

# Wikipedia カテゴリネットワークからの意外性のある関係性の抽出

A study for extracting serendipitous relations from Wikipedia category networks

野田陽平<sup>1\*</sup> 清田陽司<sup>2</sup> 中川裕志<sup>2</sup>

Yohei Noda<sup>1</sup>, Yoji Kiyota<sup>2</sup>, Hiroshi Nakagawa<sup>2</sup>

<sup>1</sup> 東京大学大学院学際情報学府

<sup>1</sup> Graduate School of Interdisciplinary Information Studies, University of Tokyo

<sup>2</sup> 東京大学情報基盤センター

<sup>2</sup> Information Technology Center, University of Tokyo

## 概要

本研究は、Wikipedia の記事の中から、複数の分野にまたがった意外性のある知識を発見することを目的としている。

Wikipedia は各記事が1つ以上のカテゴリに属しており、そのカテゴリネットワークはグラフ構造を成している。それらのグラフ上の構造を特徴量として利用し、機械学習により各記事に対して意外性を含む記事であるか否かの判定を行う。例えば、

「category:オープンソース」と「category:コーラ」という一見意味的に離れたカテゴリにも、それらのカテゴリに共通して属する「オープンコーラ」という記事が存在する。本研究では、このような意外性のある関係性をもった記事を、機械学習を用いて自動的に発見する手法を提案する。

## 1. はじめに

ブログやマイクロブログ、BBS、WikiなどはCGM(Consumer Generated Media)と呼ばれ、その記事数は日々増加している。日本においては、ブログサイト数は延べ1,600万件、総記事数は130億を超えており、1か月に約4,000万から5,000万記事が投稿されている[1]。また、日本におけるTwitterの利用者数は193万人を超え、新たなCGMとしての地位を確立している。以上のようなWeb上の膨大なCGM情報の中から有用な情報を発見・抽出するニーズが高まっている。

CGMから有用な情報を探し出す手段としては、大きく分けてクエリ型検索エンジンとディレクトリ型検索エンジンが存在する。GoogleやYahoo!などのクエリ型の検索エンジンは、ユーザから与えられたクエリを元に、そのクエリに似た文書を抽出し、提示する。一方、GoogleディレクトリやYahoo!カテゴリなどのディレクトリ型の検索エンジンは、似たテーマのウェブページを同じカテゴリにまとめ、階層的に格納することで、ユーザが探しているウェブページに辿り着きやすくするように設計されている。

クエリ型やディレクトリ型の検索エンジンは、ユーザが想定している情報を抽出する際には非常に役に立つが、有用な情報は、ユーザが想定しているものの中のみ存在するとは限らない。ユーザが意外であると感じるような情報の中にも、有用な情報は含まれている可能性がある。しかし、これらの情報は、ユーザが自ら予想して検索エンジンを用いて抽出することは困難である。本研究では、ユーザが想定しないような意外性のある知識をWeb上から抽出することを最終的な目標とする。まずは、Web上で組織的に整備されているコンテンツであるWikipediaの記事の中から、意外性のある知識を発見するという問題を扱う。具体的には、Wikipediaの記事の中から、複数の分野にまたがった意外性のある知識を発見する。

本研究では、Wikipediaのグラフネットワークの構造から特徴量を抽出し、機械学習を用いて意外性のある関係性の判定を行う。なお、抽出される情報の有用性の判定も重要な課題であるが、有用性に関しては判定の対象としない。

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。3章ではWikipediaの構造や、それがもつ特徴について説明し、本研究のアイデアを明確にする。4章では、Wikipediaのグラフネットワークの構造から抽出した特徴量と、正解データについて説明する。5章では4章で説明した特徴量を用いて行った機械学習の結果を示し、Wikipediaの中から抽出された意外性のある関係を提示する。また、6章では本稿のまとめを行う。

## 2. 関連研究

本研究に関連する先行研究として、意外性に着目した研究と、Wikipediaのカテゴリ関係を利用した研究を挙げる。

まず、意外性のある情報の提示に関する研究には、レコメンドエンジンに関する研究がある[2]。これらの推薦システムには、ユーザが予想できないような意外性のあるアイテムを推薦することが求められるため、初期の段階から意外性のあるアイテムを提示する工夫がなされてきた。ウェブ上の文書から意外性のある情報を抽出する研究には、

\*連絡先: noda@r.dl.itc.u-tokyo.ac.jp

Nadamoto らの研究がある[3][4]. BBS などのコミュニティ型コンテンツにおける議論の中で、話題には上がっていないが重要であるコンテンツを”コンテンツホール”と定義し、それらを抽出することを提案している。具体的には、コミュニティ型コンテンツ上でのテーマに関する Wikipedia の記事を取り上げ、Wikipedia には含まれているが、議論には上がっていない視点を抽出している。また、Torisawa[5], Stijin[6]らは、Wikipedia から、上位下位概念を抽出し、検索された対象を視覚的に検索できるシステムを開発している。

次に、Wikipedia のカテゴリ関係を利用した研究を挙げる。Nguyen らは、Wikipedia の記事を分析することで、Wikipedia の記事ごとの関連性を抽出している[7]。また、Strube らは、Wikipedia の各記事が属するカテゴリ情報を利用し、各概念同士の関連の度合いを算出している[8]。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[9]。

以上のように、Wikipedia のカテゴリ構造を用いて、意外性を直接扱っている研究はまだ行われていない。Nadamoto らの研究は、コミュニティ型コンテンツ上での議論の中で、欠損している視点を Wikipedia の記事を軸に抽出しているが、本研究の立場とは異なる。本研究では、異なる概念を結びつけるような関係性を抽出することで、意外性のある関係性を抽出する。人手により日々更新され密になっている Wikipedia のカテゴリネットワークの構造を特徴量として用い、意外で価値のある情報の発見を目指す。

### 3. Wikipedia の構造を利用した意外性のある関係性の抽出

Wikipedia は、誰でも編集可能な巨大なウェブ百科事典である。英語版 Wikipedia は 2009 年 11 月 14 日時点で 302 万記事、日本語版 Wikipedia は 63 万記事を越え、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。

また、記事の内容に対する多くの編集者により編集されており、ガイドラインによって中立的な観点に基づいて記述されることが求められている。「保護の方針」や「削除の方針」など、Wikipedia 自体を信頼性の高い状態に保つような方針も整えられており、万全とは言えないまでも、利用価値は高い。

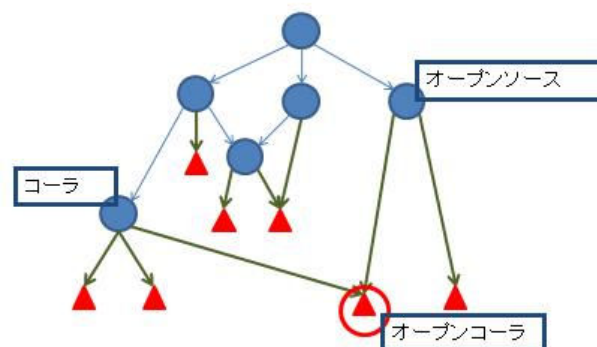
以上のように、Wikipedia は、記事の内容の分野的な網羅性と、その記事自体の信頼性が比較的高い CGM であると言える。本研究において Wikipedia を対象とした理由も、この 2 つの理由による。

ここで、Wikipedia の構造について説明する。Wikipedia

は各記事が 1 つ以上のカテゴリに属している。日本語版 Wikipedia は 9 個の主要カテゴリ<sup>1</sup>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、「オープンコーラ」という原材料が公開されているコーラに関する記事があるが、この項目は「オープンソース」というカテゴリや、「コーラ」というカテゴリに属している。「オープンソース」と「コーラ」というカテゴリは、一見関連がないものであると考えられるが、「オープンコーラ」という項目によって 2 つの概念が結びついている。本研究では、このような意外な関係を Wikipedia の中から大量に発掘することを目的とする。具体的には、図 1 のような 2 つのカテゴリに共通して表れる記事をもつ関係性をすべて取得し、ある記事が 2 つのカテゴリのつながりにおいて意外性をもつ項目であるか否かを判別する。

図 1. 2 つのカテゴリと記事の関係



### 4. グラフ構造から得た特徴量と、正解データの作成

本研究の目的は、Wikipedia の中から意外性のある関係性を含んだ項目を抽出する事である。本研究では、以下の特徴量を使用し、SVM を用いて意外性の有無を判定した。これらの特徴量の抽出には、Wikipedia のダウンロードページ[10]からダウンロードできるダンプファイルを加工したものを利用した。ダンプファイルの加工には、Wikipedia のグラフ構造をノードとエッジのデータとして表わすことができる、オープンソースソフトウェアの Wik-IE[11]を使用した。なお、ここでは、カテゴリを  $c_i$ 、記

<sup>1</sup>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

事を  $a_i$  と表記する。

・親カテゴリの子供の数

$$Child(c_i) = count(a_k | a_k \in c_i)$$

親カテゴリの子供の数は、Wikipedia の各カテゴリに属する記事の数を数え、その数を採用した。

・親カテゴリ A, B の階層

$$H(c_i) = \min |root - c_i|$$

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 とし、その root との最短距離を用いた。この階層が深ければ深い程、カテゴリが示す概念が具体的かつ専門性が高いカテゴリであり、浅ければ浅い程、より抽象的な概念を示すカテゴリであると考えられる。

・親カテゴリ A, B の共通子項目（共通インスタンス）の数

$$CoChild(C_A, C_B) = count(a_i | a_i \in C_A \wedge a_i \in C_B)$$

カテゴリ A, B の共通子項目の数は、2つのカテゴリに共通して表れる子項目の希少性を表している。共通項目数が多いければ、その2つのカテゴリのつながりは一般的なつながりであるが、共通項目数が少なければ、その2つのカテゴリのつながりは意外性の高いつながりであると考えられる。

正解データの作成方法について説明する。正解データは、3章の図1のような関係性を Wikipedia の中からすべて抽出し、人手で意外性の有無を判定した。正解データは、正例を 210 関係、負例を 180 関係の、計 390 件を作成した。

正例、負例のそれぞれのデータの平均、標準偏差は表 1, 2 の通りである。

表 1 意外性のある教師データの平均・分散

Feature	平均	標準偏差
親カテゴリ A の子供の数	273.852381	643.7357726
親カテゴリ A の階層	3.885714286	1.19773028
親カテゴリ B の子供の数	118.3380952	257.3707257
親カテゴリ B の階層	4.298076923	1.243288194
共通子項目の数	1.133333333	0.799603076

表 2. 意外性のない教師データの平均・分散

feature	平均	標準偏差
親カテゴリ A の子供の数	1077.491713	1793.346588
親カテゴリ A の階層	5.604519774	1.458354428
親カテゴリ B の子供の数	702.5138122	1757.900223
親カテゴリ B の階層	5.662857143	1.341444948
共通子項目の数	163.2983425	390.7049888

## 5. 機械学習による意外性の判定

本研究では、4章で作成した教師データを利用し、SVM による判別を行った。なお、教師データは4つに分割し、交差検定を行った。SVM は、R 言語のパッケージ kernlab の SVM 関数である、ksvm を用いた。カーネルには、線型カーネル、2次、3次の多項式カーネル、ガウシアンカーネルを用いた。それぞれのカーネルを用いた SVM による結果は表3の通りである。

どの特徴量が意外性のある情報を抽出するために効果的に効いている特徴量なのかを調べるために、各特徴量以外の特徴量を用いた実験も行った。ここでは、共通項目の数を除外した際に精度が低くなっていることから、この特徴量が意外性のある関係性を抽出する際に重要な特徴量であることが分かる。なお、全関係性について意外性の有無を判別した結果の抜粋を、表4に示す。

## 6. まとめと今後の課題

本研究では、Wikipedia の中から意外性のある関係を抽出することを目的とし、Wikipedia のカテゴリネットワークから特徴量を抽出し、機械学習を用いて意外性の判定を行う手法の提案を行った。しかし、意外性を含む関係性は、全体の中の一部に過ぎず、それを確実に抽出することは非常に困難である。また、意外性の定義は人それぞれ違うため、評価しにくい等の問題点もある。意外性の定義をより詳細に検討し、本稿では考慮されていない特徴量を導入することで、より明快な形で意外性のある関係を提示できるようにする必要がある。また、本研究においては機械学習を用いて意外性の判定を行ったが、機械学習がこの問題を解く際に有力な手法であるのかも、慎重に考える必要がある。

## 参考文献

[1] インターネット視聴率データ、ビデオリサーチインタラクティブ

- [ 2 ] M.Sarwar,G.Karypis,A.J.Konstan and J.Riedl, “Item-based collaborative filtering recommendation algorithms”, WWW01,pp22-32, 2001
- [ 3 ] A Nadamoto, E Aramaki, Yohei Murakami, “Searching for Important but Neglected Content from Community-type-content”, SITIS, pp.161-168, 2008
- [ 4 ] A Nadamoto, E Aramaki, T Abekawa, Y Murakami, “Content Hole Search in Community-type Content”,WWW2009, PosterSessions,pp1223-1224, 2009
- [ 5 ] Kentaro Torisawa, Stijn De Saeger, Yasunori Kakizawa, Jun’ichi Kazama, Masaki Murata, Daisuke Noguchi, and Asuka Sumida, TORISHIKI-KAI, an Autogenerated Web Search Directory. ISUC2008, pp.179-186,2008
- [ 6 ] Stijn De Saeger, Kentaro Torisawa, Jun’ichi Kazama, Lookingo for trouble, Proc of The 22nd International Conference on Computational Linguistics, 2008
- [ 7 ] Dat P.T. Nguyen, Yutaka Matsuo, MitsuruIshizuka.:Relation Extraction from Wikipedia Using Subtree Mining, AAAI07, pp.1414-1420, 2007
- [ 8 ] Michael Strube, Simone Paolo Ponzetto, WikiRelate!Computing Semantic Relatedness Using Wikipedia,AAAI06, pp.1419-1424 ,2006
- [ 9 ] 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
- [ 1 0 ] Wikipedia ダウンロードページ, <http://download.wikipedia.org/jawiki/>
- [ 1 1 ] Wik-IE,sourceforge, <http://sourceforge.jp/projects/wiki-ie/>

表 3. SVM による結果

	Linear Kernel	Polynomial Kernel(2)	Polynomial Kernel(3)	Gaussian Kernel
すべての feature	83.75%	<b>85%</b>	81.25%	<b>85%</b>
親カテゴリ A の子供の数以外	82.5%	77.5%	73.75%	78.75%
親カテゴリ A の階層以外	75%	75%	78.75%	76.25%
親カテゴリ B の子供の数以外	72.5%	80%	76.25%	67.5%
親カテゴリ B の階層以外	77.5%	77.5%	81.25%	82.5%
共通子項目の数	72.5%	71.25%	71.25%	70%

表 4. 抽出された正例・負例

SVM による判別結果	カテゴリ A	カテゴリ A の子供の数	カテゴリ A の階層	カテゴリ B	カテゴリ B の子供の数	カテゴリ B の階層	項目	共通子項目数
○	奇跡の水	10	3	機能水	7	3	アルカリイオン水	2
○	オープンソース	374	4	コーラ	26	6	オープンコーラ	1
○	能力開発	63	3	キャラクター	7	5	キャラ立ち	1
×	ヴァッレダオスタ州のコムーネ	46	5	イタリアのコムーネ	436	7	アルナド	8
×	日本のタレント	7617	6	東京都出身の人物	9825	5	堺正章	2088
×	静岡県の鉄道駅	34	7	葵区	140	7	静岡駅	11