3C1-OS-13a-5

DBpediaを利用した作家推薦の試み

An approach to author recommendation in DBpedia

一瀬　詩織　　大西　可奈子　　小林　一郎
Shiori Ichinose　　Kanako Onishi　　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
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

When someone finds books at a library or a bookstore, s/he often choices them written by her/his favorite authors. If s/he is recommended other writers who have close relationship with her/his favorite authors, s/he should be interested in them. The objective of our study is to propose a method to recommend books taking account of authors' relation. We use DBpedia, a semantic database constructed based on Wikipedia, as data resource for our recommendation, and propose a recommendation method for books, taking up two factors for recommendation: author's eminence and close relationship among authors.

1. はじめに

読みたい本を選ぶ時，好きな作家の別の著作など，既に興味を持っている作家を参考にして本を決定する場合がある．興味のある作家と関係のある別の作家を推薦する場合，この新しい作家に対しても，ユーザは興味を持つ可能性がある．本研究では，近年様々な広がりを見せている Linked Open Data の一部である DBpedia を用い，作家間の関係を利用した関連作家の推薦を行うことを目的とする．DBpedia より取得した作家に対し，それぞれ作家の著名度，関係の強さによる評価を用いた，2 つの推薦手法を提案する．また，実際に手法を用いて取得した推薦作家について考察を行う．

2. 関連研究

DBpedia を用いた推薦に関する研究としては，他に Passant[Passant 10] による音楽推薦システム dbrec がある．dbrec では DBpedia 上のバンドやソロアーティストのリソースに対し，直接リンクしている他のアーティストをリンク関係を用いたスコア LDSD によって評価することにより，推薦を行っている．しかしながら，DBpedia 上の作家のリソースはリソース間の直接リンクが少なく，間のメタデータも限定的なものであるため，この手法による評価を行うことは難しい．本研究ではより多くの推薦候補から作家推薦を行うため，他のリソースとの繋がりによって間接的にリンクした作家も対象とした推薦を行う手法を提案する．

3. DBpediaからの推薦候補抽出

3.1 DBpediaからの情報抽出

DBpedia は Wikipedia の持つ情報から構造化されたデータを抽出し，RDF により公開しており，現在[*1]360 万以上の物や事象に関する記述が存在する．これらのもの (以下リソースと呼ぶ) はそれぞれ異なった URI が与えられ，トリプルと呼ばれる主語，述語，目的語の 3 つの要素の組によって情報が記述されている．例えば，図 1 は “宮沢賢治は作家である” という情報を表すトリプルのグラフである．丸ノードと矢印はリソース，四角のノードは文字列のリテラルを表す．

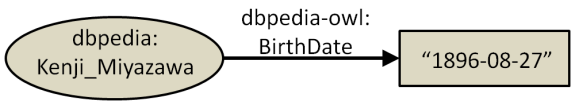

図 1: 宮沢賢治が作家であることを表すトリプル

着目しているリソースを主語か目的語に持つトリプルを DBpedia から抽出することで，対象リソースについて記述された情報を得ることができる．DBpedia のエンドポイント[*2]を通じ，SPARQL[*3]クエリによる問い合わせを行うことで，このような条件付きの情報が取得可能である．着目リソースを W としたとき，W とプロパティによって結び付いたリソースまたはリテラルの集合 $R_W = \{r_1, r_2, ..., r_n\}$ を抽出する SPARQL クエリは以下のようになる．

```
SELECT ?r WHERE {
{ <W> ?property ?r .}
UNION
{ ?r ?property <W> .} }
```

3.2 関連作家の抽出

ユーザが興味を持っている作家の DBpedia 上のリソースを着目作家 w とし，関連作家の定義を行う．DBpedia 上のリソースを a，リソース間のプロパティを $prop$，$prop'$ としたとき，関連作家 r は w と以下のいずれかの関係を持つ作家であるとする．

$$w \leftarrow (prop) \rightarrow r$$

$$w \leftarrow (prop) \rightarrow a \leftarrow (prop') \rightarrow r (w \neq a)$$

ここでプロパティの向きが両方向になっているのは，関係の方向を考慮しないことを表す．つまり，$w - (prop) \rightarrow r$，$w \leftarrow (prop) - r$ のいずれのプロパティも w と r の関係であるとみ

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　2012 年 4 月現在

∗2　http://dbpedia.org/sparql

∗3　http://www.w3.org/TR/rdf-sparql-query

なす.
また，“作家である” ことは次の関係を持つことであると定義する.

$$r - (rdf : type)^{*4} \rightarrow dbpedia - owl : Writer^{*5}$$

これらの定義を SPARQL クエリで表したものを用い，DBpedia から関連作家の抽出を行う．例として，作家 w と直接のリンク関係にある作家 x を取得するために用いた SPARQL クエリを例として以下に示す.

```
SELECT * WHERE {
{ < 作家リソース w> ?property ?x .}
UNION
{ ?x ?property < 作家リソース w> .}
?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#
type> < http://dbpedia.org/ontology/Writer> .}
```

3.3 推薦候補の決定

着目作家とプロパティによって直接結び付いているリソースの数は作家によって様々であるが，一般に数 10〜数 100 程度である．これに対し，間に別のリソースを介して間接的にリンクしているリソースの数は 1 万以上の膨大な数になる．このため本推薦においては，リソース間のプロパティが特定の種類である作家のみを推薦候補とし，関連作家の絞り込みを行った．作家リソースが持つ情報の推薦への重要性について，被験者 20 代女性 18 名にアンケートを行った結果作家の影響関係が最も推薦に有効であると判断したため，推薦候補の決定には DBpedia で用いられている “http://dbpedia.org/ontology/influenced” , “http://dbpedia.org/ontology/influencedBy” の 2 つのオントロジーを用いることとした．着目作家から推薦候補の作家まで，関係のリンクを辿る回数を今後 2 者間の ‘距離’ と呼ぶ．DBpedia において作家 “Haruki Murakami ” を表すリソース*6 と，これらのプロパティにより結び付いている作家との関係を図 2 に示す.

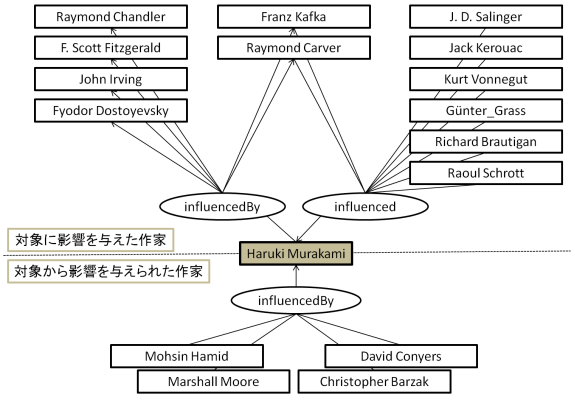

図 2: “Haruki Murakami” から距離が 1 の関連作家

4. 作家の著名度を考慮した推薦

着目作家 w と関係を持つ推薦候補のうち，著名度の高い作家を推薦する手法を提案する．評価には大西ら [Onishi 12] によって定義されている，リソースの重要性を測る指標 Hub Score を用いた．これはハイパーリンク解析に用いる Hits アルゴリズムや PageRank アルゴリズムを RDF データに適用した指標であり，他のリソースとの関係において，着目するリソースの重要性を測ることのできる指標である．着目作家 w と直接リンクした要素の集合を $R_w = \{r_1, r_2, ..., r_\alpha\}$ とし，R_w の要素 r_α とリンクした要素の集合を $\Omega = \{s_1, s_2, ..., s_\omega\}$，その内，$r_\alpha$ 以外とのリンクを持つ要素の集合を $\Psi = \{t_1, t_2, ..., t_\psi\}(\Psi \subseteq \Omega)$ とする．このとき，r_α の Authority Score，Resource Score はそれぞれ $AuthorityScore(r_\alpha) = \omega$, $ResourceScore(r_\alpha) = \psi$ と定義される．また，Ψ の各要素の Authority Score の中央値を M，標準偏差を SD とするとき，Authority Score が $M \pm 1SD$ の範囲内である要素の集合を $\Phi = \{u_1, u_2, ..., u_\phi\}(\phi \leq \psi, \Phi \subseteq \Psi)$ とすると，w の Hub Score は以下のように定義される.

$$HubScore(w) = \sum_{r \in \Phi} \frac{AuthorityScore(r_\alpha)}{ResourceScore(r_\alpha)}$$

Φ の設定は Ψ の要素の内で，Authority Score が極端に大きいリソースを除くために行っている.
Authority Score は対象リソース自身の情報の豊富さを示す指標であり，スコアが高いほど対象リソースは情報が豊富であることを示す．また，Resource Score は対象リソースと他のリソースとどれくらい関わりを持っているかという指標であり，スコアが高いほど対象リソースは多くのリソースとの関わりを持つことを示す．Hub Score が高くなるのは，対象リソースにリンクしている他のリソースが多くあり，情報を豊富に持っているにもかかわらず対象リソース以外との関係が少ない時である．このような場合，対象リソースのハブとしての意義は大きくなり，他のリソースに対する重要度が高まると考えられる.

4.1 実験 1

作家 “Haruki Murakami” から距離が 1，距離が 2 の作家について Hub Score を求め，上位 10 名を推薦対象とした．結果を表 1 および表 2 に示す*7．また，これらの推薦作家と着目した作家 “Haruki Murakami” との影響関係を図 3 に示す.

表 1: 距離が 1 における関連作家の Hub Score(評価対象 16 件)

順位	作家	Hub Score	Authority Score	Resource Score
1	Fyodor Dostoyevsky	571.33	310.0	212.0
2	Franz Kafka	423.05	286.0	182.0
3	Kurt Vonnegut	360.74	256.0	153.0
4	Jack Kerouac	345.44	215.0	137.0
5	Raymond Chandler	248.46	202.0	111.0
6	F. Scott Fitzgerald	201.52	178.0	87.0
7	J. D. Salinger	198.58	188.0	88.0
8	Raymond Carver	181.29	171.0	80.0
9	Günter Grass	176.04	154.0	67.0
10	Richard Brautigan	151.40	133.0	64.0

4.2 考察

表 1，2 より，Fyodor Dostoyevsky，Stephen King 等，一般的に良く知られた作家の Hub Score が高い傾向にあることが分かる．したがって対象作家のリソースと関係した著名な人物を推薦する場合，Hub Score は有効であると考えられる．また図 3 より，距離が 2 の著名な作家を推薦する場合，着目対象との関係の在り方による異なった推薦の可能性が存在することが分かった.

Stephen King は表 2 において最も評価の高い作家であり，着目作家とは互いに「Marshall Moore に影響を与えた」とい

*4 http://www.w3.org/1999/02/22-rdf-syntax-ns#type
*5 http://dbpedia.org/ontology/Writer
*6 http://dbpedia.org/resource/Haruki_Murakami
*7 実験で得られた結果は全て 2012 年 2 月におけるものである.

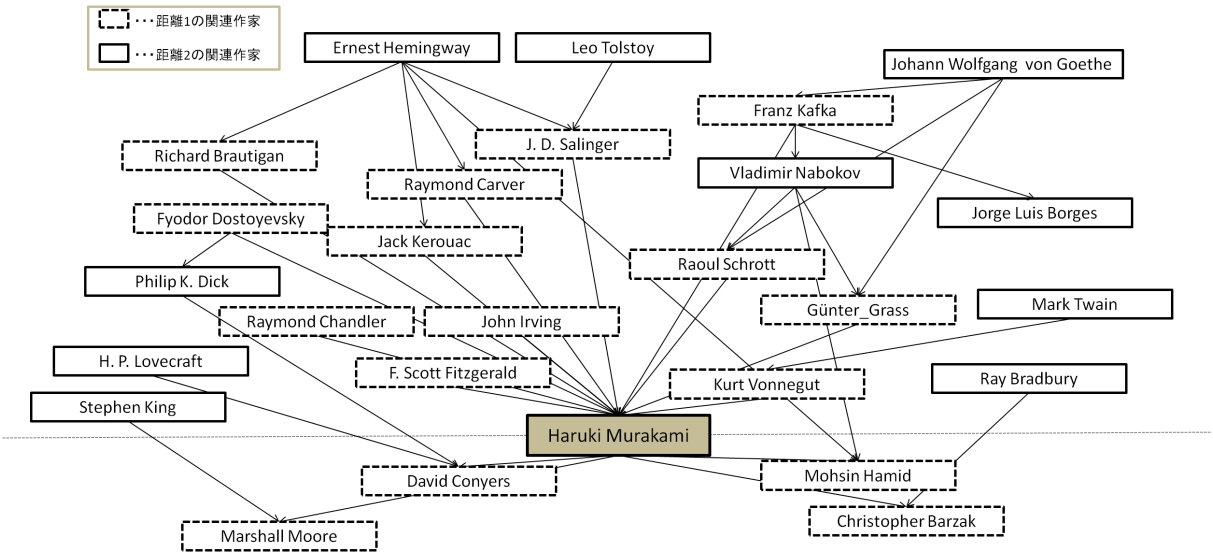

図 3: 距離が 2 の作家との影響関係

表 2: 距離が 2 における関連作家の Hub Score(評価対象 374 件)

順位	作家	Hub Score	Authority Score	Resource Score
1	Stephen King	1079.81	506.0	405.0
2	H. P. Lovecraft	657.18	362.0	260.0
3	Ernest Hemingway	613.64	333.0	224.0
4	Jorge Luis Borges	521.75	307.0	208.0
5	Leo Torstoy	513.08	275.0	186.0
6	Philip K. Dick	490.71	269.0	166.0
7	Ray Bradbury	427.66	257.0	171.0
8	Johann Wolfgang von Goethe	419.26	269.0	165.0
9	Mark Twain	411.53	294.0	162.0
10	Vladimir Nabokov	411.10	290.0	171.0

う関係を持つ．このとき Haruki Murakami と Stephen King との間に直接の影響関係はないが，互いに同じ人物に影響を与えたという観点で，両者を比較した推薦が行えると考えられる．一方，Ernest Hemingway は図 3 において 5 人の作家に影響を与えており，そのうち 4 人は「Haruki Murakami に影響を与えた」という関係を持つ．Haruki Murakami は Hemingmay から間接的に多大な影響を受けたことが推測され，着目作家の作風の源流となった人物の推薦が可能だと考えられる．

King と Hemingway を比較した場合，Hub スコアが高いのは King であるが，作家間の関係の数より，着目作家とより関係が強いのは Hemingway であると考えられる．このような着目作家と関係の強い作家を推薦するために，作家間の関係を求める実験を行った．

5. 作家間の関係数に基づいた推薦

着目作家 w の関連作家の集合を $W = \{r_1, r_2, ..., r_\beta\}$ とする．このとき，作家 w と関連作家 r_β との間には，プロパティと他のリソースによって構成された直接または間接のリンクが存在する．このようなリンクの数が多く，また種類が豊富であるほど，両者の間には強い関係があると考えられる．
作家間の関係数に基づいた推薦では，作家 [w] と関連作家 r_β の間のリンク数を求めることにより推薦を行う．リンクの構成要素を以下のように定義し，異なった要素で構成されている場合をそれぞれ別のリンクであるとする．

- 作家間のプロパティ（$\leftarrow (prop)-, -(prop) \rightarrow$）
- w と w_β との間に存在するリソース

例として，[w] $-(prop) \rightarrow r_\beta$ ，[w] $\leftarrow (prop) - r_\beta$ という直接リンクが存在したとき，二つは別々の関係である．また，二者間に存在するリソース p，$q(p \neq q)$ に対し，$w \rightarrow (prop)-p \rightarrow (prop')-w_\beta$ と $w \rightarrow (prop)-q \rightarrow (prop)-w_\beta$ は別の関係である．

作家 w と集合 W の要素 r_β との間の関係数を N_β とし，この関係数が多いほど w との関係が強いものと考え，推薦を行う．

5.1 実験 2

着目作家 ‘‘Haruki Murakami’’ と，距離が 2 以内の影響関係にある作家の集合を推薦候補とした．このとき推薦候補となる作家は 390 人で，着目作家との関係数は最大 6 であった．関係数と候補となる作家の人数について纏めたものを表 3 に示す．また，関係数 4 以上の作家を表 4 に示す．

表 3: 関連作家と関係数 (評価対象 390 件)

関係数	作家数
6	2
5	4
4	5
3	17
2	77
1	285

表 4: 関係数による推薦の結果 (関係数 4 以上)

関係数	作家
6	Cagdas Cetinkaya
6	Paul Auster
5	Albert Camus
5	Franz Kafka
5	Philip Roth
5	Ernest Hemingway
4	J. D. Salinger
4	John Irving
4	Vladimir Nabokov
4	Richard Yates
4	Fridrich Nietzche

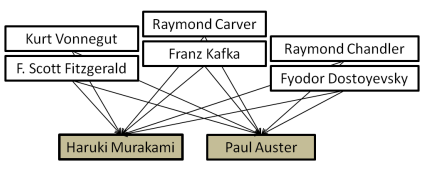

図 4: Haruki Murakami，Paul Auster 間の影響関係

5.2 考察

図 4 より，Haruki Murakami と Paul Auster とは 6 人の共通の人物から影響を受けており，両者には強い関係があると考えられる．実際に両者は年代も近く，作風も似ていると言われており，この推薦は妥当であると考えられる．
一方で，同じく 6 人の共通人物から影響を受けている Cagdas Cetinkaya は DBpedia 上に影響関係のデータが存在するにも関わらず，著作が存在しないという人物である．この推薦では 2 者間のリンク数にのみ着目していたためこのような人物が上位に来たと考えられるが，推薦結果としては明らかに不適当である．よってこのような作家が上位とならないような，評価値の改善が必要であると考えられる．著作が少ない作家はリソースの持つ情報量が少ない傾向にあると推測できるため，4 章の Hub Score をこの手法に組み合わせることによって今後評価の改善が行えるのではないかと考えている．

6. おわりに

本研究では，DBpedia を用いて作家推薦を行ための 2 つの手法を提案した．4 章では，作家間のリンク構造に基づいた作家の重要度の評価を行うことにより，ある作家と関係を持つ作家のうち著名度の高い作家の推薦が行えることを実験 1 によって示した．また 5 章では，着目作家と関連した作家と着目作家との間のリンク数に基づき，両者の関係の強さを求める手法を提案し，実験 2 で実際の推薦を行った．今後，提案手法により抽出された結果が有意な物であるか被験者実験による評価を行うことで，本手法の有効性を示したい．また両者の手法を組み合わせることにより，それぞれの手法を改善したより良い推薦手法を提案したいと考えている．

参考文献

[Passant 10] Alexandre Passant, ‘‘rec -- Music Recommendations Using DBpedia’’, The 9th International Semantic Web Conference, ISWC’10, pp209-224, 2010.

[Onishi 12] Kanako Onishi and Ichiro Kobayashi, ‘‘Information Enhancement on a Focused Object using Linked Data’’, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.16, No.1. pp4-12, 2012.