8	India	7.48E-04
9	Canada	7.10E-04
10	Italy	6.11E-04

クエリ内の Object 要素 A にはそれぞれ「Writer(作家)」,「Actor(俳優)」,「Country(国)」を定義している以下の 3 つのオントロジーを用いた．3 つのクエリはタイプを指定しているため，タイプの違いによる PageRank 値の偏りは起こらないと考えられる．

- http://dbpedia.org/ontology/Writer
- http://dbpedia.org/ontology/Actor
- http://dbpedia.org/ontology/Country

Writer，Actor，Country それぞれを用いて SPARQL 問い合わせを行った結果，取得したリソースの上位 10 件を表 3 に示す．また，取得リソースに対し，PageRank 値を用いたランキングを行った場合の上位 10 件の結果を表 4 に示す．

表 3 と 4 を比較すると PageRank 値によるランキング後はランキング前よりも “William Shakespeare”，“United States” といった著名な人物・国のリソースが上位に出現しており，PageRank 値は SPARQL クエリ検索結果のランキングにおいて，一定の効果があると考えられる．しかし表 4 では，作家のランキング結果において 6 位の「Tacitus（政治家)」，俳優のランキングにおいて 6 位の「Zhang Yimou（映画監督)」など，作家や俳優としてよりも他の属性で著名であると考えられる人物も上位にランキングされた．今回は単純な PageRank 値によるランキングを行ったが，実際のランキングではリソースの重要度だけでなく,「作家」であるリソースを探したい，といったクエリの内容とリソースの情報とがどれだけ一致しているかを考慮する必要があると考えられる．

4. SPARQL クエリによって取得したリソースの特性調査

SPARQL クエリによって取得したリソースがどのような情報を持つのか調査するため，前章で用いたものと同様の 3 つの SPARQL クエリによって取得したリソースに対し，さらにそのリソースが主語となるトリプルのプロパティ情報の取得を行った．また，比較対象として新たに A に ?x(すべてのリソース) を適用した場合についてリソースの取得を行い，同様にプロパティの取得を行った．それぞれのクエリに対する結果リソース数とプロパティ数を纏めたものを表 5 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 件のプロパティの共通出現率を纏めたものを表 6 に示す．

表 5: 各問い合わせ結果の取得リソース数とプロパティ数

A	取得リソース	プロパティ
All	1048576	41478738
Writer	13743	505540
Actor	2431	93593
Country	2710	143315

表 6: 出現頻度上位 100 件のプロパティの共通出現率

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

表 7 は A に “http://dbpedia.org/ontology/Writer” を適用して取得したリソースの出現頻度上位 20 件のプロパティとその出現率，および A に ?x を適用した場合の対象プロパティの出現率である．“http://www.w3.org/2000/01/rdf-schema#label”，“http://dbpedia.org/ontology/abstract” などの上位のプロパティは Writer を指定した場合とそうでない場合のどちらの場合でも出現頻度が高く，このようなプロパティはクエリによらない，一般的なプロパティであることが考えられる．一方で，16 位の “http://xmlns.com/foaf/0.1/surname”，18 位の “http://dbpedia.org/property/dateOfBirth” など，Writer を指定した場合とそうでない場合とで出現頻度が大きく異なるプロパティも見られた．このプロパティは取得したリソースの集合における特徴的なプロパティであり，ユーザが絞り込みを行う際にも有用な手掛かりとなることが考えられる．また表 6 で Writer と Country のプロパティの共通出現率よりも Writer と Actor のプロパティの共通出現率の方が高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，取得したリソースの持つプロパティも共通のものが多く出現することが分かった．今後はさらに調査を進め，クエリ検索におけるクエリと結果となるリソースとの一致度の評価などに生かしていきたい．

5. おわりに

本研究では 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，PageRank アルゴリズムを用いたリソースの重要度評価，リソース集合に特有のプロパティ情報の調査，の 2 つの実験を通じ，ランキング手法の検討を行った．PageRank 値を用いることで，単純な SPARQL クエリを用いた検索において重要なリソースは上位にランク付けされるが，クエリの意図を考慮したランキングを行うためには指定された要素と結果のリソースとの一致度をランキングに取り入れる必要がある．さらにプロパティ情報の調査から，タイプの違いによるプロパティの出現頻度の違いと，タイプの違いによるリソース共通のプロパティ，特有のプロパティが存在することが分かった．リソースの重要度はプロパティの頻度情報は手法は異なるものの，先行研究である Bamba ら [Bamba 04] の手法でも利用されている．今後は先行研究の手法を考慮しつつ，今回調査したリソースの重要度，プロパティ情報を取り入れたランキング手法について引き続き検討を行う．プロパティ情報についてもさらに調査を行い，SPARQL 問い合わせ結果のランク付けに利用したい．

表 3: SPARQL クエリ検索における上位 10 件の DBpedia リソース

順位	Writer	PR 値	Actor	PR 値	Coutry	PR 値
1	Ayn Rand	5.90E-6	Jet Li	2.55E-6	Alegria	1.62E-7
2	Aldous Huxley	5.76E-6	Tom Cruise	6.09E-6	Andorra	2.40E-5
3	A. E. van Vogt	6.43E-7	Ilona Staller	1.87E-7	Azerbaijan	9.62E-5
4	A. A. Milne	1.53E-6	Bruce Lee	3.61E-6	Aruba	1.23E-5
5	Allen Ginsberg	5.07E-6	Lee Armstrong	4.65E-8	Angola	3.04E-5
6	August Derleth	1.17E-6	Asia Carrera	1.30E-7	Albania	4.70E-5
7	Agatha Christie	5.64E-6	Sophie Dahl	2.64E-7	Afghanistan	7.85E-5
8	Anatole France	1.42E-6	Ginger Lynn	2.86E-7	Antigua and Barbuda	9.21E-6
9	Andr Paul Guillaume Gide	3.07E-8	Johnny Depp	4.83E-6	Anguilla	5.77E-6
10	Alan Garner	4.16E-7	Zhang Ziyi	7.08E-7	Akkadian Empire	7.07E-6

表 4: PageRank 値によるランキング結果上位 10 件の DBpedia リソース

順位	Writer	PR 値	Actor	PR 値	Coutry	PR 値
1	William Shakespeare	8.10E-5	Tom Cruise	6.09E-6	United States	2.77E-3
2	Cicero	3.42E-5	Jackie Chan	5.19E-6	France	1.07E-3
3	Charles Dickens	2.72E-5	Johnny Depp	4.83E-6	United Kingdom	1.03E-3
4	Johann Wolgang von Goethe	2.68E-5	Bruce Lee	3.61E-6	Germany	9.41E-4
5	Virgil	2.45E-5	Jet Li	2.55E-6	England	8.04E-4
6	Tacitus	2.44E-5	Zhang Yimou	1.59E-6	India	7.48E-4
7	J. R. R. Tolkien	2.42E-5	Andy Lau	1.55E-6	Canada	7.10E-4
8	Plutarch	2.40E-5	Chow Yun-fat	1.53E-6	Italy	6.11E-4
9	Thomas Aquinas	2.26E-5	Sammo Hung	1.42E-6	Japan	5.82E-4
10	Dante Alighieri	2.21E-5	Stephen Chow	1.33E-6	China	5.60E-4

表 7: A に Writer を適用した場合における出現プロパティ上位 20 件と出現率

順位	Writer		?x
	URI	出現率	出現率
1	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	1.0000	1.0000
2	http://dbpedia.org/ontology/wikiPageRevisionID	0.9991	0.9332
3	http://www.w3.org/2000/01/rdf-schema#label	0.9991	0.9797
4	http://dbpedia.org/ontology/abstract	0.9991	0.9325
5	http://www.w3.org/2000/01/rdf-schema#comment	0.9991	0.9328
6	http://xmlns.com/foaf/0.1/isPrimaryTopicOf	0.9991	0.9332
7	http://dbpedia.org/ontology/wikiPageWikiLink	0.9991	0.9332
8	http://www.w3.org/ns/prov#wasDerivedFrom	0.9991	0.9332
9	http://dbpedia.org/ontology/wikiPageID	0.9991	0.9332
10	http://dbpedia.org/property/wikiPageUsesTemplate	0.9987	0.9290
11	http://purl.org/dc/terms/subject	0.9978	0.9332
12	http://xmlns.com/foaf/0.1/name	0.9975	0.8578
13	http://dbpedia.org/property/name	0.9971	0.6827
14	http://dbpedia.org/resource/Template:Persondata	0.9861	0.2371
15	http://dbpedia.org/resource/Template:Infobox_writer	0.9574	0.0078
16	http://xmlns.com/foaf/0.1/surname	0.9288	0.2211
17	http://xmlns.com/foaf/0.1/givenName	0.9288	0.2270
18	http://dbpedia.org/property/dateOfBirth	0.8819	0.2220
19	http://dbpedia.org/property/occupation	0.8691	0.0708
20	http://dbpedia.org/property/birthDate	0.8311	0.1940

参考文献

[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)

[Ding 04] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V. C. and Sachs, J.: Swoogle: A search and metadata engine for the semantic web. Proc. 13th ACM Conference on Information and Knowledge Management, Washington D.C. (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)