The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

知識工学者のための日本語Wikipediaのカテゴリ階層構造の再整理

Reorganizing Hierarchical Category Structure of Japanese Wikipedia for Knowledge Engineers

中川嵩教 *1　　吉岡真治 *1*2
Takanori Nakagawa　　Masaharu Yoshioka

*1北海道大学　　*2理研 AIP
Hokkaido University　　RIKEN AIP

Since Wikipedia is an on-line encyclopedia that covers varieties topics, there are several attempts to extracting knowledge from its contents. Wikipedia category is one type of information that provides topical index for the articles and organized as hierarchical manner. However, since hierarchical structure of Wikipedia category have different characteristics compared to the one for knowledge representation, it may not fully and correctly utilized at this moment. In this paper, we propose a method to extract hierarchical structure that have similar characteristics with knowledge representation.

1. はじめに

Wikipedia*1 は世界最大のインターネット百科事典であり、様々な形で知識源として活用する方法が検討されている。具体的には、各記事に関する属性情報（所在地、生年月日、所属…）などを記述したインフォボックスから属性情報を抽出する DBpedia[Bizer 09] の研究や、カテゴリの情報を用いて、様々な単語間の類似性を計算する方法 [Ponzetto 07, Taieb 13] や、オントロジーの構築に役立てる YAGO2[Hoffart 13] や、日本語 Wikipedia オントロジー [玉川 10] の研究などが存在する。また、我々は、これまでに、このカテゴリの性質に関する分析 [藤原 12, Yoshioka 14] を行っており、これらの研究の中で、Wikipedia のカテゴリの持つ特殊性を指摘し、その特殊性を考慮した利用方法が必要であることを提案してきた。

本研究では、これまでの研究踏まえて、Wikipedia のカテゴリ階層を概念階層として扱う場合の問題点を、具体例を示しながら指摘するとともに、その問題点を解消するための、知識工学的観点からの Wikipedia カテゴリの再整理を行う方法を提案する。

2. Wikipedia カテゴリ

2.1 Wikipedia カテゴリの階層構造

Wikipedia において、カテゴリとは、膨大な記事群を様々な観点から分類するための索引であり、各記事には、それぞれに一つ以上のカテゴリが付与される。また、このカテゴリは、さらに詳細なカテゴリと関連付けることにより、カテゴリは階層的な構造となっている。このカテゴリ階層については、基本的には、下位カテゴリに属する記事は、上位カテゴリにも含まれるという包含関係が成立することが期待されることになっており、カテゴリ階層は、知識工学で用いられる概念階層と似た性質を持つことが期待されている。しかし、このカテゴリ階層の構造は、Wikipedia に登録される記事の増加に伴い、必ずしも、包含関係が成り立たない形でカテゴリ階層が作られるようになっている。次小節では、このような包含関係が成り立たないカテゴリが生成される原因である「分割のためのカテゴリ」について説明する。

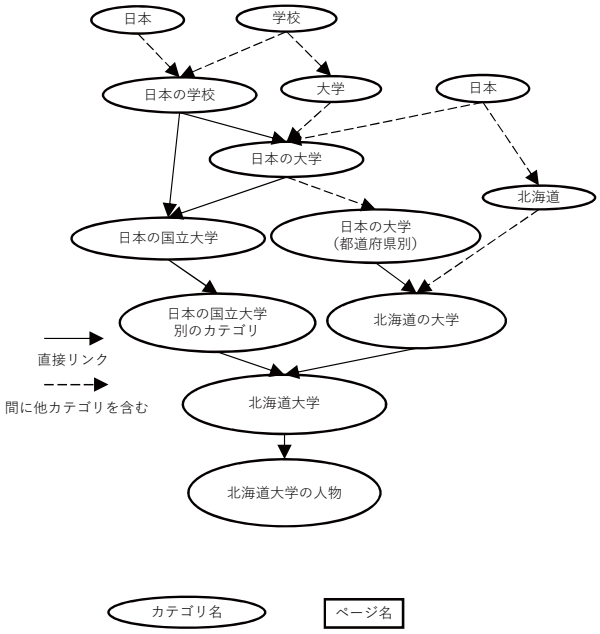

図 1: 分割のためのカテゴリを含むカテゴリ階層

2.2 分割のためのカテゴリ

Wikipedia カテゴリには、それぞれに所属するページ数に制限は無いが、一つのカテゴリに対し、大量のページが所属する場合、閲覧者の視点からすると、不都合であるため、そのカテゴリを分割するようなカテゴリが生成される。図 1 は「北海道大学」を例にとったものであるが、カテゴリ「北海道大学」が、カテゴリ「大学」の下に直接繋がっていたとすると、その他世界中の大学も同じように繋がることとなり、カテゴリ「大学」の下には数万もの大学名が所属し、閲覧者からすると意味のないものとなってしまう。そこで、カテゴリ「大学」を「日本の大学」、「日本の国公立大学」と分割していけば、閲覧者からすると、日本の国公立大学の一覧を見ることができるなどの利点を持つことができる。以上のような役割を持つカテゴリを

連絡先: 中川嵩教，北海道大学工学部情報エレクトロニクス学科，札幌市北区北 14 条西 9 丁目,011-706-7161，f-b-hawk07@eis.hokudai.ac.jp

*1 https://www.wikipedia.org/

「分割のためのカテゴリ」と呼ぶ。一方で、英語版 Wikipedia では、このようなカテゴリは「set and topic カテゴリ」（以下 SaT と略記）と呼ばれ、「分割のためのカテゴリ」でないカテゴリは、「set カテゴリ」、または、「topic カテゴリ」のどちらかに分けられる。先の例を見ると、「日本の大学」が SaT であり、「日本」が topic、「大学」が set である。ほとんどの場合、SaT カテゴリは、set と topic を共に、親（もしくは先祖）カテゴリに持つ。

3. Wikipedia カテゴリの問題点と再構成

前述の、set または topic と SaT の関係を見た時、親カテゴリの set または topic と、その子カテゴリの SaT の関係は、分割が行われているだけのため、包含関係が成り立つと考えて良い。しかし、SaT から分割でないカテゴリに到達した際に、包含関係が途絶えてしまうことがある。例を見ると、「大学」から分割が行われ、「北海道大学」までは大学についての階層であるが、その「北海道大学」の下には、「北海道大学の人物」という人物の階層が続いていく。このような問題が Wikipedia カテゴリの階層構造に多く含まれ、結果として、階層構造全体を知識工学的観点から活用する際の大きな妨げとなっている。つまり、問題の起こる原因は以下の仮説で説明することができると考えた。

仮説　上位カテゴリが topic カテゴリの場合（SaT の topic 部分など）、クラス制約を含め、包含関係が成り立たない場合が存在する。

この問題は、topic に用いられる概念については、SaT カテゴリの分割時に topic を基準に分割される場合を除いて、topic に関連する階層構造は、上位階層を参照しないということで、上記のような問題の大半が解決すると考えられる。ただし、topic 間の関係でも、地理的な包含関係などが成り立つ場合がある。この仮説の検証をするためには、以下の作業が必要となる。

1. カテゴリを set や topic といった、構成要素となるカテゴリ（以下 Single と略記）と SaT のような複合的なカテゴリ（以下 Combination と略記）に分割すること。
2. 構成要素となるカテゴリを set と topic に分割すること。
3. topic 間の階層関係がある場合には、それを分類すること。（地理的包含関係、メンバー、関連など）

4. 分析作業

今回は、前章記載の作業の 1,2 を行い、そのデータを用いて、Wikipedia のカテゴリ構造の中から、部分構造を取得し、その部分構造について考察する。作業 3 については今後の課題とする。

4.1 扱ったデータと作業方針

作業は、Wikipedia の 2017 年 10 月 20 日のダンプデータをもとに行った。このデータ中には、207,036 件のカテゴリが存在する。この実験では、スタブなどの Wikipedia のメンテナンスをするために設定されているカテゴリを除く、197,378 件から、更に、上下カテゴリに関する情報が存在しない 14,644 件を除いた、182,734 件のカテゴリを分析の対象とした。

まず、このカテゴリから分割のためのカテゴリを認識するために、以下の基準にあたるものに Combination、それ以外のものを Single として分類を行った。

- 動詞句を含むカテゴリ
- 付加情報が付くカテゴリ、また、それを上位カテゴリとして持つカテゴリ
- 「SaT」→「SaT」というペアの上下のカテゴリ
- 「set(または topic)」→「SaT」というペアの「SaT」のカテゴリ

そして、Single に分けられたものの中で、Combination → Single となっている Single に topic のラベルを付け、それ以外のものに set のラベルを付けた。そのラベルが付けられたデータを用いて、Single から始まり、Combination を辿り、topic に到達するまでの階層を部分階層として取得し、その部分階層の中に、包含関係が成り立たない関係が存在するかどうかを確認した。

4.2 分析結果

各カテゴリに対するラベル付けに関しては、57,976 件に Single、122,945 件に Combination のラベルが付与された。更に、Single の内 39,956 件に topic、18,020 件に set のラベルが付与された。

分類されたものの中から、500 個のカテゴリをランダム抽出し、誤っているものがあるかを確認したところ、Combination のラベルが付いたものの中には、Single とした方が良いものは 1 件（例：アラブの春）存在するのみであったが、Single のラベルが付いたものの中に、25 件（例：国籍別の在外人）ほど Combination とした方が良いものが存在した。前者の問題は、主に、固有名詞に関する問題なので、そのリストを並行して作成していくことにより、解消可能だと考える。後者については、Combination と判断するためのパターンが不足していることが原因である。先の誤りでは、「A」→「B の C」の「B の C」が Combination として抽出できていない。このような形は子の構成要素が、親を言い換えた形になることによって起こる。先の例では、「外国人」→「国籍別の在外人」となっている。これは「外国人≒外人」となるようなリストがあれば解決すると思われる。このように、見つかった例外事例から、さらなるパターンを作成することで、その精度の向上を目指す予定である。

set と topic に関する分類については、パターンにより topic と分類できない場合は、set とするという基準で分類を行っているために、topic として分類されたものにはほとんど誤りがないが、topic の分類に関するパターンが不足していることもあり、set の中に topic と分類したほうが良いもの（例：夏目漱石）が含まれている。こちらについても、獲得した例外事例を分析し、パターンを拡充することで、その精度の向上を目指す予定である。

ただ、両者ともに、一定レベルの精度が高いパターンが作成された場合に、それらのパターンによる分類結果に当てはまらないものをさらに分析していくという形で進めていくことによって、少しずつ分類が難しいカテゴリの特徴が明らかになっていくことを期待している。

次に、ラベル付けされたデータを用いて取得した部分階層の考察を、具体例を用いて述べていく。取得した部分階層の良い点としては、例えば、「学校」というカテゴリから、今回設定した基準で子カテゴリを辿っていくと、「学校」→「各国の学校」→「日本の学校」→「日本の大学」→「日本の国立大学」→「北海道大学」という階層が得られる。本来はこの下に「北海道大学の人物」という「人物」に関する階層が続くが、

「北海道大学」という topic カテゴリで階層が切れることにより、「学校」に関する階層のみを取得することができた。しかし、「学校」→「学校を題材とした作品」という set が追加され分割された時には、その下に付くカテゴリは、「作品」に関する階層が続いてしまっていたり、「学校」→「各国の学校」→「アメリカ合衆国の学校」→…→「カリフォルニア州の大学」→「サンフランシスコ美術大学の教員」のように Combination → Combination の際に、set の部分（あるいは set と topic 両方）の影響により、「人物」を表すカテゴリが入ってしまうことがある。前者に対しては、今回は、上（親）から下（子）へと辿っていったが、同じように、下から上に辿っていったときの階層と比較をすれば解消すると思われ、後者に対しては、Combination の構成要素に対し、それぞれが、set か topic かを断定し、set が変更された際の、許容範囲を設定すればある程度解消されるように思える。

上記を考慮して、もう 1 つの具体例として、「職業」というカテゴリから辿っていった結果を見ると、「職業」→「分野別の職業」→「メディア関連の職業」→「放送関連の職業」→「ナレーター」では（「ナレーター」が set か topic であるかの議論は今後していく必要があるが）、その下にある、「ナレータープロダクション」という企業を表すカテゴリへはいかず、職業の範囲で階層が途切れている。また、「職業」→「専門家」→「武道家」→「槍術家」という例を見ると、set 間であれば、包含関係を保ちながら、topic へと辿りつくことができるということがわかる。

4.3 想定する活用事例

まず、各カテゴリを SaT、set、topic に分類するだけでも、それぞれに注目をしたページ間の類似度を計ることができる。例えば、「北海道大学」と「札幌駅」を見ると、札幌駅は北海道大学の最寄駅となり、位置的にも、関係的にも非常に近い関係にあるが、Wikipedia カテゴリの類似度では、「北海道」→「北海道の教育」→「北海道の学校」→「北海道の大学」→「北海道大学」→「北海道大学（ページ）」と、「北海道」→「北海道の市町村」→「北海道の市」→「札幌市」→「札幌市の交通」→「札幌市北区の交通」→「札幌市北区の鉄道駅」→「札幌駅（ページ）」とで、非常に遠い。対して、「九州大学」は、位置的にはかなり遠い位置にあるが、Wikipedia カテゴリの類似度では、「日本の国立大学別のカテゴリ」→「北海道大学」→「北海道大学（ページ）」と、「日本の国立大学別のカテゴリ」→「九州大学」→「九州大学（ページ）」とで、かなり近いものとなる。どちらを近いものとしたいかは、用途によるが、「北海道」や「札幌市」といった topic を優先するか、「大学」のような set を優先するかということが、各々の類似度に関する重み付けなどによって表現可能となることが期待される。

更に、例えば「職業」というカテゴリから繋がっていくものを見ると、再整理された Wikipedia 構造では、最終的には職業名にたどり着くことが期待されるため、その一覧を辞書のような形で取得できる。その辞書は、先の例の「槍術家」のようなあまり一般的ではない職業を含むことができ、用途次第では、便利なものとなる。

4.4 データ公開に向けて

Single と Combination の分類については、もう少し基準をしっかりと付け、例外に関しては、そこまで多くの数があるわけではないので、最悪手作業で処理していけばよいと考えている。set と topic の分類については、まだいくつかの分類基準を増やしていかなければならないと思われる。現在考えているのは、形態素解析を用いた一般名詞（set）と固有名詞（topic）の同定や、言語間リンクを用いた英語版表記の確認等である。上記二つができれば、Combination の構成要素それぞれが、set か topic かを断定することができ、更に、その構成要素のつなぎとして使われる語（「の」「を題材とした」等）についても取得することができる。また、カテゴリ間の関係の同定も行っていく必要がある。

5. まとめ

本研究では、Wikipedia カテゴリをカテゴリ構造全体として扱う際の問題点についての議論を行い、簡単な基準で再整理を行い、考察した。set と topic の分類等がまだ不十分に思われたようにもかかわらず、topic で階層を止めるという単純な基準だけでも、本来包含関係が成り立たないようなを切捨て、包含関係が保たれている部分が多いような階層を取得することができた。これから、データを整理し、基準を増やしていくことで、Wikipedia カテゴリを、様々な研究に利用できるように再整理していきたい。

参考文献

[Bizer 09] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S.: DBpedia - A crystallization point for the Web of Data, *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 7, No. 3, pp. 154 – 165 (2009)

[Hoffart 13] Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, *Artificial Intelligence*, Vol. 194, No. 0, pp. 28 – 61 (2013)

[Ponzetto 07] Ponzetto, S. P. and Strube, M.: Knowledge Derived from Wikipedia for Computing Semantic Relatedness, *J. Artif. Int. Res.*, Vol. 30, No. 1, pp. 181–212 (2007)

[Taieb 13] Taieb, M. A. H., Aouicha, M. B., and Hamadou, A. B.: Computing semantic relatedness using Wikipedia features, *Knowledge-Based Systems*, Vol. 50, No. 0, pp. 260 – 278 (2013)

[Yoshioka 14] Yoshioka, M.: Analysis of Japanese Wikipedia Category for Constructing Wikipedia Ontology and Semantic Similarity Measure, in *Information Retrieval Technology 10th Asia Infomation Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014 Proceedings*, pp. 470–481, Springer-Verlag GmbH (2014), LNCS8870

[玉川 10] 玉川 奨, 桜井 慎弥, 手島 拓也, 森田 武史, 和泉 憲明, 山口 高平?F 日本語 Wikipedia からの大規模オントロジー学習, 人工知能学会論文誌, Vol. 25, No. 5, pp. 623–636 (2010)

[藤原 12] 藤原 嵩大, 吉岡 真治?FWikipedia の階層関係を分析するためのカテゴリパターンの提案, 2012 年度人工知能学会全国大会 (第 26 回) 論文集 (2012), CD-ROM 2C1-NFC2-4