2C1-NFC-2-4

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

Proposal of Category Description Pattern for Analyzing Wikipedia Category Structure

藤原 崇大[*1]
Takahiro Fujiwara

吉岡 真治[*1]
Masaharu Yoshioka

[*1] 北海道大学
Hokkaido University

There are several attempts to construct large scale ontology by using Wikipedia category information. However, there is no comprehensive analysis about the information described in the category description. In this paper, we propose a framework to analyze category description by using description pattern. By using this pattern, we can propose a new method to construct candidate pairs of class-subclass, class-instance, and whole-part relationship from Wikipedia category structure. We also discuss the issues for using Wikipedia category structure for constructing ontology.

1. はじめに

近年、インターネットの普及により、ウェブ上には膨大なテキストが存在している。これらのテキストを有効活用するためにはそのテキストの持つ意味に注目する必要があり、大規模なオントロジーを構築する研究が進められている。特に近年ではインターネット上の大規模百科事典であるWikipedia[1]からこの意味関係を抽出し、オントロジーを構築する研究が行われている。具体的には、Wikipediaが持つページの分類体系であるカテゴリを用いて、クラスの階層関係を構築する手法が提案されている[Hoffart et. al. 11, 玉川 11] 。しかし、カテゴリ情報にはクラスを表す情報だけでなく属性を表す情報が含まれているため[Nastase et.al., 08]、単純にカテゴリ名をクラスとして利用する際には問題がある。

そこで本研究では Wikipedia のカテゴリについて、その記述パターンを網羅的に分析するとともに、そのパターンを用いたカテゴリの階層関係の分析を行い、カテゴリを用いてオントロジーを構築する際の注意点などについて議論を行う。

2. Wikipedia のカテゴリ記述パターン

Wikipedia のカテゴリとは、膨大な Wikipedia のページを分類するために利用される分類体系である。このカテゴリには、階層関係が存在し、様々な観点から対応するページの絞り込みが可能となる。カテゴリからの知識抽出の研究としては、カテゴリに属するページのクラスの推定と概念間のクラス階層の抽出によるオントロジーの構築[Hoffart et. al. 11, 玉川 11]や、属性情報の抽出[Nastase et.al., 08]といった様々な研究が行われている。

しかし、これらの研究では、特定の条件にマッチするカテゴリを利用する方法を提案しているが、カテゴリ全体の構造についての分析は十分に行われていない。

そこで、本研究では、日本語Wikipediaのカテゴリ（総数100,997 件 [2]）についての網羅的分析を行う。ただし、このカテゴリの中には、Wikiepedia特有のカテゴリが含まれる。例えば、「荒らし対策」「投稿ブロック」などWikipediaのページを管理する上でつけられるカテゴリや、「曖昧さ回避」「2010 年 1 月-6 月検証が求められている記事」など、他の似たページとの注意を促す物やページの不確実性を記すカテゴリがある。これらはオントロジーを構築する際には有用でないカテゴリであるので、ひとまず削除し、残りの 95,785 個に対して分析を行う。

2.1 カテゴリ記述パターンによるカテゴリの分類

Wikipedia 固有の表現カテゴリを除いたカテゴリに対して分析を行った結果、ほとんどの Wikipedia のカテゴリは以下の組み合わせで記述されることが確認された。

- 名詞　（例：日本、選手、北海道大学）
- 修飾節　（例：かつて存在した、廃止された）
- 助詞　（例：が、の、を、に）
- 動詞句　（例：関する、舞台とする、所属した）
- 接続詞　（例：および、による）
- 付加情報　（例：（五十音別）、（都道府県別）、あ行）

この内、修飾節や付加情報はカテゴリ記述パターンの分析には有用でないと判断し、分類の際にひとまず削除して扱う。例として、「かつて存在した大学」は「大学」として、「日本の企業_(都道府県別)」は「日本の企業」と扱って分類を行う。

これらの基準でのカテゴリの記述パターンの分類を行った結果を表 1 に記す。名詞や助詞の判定には、MeCab[3]を利用した。分類の結果、これらのパターンに当てはまらないものが少数存在したが、これらについては、MeCabによる形態素解析のミスと判断できるものは、手作業により分類を行った。

表 1 の結果から、日本語 Wikipedia のカテゴリは名詞単独（以下 Noun）が約 53%で名詞＋助詞「の」＋名詞（以下「A の B」）が約 44%とこの両記述パターンで全体の約 97%を占めていることが分かる。残りの 3%のカテゴリには「文学を原作とする作品」（A を VB）や「日本に関する書物」(A に VB)といった動詞句を含むカテゴリが多く該当するが、ここで使われている動詞句の種類自体は少なく、ある動詞句に対して複数の A と B が組み合わさって、数が膨れ上がっている結果となっている。また、ここで使われている A、B の名詞のほとんど全ては「A の B」で使われている名詞の中に含まれているため、名詞の分類としては、「Noun」と「A の B」に焦点を当てて更なる分析を行う。

連絡先：藤原 崇大，北海道大学大学院情報科学研究科
fujiwara@kb.ist.hokudai.ac.jp

[1] http://ja.wikipedia.org/

[2] http://dumps.wikimedia.org/jawiki/20120206/ にある 2012 年 2 月 6 日のダンプデータを利用した。

[3] http://code.google.com/p/mecab/

表1：Wikipediaのカテゴリ表記パターンと個数及びカテゴリ例

カテゴリ記述パターン	略記	個数
名詞単独 例:「日本」「宇多田ヒカル」	Noun	41,864
名詞+助詞+名詞　　の 例:「日本の野球選手」	AのB	50,781
による 例:「DC-8による航空事故」	AによるB	86
と 例:「軍事教育と訓練」	AとB	47
名詞+助詞+名詞+助詞+名詞 の—の 例:「京都市の寺院の画像」	AのCのB	1,143
の—と 例:「言語の転写と翻字」	AのCとB	38
と—の 例:「日本と海外の合作アニメ」	AとCのB	35
名詞+接続詞+名詞　　および 例:「自動認識およびデータ取得」	AおよびB	3
名詞+及び+名詞＋の+名詞 例:「阪神タイガース及びその前身球団の選手」	A及びCのB	108
名詞+助詞+動詞句+名詞　　に 例「商業に関する学科」	AにVB	786
を 例:「鉄道を題材にした作品」	AをVB	706
で 例:「スイスで発生した航空事故」	AでVB	54
その他(動詞句を含む) 例:「格闘する擬似化キャラクター」		114
その他(動詞句を含まない) 例:「長大な音楽作品名」		20

(1) カテゴリ記述パターン「A の B」の分析

カテゴリ記述パターン「A の B」の A、B それぞれに該当する名詞の総種類と出現回数を調査した結果を表 2 に記す。

カテゴリ記述パターン「A の B」のカテゴリは 50,781 個存在する。その内 A に該当する名詞は 17,260 個、B に該当する名詞は 4,321 個あり、それらの組み合わせにより 50,781 個まで膨れ上がっている。A に該当する名詞の上位は国名、地名で占めており、特に日本の国名・地名が大半を占めている。これは、日本語 Wikipedia には日本に関する情報が詳細に分類されて記述されているためである。B に該当する名詞の上位は「画像」、「人物」、「選手」といった名詞が占めており、国名・地名に修飾される名詞が上位に来ている結果となった。

カテゴリ記述パターン「A の B」のほとんどは A が B を修飾する形となっているため、A には国名・地名やインスタンスのような名詞が多く、B にはクラスを示すような名詞が多く見られた。

(2) カテゴリ記述パターン「Noun」の分析

カテゴリ記述パターン「Noun」のカテゴリは 41,864 個存在するが、人名や作品名、地名・国名といった固有名詞から一般名詞と様々なタイプのカテゴリが混在しており、一概に分析することが出来なかった。カテゴリ記述パターン「A の B」の A の名詞と Noun の名詞が一致しているのは 7,862 個(約 46%)であった。A のみに存在する名詞としては、「千葉県出身」などといった、属性とその関係を表すような名詞と、「韓国」のようにカテゴリが作られていない固有名詞が多く存在した。また、B の名詞と Noun の名詞が一致しているのは 2,197 個(約 51%)であった。B のみに存在する名詞としては、「項目名」などの、「A の B」の組み合わせでないと意味がはっきりしない語が存在した。その他には、「公立学校」のように、カテゴリとして作成してもかまわないと考えられる名詞も多く含まれていることが確認された。

表 2:「A の B」の A に該当する名詞

A に該当する名詞	出現回数
日本	2,767
アメリカ合衆国	831
各国	532
イギリス	504
フランス	474
・・・	・・・
北海道	162
・・・	・・・
神戸市	31
・・・	・・・
千葉県出身	2
・・・	・・・

表 3:「A の B」の B に該当する名詞

B に該当する名詞	出現回数
人物	3,190
画像	2,471
選手	1,916
教員	1,822
楽曲	1,766
アルバム	1,245
鉄道駅	917
企業	759
歴史	718
・・・	・・・

3. カテゴリ記述パターンを用いた階層の分析

Wikipedia のカテゴリには図1のような階層関係が存在する。前節で述べたカテゴリ記述パターンがどのような形で階層関係に表れるかを分析することにより、カテゴリ間の階層関係の分析を行う方法を提案する。

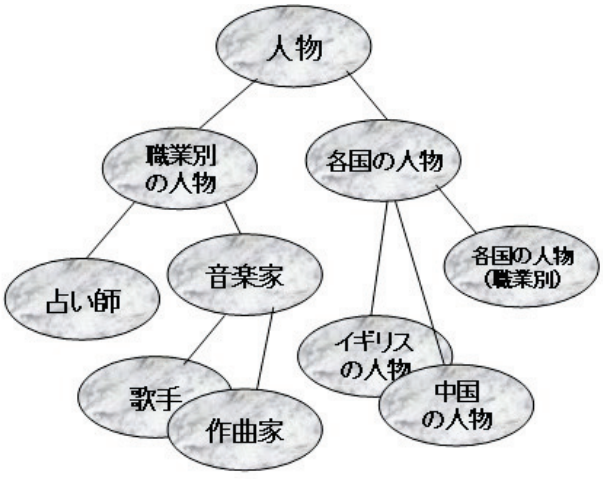

図 1: Wikipedia のカテゴリの階層関係の一例

3.1 カテゴリ記述パターンによる階層関係の分類

本研究では、Wikipedia に特有のカテゴリを除く 95,785 件のカテゴリについての階層関係 205,979 ペアについて分析を行う。表 4 に、第 2 節で定めたカテゴリ記述パターンでこれらの関係を分類した結果を記す。

表 4: Wikipedia のカテゴリが持つ階層関係の
カテゴリ記述パターンによる内訳

		上位カテゴリ		
		Noun	A の B	その他
下位カテゴリ	Noun	52,609	29,835	3,718
	A の B	27,633	78,888	6,230
	その他	1,560	3,412	2,094

表 4 にあるように、Wikipedia のカテゴリが持つ階層関係は「Noun」と「A の B」の組み合わせで約 92%を占める。カテゴリの比率に比べるとやや少なくなっているが、これは、「A に VB」や「A を VB」といった動詞句を含むカテゴリが中間階層に多く存在するのに対し、固有名詞などの Noun は、末端の階層に表れることが多いことからこのような結果となったと考えられる。

3.2 カテゴリ記述パターンを用いた階層関係の分析

本研究では、上位カテゴリと下位カテゴリの記述パターンに注目することにより、カテゴリ階層関係についての特徴を分析する方法を提案する。ただし、表2の組み合わせのうち、「Noun」から「Noun」の組み合わせについては、分析する手掛かりが少ないため、「Noun⇒A の B」、「A の B⇒Noun」、「A の B⇒A の B」に注目して分析を行う。

(1) 「A の B⇒A の B」

「A の B⇒A の B」については、上位と下位の A,B の同一性に基づき、表 5 に示す 4 パターンに分類できる 。

表 5:カテゴリ記述パターン「A の B」の
上位下位の関係パターン

上位カテゴリ	下位カテゴリ	頻度
A1 の B1 例:各国の人物	A1 の B1 例:各国の人物(職業別)	1,027
A1 の B1 例:日本の野球	A1 の B2 例:日本の野球選手	29,460
A1 の B1 例:各国の地理	A2 の B1 例:日本の地理	41,891
A1 の B1 例:ファッション界の人物	A2 の B2 例:日本の美容師	6,510

「A1 の B1⇒A1 の B1」については、基本的には付加情報が付いただけのものである。また、「A1 の B1⇒A2 の B2」については、分析の情報が不足しているため、「A1 の B1⇒A1 の B2」、「A1 の B1⇒A2 の B1」の二つに着目して、さらに分析を行った。まず、「A1 の B1⇒A2 の B1」について分析を行う。本研究では、多くの B1 に共通して存在する A1 と A2 については、意味的に関係性が近いという仮説に基づき、共通な B1 を持つ A1 と A2 のペアリストを作成した。例えば「各国の地理⇒日本の地理」ならば、B1(地理)を共通として、「A1(各国)⇒A2(日本)」のペアリストを作成した。また、その異なりをとると、15,132 件となった。また、同様に、「A1 の B1⇒A1 の B2」についても分析を行うと、その異なりが 5,371 件となった。その代表的なペアの情報を表 6,7 に示す。

表 6: B1 を共通とする A1A2 のペアリスト

A1⇒A2 のペアリスト	出現回数
各国⇒日本	448
各国⇒アメリカ合衆国	377
各国⇒フランス	311
各国⇒イギリス	303
・・・	
日本⇒北海道	126
日本⇒東京都	97
・・・	

表 7: A1 を共通とする B1B2 のペアリスト

B1⇒B2 のペアリスト	出現回数
スポーツ選手⇒サッカー選手	172
スポーツ選手⇒オリンピック選手	162
スポーツ選手⇒陸上競技選手	135
・・・	
地理⇒都市	212
地理⇒地形	180
・・・	

表 6 から、A1 と A2 のペアには、「各国⇒日本」のように、クラスとインスタンスの関係になっているものや、「日本⇒北海道」のように、地理的な包含関係を示しているものが多く存在することが確認された。このような関係は、B1 に相当するインスタンスの持つべき属性を示していることが多い。例えば、「各国の B1」の下位カテゴリである「日本の B1」では、所在地、出身などの属性が「日本」であることを示している。これは、 [Nastase et.al., 08]らが指摘している属性情報による分類が、日本語の Wikipedia にも存在していることを示している。

また、表 7 から B1 と B2 の関係には、「スポーツ選手⇒サッカー選手」といったクラスとサブクラスの関係の情報が存在するだけでなく、「地理」と「都市」といった関連概念のようなものを多く含むことが確認された。

また、一般に、A1 と A2 のペアの頻度が高いものは、A1 で分類すると共通の性質を持っているものが多く、このペアの頻度情報を用いることで、クラスとインスタンスの関係や地理的な包含関係に関する情報の構築に有用であることを確認した。

一方、共通の A1 を持つ B1 と B2 のペアの上位には、クラス階層の情報を得ることができるものも存在するが、「地理」と「都市」といった関連概念のようなものを多く含む。しかし、A1 と A2 のペアと同様に、B1 で共通するものを集めると、その性質は共通である場合が多く、手作業での概念情報抽出に有用であることが確認された。

(2) 「Noun⇒A の B」

次に、「Noun⇒A の B」について分析を行う。この組み合わせでは、A もしくは、B に Noun が付いている場合が多く、「A⇒A の B」となっているものが 14,191 件、「B⇒A の B」になっているものが、5,376 件存在し、その他が 8,066 件であった。「A⇒A の B」では、「日本⇒日本の地理」のように、特定のインスタンスに対する付加情報を追加するものが多く存在した。一方、「B⇒A

の B」には「地理⇒日本の地理」のように、対象とするカテゴリが持つべき属性を制限するようなものが多く存在した。

(3) 「A の B⇒Noun」

最後に、「AのB⇒Noun」であるが、例外(「日本の寺⇒寺_(都道府県別)」[1]など)的な 23 件をのぞく 29, 812 件でA,Bと Nounの間に一致は見られなかった。これらの関係の多くは、NounとBの間にクラスとインスタンスの関係、クラス・サブクラスの関係、地理的な包含関係であり、こちらについても、Bを中心に性質が似ていることが確認され、これらの情報を分析することで、多くのカテゴリの情報が獲得できることが示唆された。

4. カテゴリの階層関係からの情報抽出

前章で提案した分析手法を用いることで、多くの階層関係の情報が抽出できる可能性が確認できた。そこで本研究では、実際のデータからのカテゴリ階層の情報抽出を行った。「A の B⇒A の B」と「A の B⇒Noun」の情報から、その対応関係として、以下の 3 種類の情報を作成した。

- クラス―サブクラス (スポーツ⇒野球) 6,245 件
- クラス―インスタンス (アーティスト⇒上戸彩) 22,697 件
- 地理の包含関係 (北海道⇒札幌) 5,300 件

この分類の過程の中で、B を共通とする「A の B」に存在する特徴的な表現として、以下のような階層関係が見つかった。

- 年代―年代 (1980 年代の書籍⇒1981 年の書籍)
- 属性値―属性値を持つインスタンス(日本のアルバム⇒宇多田ヒカルのアルバム)

また、作成した対応関係の情報を用いて、B を共通とする「A の B」における A1 と A2 の関係の分析をしたところ、A2 が A1 のインスタンスであるものが 16,341 件(約 39%)、A2 が A1 の地理的包含関係にあるものが 14,290 件(約 34%)となり、この二つが多くをしめる。また、上記の年代の関係を表すものが 3,883 件(約 9%)となっている。一方、A2 が A1 のサブクラスであるものは 382 件(1%弱)と非常に少ないことが確認された。残りについては、上記の属性値―インスタンスなどの様々な関係が含まれるが、今後、より詳細な分析を行う予定である。

次に、A を共通とする「A の B」における B1 と B2 の関係の分析をしたところ、B1 が B2 のサブクラスと判断されるものが、15,247件(約 51%)と非常に多いが、インスタンスと判断されるものが 346 件(約 1%)、全体部分の関係にあるものが 196 件(約 0.6%)と非常に数が少なかった。一方で、法務官僚⇒官僚(司法省・刑部省)[1]といった形のクラスの逆転が見受けられた。

また、分析を行わなかった「A1 の B1⇒A2 の B2」においても、A1 と A2 の間に、地理の包含関係が 633 件、インスタンスが 247 件、年代が 102 件、クラスが 38 件と約 16%のペアについては分類が行えた。また、B1 と B2 の間には、地理の包含関係が 111 件、インスタンスの関係が 3 件、クラスの関係が 1,101 件と同様に、約 19%のペアに対して分類が行えた。

また、B1 においては、「漫画のキャラクターゲーム⇒ドラゴンボールのゲーム」のように、クラス階層の逆転が 497 件存在した。今後のクラス階層の分析の際には、上記のような問題点があることを考慮して、分析を行う必要がある

また、残りのペアについて分析をすると、「東京都の大学⇒東邦大学の教員」のように、中間の「東邦大学」のようなカテゴリが存在しない場合なども存在した。ただし、その出現には、一定のパターンが存在するため、個別に対処することで、「A の B⇒A の B」の場合の類型化はより網羅的に行えると考えている。

4.1 考察

本稿での様々な分析結果から、カテゴリにはクラスやインスタンスを表すカテゴリと属性を表すカテゴリが混ざっていることが分かった。また、各々のカテゴリとサブカテゴリの間の関係においても、対象が持つ属性の詳細化(日本の地理⇒関東地方の地理⇒東京都の地理)のような階層関係と、クラス―サブクラスの階層関係(著作家⇒随筆家)などの関係が混在している。

そのため、カテゴリの階層関係を用いた類似性を判断する際には注意が必要である。例えば、以下のカテゴリがついた三つの記事の類似性を量りたいと仮定する。

- 記事 A 「存命人物・日本の詩人」
- 記事 B 「存命人物・随筆家」
- 記事 C 「2010 年没・ドイツの詩人」

ここで、職業を基準に類似性を考えると、類似性の高い記事は AC である。しかし、全てのカテゴリを同列に扱い、階層距離によって類似性を量った場合、カテゴリの階層関係は以下であるので、

- 著作家 ⇒ 詩人 ⇒各国の詩人⇒日本の詩人
  - ┗⇒ 随筆家 ┗⇒ドイツの詩人
- 人物 ⇒ 没年 ⇒21 世紀没 ⇒2010 年没
  - ┗⇒ 存命人物

記事 AB の距離は 4 階層、記事 AC の距離は 6 階層、記事 BC の距離は 8 階層となり、AB>AC>BC と間違った結果となる。

このように、類似性を量る際にクラスカテゴリと属性カテゴリでは階層の重みが違い、同列に扱うべきでないことが分かる。

5. おわりに

本稿ではカテゴリの記述パターンの分析を行い、その分類パターンを用いたカテゴリの階層関係の分析を行ったが、階層関係の構築はまだ十分に行っていない。現在作成したカテゴリ階層の情報をより詳細に分析することにより、より多くの階層関係の情報抽出を行うと共に、[Nastase et.al., 08]らが英語の Wikipedia のカテゴリについて作成した属性情報抽出のためのパターンを日本語の Wikipedia についても作成する予定である。

参考文献

[玉川 11] 玉川 奨, 関本 有佳, 森田 武史, 山口 高平, "日本語 Wikipedia からプロパティを備えたオントロジーの構築", 人工知能学会論文誌 特集論文「近未来チャレンジ」 Vol.26 No.4 pp.504-517 (2011.7)

[Hoffart et. al. 11] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis Kelham, Gerard de Melo, and Gerhard Weikum "YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages", in the proceedings of the 20th International World Wide Web Conference (WWW 2011), 2011.

[Nastase et.al., 08] Vivi Nastase , Michael Strube, "Decoding Wikipedia Categories for Knowledge Acquisition", In Proceedings of the 23rd national conference on Artificial intelligence - Volume 2 (AAAI'08), Anthony Cohn (Ed.), Vol. 2. AAAI Press 1219-1224.

[1] 末尾の (都道府県別) 、(司法省・刑部省)は付加情報として削除するため