人工知能学会研究会資料
SIG-SWO-039-04

Wikipedia Category Consistency Checker (WC3) を用いたWikipediaカテゴリの体系的分析

Systematic Analysis of Wikipedia Category using Wikipedia Category Consistency Checker (WC3)

吉岡真治 [1*]
Masaharu Yoshioka[1]

[1] 北海道大学大学院情報科学研究科
[1] Graduate School of Information Science and Technology, Hokkaido University

Abstract: There are many researches on analyzing the quality of Wikipedia. However, most of the analysis focusing on the quality of text textual contents of the articles and pay little attention to the quality of Wikipedia categories added to the articles. In this paper, we propose a systematic method to analyze Wikipedia categories by using Wikipedia Category Consistency Checker (WC3) that supports to analyze consistency of the category information in Wikipedia by using DBpedia information.

1 はじめに

Wikipediaは、多くのボランティア編集者によって編纂されているWeb上の百科事典である[1]。このWikipediaには、多くの事項に関する記事が存在するだけでなく、階層的に定義されるカテゴリによって、これらの記事が体系的に整理されている。このカテゴリ情報は、ある種の概念階層の情報を含んでいるため、YAGO2[1]や日本語Wikipediaオントロジー[2]におけるクラス階層の定義や、クラス・インスタンスの関係情報の抽出などに用いられている。

このWikipediaの品質に関する議論は多く行われているが、これらは主に、記載されている内容の正確さを中心としたものがほとんどであり[3, 4, 5]、カテゴリ情報に関する体系的な分析は行われていない。この問題に対し、我々は、DBpediaの情報に基づいてカテゴリの付与状況の一貫性を検証するWikipedia Category Consistency Checker (WC3:WC-triple) [2]を提案している[6, 7]。このWC3では、DBpediaに存在する記事に関するメタデータを利用して、同一のカテゴリに属する記事を可能な限り過不足なく検索できるSPARQLクエリを作り、検索結果と比較することにより、カテゴリに属する可能性のある記事の候補や、誤って記事に付与されている可能性のあるカテゴリの情報を提示することが可能である。

本論文では、このWC3を用いてWikipediaのカテゴリ情報を体系的に分析する方法を提案し、具体的な分析事例を紹介する。

2 WC3(WC-triple:Wikipedia Category Consistency Checker)

Wikipedia Category Consistency Checker(WC3:WC-triple) [6]は、Wikipedia中の構造化情報の記述スタイルの一貫性(特に、同一のカテゴリに属する記事群における一貫性)をDBpediaの情報を用いてチェックするシステムである。具体的には、“Songs written by Paul McCartney”のように、クラス(song)とトピック(Paul McCartney)の組合わせで表されるようなカテゴリに対して、DBpediaのメタデータデータベースを参照しながら、そのカテゴリに属する記事をできるだけ過不足無く抽出可能なSPARQLクエリの作成を行い、その検索結果と実際の記事群を比較することにより、メタデータの不足、新しくカテゴリに追加するべき記事候補の発見を行う。

*連絡先： 北海道大学大学院情報科学研究科
〒 064-0806 札幌市北区北 14 条西 9 丁目
E-mail: yoshioka@ist.hokudai.ac.jp

[1] http://en.wikipedia.org/
[2] http://wnews.ist.hokudai.ac.jp/wc3/

本システムは、最初、英語版の Wikipedia を対象として作成されたが [6]、言語依存性がほとんどないことから、日本語版 Wikipedia への拡張も行われている [7]。この日本語版 Wikipedia では、英語版の初期のものと異なり、作成する SPARQL クエリに FILTER 構文を利用することにより、より柔軟なクエリが作成できるだけでなく、サンプルを利用した SPARQL クエリ作成の高速化を実現している。

本研究では、英語版の Wikipedia を解析するにあたり、英語版の WC3 に次小節で述べる日本語版と同様のアルゴリズムを適用したシステムを用いて分析する。

2.1 システムの動作アルゴリズム

WC3 で対象とするカテゴリは、“Songs written by Paul McCartney” や「山下達郎の楽曲」といったように、“Paul McCartney” や「山下達郎」といったトピックを表すキーワードと、“Song” や「楽曲」といったクラスを表すようなキーワードの組合わせで出来たカテゴリを対象に、DBpedia の情報を用いて、そのカテゴリーの情報を表す適切な SPARQL クエリを作成する。適切な SPARQL クエリとは、主にクラス関する制約条件と、トピックに関する制約条件の組み合わせにより作成され、カテゴリに属する記事を可能な限り過不足なく検索できるクエリを意味する。

本システムでは、下記の日本語版用に開発したアルゴリズム [7] を英語版のシステムに適用したものを利用する。

1. カテゴリを入力とし、そのカテゴリに属する記事集合から他の記事へのリダイレクトとなっている記事を除いた集合 Pc を抽出する。
2. 抽出した記事の数が pst を越える場合には、pst 件の記事を抽出し、記事数が pst 以下の場合は、全ての記事を利用する。この記事集合を Pt とする。
   (a) これらのもつ異なり属性からカテゴリに関する属性[3]を除いた全ての異なり属性について、各属性 $(a_1, ..a_n)$ が存在する記事の集合 $PP_1, ..PP_m$ を計算する。この時、$|PP_i| \geq rt \times |Pt|$ を満す属性について、全データベース中でその属性を持つ記事の集合 $PA_1, ..PA_n$ を用いて、精度 $p_i = |PP_i|/|PA_i|$、再現率 $r_i = |PP_i|/|Pt|$、F 値 (精度と再現率の調和平均) を計算する。
   (b) FILTER 構文で用いる文字列としては、トピックやクラスを表すキーワードに限定する。そのため、兄弟カテゴリとの文字列比較を行い、共通部分を除去することで、トピックやクラスを表すキーワードを作成する (例：「山下達郎の楽曲」について、兄弟カテゴリ「槇原敬之の楽曲」などを用いると、「山下達郎」を抽出し、「山下達郎のアルバム」を用いると、「楽曲」を抽出する。英語版についても同様に、“Songs written by Paul McCartney” について、兄弟カテゴリである “Songs written by Bob Dylan” などを用いて、“Paul McCartney” を抽出する)。先の手順の属性において、そのトリプルの目的語にこれらの文字列を含む場合は、上記の属性の目的語部分を変数化して、FILTER 構文と組み合わせたものを制約の候補として追加する。これらの候補についても、F 値を計算する。
3. 属性、もしくは、目的語を変数化して FILTER 構文と組み合わせたもののうち、F 値の大きな方から上位 10 件を組合わせのために用いるクエリの要素の候補とする。
4. 3 で求めた属性や FILTER 構文の結果では精度と再現率のバランスを考慮するために、クラスを表すような一般的な属性が候補に含まれない可能性がある。そのため、利用する記事のうち、述語として、http://www.w3.org/1999/02/22-rdf-syntax-ns#type を持つものに限定し、その再現率を計算する。再現率が trt(type recall threshold) 以上のものに限定し、F 値の高いもの 2 件を組み合わせの候補とする。
5. 3 で作成した候補集合と、クラスの候補となる属性を組み合わせて候補となる SPARQL クエリを作成する。クラスの候補となる属性が存在しない場合には、3 で作成した候補集合を単独で利用する。これらの候補について、同様に、精度、再現率、F 値を計算し、F 値の最も高いものをクエリの候補とする。

このようにして作成したクエリを満たす記事の集合と、対応するカテゴリの記事について、比較することにより、以下の 3 種類の記事の情報を収集する。

Found クエリにより見つけられたカテゴリの記事

NotFound クエリにより見つけられなかったカテゴリの記事
DBpedia により抽出されたメタデータが不足している、あるいは、間違っている記事

[3]カテゴリに関する参照を行っている属性と、カテゴリを情報源として作成されている Yago の情報については、候補から除外している。

Error クエリにより見つけられたがカテゴリに属さない記事
適切なクエリが作成されていない場合には、その問題を分析するための情報となり、適切なクエリが作成されている場合には、カテゴリに追加する候補として考えられる記事

3 WC3による英語版Wikipediaのカテゴリの体系的分析

3.1 体系的分析の方法

本研究では、前節で紹介したWC3を用いてWikipediaのカテゴリの付与状況についての体系的分析を行う方法を提案する。

一般の百科事典において、Wikipediaのカテゴリに相当するようなインデックスを作成する場合を考えると、インデックスの適切性 (記事に対してインデックスを付与することが適切であるかどうか) だけでなく、網羅性 (インデックスを付与すべき記事に全てインデックスを付与しているかどうか) についても考慮されることが求められる。しかし、ボランティア編集者が主体のWikipediaでは、多くの新しい記事が継続的に追加されること、これらの記事を追加する編集者が全てのカテゴリについての付与基準などを把握しているわけではないことなどから、インデックスの網羅性に関する議論などは行われていない。

この様な問題に対して、WC3はWikipediaカテゴリに対応づけられる形で作成されたSPARQLの検索式を満す記事群を検索することができ、この検索式がWikipediaカテゴリの説明文と比較して妥当である場合においては、WC3を用いることにより、Wikipediaカテゴリが、対象となる記事群に対して、どれくらい網羅的に付与されているかを議論することが可能となる。

3.2 分析事例

本研究では、2015年版のDBpediaの情報に基づき、前節で述べたアルゴリズムを用いて分析を行うWC3を利用し分析を行う。このシステムでは、パラメータとして $pst = 50$,$rt = 0.2$, $trt = 0.6$ を利用している。

本研究では、分析の具体事例として、生年に関するカテゴリである “1970 births” と、出身地に関するカテゴリ “People from Tokyo” を例にとり、分析方法について述べる。

利用するWC3のシステムとしては、上記の改良したアルゴリズムに基づく英語版WC3を2015年版のDBpediaのデータを利用して作成した。

まず、最初に、生年に関するカテゴリとして、“1970 births” を対象として分析を行った。この時、作成されたSPARQLクエリを、以下に示す。このカテゴリに関する自然言語による説明は “People born in 1970” であり、妥当なクエリが作成されたと考えている[4]。

```
SELECT DISTINCT ?s
WHERE {
?s rdf:type foaf:Person .
?s dbo:birthDate ?o1 .
FILTER regex (?o1, ‘‘1970’’)
}
MINUS { ?s dbo:wikiPageRedirects ?o . }}
```

“1970 births” のカテゴリの付与された記事は、10,271記事であり、そのうち、Foundに分類されたものが9,963記事 (NotFoundが、308記事)、Errorに分類されたものが、392記事となった。このSPARQLクエリの精度、再現率、F値は、それぞれ0.96,0.97,0.97であった。

もし、適切なSPARQLを満す全ての記事をカテゴリを付与すべき記事だと考えるのであれば、WC3の結果を分析する際に重要となる評価項目は、先ほどの指標の中の精度が、網羅性を示す指標となる。

このErrorの内容について分析をすると、いくつかのパターンがあることが確認された。

- 複数の人間に対応する記事の存在
  “List of Playmates of 1989” の様に、記事自体が複数の人間に対応する場合に、記事に対するカテゴリとして、1970年生れの人物に関する記載はあるものの、“1970 births” のカテゴリが付与されていない。

- 複数のdbo:birthDateが登録されている記事の存在
  一人の人間に対して、複数のdbo:birthDateが登録されている記事が存在し、対応するデータが間違っている場合にErrorと判定される (“Loren Bouchard” “Ben Chaplin” など)。多くの場合、適切なdbo:birthDateのデータに対しては、適切なカテゴリが付与されている。

- 対応するカテゴリが存在しない。
  適切なカテゴリが存在しない記事で、例としては、birthsに関連するカテゴリが全く付与されていない “Takeharu Ishimoto” やbirthsに関連するより抽象的なカテゴリ (“1970s births”) が付与されている “Eileen Catterson” などがある。

[4]本論文では、“http://dbpedia.org/ontology”, “http://dbpedia.org/property”, “http://www.w3.org/1999/02/22-rdf-syntax-ns#type”, “http://xmlns.com/foaf/0.1” “http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#” について、“dbo”,“dbp”,“rdf”, “foaf”, “dul” で参照する

今回の分析対象とした生年のように、本来、一つの記事で表わされる個人に対しては、一つの属性しか持たないといったような制約を導入して分析を進めるとより詳細な分析が可能となると考えられる。

また、“1901 births” から “2000 births” までの 100 のカテゴリについて分析を行った場合の精度、再現率、F 値の平均は、それぞれ 0.97,0.96,0.97 であった。精度は十分に高く、生年のような比較的理解しやすいカテゴリについては、かなり網羅的にカテゴリ情報が付与されていることが確認された。

次に、出身地に関するカテゴリとして、“People from Tokyo” を対象として分析を行った。この時、作成された SPARQL クエリを、以下に示す。このカテゴリに関する自然言語による説明は “This category list people who were born in or resident of Tokyo, Japan” であり、people という意味からすると、抽象度の高いクラスが選択されている。

```
SELECT DISTINCT ?s
WHERE {
?s rdf:type dul:Agent
?s dbp:placeOfBirth ?o1 .
FILTER regex (?o1, ‘‘Tokyo’’)
MINUS { ?s dbo:wikiPageRedirects ?o . }}
```

これは、このクエリと、タイプの部分を生年の時と同じように、「`?s rdf:type foaf:Person .`」と置き換えたものの結果が同一であるために、起きた問題であるが、結果が全く同一であるため、WC3 による分析結果としては、東京生れの人間を検索した場合と同等と考えられる。

“People from Tokyo” のカテゴリの付与された記事は、1,402 記事であり、そのうち、Found に分類されたものが 827 記事 (NotFound が、575 記事)、Error に分類されたものが、1,283 記事となり、この SPARQL クエリの精度、再現率、F 値は、それぞれ 0.39,0.59,0.47 であった。再現率が低いのは、生成されたクエリが、“This category list people who were born in Tokyo, Japan” に対応し、residents に対応する項目を持っていないところに問題がある。現状の WC3 では、この様な複数の条件の or で組合わせで表現されるような制約に対応できないという問題点が確認された。

一方、精度に関する分析についてであるが、この SPARQL クエリで発見された Error の中には、“Ayako Sono” のように、“Writers from Tokyo” という “People from Tokyo” のサブカテゴリが付与されているものが存在し、Error として分類することが不適切であるものが含まれていた。そこで、Error の中に含まれるサブカテゴリに属する記事については、Error for Children Category というグループと、その残りに分類する事とした。この場合に、Error for Children Category に属する記事は、562 記事であり、残りの Error は 721 記事となった。この Error for Children Category に属する記事を正解と考えた場合の精度は、0.53 となった。

この Error の内容について分析をすると、いくつかのパターンがあることが確認された。

- 出身地に関するカテゴリを含まない記事
  “Takada Akemi” のように、出身地に関するカテゴリを全く含まない記事が多く存在する。
- 異なる出身地を含む記事
  “DJ Krush” のように、 “Musicians from Ibaraki Prefecture” といった異なる出身地が記述されている記事が存在した。これは、出身地のカテゴリを付与する編集者は、付与可能なカテゴリが複数存在する (出生地、育ったところ、有名になった時点で活動していたところ) 場合に、全てを付与するのではなく、主観的に選んだ場所に関するカテゴリを付与している可能性が高いと考えられる。

この様に、Wikipedia のカテゴリには、本来の定義とは別に、編集者の主観によって、カテゴリの情報が付与されたり付与されなかったりするといった状況が存在することが確認された。

3.3 分析結果の考察と今後の課題

上記の事例から、Wikipedia のカテゴリには、生年に関するカテゴリのように、かなり網羅的に付与されているカテゴリから、出身地のようにカテゴリ付与の基準に曖昧性があったり、そもそも網羅的なカテゴリ付与が行われていない記事が多いようなカテゴリが存在することが確認された。

この様な情報を Wikipedia のカテゴリ情報をメンテナンスしているボランティア編集者にフィードバックすることにより、より網羅性の高い一貫したインデックスとして利用可能な Wikipedia カテゴリを記事に付与することができるようになると考えている。

一方、WC3 で生成する SPARQL クエリが、必ずしも適切なクエリを生成できない場合があることも確認された。この問題への対応方法としては、全て自動で行うのではなく、分析に成功したクエリを保存する枠組を作る方法を考えている。このようなクエリは、同じカテゴリを分析するときに有用なだけでなく、兄弟カテゴリの分析時においても、トピックに相当する部分を書き換えるだけで、利用可能となる。また、この様な SPARQL クエリを自然言語によるクエリと対応づける形で保存することにより、より、曖昧性の少ないカテゴリの定義に関する議論が可能となると考えている。

4 おわりに

本論文では、Wikipedia のカテゴリ情報の付与がどれくらい網羅的に行われているかといった体系的に行うために WC3 を利用する方法を提案した。また、具体的な分析事例を通して、Wikipedia のカテゴリ情報には、生年のような曖昧性もあまりなく、網羅的に付与されているようなカテゴリだけでなく、出身地のようなカテゴリ付与の基準が曖昧なカテゴリも存在することが確認された。今後は、SPARQL クエリを共有し再利用を行う枠組を構築することなどによって、より適切なクエリを容易に利用できる環境を作り、ボランティア編集者に使ってもらえるツールとして提供していくことを考えている。

謝辞

また、本研究の一部は、科研費基盤研究 (B) 25280035 により行われた。ここに記して、謝意をあらわす。

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] 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.

[3] Jim Giles. Internet encyclopaedias go head to head. *Nature*, Vol. 438, pp. 900–901, 2005.

[4] Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 12, pp. 1720–1733, 2007.

[5] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. Measuring article quality in Wikipedia: Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pp. 243–252, New York, NY, USA, 2007. ACM.

[6] Masaharu Yoshioka and Rhett Loban. WC3: Wikipedia Category Consistency Checker based on DBPedia. In *Proceedings of 11th International Conference on Signal-Image Technology & Internet-Based Systems*, pp. 712–718, 2015.

[7] 吉岡真治. 日本語版 wc3(wikipedia category consistency checker) - 日本語版 wikipedia のカテゴリに所属するページのメタデータの一貫性の分析 -. 人工知能学会第 25 回セマンティックウェブとオントロジー研究会, 2015. SIG-SWO-037-04.