Wikipediaカテゴリの構成要素に注目したカテゴリ階層の分析

中川 嵩教　小板橋 佳晃
北海道大学院情報科学研究科
{f-b-hawk78,yk-wwm-aa}@eis.hokudai.ac.jp

吉岡真治
北海道大学院情報科学研究科, 理研 AIP
yoshioka@ist.hokudai.ac.jp

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、その特徴として、ページ、インフォボックス、カテゴリといった構造化がなされている。このように構造化された情報を用いて、DBpedia[1] では、ページの情報に関するメタデータの抽出や、カテゴリ階層をオントロジーの構築に利用する YAGO2[2] や、日本語 Wikipedia オントロジーの研究 [4] などが行われている。しかし、Wikipedia カテゴリには、単純な包含関係ではない階層関係が存在し、単純に Wikipedia カテゴリの階層全てを概念階層として扱うと不都合が生じる。そのため、既存の研究 [2, 4] では、アドホックに、その一部のみを利用していた。この問題に対し、本研究では、過去の Wikipedia カテゴリに関する分析 [3, 6] や Wikipedia カテゴリの階層構造の再整理の提案 [5] を踏まえ、それぞれのカテゴリの種類や、カテゴリ間の関係を分類することで、Wikipedia カテゴリ階層から目的に応じた階層関係を抽出できるような方法を提案する。

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、上位カテゴリの性質も含むという包含関係が成立することが期待される。よってカテゴリ階層は、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、既存の階層の中に便宜上のカテゴリが作られ、それにより、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。

[1]https://www.wikipedia.org/

次小節では、包含関係に注目した際に注意すべきであるカテゴリの種類について述べていく。

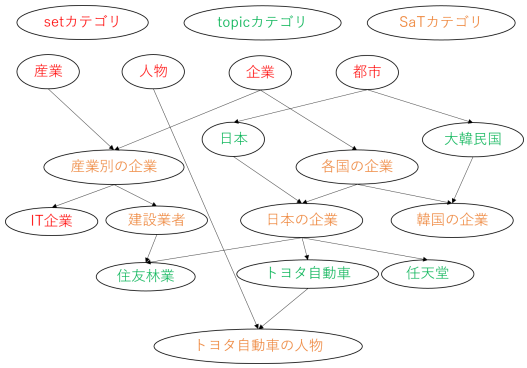

図 1: 分割のためのカテゴリによるカテゴリ分割

2.2　カテゴリの種類

Wikipedia カテゴリは、ページを分類する基準であり、「企業」、「人物」といった概念の種類を表すようなカテゴリだけでなく、「日本」「トヨタ自動車」といった具体的な事象を表すようなカテゴリが存在する。英語版 Wikipedia では、前者を set カテゴリ、後者を topic カテゴリと読んでいる。また、Wikipedia では、一つのカテゴリに大量のページが所属するとそのリストを閲覧することが困難になることから、このようなカテゴリは、様々な基準により、より詳細なカテゴリに分割することが求められている。この時できあがるカテゴリの多くは、前述の Set と Topic の組合わせであることから、このようなカテゴリは、Set-and-Topic(SaT) カテゴリと呼ばれている。例えば、図 1 の例ではオレンジ色の「日本の企業」「産業別の企業」などが SaT カテゴリに該当する。

3　Wikipedia カテゴリとその階層関係の分類

3.1　Wikipedia カテゴリの分類

これまでの研究 [5] では、Wikipedia の提案する分類に従い、Wikipedia カテゴリの分類を、概念の種類

を表す set カテゴリと、具体的な事象を表す topic カテゴリ、その組合わせである SaT カテゴリの 3 種類に分けて分類を行ってきた。しかし、網羅的な分類を進める中で、「産業別の企業」のように set と set の組み合わせと考えた方が良いカテゴリや「2017 年の日本」のようなトピックとトピックの組み合わせと考えた方が良いカテゴリが多い事が確認された。特に、topic と topic の組合わせの場合には、topic カテゴリの性質を持つため、これらを別のグループとして扱うこととした。また、set に対する制約条件が set で与えられることも多いことが確認されたため、従来の SaT のうち、set を構成要素として持ち、set の性質を持つもの (多くの場合、最後に来る set の名詞を分割したもの) を制約つき set(ConstraindSet) と呼び、CS と表記することとした。topic 同士の組合わせで出来ているカテゴリに関しては、現段階でははは、topic として扱うこととした。

3.2 Wikipedia カテゴリ階層の分類

Wikipedia のカテゴリ階層には、先ほどの set,topic,CS が様々な形で接続されており、特に、CS(例えば、「トヨタ自動車の人物」) では、制約を与える前の set(「人物」) や topic(「トヨタ自動車」) を親 (もしくは祖先) に持つ。この時、set と CS の関係は、分割の関係であり包含関係が成り立つが、後者の topic と CS の関係では、topic の上位概念 (「企業」) の下に、set(「人物」) が属するという形になり、概念の包含関係として不適切なものを含む事になる。このような問題は Wikipedia カテゴリの階層構造に多く含まれ、結果として、階層構造全体を概念階層として活用する際の大きな妨げとなっている。

この問題を解決するために、本研究では、日本語 Wikipedia のカテゴリの全ての階層関係について、網羅的にカテゴリの分類と、それを考慮したカテゴリ階層関係の分類を行う。この結果に基づいて、用途に応じて適切なカテゴリ階層関係の分類を選択することで、Wikipedia カテゴリの階層関係中に埋め込まれている意味的なずれ (上記のように、企業の下に人物のカテゴリを含む) などを起こさない概念階層構造や、topic とそれに関するカテゴリなどを抽出できる枠組を提案する。

4 日本語版 Wikipedia のカテゴリの分析

4.1 扱ったデータと作業方針

作業は、日本語版 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。

4.2 カテゴリの分類

set,topic のカテゴリと CS などの複数のカテゴリの組み合わせで表現されるカテゴリを分類するために、複数のカテゴリの組み合わせの際によく用いられている「A の B」という形といった、定型の記述が存在する [6] ことを踏まえ以下の流れで行った。Wikipedia の組合わせによる作られるカテゴリとそれ以外のカテゴリを分類し、さらに、set,topic に分類するために、以下の手順で分類を行った。

1. 一番頻出なパターンである「A の B」の形になっているカテゴリを抽出。 それぞれ「A」と「B」の出現回数を数え、それぞれが一定数以上の回数出現しているものを SaT とする。出現が少ないものは、目で見て判断した。[2]頻出であった「A」と「B」をリストにしておく。

2. 1 の「A」と「B」のリストを用いて、「A の B」以外の「A〜B」となっている SaT を抽出する。

3. 2 の「〜」の部分を抽出し、「A の B」では現れなかった「A」と「B」を抽出。

4. 上記に当てはまらなかったものは set か topic と考え、「A」のリストに存在するカテゴリは topic である可能性が高く、「B」のリストに存在するカテゴリは set である可能性が高いということと、主観により set と topic に分類をした。迷ったものに関しては、ひとまず topic にした。

5. topic リストを用いて組合わせになっているカテゴリの中で topic の組み合わせになっているものを topic に分類し、残りを CS とした。

[2]「A」と「B」の出現回数を数えたのは、「となりのトトロ」などの、助詞の「の」を含む固有名詞を SaT としないようにするためである。

set、topic に関しては、基本的には、1 つの用語に対して、どちらかの分類に割り振る予定であったが、「カエル」は「両生類」のインスタンスにも見えるが、「カエル」の下位には「アオガエル科」「アカガエル科」などカエルの種類が並びこれを見るとクラスのようにも思えるといった問題があり、階層関係での役割を考慮して、異なる分類を設定した用語も存在する。その結果、カテゴリ総数 182,734 件の内、129,599 件を CS に、14,117 件を set に、40,637 件を topic に分類した。

4.3 カテゴリ間の階層関係の分類

次にカテゴリ間の親子関係 (451,074 件) について、親子が属するカテゴリ分類に注目して分析を行った。その結果、次のような異なる代表的な関係が存在した。

- 「制約詳細化」: CS → CS である set 部分を共通として topic が詳細化される関係である。例としては「アジアの企業」→「日本の企業」などである。制約付加と同じように set 部分は変わらないので包含関係は満たされる。
- 「クラス-サブクラス」: set 部分を見たときに、同じではないが、概念としては同じとなるような関係である。set → set 関係の「作家」→「著作家」のようなものや、CS → CS 関係の「日本の企業」→「日本の多国籍企業」のような、topic 部分がないまたは共通として、set 部分がクラス-サブクラスとなっている関係が主だが、「SF 作品」→「未来を題材とした作品」や、「日本のクラブに所属するサッカー選手」→「ベガルタ仙台の選手」のように、set 部分だけを見ると逆転が起こっているが全体を見るとクラス-サブクラスとなっているような関係も一定数存在した。概念としては同じということで、この関係においても包含関係は満たされる。
- 「クラス付加」: あるカテゴリに新たな set が付け加わり CS となる関係である。多くは「トヨタ自動車」→「トヨタ自動車の人物」のように topic に set が付け加わる関係が多いが、「アニメ」→「アニメに関する企業」や「歴史の人物」→「歴史の人物を題材とした作品」のように、set 部分の後ろに新たな set が付け加わり、概念が変わってしまう set → CS 関係や CS → CS 関係も存在した。この関係においては上位カテゴリ（topic ならその上位カテゴリ）と下位カテゴリの間では概念が異なるため、包含関係は満たされない。
- 「Instance of」: 関係は「格闘技漫画」→「北斗の拳」や「日本の国公立大学」→「北海道大学」といった、下位カテゴリが上位カテゴリの概念の具体例となるような関係である。
- 「topic 包含関係」:「イギリス」→「イングランド」地理的な包含関係があるものが存在する他、「AKB48」→「前田敦子」のようなメンバー関係など、概念の記述はないが、topic 同士で包含関係があるような関係である。

多くのカテゴリについては、上記の分類で説明することができたが、これに当てはまらないものも、まだ多数存在している。

具体的には、「日本」→「日本関連一覧」のような一覧を表示させるためのカテゴリや、「スポーツ施設」→「スポーツ施設の画像」のような画像を表示させるためのカテゴリなど、Wikipedia 特有と思えるカテゴリ名を含むカテゴリが存在する。また、歴史や二国間関係等、サブクラスやインスタンスを広く取るものも例外として扱うこととした。前者を例にとると、「日本の歴史」の下位カテゴリには「日本史の人物」という人物を表すカテゴリと、「日本の古都」→「京都」のように「都市」を表すようなカテゴリが横並びになってしまい、概念階層として扱うには適しているとは言いがたい。他には、化学物質や生物分類等、専門的な知識がないとカテゴリ間の関係が正しく決められなさそうな関係が、主に、CS → CS 関係や topic → topic 関係に約 10,00 件程度存在した。例外の数としては、topic カテゴリが関わる関係は固有名詞ということもあり、関係の判断がつきづらいので、多くなっている。CS → topic 関係は類似した topic が横並びになることで判断が付きやすく、topic → CS 関係は topic 部分が共通する場合が多いため判断が付きやすいので例外の数としては少ない。CS → CS 関係でも topic 部分で判断に迷うようなものが多いことに加え、「ニューヨーク州の大学」→「ニューヨーク大学の教員」のように「ニューヨーク大学」という topic カテゴリや、「イギリスの映画」→「イギリスの俳優」のような「イギリスの映画に関わる人物」という CS カテゴリなど、望まれるカテゴリが存在せず、異なる概念が繋がる関係が例外として多く含まれた。

これらの階層関係の分類と、親子カテゴリが属するカテゴリの分類 (set,topic,CS) を用いた分類との関係を表 1 に示す。

表 1: カテゴリ間の関係

親カテゴリ→子カテゴリ	制約付加	制約詳細化	クラス-サブクラス	クラス付加	Instance of	topic 包含関係	その他
set → set	0	0	12,479	0	0	0	2,045
set → topic	0	0	0	0	4,883	0	4,704
set → CS	7,460	0	4,611	1,956	0	0	1,397
CS → set	0	0	6,165	0	0	0	1,891
CS → topic	0	0	0	0	57,608	0	3,657
CS → CS	0	108,494	86,562	975	0	0	48,675
topic → set	0	0	0	0	0	0	2,071
topic → topic	0	0	0	0	0	7,515	14,282
topic → CS	0	0	0	45,635	0	0	2,336

4.4 考察

上記の分類のうち、「制約付加」「制約詳細化」「クラス-サブクラス」の関係には包含関係を満たす関係性であり、「クラス付加」の関係では、基本的にその上位までの概念とは異なる概念が付加されるため、包含関係は満たさない。「Instance of」や「topic 包含関係」に関しては、トピックの関係から階層関係を分析する際には有用であると思われるが、概念階層の観点から分類する場合には、不適切な場合もあるので、扱いを気をつける必要がある。また、単純に、CS → CS については包含関係が成り立つと思っていたが、間にカテゴリ望まれるカテゴリが生成されず、異なる概念が繋がる関係が多く存在することが確認された。この問題については、今回提案した分類を各親子関係に付与することで、利用者は適切な関係のみに注目し階層関係を抽出して利用できると考えられる。

5 まとめ

本研究では、Wikipedia カテゴリを set,topic のような一つの用語で表されるようなカテゴリと複数 (多くの場合:2 個) の構成要素から構成されるカテゴリについても分類を行い (制約つき set:CS の分類を導入) に注目し、カテゴリを分類すると共に、カテゴリ階層についての分類も行った。その結果、親子のカテゴリの分類がカテゴリの階層関係の分類と大きく関係あることが確認されたが、単純なカテゴリ分類を用いただけでは、階層関係の分類には不十分であることが確認された。本研究で作成した分類データについては、Linked Open Data の形で公開を予定している。今後の課題としては、現在、その他となっている分類についても、より類型化を進めて、詳細な分類を付与すると共に、少数の人間で行っているカテゴリ分類の妥当性などについても、複数の人間からのチェックを受けながら、洗練していきたいと考えている。

参考文献

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 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.

[2] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, Vol. 194, No. 0, pp. 28 – 61, 2013.

[3] Masaharu Yoshioka. 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.

[4] 玉川奨, 桜井慎弥, 手島拓也, 森田武史, 和泉憲明, 山口高平. 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.

[5] 中川嵩教, 吉岡真治. 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. 人工知能学会全国大会論文集, Vol. JSAI2018, pp. 2F402–2F402, 2018.

[6] 藤原嵩大, 吉岡真治. Wikipedia 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.