知識工学者のためのWikipediaカテゴリ構造の再整理の構想

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

中川 嵩教
北海道大学工学部
f-b-hawk07@eis.hokudai.ac.jp

1 はじめに

Wikipedia[1] は、ウィキメディア財団が運営している世界最大のインターネット百科事典である。この百科事典の特徴としては、数多くのボランティアエディタにより、相互に内容がチェックされながら、メンテナンスされている点や、インフォボックスやカテゴリなどの容易にメタデータを抽出可能な情報の存在があげられる。また、この様なメタデータは、DBpediaにより、Linked Open Dataで用いられるRDF形式で整理提供されるだけでなく、YAGO2[1] や日本語Wikipediaオントロジ [5] など、カテゴリ情報を活用したデータ構築も行われている。しかし、この様なオントロジ構築において、YAGO2では、Wikipediaカテゴリの階層関係の情報は、一切用いずに、階層構造としては、wordnetを利用し、日本語Wikipediaオントロジでは、特定のパターンにあてはまるカテゴリーサブカテゴリの関係が用いられているのみであり、ウィキペディアのカテゴリ構造を全体として利用していない。これは、ウィキペディアのカテゴリ構造の特殊性を考慮した際に、単純に、そのカテゴリ情報をオントロジで用いる概念階層に結び付けるのが困難であることに起因している。

本研究では、これまでに行ってきたWikipediaカテゴリに関する分析の研究結果を踏まえ [4, 6]、Wikipediaカテゴリの問題について整理すると共に、知識工学的観点からWikipediaカテゴリを再整理する手法について述べる。現在は、整理中のデータについては、Linked Open Dataとして公開予定である。

2 Wikipediaカテゴリの特徴

Wikipediaカテゴリを知識工学の観点から用いる場合の問題点について、これまでの研究 [4, 6] を踏まえて、概略を述べる。

[1]https://wikipedia.org

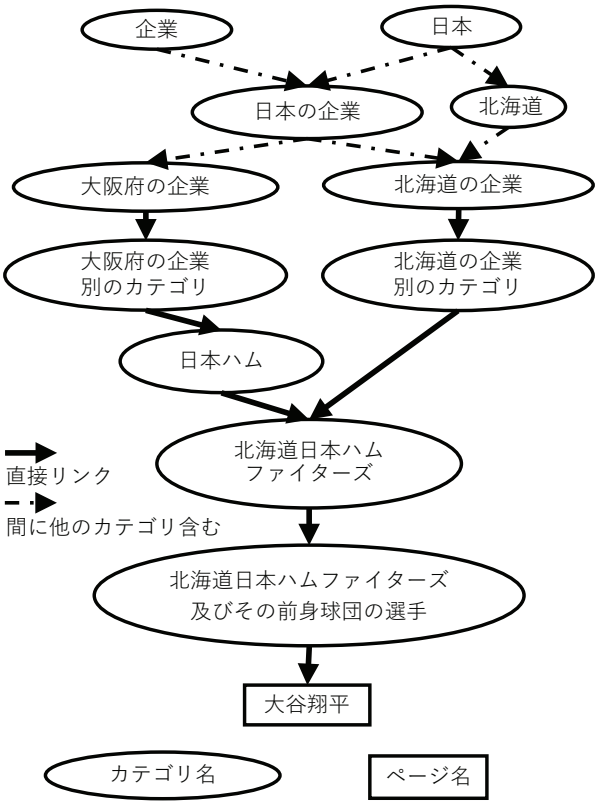

図 1: 実際のカテゴリ階層構造の一部 (抜粋)

Wikipediaのカテゴリとは、本来、下位カテゴリに属するページは、上位カテゴリに属することが想定されるといった知識工学的にも扱い易いものとして、定義され、運用が始まった。ところが、この様なカテゴリ分類では、増え続けるページ群を十分に整理できない (一つのカテゴリに1万以上あるようなページは、とても閲覧不可能である[2])。そのため、この様なカテゴリに対して、分割を行うことが推奨されている。この際の分割の結果生まれるカテゴリが、set-and-topic(以下では、SaTと略記) と分類されるタイプのカテゴリである。

具体的には、「北海道の企業」というカテゴリを

[2]英語版のWikipediaには少ないながら、そのようなページは存在する、例えば、”1991 birts”など

例にとると、「北海道」が topic で、「企業」が set である。また、ほとんどの場合、この SaT のカテゴリは、set と topic を共に、親 (もしくは先祖) カテゴリに持つ。この時、set の親カテゴリと SaT のカテゴリの関係は、本来のカテゴリの包含関係を持つと考えて良い。また topic は、このカテゴリを分割した基準を示すものであり、基準を満しているページなので、topic との関連性が存在すると考えるため、こちらも包含関係があると考えても良い。

ただし、この分割のためのカテゴリが分割ではないカテゴリに到達した際に、問題が生じる。図 1 は、「北海道日本ハムファイターズ」というカテゴリを中心として、関連するカテゴリー階層を抜粋して描いたものである。「北海道日本ハムファイターズ」は、直接、大阪とは関係なく、「大阪の企業」のリンクの下に存在するのは不自然であり、大谷翔平というスポーツ選手が「企業」を親カテゴリに持つというのも不自然である。

同様の問題が、Wikipedia の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になっている。

3 知識工学観点からの Wikipedia カテゴリ構造の再整理

前章で述べたように、既存の Wikipedia カテゴリの構造は、知識工学的観点から、単純に、概念階層のように扱うには問題がある。本研究では、この問題の発生する原因について、次のような仮説で説明できると考えた。

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。

一番簡単な例は、「北海道の企業」～(中略)～「北海道日本ハムファイターズ」～(中略)～「北海道日本ハムファイターズ及びその前身球団の選手」という (SaT → topic1 → SaT) 例で、北海道日本ハムファイターズは、「企業」ではあるが、それをトピックとした SaT のカテゴリは、その set(ここでは「選手」) に依存する。また、同様の問題が、topic と topic の関係 (先ほどの「日本ハム」と「北海道日本ハムファイターズ」の関係) についても発生する可能性がある。

この問題は、トピックに用いられる概念については、SaT カテゴリの分割時にトピックを基準に分割されるような事例をのぞいて、トピック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、トピック間通しの関係でも、地理的な包含関係を含むような場合には、状況に応じて包含関係がなりたつ場合がある。

この仮説を検証するためには、まず、以下の作業が必要となる。

- カテゴリを set や topic といった、構成要素となるカテゴリと SaT のような複合的なカテゴリに分割すること。
- 構成要素となるカテゴリを set と topic に分類すること。
- 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)

その上で、例えば、以下のような検証を行うことによって、本仮説の有用性が検討できると考えている。

- topic のカテゴリについては、上位カテゴリを参照しないと考えた場合のカテゴリから抽出される set の情報と、全てのカテゴリを参照した場合の情報の比較
- set だけで作った階層構造は、概念の抽象具体の階層として検証

4 データ構築

現在、前章で述べた仮説を検証するためのデータを Wikipedia の 2017 年 10 月 20 日のダンプデータをもとに構築中である。データについては、論文投稿時に間に合わなかったため、詳細については、発表時に譲ることとするが、以下のような基準を用いることで、データ構築を行っている。

- 分割が行われているカテゴリの認識
  カテゴリの分割が行われる際には、set,topic から SaT、もしくは、SaT から SaT の形で行われる事がほとんどである。また、SaT から SaT の形の場合には、set と topic の両方が同時に置き換わることは、ほとんどなく、set を共通にして、topic を変更する、もしくは、topic を共通にして set を変更するという事がほとんどである。よって、サブカテゴリに存在する共通文字列を抽出し、その妥当性を検証することで、少なくとも、その共通文字列を含むサブカテゴリが対象カテゴリの分割

のためのカテゴリであり、SaT 型であると考えられる。

- Set と topic の認識
  現在、様々な形でテスト中であるが、形態素解析の結果を利用した一般名詞 (set) と固有名詞 (topic) の同定や、言語間リンクを用いた英語版での表記の確認 (Set は複数形、ただし、複数形の固有名詞などもあるので、単数形なら topic とのみ判定) などを想定している。また、数少ない例外 (日本にしかない職種に対して存在「日本の公務員」→「国会議員政策担当秘書」など) を除けば、SaT の下に、set が来ることはないと考えられるため、そのような情報の活用も検討している。ただ、一度、作業をすれば、同種の作業を繰り替える必要がないことが想定されるため、現時点で機械学習などの枠組を導入する予定はない。

5 議論

上記のような作業の結果、全てのカテゴリが SaT、set、topic に分類されることにより、次のような議論ができるようになる。

- あるカテゴリに属するページに関する set に関する情報を上位カテゴリを含めて、適切に調べることができるようになる。現在のカテゴリでは、この保証がない。

- set に注目した類似度と、分割の基準がどれだけ似ているかという類似度を、各々の階層構造の近さを基準として、独立に計算可能となる。現在のカテゴリ構造を単純に用いると、topic が異なる「世田谷区の建築物」と「狛江市の建築物」は、お互い隣接している地区の建築物であるが、「東京都の建築物」→「東京都区部の建築物」→「世田谷区の建築物」と、「東京都の建築物」→「東京都の建築物 (市町村別)」→「狛江市の建築物」となる。一方、set が異なる「世田谷区の建物」と「世田谷を舞台とした作品」が「世田谷区」→「世田谷区の建物」と「世田谷区」→「世田谷を舞台とした作品」であるため、階層構造的には非常に近いという事になる。どちらの方がより類似していると考えるのかは、応用事例によって異なると考えられるが、今回の提案した set,topic,SaT の分類を考慮することにより、類似性を判断するさいに set の類似度を優先するのか、topic の類似度を優先するのかといったことが、各々の類似度に関する重み付などによって表現可能になると考えられる。

これらを踏まえて、これまでの、単語の類似度をはかるために、Wikipedia のカテゴリ全体の情報を使うという研究 [2, 3] について考える。これらの既存の手法では、Wikipedia のカテゴリは初与のものとして扱い、上記のような影響を低減させるような方法が検討されている。今後の課題としては、これらの方法についても、本研究で提案する Wikipedia カテゴリの整理手法を用いた場合の性能などについて、議論をしていく必要があると考える。

6 おわりに

本研究では、知識工学者が Wikipedia カテゴリをカテゴリ構造全体として使う際の問題点について議論を行うと共に、知識工学者が Wikipedia カテゴリ構造を使うための再整理について議論を行った。現在、データ作成の途上であるが、発表までには、日本語 Wikipedia に対応するデータを公開できるようにする予定である。

参考文献

[1] 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.

[2] Simone Paolo Ponzetto and Michael Strube. Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Int. Res.*, Vol. 30, No. 1, pp. 181–212, October 2007.

[3] Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, and Abdelmajid Ben Hamadou. Computing semantic relatedness using wikipedia features. *Knowledge-Based Systems*, Vol. 50, No. 0, pp. 260 – 278, 2013.

[4] 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.

[5] 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.

[6] 吉岡真治. Wikipedia を中心とした Linked Open Data に関する一考察. 情報処理学会デジタルドキュメント研究会, 2012-IFAT-107, 2012. IFAT-107-1.