意外性のある知識発見のための Wikipedia カテゴリ間の関係分析

The Analysis of Wikipedia categories for detecting unexpected knowledge

野田陽平 [1]　清田陽司 [2]　中川裕志 [2]

Yohei Noda[1], Yoji Kiyota[2], Hiroshi Nakagawa[2]

[1] 東京大学大学院学際情報学府

[1]Graduate School of Interdisciplinary Information Studies, University of Tokyo

[2] 東京大学情報基盤センター

[2] Information Technology Center, University of Tokyo

Abstract: Articles in Wikipedia are classified from a lot of standpoints by the category system of Wikipedia. Using this property, we can detect unexpected information from Wikipedia articles. For example, the article “Taro Aso” belongs not only to the category “Prime Ministers of Japan”, but also to the category “Olympic shooters of Japan”. In this study, we focus on such relations of categories, and processed the graph network of Wikipedia categories statistically. Finally, we detected unexpected information by using the results of statistical processing.

1. はじめに

Wikipedia は，誰でも編集が可能な巨大なウェブ百科事典である．英語版 Wikipedia は 2008 年 8 月 11 日に 250 万記事，日本語版 Wikipedia は 2008 年 6 月 25 日に 50 万項目を超え，膨大な情報量を誇る百科事典として広く認知されている．

日本語版 Wikipedia は 9 個の主要カテゴリ[1]の下に，サブカテゴリ，記事が関連付けられており，大規模なグラフ構造を成している．各項目はそれぞれ複数の親カテゴリを持っており，また，同義語はリダイレクトとして関係付けられている．

Wikipedia の記事は，カテゴリシステムによってさまざまな観点からの分類がなされている．この特徴をうまく用いると，個別の記事からだけでは得られない意外な知識の発見につなげることができる．例えば，「麻生太郎」は「日本の内閣総理大臣」というカテゴリに属しているが，一方で「オリンピック射撃競技日本代表選手」というカテゴリにも属している．本研究では，このような意外な知識を Wikipedia から大量に発掘することを目的に，Wikipedia カテゴリネットワークに関する統計処理を行い，その結果を分析した．

本稿の章構成を説明する．２章では関連研究について述べる．次に，３章では Wikipedia カテゴリネットワークの統計処理の目的と手法について説明する．４章では３章で説明した統計処理の結果を提示し，分析する．５章では本稿のまとめを行う．

[1]主要 9 カテゴリ：学問, 技術, 自然, 社会, 地理, 人間, 文化, 歴史, 総記

2. 関連研究

Wikipedia のカテゴリを利用した研究には，オントロジーを構築する研究がある．DBPedia[1]は，Wikipedia の Infobox やカテゴリ情報などを RDF に変換し，データベースから情報を検索できるようにしている．YAGO[2]は，Wikipedia のカテゴリ関係を利用して，WordNet を拡張している．また，清田らは，Wikipedia カテゴリ間の関係を図書館の分類体系と対応付けて，両者が持つ利点を生かし，情報探索のインフラとして活用している[3].

Nguyen らは，Wikipedia の記事を分析することで，Wikipedia の記事ごとの関係性を抽出している[4]. また，Strube らは，密な Wikipedia の各記事が属するカテゴリの情報を利用し，各概念同士の関連の度合いを算出している[5]．また，これに対し中山らは概念ごとの関連の度合いだけでなく，その概念同士がどのような関係性にあるのかなどの意味関係を定義したオントロジの構築法を提案している[6].

本研究では，人手により日々更新され密になっている Wikipedia のカテゴリ関係を用いることで，Wikipedia がもつフォークソノミーの利点を生かし，意外で価値のある情報の発見を目指す．

3.Wikipedia カテゴリネットワークに関する統計処理

3.1. 統計処理の目的

本研究の目的は，Wikipedia のカテゴリネットワークの情報を用い，意外性のある情報を発見することである．ここでは，意外性のある情報を含んだ項目を下記のように定義した．

まず，カテゴリ C に項目 Z が含まれることを，$Z \in C$ と書く．カテゴリC_i, C_jが存在したとき，共通の子項目を持つカテゴリC_i, C_jを項目 Z に関する共起カテゴリセットと呼び，C_{ij}と表すことにする．カテゴリCi, Cjに同時に含まれる項目，つまり，共起カテゴリセットの子項目 Z の集合 S は，下記のように表わされる．

$$S_{ij} = \{Z|Z \in C_i \wedge Z \in C_j\} = \{Z|Z \in C_{ij}\}$$

本研究では各 Z が含まれるS_{ij}に関する統計情報を用いることで，意外性のある知識情報を発見できると考え， $F(S_{ij})$をS_{ij}に含まれる Z の個数とした．

$$F(S_{ij}) = count(S_{ij})$$

意外性のある知識情報を含む項目 UI は，下記のように定義した．

$$UI = \{Z|F(S_{ij}) < \theta\}$$

本研究では，子の数が少ない共起カテゴリセットに属する項目 Z の中に，カテゴリC_i, C_jにまたがった意外性のある知識が含まれているであろうと仮定した．ここで，上記の仮定に加え，カテゴリC_i, C_j同士が子孫関係にあるかどうかを考慮した．子孫関係の定義については，３．２にて後述する．子孫関係を考慮した理由は，$C_i \in C_j$のようにカテゴリC_jがカテゴリC_iの子孫である場合，C_iとC_jは同じ分野に関するカテゴリであり，多くの分野にまたがる意外性のある知識情報を含む項目を発見するという本研究の目的から外れるからである．

3.2. 手法

Wikipedia の全データは，Wikipedia のダウンロードページから XML ファイルの形式でダウンロードすることができる[7]．本研究では前処理として，Wikipedia カテゴリ関係を TSV 形式のファイルに再現できるオープンソースソフトウェア Wik-IE[8]を使用し，Wikipedia カテゴリネットワークを表すデータを作成した．データは，Wikipedia のグラフ構造を edge と node の２つのデータで表現している．作成した edge ファイルと node ファイルは，NLP 若手の会有志によって実装された汎用シソーラス探索ライブラリ[9]のフォーマットに準拠している．なお，Wikipedia の全データには 2008 年 12 月 10 日付けの日本語版 Wikipedia のデータを用いた．出力された edge ファイル，node ファイルの抜粋例を表１，表２に示す．

表１．edge.tsv

id-from	id-to	Relation
52220	680877	hypernym
54817	52220	redirect
362802	52220	redirect
410798	52220	redirect

表２．node.tsv

Id	title	kind
52220	IPod	leaf
54817	Ipod	redirect
362802	アイポッド	redirect
41078	IPod photo	redirect
680877	category:IPod	node

edge ファイルの hypernym はカテゴリ関係，redirect は同義語などを関連付けるリダイレクト関係を表す．また，node ファイルの leaf は各記事の項目，node はカテゴリ，redirect はリダイレクトの項目を表す．上記で作成したデータを用いて，以下のデータを作成し，平均・標準偏差を計算した．

- 各項目が属するカテゴリの数(親カテゴリの数)
- 各カテゴリが持つ子項目の数
- 各項目に関する，共通の親を 1 つ以上持つ項目の数(兄弟項目の数)
- 各カテゴリセットが持つ子の数

なお，本研究ではリダイレクト関係は無視して計算している．

次に，3.1 で述べた子孫関係の調査手法について述べる．子孫関係は，Wikipedia の各カテゴリが位置する階層$H(C_i)$を調べ，各カテゴリが属する階層と同じ階層化，それよりも深い階層へ向かってのみエッジをたどるようにして調べた．Wikipedia のカテゴリ

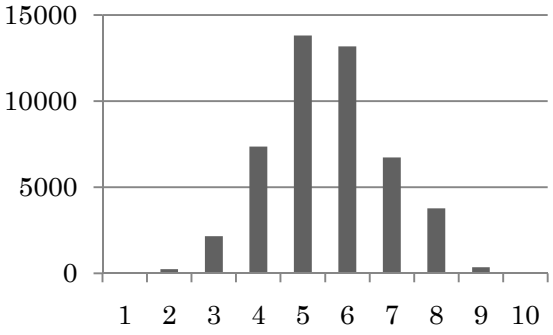
は”Category:共通カテゴリ”を起点としグラフ構造を成しており，本研究では Wikipedia カテゴリネットワークにおいてルートノードにあたる”Category:主要カテゴリ”からの距離（ホップ数）を階層とした．各カテゴリの階層$H(C_i)$は Wikipedia カテゴリの起点である”Category:主要カテゴリ”を root とすると，下記のように表わされる．

$$H(C_i) = \min|\mathrm{root} - C_i|$$

各階層の合計カテゴリ数を，図１に示す．

カテゴリC_i, C_jが子孫関係になっているか否かは，上記で調査した各カテゴリが属する階層の情報を用い，調査対象のカテゴリよりも同じ階層かそれよりも深い階層へ向かって探索して判定した．

図 1.　階層別合計カテゴリ数



3.3.　集計結果

3.2 において提示した各統計処理の結果は，以下の通りである．

表３．各項目に対する親カテゴリの数

平均	2.6759 カテゴリ
標準偏差	1.8642

表４．各カテゴリに対する子項目の数

平均	28.2715 項目
標準偏差	110.4709

表５．各項目に対する兄弟項目の数

平均	1279.6802 項目
標準偏差	2534.3208

表６．カテゴリセットに対する子の数

平均	28.8968
標準偏差	122.9076

各項目はそれぞれ平均で 2.6759 カテゴリに属しており，各カテゴリはそれぞれ平均で 28.2715 個の子項目を持っている．

4. 分析と応用

Wikipedia のカテゴリネットワークの統計処理を行うことで，各カテゴリセットに属する項目の数や，各カテゴリに属する子孫の数，カテゴリセットが互いに子孫関係にあるか否かの情報を得ることができた．以上の情報を用い，本研究の目的である意外性のある知識情報について分析する．

表７に，共起カテゴリセットの子項目数$F(S_{ij})$の値が大きいものと小さいものの例を挙げる．$F(S_{ij})$の値が小さくなるほど，カテゴリ同士の関係性が意外なものであることがわかる．たとえば，麻生太郎は”Category:日本の内閣総理大臣”に属しているが，一方で”Category:オリンピック射撃競技日本代表選手”にも属している．麻生太郎がオリンピックの射撃選手であることは一般的にあまり知られておらず，意外性のあるカテゴリ関係を含んだ項目と言える．また，”Category: 呪術”と”Category:アメリカ合衆国の大統領”に共通して現れる唯一の項目である”テカムセの呪い”は，1840 年から 1960 年までの 20 で割り切れる年に選出された大統領が皆在職中に死去した事実に基づいて語られている呪いであるが，両カテゴリ間の意外な関連をあらわす項目として興味深い項目である．一方，$F(S_{ij})$の値が大きなものは，カテゴリ同士の関係に特徴的な意外性はみられなかった．また，カテゴリC_i, C_jが子孫関係にあるか否かも重要な視点である．カテゴリC_i, C_jが子孫関係にある場合も，カテゴリ同士の関係に特徴的な意外性はみられなかった．たとえば，”漢方医学”は”Category:医療”と”Category:伝統医学”に属しているが，”Category:伝統医学”は”Category:医療”の子孫に該当するカテゴリである．

5. おわりに

本研究では，Wikipedia のカテゴリ関係を分析することで，多分野にまたがる意外性のあるカテゴリ関係をもつ項目を発見することを提案した．単純なカテゴリ関係の処理結果を用いることで，興味深い結果を得ることができた．意外性の定義は人それぞれ異なるものの，$F(S_{ij})$の値とカテゴリ間関係の意外性にはある程度の相関がみられた．

しかし，$F(S_{ij})$の定義や UI を決定する閾値 θ など，工夫すべき点が残されている．したがって，これらの定義をより詳細に検討し，意外性について定式化

表７．共起カテゴリセットとその子項目の例

Ci	H(Ci)	Cj	H(Cj)	子孫関係	Z	F(S_{ij})
category:日本の内閣総理大臣	5	category:オリンピック射撃競技日本代表選手	8	なし	麻生太郎	1
category:弁当	4	category:キャラクター	5	なし	キャラ弁	1
category:呪術	4	category:アメリカ合衆国の大統領	6	なし	テカムセの呪い	1
category:日本の経済学者	5	category:オリンピックサッカー日本代表選手	8	なし	堀江忠男	1
category:コンピュータウイルス	5	category:福田康夫	4	なし	福田ウイルス	1
category:日本の国会議員	5	category:銅像	4	なし	小渕恵三	1
category:呪術	4	category:アメリカ合衆国の大統領	6	なし	テカムセの呪い	1
category:イギリスの企業	5	category:イギリスの鉄道事業者	6	あり	ユーロトンネル会社	1
category:医療	2	category:伝統医学	3	あり	漢方医学	1
category:アメリカ合衆国の映画作品	6	category:恋愛映画	4	なし	卒業 (1967 年の映画)	55
category:アメリカ合衆国のオリンピック選手	6	Category:アメリカ合衆国のオリンピック金メダリスト	7	なし	ビーナス・ウィリアムズ	55
category:日本の俳優	6	category:東京都出身の人物	5	なし	木村拓哉	2088

する必要がある．また，機械学習により，意外性のあるカテゴリの特徴を発見する手法も考えられる．今後は，本研究で得られた統計情報をパラメータとして用い，意外性のあるカテゴリ関係をもつ項目を発見することを目指す．

参考文献

[１] Auer, S. and Bizer, C. and Kobilarov, G. and Lehmann, J. and Cyganiak, R. and Ives, Z.:DBPedia: A Nucleus for a Web of Open Data, Lecture Note in Computer Science, vol.4825, pp.722 (2007)

[２] Suchanek, F.M. and Kasneci, G. and Weikum, G.:Yago: a core of semantic knowledge, Processing of the 16th international conference on World Wide Web, pp.697-706 (2007)

[３] 田村悟之，清田陽司，増田英孝，中川裕志: Reference Navigator:異種オントロジーの統合ブラウジングツール～図書館の分類体系と Wikipedia カテゴリの対応付け～，言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」論文集, pp.35-38 (2007)

[４] Dat P.T. Nguyen, Yutaka Matsuo, Mitsuru Ishizuka.:Relation Extraction from Wikipedia Using Subtree Mining, AAAI07, pp.1414-1420 (2007)

[５] Michael Strube, Simone Paolo Ponzetto, WikiRelate! Computing Semantic Relatedness Using Wikipedia, AAAI06, pp.1419-1424 (2006)

[６] 中山浩太郎，原隆浩，西尾章次郎: 自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築に関する一手法, DEWS2008 (2008)

[７] Wikipedia ダウンロードページ，http://download.wikipedia.org/jawiki/

[８] Wik-IE,sourceforge, http://sourceforge.jp/projects/wiki-ie/

[９] 清田陽司，阿辺川武，吉田稔，田村悟之，坂井哲: 汎用シソーラス探索ライブラリの開発，言語処理学会第 14 回年次大会発表論文集(PD1-3)，pp.257-260 (2008)