Vol.2012-DBS-154 No.1
Vol.2012-IFAT-107 No.1
2012/8/1

Wikipediaを中心としたLinked Open Dataに関する一考察

吉岡 真治[1,a)]

概要：現在、Wikipedia の情報を中心に様々な Open Data を関連づけて活用する Linked Open Data の研究が行われている。本発表では、Wikipedia のページが表現する情報の粒度が与える影響について考察を行う。

1. はじめに

近年、Web 上に公開された多くの有用なコンテンツを活用するために、複数のコンテンツの間の関係 (Linked Data と呼ぶ) を記述することにより、よりその有用性を高めようとする Linked Open Data (LOD)[1] と呼ばれる活動が盛んになっている。この LOD では、数多くのインスタンスの情報やカテゴリの階層構造を持つ Wikipedia[*1]から様々な関係を取り出して利用可能とした DBPedia[*2] のデータを中心に様々なデータが関連づけられている。

我々は、これまでに、Wikipedia に書かれている情報を活用するための研究として、Wikipedia と GeoNames[*3]の間の自動的リンク発見とメンテナンスの手法 [2] や Wikipedia のカテゴリ情報の分析 [3] を行ってきた。

本稿では、これまでの研究を踏まえ、Wikipedia のページやカテゴリが持つ性質が、LOD の中心として Wikipedia を使う場合の影響について議論を行う。

2. Wikipediaが持つ情報とその活用

Wikipedia には、現在、日本語版に 812,296 件、英語版に 3,983,274 件の記事[*4]があり、これらのページが階層的なカテゴリによって、分類されている。また、この階層的なカテゴリの性質を利用した Wikipedia オントロジーの構築が行われている。

本章では、Wikipedia に記述される情報の中で、カテゴリとページに注目して、概観すると共に、これらの情報を利用した Wikipedia オントロジーについて紹介する。

2.1 Wikipediaのカテゴリ

ここでは、日本語版 Wikipedia のカテゴリに関する説明ページ「Wikipedia:カテゴリ」、「Wikipedia:カテゴリの方針」をもとに、カテゴリがもつ性質について概観する。

Wikipedia におけるカテゴリの定義は、「カテゴリとは、記事を分野別にまとめた索引」である。これらのカテゴリは、「総記」、「学問」、「技術」、「自然」、「社会」、「地理」、「人間」、「文化」、「歴史」の 9 個の主要カテゴリのいずれかに分類される。

さらに、Wikipedia においてはこれらのカテゴリが、図 1 に示す基準により作成されるサブカテゴリが作成され、各々のカテゴリ間の関係については、図 2 に示す基準で階層化が行われる。

また、カテゴリに包含する記事数が増えた場合には、より具体的なサブカテゴリが作成される。このような基準で作成されたサブカテゴリには、上位カテゴリの内容をおおむねカバーしているようなサブカテゴリの組 (例えば、「アジアのサッカークラブ」に対して、「日本のサッカークラブ」、「シンガポールのサッカークラブ」など) が作成できる場合と、「映画作品」に対する「アカデミー受賞作品」の様に、「アカデミーを受賞しなかった映画作品」といった意味のないサブカテゴリを作らないとこの様な組が出来ない場合がある。ただし、後者の場合には、意味のないカテゴリを作らないため、サブカテゴリの網羅性が保証されない。

この前者のように、上位カテゴリの内容をおおむねカバーしているようなサブカテゴリの組を、「分割として機能するカテゴリ」と呼び、後者を「分割として機能しないカテゴリ」と呼ぶ。また、「分割として機能するカテゴリ」

[1] 北海道大学
Hokkaido University, N14 W9, Kita-ku, Sapporo-shi, Hokkaido, 060-0814, Japan
[a)] yoshioka@ist.hokudai.ac.jp
[*1] http://wikipedia.org/
[*2] http://dbpedia.org/
[*3] http://www.geonames.org/
[*4] 2012 年 6 月 26 日現在

(1) カテゴリは第一義として、「分類」を示すものです。
「xx は YY のひとつである」と言うことができれば、「分類」を示すと言えます。項目 xx はカテゴリ YY に属するべきです。反例として、北朝鮮と韓国は関連がありますが、どちらかがどちらかを包含する関係ではありません。
(2) 上記に加えて、ウィキペディアのカテゴリとしては「関連が深いキーワード」を示すことができます。
「分類」より「キーワード」を指向しているカテゴリも存在します。記事 xx が「YY 関連用語」であるという意味合いでカテゴリ YY に属することが期待される場合があります。例として、学術用語と Category:学問の関係など。この場合も、カテゴリはより上位の概念であることが求められるため、逆の関係ではあり得ません。
(3) また、カテゴリはウィキペディアの骨組みの意味を持ちます。
カテゴリ機能の普及によって、カテゴリの構造がウィキペディアの全体構造を示すこととなりました。カテゴリ同士の関係もウィキペディア全体を意識した一貫性や無矛盾性が求められ、よいカテゴリ構造を作ることが、わかりやすいウィキペディアを作ることにつながります。似た意味合いのカテゴリや大きく重複するカテゴリがある場合は、なるべく内容をすり合わせ、統合を検討しましょう。併存させる場合も、明確な使い分けの方針を決めましょう。そうしなければ混乱が永続することになります（例：「文房具」と「事務用品」など）。

図 1 カテゴリづけの方針

多くのカテゴリは一つ以上の親カテゴリを持ちます。例えば、Category: 日本の作家は Category:各国の作家と Category:日本の人物 (職業別) の両方に含まれています。あるカテゴリを他のカテゴリのサブカテゴリとする場合、前者のカテゴリの内容が（ある程度の例外はありえますが）後者のカテゴリの内容として含まれうるものであることを確認してください。カテゴリの上下関係は親子関係であり、ループ構造にならないように注意してください。ある二つのカテゴリ同士に深い関係があり、しかし上下関係を作らないような場合は、カテゴリの本文で関連づけるに留めてください。

図 2 カテゴリの構造

については、ページに親カテゴリを重複して付与しないことになっている。

2.2 Wikipedia のページ

ここでは、日本語版 Wikipedia のページ (記事) に関する説明ページ「Help:記事とは何か」、「Wikipedia:記事名の付け方」、「Wikipedia:ページの分割・統合」、「Wikipedia:リダイレクト」、「Help:リダイレクト」をもとに、ページのタイトルの付け方や、その単位に関する議論をページがもつ性質について概観する。

Wikipeida における記事 (article) とは、「百科事典としての情報が記載されているページ」のことである。これらのページ (記事にはひとつ題名を付ける必要があり、その題名については、図 3 のような基準が提案されている。

- 認知度が高い - 信頼できる情報源において最も一般的に使われており、その記事の内容を表すのに最も著名であると考えられるもの。
- 見つけやすい - 読者にとって記事の中で見つけやすいもの（そして編集者にとって最も自然に他の記事にリンクできるもの）。
- 曖昧でない - その記事の内容を曖昧さなく見分けるのに必要な程度に的確な名称であること。
- 簡潔 - 短く、要点を突いているもの（曖昧さ回避の場合でも、カッコ内を短く保つことは必要です）。
- 首尾一貫している - 他の似たような記事においても、同じように使われているもの。

図 3 記事名の付け方

- 次のようなページ名を入力しても目的地にたどりつけるため
  - ページの主題の別名
  - ページの主題に関する副次的な話題（この場合、セクションへのリダイレクトにしてもよい）
  - 大文字・小文字、ハイフネーションなどの違う表記
  - 漢字の字体の違う表記
  - つづりや送り仮名などの違う表記
  - よくある綴り間違い、誤字など
- あるページに簡便にたどり着くため（ショートカット）
- ページを移動した後に、リンク切れを防ぐため（内部リンクは修正できますが、外部からのリンクのことも考慮しましょう）

図 4 リダイレクトの作成基準

一方、あまりに、多くの内容を含むページや、必要以上に細切れにされたページを作らないための基準が提案されている。

また、各ページには、タイトルと違っても、その内容を表示することが適切があるページが存在するため、図 4 に示すリダイレクトという枠組みが用意されている。

2.3 Wikipedia オントロジーの構築

前章で述べた Wikipedia の性質に基づいた Wikipedia オントロジーの構築が行われている。代表的なものとしては、英語版 Wikipedia に基づいた DBPedia[4] や YAGO2 (Yet Another Great Ontology 2) [5] がある。また、日本語についても、同様の試み [6] が行われている。

これらの研究では、Wikipedia のカテゴリーが持つ図 1 の 1 の分類としての役割に注目し、カテゴリ情報からクラスの情報を生成し、そのカテゴリに属するページをインスタンスとして分類するという形で、多くのインスタンスを含む大規模オントロジーの構築を行っている。

さらに、Infobox と呼ばれる構造化データからの属性情報の抽出などを行い、SPARQL を用いて、特定の条件を満たすインスタンスの検索を行うシステムなどが構築されている。

3. Wikipedia を中心とした LOD

3.1 Wikipedia オントロジーに関する考察

前節で述べたように、Wikipedia オントロジーでは、カテゴリの持つ「分類」としての側面に注目して、クラス階層の構築を行っている。しかし、カテゴリは、あくまでも、ユーザが Wikipedia を閲覧するためのナビゲーションを支援するために作られているため、必ずしも、全ての親子のカテゴリの間に分類関係が成り立っているわけではない。

我々は、2012 年 2 月 6 日のダンプデータに存在した 100,997 件から「スタブ」などの Wikipedia 固有のカテゴリを除去した 95,765 件のカテゴリとそのカテゴリ間の関係 203,975 ペアについて、その表記パターンと階層関係についての分析を行った [3]。

この研究では、Wikipedia のカテゴリを構成する要素に基本的なパターンがあることに注目した分類を行った。この研究では、一般的な、「名詞」、「助詞」、「動詞句」、「接続詞」という品詞分類に加え、「修飾節 (例えば、『かつて存在した』)」と「付加情報 (例えば、曖昧性回避のための文末の () 表記『(業種別)』) の組合わせでカテゴリの分類を行った。例えば、「かつて存在した日本の企業 (業種別)」というカテゴリは、「かつて存在した { 修飾節 } ＋日本 { 名詞 } ＋の { 助詞 } 企業 { 名詞 }(業種別){ 付加情報 }」の組み合わせと判断される。

ここで、「修飾節」と「付加情報」は、概念の階層構造を分析するためのパターンとしては、あまり有用でないと考え、これらの項目を無視して、パターンの数を数えたところ、表 1 のような件数となった[*5]。

表 1 Wikipedia のカテゴリの表記パターンによる分類

カテゴリのパターン	件数
名詞単独 「日本」、「宇多田ヒカル」	42,044
名詞＋の＋名詞 「日本の野球選手」	50,643
名詞＋の＋名詞＋の＋名詞 「京都市の寺院の画像」	1,141
名詞＋に＋動詞句＋名詞 「商業に関する学科」	785
名詞＋を＋動詞句＋名詞 「鉄道を題材にした作品」	706
その他	446

ここで、品詞の判定には、MeCab を利用したが、明らかに固有名詞と判断できるもの (例：天空の城ラピュタ) については、MeCab の結果ではなく、固有名詞と判断することとした。この結果、カテゴリの内、約 44%が「名詞単独」であり、53%が、「名詞＋の＋名詞」の形で表されていることが確認された。よって、「名詞単独」と「名詞＋の＋名詞」のみに注目するだけで、大部分の Wikipedia のカテゴリの情報を利用できると考えられる。

また、「名詞単独」→「名詞 A ＋の＋名詞 B」の形式では、27,356 件中、親の名詞と名詞 A が同じ (「日本」→「日本の人物」) ペアが 14,051 件、名詞 B が同じ (「作家」→「日本の作家」) ペアが 5,378 件と大部分を占めた。これは、「分割として機能するカテゴリ」の上位では、分割の元になるカテゴリと分割の基準となるカテゴリを合わせて親に持つことが多いためだと考えられる。

次に、「名詞 A ＋の＋名詞 B」→「名詞単独」の関係では、名詞 B と子の名詞の間の関係が強く、「日本の歌」→「演歌」のようなクラス・サブクラスの関係や、「大阪府の大学」→「大阪大学」のようなクラス・インスタンスの関係が存在した。

これらの分析結果から、Wikipedia のカテゴリ階層をオントロジーの概念階層として利用する場合の問題点を以下に述べる。

(1) Wikipedia のカテゴリには、インスタンスの情報が存在する。
Wikipedia のカテゴリには、クラス以外に、特定のインスタンスに関連する概念を関連づけるためのカテゴリが存在する。

(2) Wikipedia のカテゴリには、概念階層として不適切なものを含む
「日本」→「日本の人物」という親子関係は、「日本」が「地理」を主要カテゴリとして持ち、「人物」が「人間」を主要カテゴリを持つことから、一般的な概念階層としては、不適切であると考えられる。

(1) については、カテゴリを利用する際には、そのカテゴリがインスタンス (固有名詞) であるかどうかを判断する必要があることを示している。また、(2) については、「札幌市」→「札幌市の企業」→「北海道日本ハムファイターズ」というカテゴリが存在し、「名護市」にキャンプ地という関係で、「北海道日本ハムファイターズ」というカテゴリが付与されている場合に、「札幌市」→「名護市」という関係を付与するといったカテゴリの意味的ドリフトを起こす原因となる。

Wikipedia オントロジーの構築では、これらの問題に対処するために、様々なヒューリスティックスが用いられている。しかし、うまく行かない反例が見つかる毎にヒューリスティックスを修正するような方法では、どこまで作業を進めれば良いかが不明となる。そこで、上記のようにパターンを分類していくと共に、既存の獲得した情報を利用して、これまでの手法で、うまく利用できている親子関係と、利用できていない親子関係を分類することが有用であると考えている。

[*5] 表の値は、論文発表後に再検討を行った結果を反映しているため、[3] とは異なる。

また、リダイレクトによる同義語の収集という方法も提案されているが、図 4 にあるように、リダイレクトには、異表記以外の使い方もあり、これを分離することが必要である。ただし、情報検索などのように、関連概念も含めて検索語拡張をするという立場であれば、リダイレクトを使った検索語拡張というのは、問題が少ない可能性もある。

一方、Wikipedia のカテゴリの現状にも、問題があると考えている。例えば、分割のために機能するカテゴリで用いられるカテゴリの多くは、特定の属性を持つことによって分類されている。ところが、この関係の一貫性を手動で保つためには、大きな労力がかかる。例えば、1976 年生というカテゴリには、生年月日が 1976 年の人が全て網羅されていることが期待されるが、必ずしも、全ての人にカテゴリがついている保証はない。DBPedia の様な形でデータを整理することができれば、この様なカテゴリは、SPARQL の検索クエリと対応づけることが出来るはずであり、今後は、お互いの協調が求められると考えている。

3.2 Wikipedia と GeoNames のリンク発見とメンテナンス

Wikipedia のカテゴリ情報には、先に述べたように、「分割として機能するカテゴリ」が存在し、その多くが、地理的な情報 (国名や地域名) などで、分割をされている。我々は、この性質を利用して、異なる Open data である英語版の Wikipedia と GeoNames の間のリンク発見の方法を提案している [2]。

GeoNames は、Creative Commons attribution ライセンスで開発されている地名情報に関するデータベースであり、各地名には、「ID」、「名前」、「別名」、「国名」、「行政単位」、「地名のタイプ」、「座標」などの属性情報が付加されている。この GeoNames には、2012 年 2 月 1 日時点で、8,105,590 件の世界中の地名の情報が存在している。

本手法では、Wikipedia のカテゴリの情報から、「国名」や「行政単位」の情報を抽出すると共に、「地名のタイプ」を推定することによって、座標の情報を用いない Wikipedia と GeoNames の間のリンク発見の方法を提案した。具体的には、Wikipedia の地理情報が、「Geography of ¡国名 or 地域名¿」(例：Geography of Japan, Geography of Ohio) のサブカテゴリに存在することに注目し、そのカテゴリの階層関係から対応するページがどの国のどの地域の情報を表しているかを推定すると共に、「地名のタイプ」に対応する文字列が、カテゴリ中に存在するか否かによって、リンクの発見を行った。

この結果、いくつかのヒューリスティックスを用いることによって、かなり高精度 (97%) に、GeoNames と Wikipedia の間のリンクを発見することが可能となった。

一方、GeoNames に既に存在している Wikipedia のリンクと比較することによって、本手法が、手法自体のエラーを発見するだけでなく、既に、存在している不適切なリンクを発見できることを示した。

この過程において、GeoNames と Wikipedia のリンクの間に、LOD の論文 [1] では、指摘されていない問題に直面した。[1] の研究では、GeoNames と Wikipedia の間で、同一の地名を表しているものに owl:SameAs で関連づけることを提案してる。しかし、以下のような場合に、GeoNames と Wikipedia の対応関係が 1 対 1 とならずに、問題が発生する。

- 複数の地点を設定できる地名
  川や、山脈、都道府県など、広がりをもつ地名は、同一の地名に対して、複数の地点を設定することが可能である。
- 対応する GeoNames の情報が複数存在する地名
  GeoNames では、複数の役割を果たす地名 (例えば、「街」であると共に、「首都」である) があった場合に、GeoNames では、同一名称で、複数のエントリが作成される。この様な場合の多くは、Wikipedia のページとどちらのエントリも対応すると考えられる。
- 複数の地点を含むページ
  Wikipedia のページでは、必要以上に細切れにされたページを作らないために、複数の地名の情報が一つのページに記述される (例えば、山脈のページに山の一覧が含まれる) ことがある。

このようなエントリ間を owl:SameAs でつなぐことは、問題があり [7]、適切な処理を行う必要がある。

3.3 Wikipedia を中心とした LOD の問題点

これまでに議論してきたように、Wikipedia の編集方針は、Wikipedia オントロジーや、それらをつなぐ LOD での利用を考慮したものになっていないため、幾つかの点で不整合が生じている。特に、一番大きな問題は、Wikipedia のページという単位が、表現したいインスタンスやクラスの粒度が一致しないという問題である。

一つの解消の仕方としては、リダイレクトを積極的に使い、複数のインスタンスについて述べているページでは、必ず、個々のインスタンスについて、リダイレクトを作成し、そのリダイレクトをページのセクションなどの特定の場所に対応づけるという方法である。この場合、基本的な Wikipedia の規約は変更せずに、粒度を揃えることができ、やがて、そのセクションがページとして独立した場合でも、その一貫性を保つことができる。ただし、どのようなレベルのインスタンスをリダイレクトとして作成すべきなのかと言った問題が発生したり、手間がかかると行った問題点がある。また、DBPedia などによるデータの抽出が行えないというのも欠点である。

4. まとめ

本研究では、Wikipedia におけるページやカテゴリの作成基準と Wikipedia オントロジーや LOD での利用で想定している状況を比較することにより、その不整合について考察を行った。

これらの不整合の解消については、その対応方法は考えられるものの、実際に行うとなると、それなりの手間がかかることが想定される。

ただ、これらの不整合を無視した形で行っている研究の多くにおいて、それなりに有用な結果が得られていることから、この様な不整合が、どのようなときは影響が少なく、どのようなときには影響が大きいのかということについて、より詳細な考察をしていく必要があると考えている。

参考文献

[1] Bizer, C., Heath, T. and Berners-Lee, T.: Linked Data - The Story So Far, *International Journal on Semantic Web and Information Systems*, Vol. 5, No. 3, pp. 1–22 (2009).

[2] 吉岡真治，劉　亦奇，神門典子：Wikipedia カテゴリを用いた Wikipedia と GeoNames 間のリンク発見とメ ンテナンス，情報処理学会論文誌データベース (TOD)， Vol. 5, No. 3 (2012).
採録決定.

[3] 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.

[4] 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).

[5] Hoffart, J., Suchanek, F., Berberich, K. and Weikum, G.: YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia, *Artificial Intelligence* (2012). (to appear) http://www.mpi-inf.mpg.de/yago-naga/yago/publications/aij.pdf.

[6] 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).

[7] Ding, L., Shinavier, J., Shangguan, Z. and McGuinness, D. L.: SameAs networks and beyond: analyzing deployment status and implications of owl:sameAs in linked data, *Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I*, ISWC'10, Berlin, Heidelberg, Springer-Verlag, pp. 145–160 (2010).