

リソース間の有意な関係を発見する Linked Data解析手法への取り組み

大西可奈子

小林一郎

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

1 はじめに

RDF データにおいて、リソース間にはプロパティを用いてたくさんの関係が記述されている。リソース間の類似度を測る手法は多く研究されているものの、リソース間にある有意な関係を抽出することについてはまだ多くの研究の余地がある。Philipp ら [1] は、2つの物事の間にある関係を視覚的に表示するシステムを構築した。このシステムでは多くの関係が見つかる一方で、多く見つけられ過ぎるが故に各関係が持つ意味はユーザには理解し難い。そこで本研究では、RDF で記述されたリソース間関係から、類似度では計れない有意な関係を抽出することを目的とする。

2 リソース間関係の取得

RDF 形式で記述されている物事はリソースと呼ばれ、属性（以下、プロパティ）を介し、他のリソース、文字列、数値等と結びつくことによって、リソースに関する様々な情報が記述されている。例えば、“東京は日本の首都である”という知識の記述は、“東京”というリソースが“日本”というリソースに、“首都である”というプロパティを介して結びつくことにより表現される。“首都である”というプロパティは例えば DBpedia[2] では、*dbpprop:capital* と表される。このように、プロパティは方向を持ち、リソース間の関係を表現するのに重要な要因となる。

Philipp ら [1] は、リソース間に存在する関係のうち、リソースとリソースの間に複数のリソースを含む関係も取得している。一方で、そのような複雑な関係はユーザが理解し難く、取得する際の計算量も増えてしまうことから、プロパティの方向の変更は一度のみ認めるものとしている。本研究においても、複雑な関係はユーザが理解し難く大きな計算コストが必要となることから、2つの異なるリソースを $r1$, $r2$, 任意のプロパティを $prop$, $prop'$, リソースまたは任意の値を a , プロパティの方向を $\leftarrow$, $\rightarrow$ とする時、リソース間の関係を以下の6つのパターンで取得した。

$$\begin{aligned} &r1 \leftarrow prop \rightarrow r2 \\ &r1 \leftarrow prop \leftarrow r2 \\ &r1 \leftarrow prop \rightarrow a \leftarrow prop' \rightarrow r2 \\ &r1 \leftarrow prop \leftarrow a \leftarrow prop' \leftarrow r2 \\ &r1 \leftarrow prop \rightarrow a \leftarrow prop' \leftarrow r2 \\ &r1 \leftarrow prop \leftarrow a \leftarrow prop' \rightarrow r2 \\ &(r1 \neq r2) \wedge (r1 \neq a) \wedge (r2 \neq a) \end{aligned}$$

An Approach to Linked Data Analysis for Discovering Meaningful Relation between Resources

Kanako ONISHI(onishi.kanako@is.ocha.ac.jp),

Ichiro KOBAYASHI(koba@is.ocha.ac.jp)

Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

この時、リソース $r1$, $r2$, および任意の値 a はすべて異なるものとする。 $prop$ と $prop'$ は同一のプロパティでも可とする。なお、関係を形成するプロパティとリソースまたは値が順番も含めて同一の場合は、それらの関係は同じものと見なしひとつに纏める。

3 有意な関係の定義

他のリソースとの関連性を考える時、特異と思われる関係を“有意な関係”と定義し、2つの異なるリソース $r1$ と $r2$ の間に存在する関係の性質を評価する指標を定義する。

1. リソース間距離

$r1 \leftarrow prop \rightarrow r2$ や $r1 \leftarrow prop \leftarrow r2$ のようにリソース間に他のリソースを含まないものを、距離が1の関係と定義する。すなわち、 $r1 \leftarrow prop \rightarrow a \leftarrow prop' \rightarrow r2$ や $r1 \leftarrow prop \leftarrow a \leftarrow prop' \leftarrow r2$ のようにリソース間に他のものを1つ含むものは、距離が2の関係となる。リソース $r1$ と $r2$ がより近い距離に多くの関係が存在する時、リソース $r1$ と $r2$ は親密な関係であると仮定する。

2. プロパティの方向

$r1 \leftarrow prop \rightarrow r2$ と $r1 \leftarrow prop \leftarrow r2$ や、 $r1 \leftarrow prop \rightarrow a \leftarrow prop' \rightarrow r2$ と $r1 \leftarrow prop \leftarrow a \leftarrow prop' \leftarrow r2$ は、プロパティの向きが一定である。すなわち、これらの間に有意性の差はないものと考えられる。

一方で、 $r1 \leftarrow prop \rightarrow a \leftarrow prop' \leftarrow r2$ と $r1 \leftarrow prop \leftarrow a \leftarrow prop' \rightarrow r2$ は、リソース間のプロパティの向きが一定ではない。このような関係で表される $r1$ と $r2$ の関係は、プロパティの向きが一定のものに比べて、距離が同じでも有意性は低いものと考えられる。

3. リソース間に含むリソースの Hub Score

HITS アルゴリズム等を基に Onishi ら [3] によって提案された、RDF データにおけるリソースの性質を評価する3つのスコアを用いる。本研究では、提案された3つのスコアの内、Hub Score を利用する。Hub Score は対象とするリソースが関わる他のリソース群について、どの程度、情報が記述されているかを示す指標であり、他の2つのスコア (Authority Score, Resource Score) を利用して定義される。ここで、Authority Score とは、対象とするリソースについて、どの程度、情報が記述されているかを示す指標で、リソースがプロパティを介して結びつくリソースまたはリテラル等の総数で表される。また、Resource Score とは、注目リソースがどの程度、他のリソースと関わっているかを示す指標で、リソースがプロパティを介して結びつくリソースの数で表される。上記のことから、Hub Score は式 (1) のように表される。

$$HubScore = \sum \frac{AuthorityScore}{ResourceScore} \quad (1)$$

表 1: 各距離に存在する関係の数

関係 \ 距離	1	2 (一定)	2
O と C	0	2(1)	9(9)
O と Y	0	1(1)	12(11)
O と H	0	0	7(7)
C と Y	0	0	5(5)
C と H	0	6(2)	5(5)
Y と H	0	0	15(13)

Hub Score が大きいリソースは小さいリソースよりも、周知のリソースである傾向にある [3]. すなわち、距離が 2 の関係において Hub Score が小さいリソースを挟む関係の方が、注目している 2 つのリソースにとって、より r_1 と r_2 に特化した固有の関係であると考えられる。

4 実験

本手法を利用し、イギリスと日本の作家 4 名 (Oscar Wilde, Charles Dickens, Yukio Mishima, Haruki Murakami) の間にある複数の関係から有意な関係を抽出する。関係を抽出するデータには DBpedia[2] を利用し、問い合わせは SPARQL* を利用して、エンドポイント† から行った。4 人の間にある関係の数を表 1 に示す。括弧内の値は、リソース間に含むリソースの種類の数を示し、O は Oscar Wilde, C は Charles Dickens, Y は Yukio Mishima, H は Haruki Murakami の略称とする。また、距離の 2 (一定) とはプロパティの方向が一定である距離が 2 の関係を示し、括弧書きがない距離の 2 はプロパティの方向が一定ではない距離が 2 の関係を示す。

4.1 リソース間の有意な関係抽出

Charles Dickens と Haruki Murakami の間にある関係で、有意と思われる関係の抽出を行う。2 つのリソースが持つ 11 個の関係のうち、間にリソースを含む関係は 7 つであった (表 1 参照)。それらの内訳は表 1 に示す通りである。Charles Dickens と Haruki Murakami を結びつける関係の間に含むリソースは Franz Kafka と John Irving の 2 つで、以下のような関係となっている。

Charles Dickens $\leftarrow$ *influencedBy* — **Franz Kafka** $\leftarrow$ *influences* — Haruki Murakami
 Charles Dickens $\leftarrow$ *influencedBy* — **John Irving** $\leftarrow$ *influencedBy* — Haruki Murakami

ここで、Franz Kafka の Hub Score は 423.05, John Irving の Hub Score は 132.18 である。Hub Score はそのリソースの周知度を表すことが示されているので、Hub Score がより小さい John Irving を間に含む関係の方が、2 つのリソースにとって固有のものと考えられる。反対に、Franz Kafka を含む関係は、より周知の関係であると考えられる。実際に Franz Kafka は Charles Dickens や Haruki Murakami に限らず多くの人間に影響を与えており、 $r_1 \rightleftharpoons$ Franz Kafka $\rightleftharpoons$ r_2 という関係が成り立つリソース r_1 , r_2 は、John Irving を介するものよりも多く存在している。

*<http://www.w3.org/TR/rdf-sparql-query/>

†<http://dbpedia.org/sparql>

4.2 2つのリソース間関係の比較

リソース間に存在する関係の距離及び関係の数が近い、Oscar Wilde と Charles Dickens (O と C), 及び、Oscar Wilde と Yukio Mishima (O と Y) に着目する。彼らの間にある関係を比較し、Oscar Wilde とより特異な関係にある人物はどちらなのか抽出する。表 1 に示した通り、プロパティの方向が一定の距離が 2 の関係において、それぞれの間には 1 つのリソースが含まれ、以下のような関係で結びついている。

Oscar Wilde — *influencedBy* → **Victor Hugo** — *influenced* → Charles Dickens

Oscar Wilde — *influences* → **Victor Hugo** — *influenced* → Charles Dickens

Oscar Wilde — *influenced* → **André Gide** — *influenced* → Yukio Mishima

ここで、Oscar Wilde と Charles Dickens を結びつける存在である Victor Hugo の Hub Score は 301.21, Oscar Wilde と Yukio Mishima を結びつける存在である André Gide の Hub Score は Hub Score 138.63 である。従って 4.1 節で述べた通り、Oscar Wilde は Yukio Mishima と、より特異な関係を持つことがわかる。実際に、Oscar Wilde の作品や生き様は、Yukio Mishima のそれに通じるものがあると言われている。一方、Oscar Wilde と Charles Dickens は同じイギリス人という共通点はあるものの、彼らのイギリスでの捉われ方には大きな差がある‡。

5 おわりに

本研究では、RDF で記述されたリソース間関係から、類似度では計れない有意な関係を抽出するため、リソース間の距離、プロパティの向き、RDF データにおけるリソースの Hub Score[3] 等を利用し、より有意な関係であると考えられる条件を定義し、関係抽出を行った。

今後は、有意な関係の抽出手法をさらに熟慮すると共に、本手法によって抽出された関係が有意なものかどうか被験者実験を行い、その結果に対して適切な検定を行い、その有効性を示す。

参考文献

- [1] H. Philipp, H. Sebastian, L. Jens, L. Steffen and S. Timo. RelFinder: Revealing Relationships in RDF Knowledge Bases, Proceedings of the 4th International Conference on Semantic and Digital Media Technologies: Semantic Multimedia, SAMT '09, Berlin, Heidelberg, pages 182–187. Springer-Verlag, 2009.
- [2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives. DBpedia: a nucleus for a web of open data, Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, ISWC'07/ASWC'07, pages 722–735. Springer-Verlag, 2007.
- [3] K. ONISHI and I. KOBAYASHI: Information Enhancement on a Focused Object using Linked Data, Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.16 No.1 2012, to appear.

‡ 著者の知人の文学に造詣の深い人の話に基づく