1G4-2

DBPediaの情報に基づくWikipediaのカテゴリ情報の一貫性の分析

Consistency Analysis of Wikipedia Category based on DBPedia information

吉岡 真治[*1] Rhett Loban [*2]
Masaharu YOSHIOKA Rhett Loban

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

[*2]Queensland University of Technology
Queensland University of Technology

Wikipedia is a free encyclopedia on the Internet that is maintained by large numbers of voluntary editors. There are several researches that analyzes quality of the textual contents in the Wikipedia, but there is no specific analysis on the quality of structured data (e.g., category structure, attributes in infobox) in Wikipedia. DBPedia is a database that extracts structured information from the Wikipedia and widely used as a core part of Linked Open Data. In this paper, we propose a system WC3 (Wikipedia Category Consistency Checker) that supports to evaluate consistency of the category information in Wikipedia by using DBPedia information.

1. はじめに

Web 上の百科事典である Wikipedia[*1]には、多種多様な事象に関するページが存在している。このページの多くには、Infobox と呼ばれるページの内容のタイプに特有の構造化情報を表示する部分や、分類を表すためのカテゴリの情報などが付与されている。DBPedia[Bizer 09b] は、各々のページからこの構造化された情報を抜出し、大規模な事象に関する構造化情報のデータベースを構築している。また、この DBPedia は、Linked Open Data[Bizer 09a] の中心として、様々なデータと関連づけられて利用されている。

この DBPedia の情報の品質は、Wikipedia の記述に依存するが、その記述の品質については、編集者に依存する。この Wikipedia の記述に関する信頼性については、その記述内容についての分析 [Giles 05, Stvilia 07] や、ページの編集に携わった人々に関する属性を用いた分析 [Hu 07] などが行われている。また、DBPedia の情報については、他の Linked Open Data などと比較した分析なども行われている [Kittur 08, Mendes 12]。この他にも、DBPedia の情報の信頼性を検証するために、Wikipedia の情報元を分析する手法 [Orlandi 11] などが提案されている。しかし、DBPedia の情報を用いて、Wikipedia の情報を分析し、Wikipedia の品質向上につなげようという研究はほとんど行われていない。本研究では、DBPedia により抽出された構造化情報を用いて、Wikipedia のカテゴリ構造の一貫性を分析する方法を提案するとともに、その分析結果を紹介する。また、この考え方を用いたカテゴリ情報の一貫性を検証するためのツールについても紹介する。

2. Wikipedia と DBPedia

2.1 Wikipedia の構造化情報と DBPedia

Wikipedia のページには、特定の項目に関する説明が、自然言語の文書により記述されるだけでなく、関連するページをまとめて扱うためのカテゴリや、ページ間のリンクなどを用いることで、他の項目との関係が記述される。また、特定の属性 (DBPedia で付与されるメタデータ) を持つことが期待されるようなカテゴリ (例えば、大学、小説、映画など) については、それらの属性を整理して表示するためのテンプレートが利用され、Infobox という形で右上に表示される。

DBPedia[Bizer 09b] は、Wikipedia から、この構造化された情報や、ページ間の関係の情報を抽出し、整理することにより、様々な項目に対する構造化情報のデータベースを構築している。DBPedia では、全ての Wikipedia のページから、機械的にデータベースを構築しており、世の中の様々な事象について網羅する大規模なデータベースとなっている。しかし、Wikipedia の記述に一貫性がないと、DBPedia における記述も一貫しないという問題がある。

2.2 Wikipedia のカテゴリ情報

Wikipedia のカテゴリ情報は、主に、ページの閲覧性の向上を目的として、類似した内容を含むページを、その内容を表す名前を持つカテゴリことを目的として付与されている。このカテゴリは、ある種の包含関係を考慮した親子関係により、束上の階層関係を構成している。このカテゴリには、「日本」、「ポール・マッカートニー」といったトピックを表すようなカテゴリ、「作家」、「歌」などのクラスを表すようなカテゴリ、「日本の作家」などのトピックとクラスの組み合わせによりあらわされるカテゴリが存在する。特に、クラスに関連するカテゴリ (例えば、「作家」) については、一つのカテゴリに、あまりに多くのページが属する場合に、このカテゴリを分割した形であるトピックとクラスのカテゴリ (例えば、「日本の作家」「英国の作家」) を作ることが推奨されており、多くのページを持つようなクラスのカテゴリに関しては、様々なトピックとクラスの組合わせのカテゴリが作られている。

日本語版の Wikipedia では、このトピックとクラスの組み合わせと考えられるようなカテゴリは、「名詞 (日本) ＋の＋名詞 (作家)」の形で記述されることが多く、このカテゴリがの約 53%であり [吉岡 14]、これらのカテゴリでこのようなカテゴリの多くは、トピックやクラスを表すようなカテゴリを親もしくは先祖 (親の親、親の親の親など) に持つこととなる。英語版の Wikipedia については、明示的に、この様な二つのカテ

連絡先: 吉岡真治, 北海道大学大学院情報科学研究科, 札幌市北区北 14 条西 9 丁目,011-706-7107, yoshioka@ist.hokudai.ac.jp

本研究の一部は、Rhett Loban の北大でのインターン滞在中に行われた。

*1 http://en.wikipedia.org/

ゴリに組合わせによる新しいカテゴリの作り方についての議論がなされている[*2]。

一方、この様なカテゴリをチェックするためのツールについては、特定のテンプレートを使っているページや、特定のWikipedia のページへのリンクを持つようなページを探すためのツールである CatScan[*3]などが存在するが、基本的には、Wikipedia の検索システムである。そのため、カテゴリの一貫性の管理は、主に、カテゴリに関係する編集者の努力によって行われている状態である。

3. DBPedia を用いた Wikipedia のカテゴリ情報の分析

3.1 対象とする Wikipedia のカテゴリと分析手法

2. 節で述べたように、同一のカテゴリに属するページにおいては、共通する内容が存在することが期待される。特に、トピックとクラスの組み合わせであらわされるようなカテゴリ(例えば、「ポール・マッカートニーが書いた歌」) のページからは、トピックやクラスに関係する共通する属性 (例えば、「作者がポール・マッカートニー」や「歌」) が DBPedia のデータベースに存在することが期待される。

本研究では、この考えに基づき、トピックとクラスであらわされるようなカテゴリについて、DBPedia のデータを用いることにより、その一貫性を検討する方法を提案する。具体的には、カテゴリに属するページが共通して持つ属性を用いて、カテゴリに属するページをできる限り過不足なく検索できる SPARQL のクエリの作成を行う。このような、SPARQL のクエリにより、カテゴリに属するページを過不足なく検索できる場合には、一貫した構造化情報が各ページに存在することが確認できる。

一方、完全に過不足ないクエリが作れない場合には、クエリの妥当性について検討するとともに、クエリを満たすがカテゴリに対応しないページや、カテゴリに属するがクエリを満たさないページについての分析が必要である。前者のページについては、本来、カテゴリに属すると判断してよいページに、適切なカテゴリが付与されていない可能性があり、後者のページについては、DBPedia が抽出可能な適切な構造化情報が存在しない可能性がある。これらの情報は、カテゴリ付与の一貫性を検証するための有用な情報となると考えている。

3.2 予備実験

本手法の妥当性を検証するために、具体的なカテゴリについて、DBPedia の 2014 年版のデータに基づく endpoint[*4] を利用し、SPARQL のクエリを手作業で作成し、本手法の妥当性の検証を行った。この結果、次のような問題があることが判明した。

- リダイレクトのページ
  Wikipedia のカテゴリ情報は、他のページのリダイレクトの役割を果たすページについても、付与可能である。例えば、「Songs_written_by_Paul_McCartney」には、525 ページが属している。しかし、その多く (394 ページ) は、曲を表すページではなく、その曲を含むアルバムへのリダイレクトのページであった。この様なリダイレクトのページについては、Infobox のような構造化情報が存在しない。結果として、DBPedia では、これらのページから、カテゴリに関する情報は抽出しているが、Infobox に存在するような構造化情報を持っていない。よって、これらのページについては、妥当性を検証するための十分な情報がページ自体に存在しないため、今回の分析対象から外すこととした。

- まとめや関連情報のページ
  今回の SPARQL のクエリで分析の対象とするトピックとクラスの組み合わせで表現されるようなカテゴリの多くは、主に、個物 (インスタンス：本、歌、人など) の分類を行うカテゴリが多くを占めていると考え、全てのページに共通する属性が存在すると考えた。しかし、いくつかのカテゴリ (例えば、「Presidents_of_the_United_States」) には、そのカテゴリに含まれる個物のリストを記述したページ「List_of_Presidents_of_the_United_States」や、その関連情報「U.S._presidents_on_U.S._postage_stamps」が存在する。これらのページは、SPARQL クエリを満たさないが、不適切なページではない。よって、このようなページが存在することを考慮した SPARQL クエリの構築を行うことが必要である。

3.3 カテゴリを表す SPARQL クエリの自動構築

予備実験の結果を踏まえ、以下のような手順で、カテゴリを表す SPARQL クエリを作成するシステムを構築した。

1. カテゴリを入力とし、そのカテゴリに属するページ集合から他のページへのリダイレクトとなっているページを除いた集合 Pc を抽出する。

2. 抽出した全てのページがもつ異なり属性からカテゴリに関する属性[*5]を除いた全ての異なり概念について、各属性 $(a_1, a_2, ...a_n)$ が存在するページの集合 $PP_1, PP_2, ...PP_n$、全データベース中でその属性を持つページの集合 $PA_1, PA_2, ...PA_n$ を用いて、精度 $p_i = |PP_i|/|PA_i|$、再現率 $r_i = |PP_i|/|Pc|$、F 値 (精度と再現率の調和平均) を計算する。

3. F 値の最も高い属性をクエリの候補とするとともに、上位 10 件を組み合わせのために用いる属性の候補とする。

4. 3 で求めた属性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、網羅性を考慮した属性の候補を、各親カテゴリについて次のような手順で作成し、組合わせ属性の候補として追加した。

   (a) 各親カテゴリについて、共通の親を持つ兄弟カテゴリ (例えば、「Songs_written_by_Paul_McCartney」の親カテゴリ「Songs_by_songwriter」に関する兄弟カテゴリ「Songs_written_by_Bob_Dylan」など) を 5 つランダムに抽出し (兄弟カテゴリが 5 以下の場合は全てを利用)、ページの集合を作成する。

   (b) 2 の手順と同様に、このページ集合が持つすべての異なり属性について、精度、再現率、F 値を計算する。この時、網羅性を考慮して、再現率が 0.9 以上 (間違いや、個物を表さないページなどがある場合

∗2 http://en.wikipedia.org/wiki/Wikipedia:Category_intersection

∗3 http://en.wikipedia.org/wiki/Wikipedia:CatScan 2015 年 3 月 18 日現在では、アクセス不可能

∗4 http://dbpedia.org/sparql

∗5 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。

を考慮して、多少の検出漏れは許容する) の属性のうちで、F 値の高いもの 2 件を組み合わせの候補として追加する。

5. 候補となった属性を 2 つ組み合わせたクエリを作成し、同様に、精度、再現率、F 値を計算し、F 値の上位をクエリの候補とする。ただし、RDF のトリプルで表されている属性のうち、対象を共有するものについては、主に、同一のトピックに関する属性の組み合わせになるため、組み合わせの候補から除外している。

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

Found クエリにより見つけられたカテゴリのページ

NotFound クエリにより見つけられなかったカテゴリのページ
二つの属性の組み合わせのクエリの場合には、不足している属性の情報を合わせて示す。

Error クエリにより見つけられたがカテゴリに属さないページ
このようなエラーページを排除するためのクエリを作成するための情報として、複数の Error ページに共通し、カテゴリに属するページには、ほとんど含まれない属性の情報を示す。

3.4 カテゴリ分析システムの構築

3.3 節で述べた方法で、カテゴリを表す SPARQL を構築し、作成したクエリに基づいて Wikipedia のカテゴリの情報を分析するシステム WC3(WC-triple:Wikipedia Category Consistency Checker) を作成した。本システムでは、基盤となるデータベースとして、DBPedia の 2014 年版のデータを用いて作成された virtuoso データベース[*6] と Wikipedia の 2014 年 11 月 6 日版のダンプデータを用いて構築した。図 1 に、WC3 の実行例を示す。本システムでは、カテゴリ名を入力することにより、前節で述べた一つの属性によるクエリと二つの属性の組み合わせのクエリの各々について、もっとも F 値の高いものについてを表示するとともに、そのクエリに対する Found, NotFound, Error の情報を表示する。

図 1 は、「Songs_written_by_Paul_McCartney」(リダイレクトを除くページ数：131) に対する結果であり、構築された SPARQL クエリは、以下に示すように、作者が Paul McCartney で音楽作品であるというものとなった。

```
SELECT ?s
WHERE {?s http://dbpedia.org/ontology/writer
http://dbpedia.org/resource/Paul_McCartney .
?s http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://dbpedia.org/ontology/MusicalWork .
MINUS { ?s
<http://dbpedia.org/ontology/wikiPageRedirects>
 ?o . }}
```

131 ページ中の 121 ページがこのクエリを満し (Found:121 ページ、NotFound:10 ページ)、クエリを満すが、カテゴリに属さないページが 10 ページ (Error:10 ページ) 存在した[*7]。

NotFound のうち、全てのページは、”?s http:dbpedia.orgontologywriter”属性を持たないものであり、類似のページと比較して、Infobox 中に Writer(s) に相当する記述がない事が確認された。こちらについては、他のページと同じように、Infobox 中に Writer(s) の記述を与えることが一貫性を向上させるためには望ましいと考えられる。

また、Error のページについて分析すると、「I_Lost_My_Little_Girl」のように、文面から、Paul McCartney が作ったことが分かるが、カテゴリのラベルが付与されていないようなものが見つかった。一方で、「Goodbye_(Mary_Hopkin_song)」のように、「Songs_written_by_Paul_McCartney」のサブカテゴリである「Songs_written_by_Lennon-McCartney」のラベルを持っているページが存在したため、間違いとは言えないとも考えられる。ただし、ほとんどの「Songs_written_by_Lennon-McCartney」に属するページでは、Writer として、Lennon-McCartney の名前が用いられており、全体の一貫性という観点からは、「Goodbye_(Mary_Hopkin_song)」については、「Songs_written_by_Paul_McCartney」のラベルを与えるよりも、Writer(s) の内容を修正する方が適切であると考えられる。

3.5 考察

「Songs_written_by_Paul_McCartney」に対する分析結果が示すように、WC3 を用いることにより、Wikipedia のカテゴリ構造の一貫性を分析するために有用な情報が得られることが確認された。本システムをいくつかのカテゴリに適応したところ、トピック部分に包含関係を含む (上記の「Lennon_McCartney」の例や、地理的な包含関係を考慮する必要がある例) 場合には、必ずしも適切な評価が行えなかった場合があるが、人や組織が作成した作品などの情報については、有用な分析を行うことができる事を確認した。

本システムは、http://wnews.ist.hokudai.ac.jp/wc3 にて公開予定である。この様なツールが Wikipedia の編集者によって用いられることは、DBPedia で用いるデータの一貫性の向上にも貢献することが期待される。

4. おわりに

本研究では、主に、クラスとトピックの組合わせで表されるような Wikipedia のカテゴリを対象にして、DBPedia の情報を用いて分析する方法を提案した。具体的には、特定の Wikipedia のカテゴリに属するページ集合を可能な限り過不足なく見つけることができるような DBPedia の情報を用いたクエリを作成し、その検索結果と実際のページ集合を比較する事によって、構造か情報の記述の一貫性の観点から問題のあるページを見つけるためのシステムを提案した。本システムを用いることにより、DBPedia の活動が Wikipedia の品質向上に貢献し、結果として、DBPedia で利用するデータの品質の向上につながるという正のフィードバックを実現することが期待される。

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。また、システムのデータ作成には、博士課程学生の Dieb Thaer 君に協力していただいた。ここに記して、謝意をあらわす。

参考文献

[Bizer 09a] Bizer, C., Heath, T., and Berners-Lee, T.: Linked Data - The Story So Far, *International Journal on*

*6 https://joernhees.de/blog/2014/11/10/setting-up-a-local-dbpedia-2014-mirror-with-virtuoso-7-1-0/

*7 全ての Wikipedia のページについては、2015 年 3 月 18 日にアクセスして確認

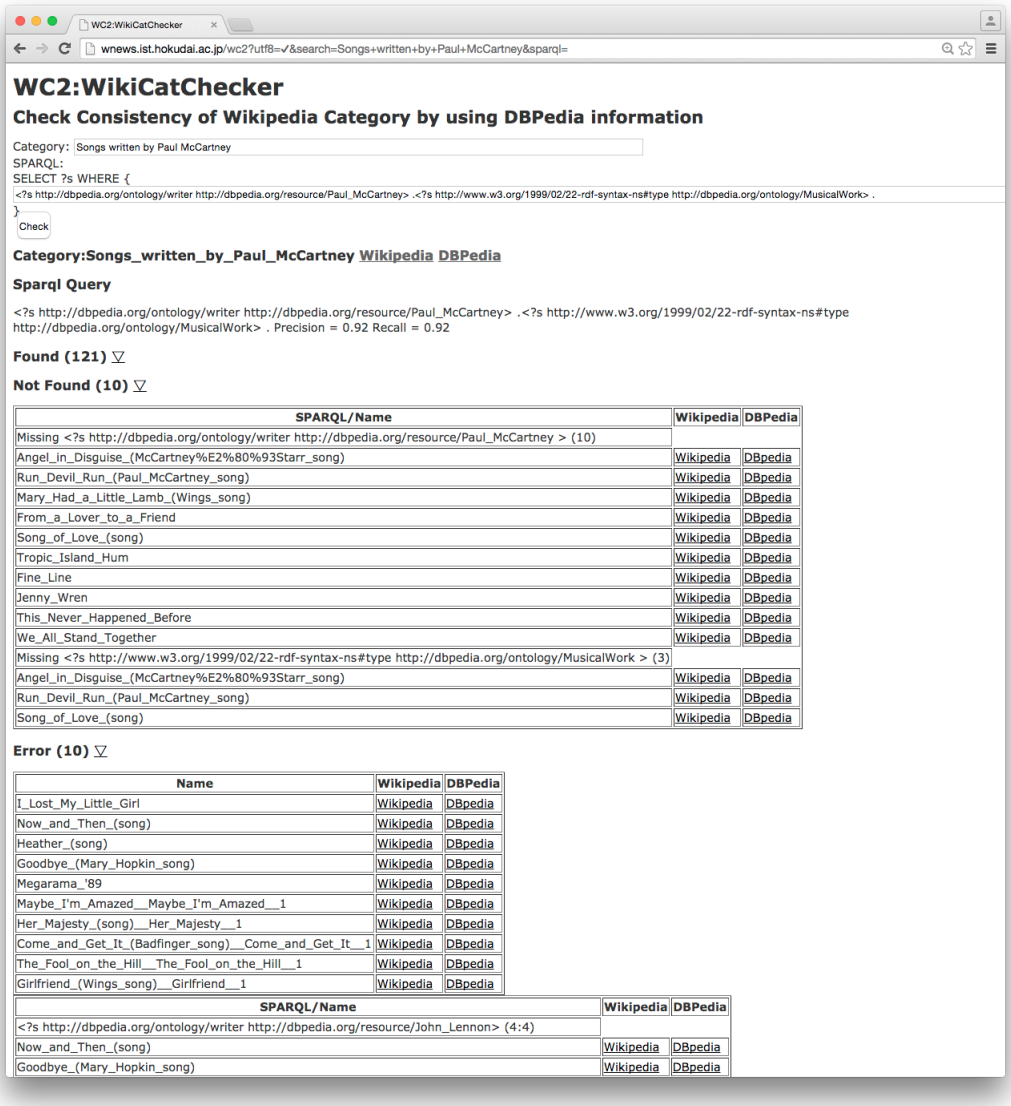

図 1: WC3(Wikipedia Category Consistency Checker) の実行例

Semantic Web and Information Systems, Vol. 5, No. 3, pp. 1–22 (2009)

[Bizer 09b] 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)

[Giles 05] Giles, J.: Internet encyclopaedias go head to head, *Nature*, Vol. 438, (2005)

[Hu 07] Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q.: 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

[Kittur 08] Kittur, A. and Kraut, R. E.: Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination, in *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pp. 37–46, New York, NY, USA (2008), ACM

[Mendes 12] Mendes, P. N., Mühleisen, H., and Bizer, C.: Sieve: Linked Data Quality Assessment and Fusion, in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pp. 116–123, New York, NY, USA (2012), ACM

[Orlandi 11] Orlandi, F. and Passant, A.: Modelling provenance of {DBpedia} resources using Wikipedia contributions, *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 9, No. 2, pp. 149 – 164 (2011), Provenance in the Semantic Web

[Stvilia 07] Stvilia, B., Gasser, L., Twidale, M. B., and Smith, L. C.: A framework for information quality assessment, *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 12, pp. 1720–1733 (2007)

[吉岡 14] 吉岡 真治：Wikipedia のカテゴリー階層関係の分類を用いた日本語 Wikipedia オントロジーの分析, 2014 年度人工知能学会全国大会 (第 28 回) 論文集 (2014), CD-ROM 2J3-4