

Wikipedia のカテゴリを用いた編集者の得意分野特定

近藤 弘隆[†] 鈴木 優^{††} 石川 佳治[†]

[†] 名古屋大学大学院情報科学研究科 〒464-8601 名古屋市千種区不老町

^{††} 奈良先端科学技術大学院大学情報科学研究科 〒630-0192 奈良県生駒市高山町 8916 番地の 5

E-mail: [†]kondou@db.ss.is.nagoya-u.ac.jp, ^{††}ysuzuki@is.naist.jp, ^{†††}ishikawa@is.nagoya-u.ac.jp

あらまし 本研究では Wikipedia のカテゴリを用いて、編集者に関係しておりかつ分野としてふさわしい単語を選択することで、編集者の得意分野特定を行う。Wikipedia のカテゴリは記事の分類を行うものであり、編集した記事のカテゴリを見れば、記事の位置づけを知ることができる。しかし、編集者の得意な分野というのは、記事のカテゴリのように明示されているわけではない。そこで、編集者の得意分野となりえる単語、すなわち、分野としてふさわしい単語をカテゴリ名に含まれている単語を利用して取得し、さらに編集者の得意分野を特定する。カテゴリ名にはそのように得意分野を表すことには向いていない単語も使用されている。Wikipedia のカテゴリは Wikipedia の全体構造を示すため、うまく利用することで分野としてふさわしい単語を得られる。また、編集者は記事を編集しているため、編集した記事の傾向をカテゴリの構造から得る。よって、カテゴリ名に含まれている単語が分野としてふさわしいかどうかを、Wikipedia のカテゴリ構造やカテゴリ名の使われ方を利用して推定する。また、カテゴリ群を単語に分解し、単語が含まれるカテゴリから編集した記事への距離を用いて関係の強い単語を取得する。これらを組み合わせ、編集者が得意であると思われる分野の単語を取得し、得意分野の特定を行う。

キーワード Wikipedia, 編集者, 得意分野, カテゴリ

1. はじめに

UGC (User Generated Content) と呼ばれるコンテンツは、一般の利用者によって形成されるコンテンツである。コンテンツ作成を行うことができる人数が多いため、幅広く、豊富な情報が存在する。そのような理由などにより、UGC は様々な分野から重要視されている。巨大で有用な UGC の一つとして Wikipedia^(注1) が挙げられる。しかし、利用者によって作成されたコンテンツは、コンテンツ制作者の力量などが不明であることなどから、コンテンツの質が保障されていないなどの問題が存在している。Web 上の巨大な百科事典である Wikipedia にも同様の問題がある。また、Wikipedia では様々な言語版が作成されているが、本稿では日本語版^(注2) を用いる。

Wikipedia の特徴として誰でも編集可能であるという点がある。現在、英語版 Wikipedia には 500 万件以上、日本語版 Wikipedia には 100 万件以上の記事が存在している。このように、Wikipedia に記述されている情報は多い。また、秀逸な記事 (Featured articles) や良質な記事 (Good articles) の選定を行う、良い記事とするための基準を示すなど、百科事典としての質の高めようとする努力が行われている。

しかし Wikipedia には記事が多い一方で、未作成の記事や、内容が不十分な記事も存在する。未完成、未作成の記事を完成させたい場合、Wikipedia では編集者が編集者に直接呼びかけ、追記を促すということが行われる。作成されていない記事の一

覧を参照する機能^(注3) が存在している。また、不特定の編集者に呼び掛ける手段としては「Wikipedia:執筆依頼」^(注4) というページで募ることが可能である。そして他言語で書かれた記事は存在するが、日本語の記事が存在しない記事に対しては翻訳依頼^(注5) というページで編集できる編集者を募集することが可能である。このように、Wikipedia には様々なページで記事の完成度、そして Wikipedia 自体の百科事典としての完成度を上げようとする動きが存在している。

しかし、記事を追記、完成させることができる編集者を探すことは利用者には困難である。なぜなら、Wikipedia は不特定多数の人物が編集する仕組みであるため、編集者が他の編集者の Wikipedia 外の情報を持っていたり、交流することは少ない。そして、Wikipedia 上で編集者同士がコミュニケーションを取る場合は、基本的に編集者ごとの会話ページや各記事のノートページなど一部の場所に限られるためである。また、議論や会話などに参加せず、編集のみを行う編集者も存在する。編集者が自分で得意分野を記入するページ「Wikipedia:ウィキペディアン/分野別」^(注6) も存在しているが、自己申告であり、Wikipedia を頻繁に編集する編集者であっても記入していない編集者も多い。そして「Wikipedia:ウィキペディアン/分野別」には現在活動していない編集者が含まれている可能性もある。

そこで我々は Wikipedia のカテゴリを利用することにより、

(注3) : <https://ja.wikipedia.org/wiki/特別:ページが存在しないリンク>

(注4) : <https://ja.wikipedia.org/wiki/Wikipedia:執筆依頼>

(注5) : <https://ja.wikipedia.org/wiki/Wikipedia:翻訳依頼>

(注6) : <https://ja.wikipedia.org/wiki/Wikipedia:ウィキペディアン/分野別>

(注1) : <https://www.wikipedia.org/>

(注2) : <https://ja.wikipedia.org/wiki/メインページ>

編集者の得意分野の把握を支援することを目的として研究を行った。本研究では Wikipedia のカテゴリを用いて編集者の得意分野を推定する手法について取り組む。Wikipedia のカテゴリを用いて、編集者に関係しておりかつ分野として相応しい単語をカテゴリ名から選択することにより、編集者の得意分野特定を行う。この論文において分野として相応しい単語とは、編集者の得意分野となりえる単語である。いくつかの単語、概念、事象をまとめた単語は得意分野として相応しい単語であると考えられる一方で、分野をまたぐ事柄を含む単語は得意分野を表す単語として向いていないと考えられる。

Wikipedia 編集者が編集したノートページとカテゴリの情報を用いて、編集者の得意分野をカテゴリ名から抽出した単語で表す。例えば、日本の戦国時代に詳しい編集者に対して編集履歴から「戦国」、「歴史」、「日本」といったような単語を取得することを目的とする。Wikipedia のカテゴリは基本的に分野分けの為に用いられているものであり、カテゴリ名を使えば得意分野を特定する手がかりになると考えられる。「真田信繁」という記事は「戦国武将」、「真田氏」などのカテゴリに属しており、カテゴリ「真田氏」は「戦国大名を輩出した氏族」などのカテゴリに属していることから、真田信繁は「戦国」時代と関わりが深いことが推測できる。編集者の得意分野を特定することができれば、記事の追記の依頼などがしやすくなると考えられる。そして、Wikipedia の百科事典の質を高めることができると考えられる。しかし、Wikipedia のカテゴリ名にも得意分野を表す分野名として適切なものと不適切なものが存在する。記事「三種の神器」は「Category:三種の神器」、「Category:名数3」、「Category:皇位」、「Category:皇位継承法」、「Category:日本の国の象徴」というカテゴリに属している。このうち、「名数3」というカテゴリは3つ何かを集めたものに対してつけられるカテゴリである。このカテゴリ「名数3」は、含まれている記事の共通点は3つの物で構成されているというだけであり、地名から法律、集団名など様々な記事が含まれている。そのため、得意分野を表すためのカテゴリ名として利用することは不適切であると考えられる。そこで本手法では、カテゴリ名を分解し、その単語に対してカテゴリの構造とカテゴリ名を利用し、単語の分野らしさを求めることにより解決しようとしている。

2. 関連研究

Wikipedia は成功した共同執筆システムである。そのため、Wikipedia 上には様々な分野の知識が大量に存在している。そういった理由などから様々な観点で Wikipedia に関する研究が行なわれている。

Wikipedia は誰でも編集可能であるため、既存の辞書に対して専門用語や新語も多く定義されている。そのことを利用し、白川ら [1] は Wikipedia のカテゴリ構造を利用し、Wikipedia の記事を確率的に分類し、語句を分類する概念辞書を作成するということに取り組んでいる。ランダムウォークを用いて記事がどのカテゴリに属しているかを確率的に算出している。いくつかの最終的に属するカテゴリを決定し、そのカテゴリにどれだけの確率で属するか算出することを目的としているが、本手

法では編集者が得意な分野を単語で表すことを目的としているため、繋がりのあるカテゴリ名を全て使用している。

中谷ら [2] は Web 検索の結果検索されたページが専門的かどうかを特定するために、検索語に関連する専門用語を抽出するために Wikipedia カテゴリを利用している。カテゴリ情報を集約し、検索語の分野の算出、その分野の専門用語の特定を行っている。この手法ではカテゴリに含まれる記事数など、カテゴリから直接得られる情報のみを使用しているが、本手法ではカテゴリの繋がりも考慮した手法となっている。

Jiang らは [3] 文脈に沿ったカテゴリのランク付けを行うために Wikipedia のリンクとカテゴリを用いたシステム Wiki3C を作成した。Schönhofen [4] はカテゴリを利用して、文書などに隠されたトピックを見つけるのに利用する研究を行っている。記事名と記事が属するカテゴリを用いてカテゴリの重みを算出し、文書にあったカテゴリを発見している。Kapanipathi ら [5] は Wikipedia を用いて Hierarchical Interest Graph というグラフで背景知識を構築し、Twitter ユーザの興味を少ないテキストから特定する方法を考案している。胡ら [6] は Wikipedia のカテゴリ構造を用いて Twitter ユーザの関心分野を抽出する手法を提案している。複数の意味を持つ単語を Wikiepdia のカテゴリ「曖昧さ回避」を使うことにより解決し、さらに記事から特定のカテゴリまでの距離を利用して得意分野を算出している。Subramaniaswamy らは [7] 文書にタグ付けをすることを目的として知識源に Wikipedia などを利用している。本研究も編集者に適切な単語を与えるという点から見れば、対象が編集者か文書かといったような違いといえる。これらの研究は、主に Wikipedia 外の文書やユーザに対してのトピック等を求める物である。それに対し、本研究は Wikipedia 編集者の得意分野特定を目的としている。よってより多くの Wikipedia 上の情報を使用し、

Wikipedia にトピックモデルを適用し、記事などの話題を特定する研究として林ら [8] の研究などが存在する。トピックモデルは文書が作られる過程をモデル化するものであり、そのモデルを用いて文書が作られた時の文書の話題を特定するといったことに使用される。この研究はカテゴリ構造を用いて正確にトピックを特定し、検索精度を向上することなどを目的としている。この研究はカテゴリ情報が含まれた文書に適用する手法であるため、文書とカテゴリが直接結びつかない編集者の得意分野特定を行っている本研究とは異なる。

Wikipedia の編集者に注目した研究として West ら [9] の研究が存在する。Yahoo ツールバーなどから得た情報から Wikipedia 編集者がそれ以外のひとどのように違うのかなどの分析を行っている。その結果、Wikipedia 編集者は平均的な人と比較し興味ある分野により精通しているということになった。この研究はしかし、未知のユーザの得意分野を特定しようとしているわけではない。Welser ら [10] は Wikipedia における編集者の社会的役割を特定する研究を行っている。編集者の得意分野を特定することを目的とした本研究とは違い、Welser らの研究は編集者の編集傾向から荒らしに対処する役割がメインの編集者かどうかなどを特定している。

本手法では、編集者がより知識を持つと考えられる記事を利用するためにノートページを利用し、より細かい分野を特定することを目的としてカテゴリ名をそのまま使用するのではなく単語に分解して扱うというところを行っている。

3. 提案手法

本手法では、Wikipedia 編集者が編集したノートページとカテゴリの情報をを用いる。そして、カテゴリ名から単語を抽出し、その単語を用いて編集者の得意分野を表す。

提案手法ではまず、得意分野を知りたい編集者を指定し、その編集者の編集履歴を取得する。その編集履歴から Wikipedia の記事に付随する議論ページであるノートページを編集した記録を利用する。ノートページは 3.1.3 で説明する通り、Wikipedia の議論の場である。ノートページは記事に詳しい編集者が利用していると考えられるため、単に編集者の投稿記録を用いるよりも正確に得意分野を特定することを目的として、ノートページへの投稿記録を使用している。ノートページに対応する記事本体は、なんらかのカテゴリに属している。カテゴリは記事を分類するための物である。そのカテゴリの構造、カテゴリ名を用いて得意分野の特定を行う。このとき、編集者の持つ知識の得意分野というのは、正確に一つの分野に分けることができないと考えられる。記事の分野はその記事が属するカテゴリを用いて推測することが可能である。編集者は何らかの記事を編集しており、編集した記事に関して編集者は知識を持っている。つまり編集した記事の分野というのは編集者の得意分野の一部であると考えられる。それら編集した記事ごとの得意分野をうまく纏めることにより編集者得意分野を特定できると考えられる。

例えば、戦国時代の武将と鎌倉時代の武将に詳しい場合、「歴史上の人物」を得意分野であるとするのは幅が広すぎると考えられる。この場合は「戦国」、「鎌倉」を得意分野とすべきであると考えられる。このように、得意分野とは細かい分野の集合である。そして、Wikipedia のカテゴリは大量に存在するが、得意分野として適当な分野がカテゴリ名に存在しない場合も考えられる。そこで Wikipedia のカテゴリを単語に分割し、分野として相応しい単語を特定する。

得意分野を表す単語を特定するために、大きく分けて 4 つの手法を提案した。1 つ目の手法は編集したノートページからのカテゴリの距離を算出し用いる基本手法。2 つ目の手法は、頻出な単語は分野を表す単語として相応しくなく、頻出な単語の順位を下げれば上手くいくだろうと仮定し、カテゴリ名に含まれる名詞の出現頻度を用いる希少単語手法。3 つ目の手法は、分野として相応しい単語というのはカテゴリにある単語であり、深ければ深いほど、その人の得意分野をうまく表すという仮定を用いたカテゴリ深度法。4 つ目の手法は分野として相応しい単語を探す際に、カテゴリ名が重要ではないかという仮定から成り立つカテゴリ分野法である。また、これらの手法の組合せについても説明を行う。

3.1 Wikipedia の特徴

3.1.1 カテゴリ

Wikipedia にはカテゴリという機能が存在する。カテゴリは記事を分類するための機能で、Wikipedia 上のページをカテゴリに属させることが可能である。記事を編集する際に記事内に「[[Category:○○]]」と記述することにより、記事やカテゴリをカテゴリ ○○ に属させることができる。そして記事の下部に「Category:○○」ページへのリンクが作成される。

ページ名が「Category:」から始まるページはカテゴリページとして扱われる。カテゴリページではそのカテゴリに属している記事、カテゴリの一覧を見ることができる。また、カテゴリは記事と同じように編集することが可能であり、カテゴリに関する説明などを記述できる。本稿では、あるカテゴリに含まれているカテゴリをサブカテゴリ、あるカテゴリが属しているカテゴリを親カテゴリとする。カテゴリに含まれるページを取得することを下に辿る、含んでいるカテゴリを取得することを上に辿ると表記する。

Wikipedia のカテゴリに関する方針は「Wikipedia：カテゴリの方針^(注7)」に記述されている。その方針によれば、カテゴリは Wikipedia の骨組みであることが求められており、よりよい構造を作ることが求められている。最上位のカテゴリの役割を果たすカテゴリとして「Category:主要カテゴリ」が存在している。「Category:主要カテゴリ」が属するカテゴリは存在しない。そして、カテゴリの方針によれば「Category:主要カテゴリ」のみが親カテゴリを持たなくてよいことになっている。そのため、基本的にはどのカテゴリも親カテゴリをたどっていくことにより「Category:主要カテゴリ」に到達する構造となっている⁽¹⁾。また、図 1 のように、カテゴリを上に辿っていくと広い分類になり、下に辿ると、専門的、狭い範囲の分類になるよう作成されている。Wikipedia の方針に従った構造になるよう Wikipedia は編集されているため、記事が属する分野はカテゴリから推定することが可能であると考えられる。例えば、記事「三種の神器」は「Category:三種の神器」、「Category:名数 3」、「Category:皇位」、「Category:皇位継承法」、「Category:日本の国の象徴」というカテゴリに属している。このことや、これらカテゴリの親カテゴリの親カテゴリなどを見ることにより、三種の神器は天皇、皇室などに関係している物であることを知ることができる。

しかし、Wikipedia のカテゴリ方針が守られていないカテゴリも存在する。例えば上位カテゴリがないカテゴリである。この上位カテゴリがないカテゴリはごく少数（2015 年 1 月 25 日時点で 6 つ（Category:主要カテゴリ含む）存在）であり、Wikipedia のシステムによって検知されている^(注8)。また、ループになっている構造、例えば図 2 のような「Category:データ構造」の親カテゴリに「Category:アルゴリズム」があり、「Category:アルゴリズム」の親カテゴリに「Category:データ構造」があるという構造が存在する。また、記事「三種の神器」

(注7) : <https://ja.wikipedia.org/wiki/Wikipedia:カテゴリの方針>

(注8) : <https://ja.wikipedia.org/wiki/特別:カテゴリ未導入のカテゴリ>

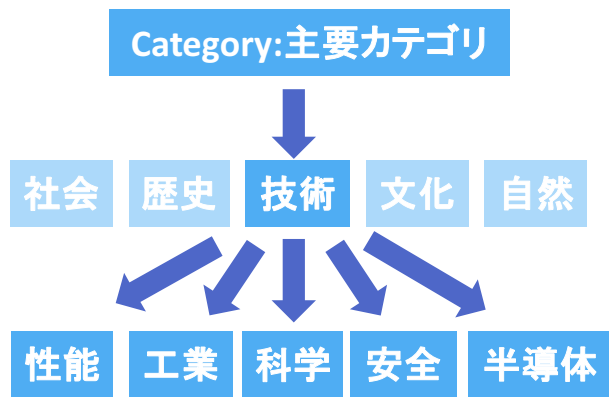


図 1 カテゴリ構造

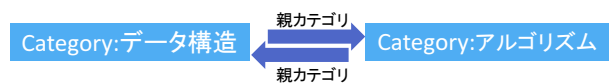


図 2 カテゴリのループ構造



図 3 投稿記録の画面

に含まれているカテゴリ「名数 3」はカテゴリ「技術」やカテゴリ「数学」に属するカテゴリである。これらの例のように、単純にカテゴリを上に見ていくだけではその記事を表す分野を特定できないことがある

3.1.2 編集履歴と編集者

Wikipedia では Wikipedia が立ち上げられてから現在までの編集履歴が全て残されている。そして Wikipedia では記事がいつ、誰によって、どのように編集されたのかという履歴や、編集者がどの記事をいつを編集したのかという編集者個人の履歴、すなわち投稿記録 (図 3) の取得が可能である。

Wikipedia の編集者は利用者名 (アカウント名) によって区別される。Wikipedia では利用者名は固有のものである。つまり同じ編集者名の編集者は複数存在することはない。また、利用者は自分のアカウントを削除することもできない。Wikipedia の編集者に関する情報は Wikipedia が開始してから現在に至るまで全ての記録が残る。そして、投稿記録など編集者に関する様々な記録を取得することが可能である。

編集者の得意分野に関してはカテゴリのように分野を特定することは難しい。なぜなら、編集者に関する情報は少ないからである。編集者はそれぞれ個人ページを作成することが可能だが、個人ページを作成する編集者は稀である。さらに個人ページに自分の得意分野を詳しく書いている編集者はさらに減る。



図 4 ノートページ

しかし、基本的に編集者は何らかの編集活動に関わっている。そして特に詳しい分野に関してはノートページで議論を行う。本研究では編集者のノートページでの編集を利用し、編集者の得意分野を特定する。

3.1.3 ノートページ

ノートページは図 4 のような Wikipedia において議論に使われるページである。ノートページのページ名は「ノート:○○」という「ノート:」から始まる。ノートページ「ノート:○○」は記事○○についての内容について議論するノートページである。ノートページも通常の記事と同じように編集が可能であり、編集者個人の投稿記録にも同様に残る。よって、ノートページへの投稿記録が編集者にあれば、編集者はそのノートページで行われた議論に参加したと見ることができる。

ノートページでは記事の内容やその記事の編集方針について議論が行われる。ノートページで議論可能な編集者であるということは、議論内容についての知識を持っているということであり、そのノートページの本体の記事について、特に詳しいと考えられる。

3.2 手法 1: 基本手法

上記の Wikipedia の特徴から、本手法では編集者が編集したノートページと対応する記事周辺のカテゴリに注目し、編集者の得意分野特定を行う。基本となる手法ではまず、得意分野を知りたい編集者の投稿記録を取得する。そしてその投稿記録の中から編集したノートページを取得する。ノートページで議論していれば、その記事の内容に詳しいのではないかと、そしてページの分野についてはページのカテゴリから分野を特定できると考えられる。よって、得意分野を以下のように算出する。

まず取得したノートページ群に対し、以下の操作を行う。

- (1) 対象のページの親カテゴリを取得する。
 - (2) 対象のページから何回の上へカテゴリを辿ることでのどりに着いたのか記録する。
 - (3) 取得したページそれぞれに対し、(1) から繰り返す。
- この操作を (1) から繰り返した回数を編集したノートページからカテゴリを上へ辿った回数とする。

この操作では幅優先探索で上にあるカテゴリを取得し、上へ辿った回数を取得する。幅優先探索である理由は、複数のページの親カテゴリを取得しているためと、カテゴリは複数の親カテゴリを持つことが可能であるためである。これらの理由により、親カテゴリを辿っていった場合、同じカテゴリに何度も到達することがある。そのため、本手法ではあるカテゴリに 2 度目以降にたどり着いた場合、それ以降の構造は 1 度目の到達と同じ構造であるため、そのカテゴリに関しては見ない。上へ辿った回数も 1 度目に到達した時点での回数 (一番低い回数) を使用する。この上へ辿った回数を利用し、一人の編集者につき、一つのカテゴリに一つの重みを与える。例えば、図 5 のよ

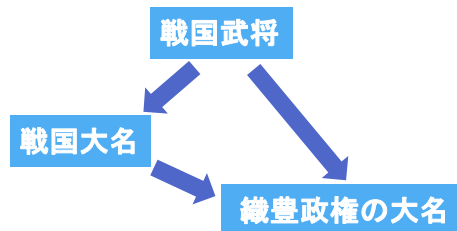


図 5 複数の親カテゴリ

うに、カテゴリ「織豊政権の戦国大名」は「戦国大名」と「戦国武将」を親カテゴリに持つ。そして、「戦国大名」も「戦国武将」を親カテゴリに持っている。このとき、本手法では「織豊政権の戦国大名」から「戦国武将」への上へ辿った回数は 1 回とし、戦国大名を経由するルートの上へ辿った回数は考慮しない。

カテゴリ C_a に含まれる単語 t の重み $d_{C_a,t}$ を以下のように算出する。

$$d_{C_a,t} = M - J_{C_a,t} \quad (C_a \text{ のカテゴリ名に単語 } t \text{ を含む場合}) \quad (1)$$

M は編集したノートページからカテゴリを上へ辿った回数に 1 を加えた数である。編集したノートページ群から何回上へ辿れば C_a が出現するかをカウントし、回数が多くかかるほどカテゴリの重みが低くなるようにしている。 $J_{C_a,t}$ は編集した記事から単語 t を含む C_a が出現するまでの上へ辿った回数である。まず編集した記事から上へ辿ることによりたどり着くカテゴリを算出し、カテゴリを形態素解析することにより単語を得る。その後、カテゴリ C_a における単語 t の重さ $d_{C_a,t}$ を算出する。

算出したカテゴリにおける単語の重みを利用し、編集者にとっての単語 t の得意分野度 W_t を算出する。

$$W_t = \sum_{a=1}^N d_{C_a,t} \quad (2)$$

単語 t が出現する N 個のカテゴリに対してその重みを合計した値が編集者の単語 t の得意分野度合 W_t となる。

このとき、記号は分野名として不適当と考えられる。また、今回は名詞のみを分野を表すことができる単語と考え、抽出の対象を名詞に絞った。

例えば、ある編集者の「歴史」という単語の得意分野度を求める場合を考える。編集者が編集したノートページをもとに取得したカテゴリの内「歴史」という単語を含んだカテゴリが「Category:江戸の歴史」、「Category:日本の歴史」と「Category:歴史の人物」で出現しており、編集したノートページからそれらのカテゴリにたどり着くための上へ辿る回数がそれぞれ 2 回、3 回、2 回、 X を 4 とした場合、単語「歴史」のその編集者の得意分野度は 5 となる。

3.3 手法 2：希少単語法

Wikipedia にはカテゴリ名の大部分が共通するカテゴリが多く存在する。例えば「Category:1800 年代生」と「Category:1900 年代生」といったように年代生という部分が共通し、それ以外の一部が違うといったカテゴリである。このような年代といった単語は出現回数が多いが、様々な分野をまたいで登場するた

め、分野を表す能力は低いと考えられる。よって、カテゴリで使われる単語が希少であるほど、編集者の得意分野を表すのに適した単語であるという仮定をたてる。つまり、カテゴリに頻出の単語は得意分野を表す単語として不適当であり、頻出な単語の順位を下げれば得意分野特定が上手くいくのではないかと仮定である。そこで、手法 1 に加えてカテゴリ一覧を用いて単語の希少性を考慮する手法、希少単語法を提案する。

Wikipedia ではカテゴリ名一覧が公開^(注9)されている。カテゴリ一覧からカテゴリ名を取得する。取得したカテゴリ一覧を用いて単語 t の希少性を考慮する。

$$W_t = \log \frac{Z}{f_t} \sum_{a=1}^N d_{C_a,t} \quad (3)$$

式 (2) での定義に加えて、 Z を全カテゴリ数、 f_t を全カテゴリのうち、単語 t が含まれていたカテゴリ数とする。単語の希少性を表す $\log \frac{Z}{f_t}$ は単語 t が出現するカテゴリが多いほど小さくなり、少数のカテゴリに出現するほど大きくなる。つまり、一般的な単語でないほどその人の得意分野として相応しい単語として扱うという操作をしている。

例えば「年」という単語に関して、基本的な手法 (3.2 節) で算出した得意分野度が 2 であった場合、全 239936 カテゴリ中 29064 カテゴリに存在するため、編集者の「年」の得意度は $2 \log 8.255$ となる。

3.4 手法 3：カテゴリ深度法

3.1.1 節で示したように、Wikipedia のカテゴリは階層構造を構築している。カテゴリを下に辿るほど、そのカテゴリが含む範囲は狭いものとなっていく。得意分野を示すためには編集者が得意な分野を過不足なく表す単語が必要である。例えば、平安時代に関する記事を編集していた編集者に対しては、得意分野を表す単語「歴史」よりも「平安」の方が適当であると考えられる。しかし、「歴史」を含むカテゴリが 2976 個存在するのに対し、「平安」を含むカテゴリは 65 個である。よって編集したノートページからカテゴリを辿っていった結果、「歴史」の出現回数が多くなり、得意分野度の順位が高くなりすぎる結果になる可能性がある。得意分野を表す単語については編集者の得意分野を過不足なく表せる分野の単語を優先すべきであると考えられる。なぜなら平安時代が得意な人がその他の時代の歴史について詳しいとは限らないからである。つまり、得意な分野を表す単語というのは得意な分野外の部分はカバーせず、最大限詳しい部分を表す単語であるという考えである。よって、分類の細かい深い部分で使われている単語ほど相応しい分野を表せると仮定している。そこで、手法 1 にこの仮定を加えた手法、カテゴリ深度法を提案する。

「Category：主要カテゴリ」から下に辿り、カテゴリにたどり着くまでの下に辿る回数を計測する。カテゴリ C_x について最上位のカテゴリである「Category:主要カテゴリ」から数えたカテゴリの深さ D_{C_x} （主要カテゴリから何回下へ辿れば C_x にたどり着くか）を算出する (図 6)。

(注9) : <https://ja.wikipedia.org/wiki/特別:カテゴリ一覧>

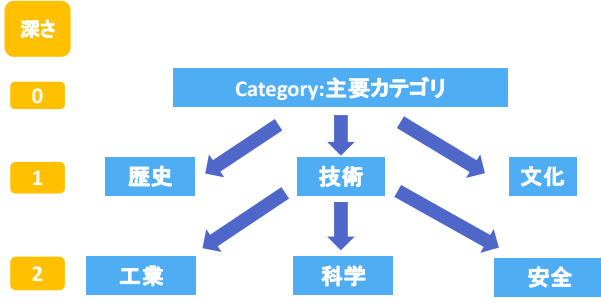


図6 カテゴリの深さ

上述のようにカテゴリは複数の親カテゴリを持つことができる。そのため、カテゴリを上から下へ辿っていくと、以前に辿ったカテゴリと同じカテゴリにたどり着くことがある。例えばカテゴリ「戦国武将」は「戦国大名」をサブカテゴリに持ち、カテゴリ「戦国大名」は「織豊政権の戦国大名」をサブカテゴリに持つ。また、カテゴリ「戦国武将」は「織豊政権の戦国大名」をサブカテゴリとして直接持っているという場合である。今回も同様に幅優先で探索し、最初にたどり着いた場合を用いる。つまり、本手法では同じカテゴリに複数回たどり着いた場合、深さの一番浅い回を使用し、2度目以降にたどり着いた場合はそこからさらに辿らない。このようにして下へ辿った回数を計算する。

この重みを利用して、以下の式のように取得した単語の得意分野度を算出する。単語 t の重みは、カテゴリ名に単語 t と一致するカテゴリ C_x が存在した場合、

$$W_t = D_{C_x} \sum_{a=1}^N d_{C_a,t} \quad (4)$$

$d_{C_x,t}$ は式 (2) での定義と同様である。カテゴリ名に単語 t と一致するものが無い場合の得意分野度は基本的な手法と同じとする。

「Category:歴史」は「Category:主要カテゴリ」の直接のサブカテゴリであるため、深さ D_{C_x} は 1 である。よって、基本的な手法 (3.2) で単語「歴史」の得意分野度が 5 であったとき、カテゴリ深度法による単語「歴史」の得意分野度は 5 となる。

3.5 手法4：カテゴリ分野法

前節では、編集者の得意分野を細かく正確に表すために、主要カテゴリからの距離というカテゴリの構造に着目した。本節では、分野として相応しい単語を探すカテゴリ名の使われ方に着目した手法を提案する。カテゴリに含まれているカテゴリはカテゴリの分類、分割を目的として使用されている。このとき、カテゴリ名が大きな分野を表す単語であれば、その分野をさらに分類、分割するカテゴリが多く作られると考えられる。例えば、「Category:物理学」は物理学のどの分野なのか分割するために、「Category:熱力学」などのカテゴリが作成され、分割されてる。そして、特に分野として大きい場合、例えば歴史といったカテゴリに含まれる「日本の歴史」といったように、「〇〇の”カテゴリ名”」といったような形でカテゴリ名が入ったカテゴリを新たに作り、それをカテゴリに含めることにより分野の分割が行われると考えられる。

つまり、あるカテゴリ C_x について、 C_x のサブカテゴリ群

のカテゴリ名に C_x のカテゴリ名が使われていた場合、 C_x のカテゴリ名はより分野らしい単語であるだろうという仮定である。例えば、「歴史」という単語のような「Category:歴史」に「Category:各国の歴史」、「Category:歴史の人物」といったカテゴリ名「歴史」をサブカテゴリに含むカテゴリを分野らしい単語であるとする。

この手法、カテゴリ分野法では手法1に加えて以上の仮設を用いて単語の得意分野度 W_t を以下のように算出する。カテゴリ C_x について、 C_x のサブカテゴリ群 S に含まれるカテゴリのカテゴリ名に C_x のカテゴリ名 N_x を含むカテゴリがいくつ入っているかという値 b_{C_x} を算出する。

そして単語 t の分類度 P_t を以下のように算出する

$$\begin{cases} P_t = b_{C_x} & (t \text{ と } N_x \text{ が一致した場合}) \\ P_t = 1 & (b_{C_x} = 0 \text{ の場合}) \end{cases} \quad (5)$$

b_{C_x} は前述のようにカテゴリ名がサブカテゴリに入っている数である。単語 t と同じカテゴリ名を持つカテゴリが存在しなかった場合などは分類度 P_t を 1 とする。

その後、編集したページから上へ辿った結果たどり着いたカテゴリに含まれていた単語 t の得意分野度を求める。

$$W_t = P_t \sum_{a=1}^N d_{C_a,t} \quad (6)$$

$\sum_{a=1}^N d_{C_a,t}$ は式 (2) と同様に求める。

例えば「Category:歴史」には「Category:日本の歴史」のような「歴史」という文字が含まれるカテゴリがサブカテゴリに 21 件存在している。よって、この手法での単語「歴史」の得意度は 21×5 となる。

3.6 組合せ

前節までの提案手法を組み合わせた手法も考えられる。手法1はもともと全ての手法に組み込まれているため、手法2,4の組合せは4種類存在する。

それぞれ、単語の得意度を以下のように算出する。

$$W_t = \log \frac{Z}{f_t} D_{C_x} \sum_{a=1}^N d_{C_a,t} \quad \text{手法2,3の組合せ} \quad (7)$$

$$W_t = \log \frac{Z}{f_t} P_t \sum_{a=1}^N d_{C_a,t} \quad \text{手法2,4の組合せ} \quad (8)$$

$$W_t = D_{C_x} P_t \sum_{a=1}^N d_{C_a,t} \quad \text{手法3,4の組合せ} \quad (9)$$

$$W_t = \log \frac{Z}{f_t} D_{C_x} P_t \sum_{a=1}^N d_{C_a,t} \quad \text{手法2,3,4の組合せ} \quad (10)$$

となる。 $d_{C_a,t}$, $\log \frac{Z}{f_t}$, D_{C_x} , P_t は、前述の定義と同様である。

4. 実験

4.1 実験方法

提案手法について検証を行う。本手法の目的は編集者の得意分野を特定することであった。そのため、実際の編集者の得意

分野と提案手法によって算出される得意分野がどの程度一致するのか調べたい。

そこで、以下のような手順で提案手法で特定した単語の順番が編集者の得意分野を特定できているか実験を行った。実験対象となる編集者を節 4.2 の方法で選択する。そして、それぞれの手法を用いて、編集者の得意分野を表す単語とその得意分野度を算出し、単語の順位を算出した。使用した実験データについて説明する。手法によって算出されるのは、編集者ごとの得意分野を表す単語と単語の得意分野度である。本実験では、10 人の編集者に対し、4 つの提案手法それぞれによって単語とその単語の得意分野度を特定する。編集履歴のデータは 1 月 20 日時点のものを使用した。6 人の判定者によって編集者の得意分野を判定し、正解データを作成した。正解データ作成の過程については、4.3 節で詳しく述べる。

編集者が持つ知識というのは得意分野を中心に、関係ないものに行けば行くほど、よく知らないものになっていくと考えられる。つまり、編集者の得意分野というのは、分野とその分野の得意度合で表せるものである。

手法によって算出された単語の得意分野の順番と、判定者によって算出された得意分野の順番が似ていれば似ているほど、その手法は上手く得意分野を表していると考えられる。そこで手法の単語の順番と判定者によって求められた単語の順番を順位相関係数を用いて算出する。

4.2 編集者の選別

以下の基準で実験に使用する編集者 10 人の選別を行った。Wikipedia には論争のあるページというテンプレートが存在している。このテンプレートは編集合戦などが起きたり、ノートページで論争が起きているページにつけられるテンプレートである。論争のあるページでは、特にそのページに詳しい人物が Wikipedia に書く内容などを巡って話し合いを行っていると考えられる。つまり、論争のあるページで議論している編集者は Wikipedia 内の記事において何らかの得意分野が存在し、議論に参加している可能性が高いと考えられる。

そこで、テンプレート「論争のあるページ」が表示されているページを編集したことのある編集者の内、ノートページを 20 記事以上 50 記事以下編集していた編集者からランダムで 10 人選択した。選択された編集者名は、ごまふあぎ、Bokemiann、如月の弥、グローバライ、矢刺幹人、みしまるもも、Naruhodou、泉州大夫惟宗朝臣、Northface、青鬼よしの 10 編集者である。これら 10 人の編集者それぞれに対し、基本的な手法、希少単語法、カテゴリ深度法、カテゴリ分野法、希少単語法とカテゴリ深度法の組合せ、希少単語法とカテゴリ分野法の組合せ、カテゴリ深度法とカテゴリ分野法の組合せ、希少単語法とカテゴリ深度法とカテゴリ分野法の組合せの各 8 つの手法で得意分野の単語の順番を算出し、評価実験を行った。

4.3 正解データ

本節では実験の正解データの作成方法について説明を行う。手法によって選択された単語が実際の得意分野としてふさわしいかどうか判定したい。そこで、得意分野の正解データを以下のように作成した。

表 1 手法によって算出された得意分野を表す単語上位 4 つ

編集者名	基本	希少	深度	分野
ごまふあぎ	日本、時代、別、歴史	日本、時代、史、別	日本、歴史、年、スタブ	日本、人物、世紀、歴史
Bokemiann	宗教、アニメ、学、日本	宗教、戦、キリスト、学	宗教、アニメ、日本、作品	宗教、アニメ、文化、作品
如月の弥	日本、関係、戦争、アメリカ合衆国	関係、戦争、軍事、日本	日本、戦争、年、アメリカ合衆国	戦争、日本、軍事、アメリカ合衆国
グローバライ	スタブ、関連、政治、日本	スタブ、関連、政治、項目	スタブ、日本、交通、建築	政治、日本、交通、社会
矢刺幹人	日本、関係、戦争、史	関係、日本、戦争、史	日本、戦争、年、世紀	日本、戦争、世紀、人物
みしまるもも	日本、作品、小説、別	小説、作品、日本、別	日本、作品、小説、映画	作品、日本、小説、映画
Naruhodou	漫画、作品、アニメ、題材	漫画、作品、アニメ、題材	漫画、作品、アニメ、年	漫画、作品、人物、アニメ
泉州大夫惟宗朝臣	人物、政治、別、外国	政治、人物、別、主義	人物、年、政治、世紀	人物、政治、アジア、世紀
Northface	日本、年、アメリカ合衆国、関係	戦争、日本、関係、アメリカ合衆国	日本、年、戦争、アメリカ合衆国	日本、戦争、世紀、アメリカ合衆国
青鬼よし	日本、人物、別、関係	日本、人物、別、関係	日本、議員、人物、年	人物、日本、アジア、政治

表 2 順位相関係数

編集者名	順位相関係数							
	基本	希少	深度	分野	希+深	希+分	深+分	希+深+分
ごまふあぎ	-0.318	-0.121	-0.066	-0.203	0.162	0.022	-0.049	0.162
Bokemiann	0.112	0.091	0.303	0.153	0.260	0.112	0.261	0.221
如月の弥	0.476	0.587	0.571	0.575	0.669	0.646	0.575	0.651
グローバライ	0.063	-0.081	0.267	0.324	0.261	0.326	0.383	0.366
矢刺幹人	0.228	0.338	0.317	0.372	0.387	0.451	0.393	0.452
みしまるもも	0.324	0.466	0.580	0.490	0.643	0.586	0.590	0.657
Naruhodou	0.180	0.242	0.440	0.348	0.545	0.454	0.425	0.511
泉州大夫惟宗朝臣	0.445	0.539	0.542	0.539	0.586	0.628	0.517	0.547
Northface	0.676	0.748	0.600	0.616	0.676	0.720	0.587	0.661
青鬼よし	0.042	-0.051	0.236	0.091	0.175	0.121	0.141	0.124
全体	0.223	0.276	0.379	0.331	0.436	0.407	0.382	0.435

まず、基本的な手法（3.2 節）を用いて、得意分野となりえる単語を取得した。続いてその中から評価用の単語を抜き出す。評価用の単語は取得した単語上位 100 単語の内、ランダムに 20 単語選択した。6 人の判定者によって単語を編集者が得意としている分野である順に並び替える。判定者それぞれが並び変えた単語ごとに順位の平均を求め、20 単語の得意分野である順番を作成し、正解データとした。また、得意分野を並び替える際の評価基準としては、「その単語が表す内容について知識を持っていそうか（その単語の内容に関する編集を行っているか）」ということを目安に判定を行ってもらった。

4.4 結果

選択した 10 人の編集者に基本的な手法、希少単語法、カテゴリ深度法、カテゴリ分野法を使用した場合、得意分野度の高い上位三つの単語は表 1 のようになった。基本的な手法によって特定された単語の中で、得意分野を表す単語ではないと判定されたものは「別」や「史」など 1 文字であるものが多かった。これは、カテゴリを単語に分解したことにより、単体ではあまり意味を持たないことになってしまったためであると考えられる。カテゴリ分野法（表 1 の「分野」の列）によって算出された得意分野を表す単語にはそのような単語の得意分野度を相対的に低くすることができていると考えられる。

10 人の編集者の相関は表 2 のようになった。表 2 の「全体」は 10 人の編集者の結果を同順位のある順位相関係数を用いてまとめたものである。表 2 の「全体」の行を見ると、単体の手法ではカテゴリ深度法が一番高くカテゴリ分野法が 2 番目に高い。手法を組み合わせた手法も含めると希少単語砲とカテゴリ深度法を組み合わせた手法が一番相関係数が高い手法である。

4.5 考察

本実験により単体の手法ではカテゴリ深度法とカテゴリ分野法が、他の手法と比較して高い相関を示す結果となった。これは、Wikipedia のカテゴリを単語に分解した場合、年や史など、1 単語ではあまり分野としての意味を持たない単語が多く含まれていたからであると考えられる。そのため、カテゴリ名になっていない年や史の重みを相対的に下げることのできるカ

テグロ深度法とカテゴリ分野法の結果が良い物となったと考えられる。より得意分野を正確に表すための改良としては、カテゴリの網羅率、すなわち、あるカテゴリに含まれる記事をどの程度編集していたか、という値を使用することにより、より得意分野特定の精度を上げることができると考えられる。

Wikipedia の編集者には記事を編集しているが、編集内容が内容に関わるのではなく、簡単な追記やカテゴリの追加にとどまる編集者も存在する。ノートページの編集に関しても内容ではなく、無署名の議論に署名を追記する、ログを整理するといったような内容に関わらない編集を行う編集者が存在する。そういった編集者が編集した記事と真にその記事、分野に詳しい編集者を本手法では区別できない。そのため、同じ手法でも相関係数に大きく差が出たのではないかと考えられる。

5. おわりに

本稿では、Wikipedia 編集者の得意分野の特定に取り組んだ。Wikipedia 編集者の得意分野特定を用いることにより、Wikipedia の記事をより良くすることに役立つと考えられる。例えば、Wikipedia 編集者の得意分野を特定することにより、存在することが望ましいが作成されていない記事、追記が必要な記事を編集可能な編集者の探索を容易にすることが可能であると考えられる。また、得意分野が似た編集者を効率的に探し、それら編集者を協力させることにより、より専門的な記事の作成に役立てることなどもできると考えられる。Wikipedia において編集することなく埋もれてしまっている編集者に、得意分野の記事を書いてもらうことができれば、Wikipedia の百科事典としての質を高めることができると考えられる。また、Wikipedia 外において、知らないこと、知りたいことが発生した場合、本手法を用いてその分野のスペシャリストを探し、直接質問をするなどの利用法が考えられる。

提案手法では編集者の得意分野を特定するために、Wikipedia カテゴリと編集者の投稿記録を用いた。得意分野を特定したい編集者のノートページへの投稿記録を取得、カテゴリ構造を利用し、得意分野の特定を行った。また、より細かい分野を特定するために Wikipedia のカテゴリ名に注目し、分野として相応しい単語の特定を行った。実験の結果、得意分野の特定に与える影響としてはカテゴリ名として成立している単語の影響が大きいと考えられる。

今回、一つのノートページから辿り着いたカテゴリも複数のノートページから辿りついたカテゴリも同様の扱い方をした。しかし、複数の記事、ノートページから辿り着いたカテゴリは得意分野特定の際により重要であると考えられる。また、カテゴリに含まれる記事の内、編集者が編集した記事の割合を考慮することにより、そのカテゴリについてどの程度の知識を持っているか判断することができると考えられる。そのようにカテゴリの重みづけを改良することも考えられる。

本手法では、編集したことがある記事を同列に扱った。編集の残存度などを用いて記事ごとに編集者の質を算出する方法を用いることにより、より正確に得意分野を特定することができるのではないかと考えられる。

Wikipedia 編集者の得意分野を表すにあたって、本手法の出力は単語とその単語が編集者にとってどれくらい得意かという順位であった。そのため、出力全体として見れば、編集者が何について編集しているか理解できるものの、実験での判定などの際に単語一つなど部分として見た場合、その単語が何を意味しているか分野としてわかりにくいものとなった。編集者の得意分野を視覚的に表す方法や、出力された単語に Wikipedia のページを関連付けるなどで、編集者の得意分野を理解できるものになると考えられる。

謝 辞

本研究の一部は科研費 25280039,23700113 の助成による。

文 献

- [1] 白川真澄, 中山浩太郎, 原隆浩, 西尾章治郎. Wikipedia のカテゴリグラフ解析による語句の確率的分類とその応用. 情報処理学会論文誌データベース (TOD), Vol. 5, No. 3, pp. 51–63, sep 2012.
- [2] 中谷誠, Adam Jatowt, 大島裕明, 田中克己. Wikipedia のリンク構造とカテゴリ構造を用いた検索語からの専門語の抽出. 情報処理学会研究報告データベースシステム (DBS), Vol. 2008, No. 88, pp. 253–258, sep 2008.
- [3] Peng Jiang, Huiman Hou, Lijiang Chen, Shimin Chen, Conglei Yao, Chengkai Li, and Min Wang. Wiki3c: Exploiting wikipedia for context-aware concept categorization. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13*, pp. 345–354, New York, NY, USA, 2013. ACM.
- [4] Peter Schonhofen. Identifying document topics using the wikipedia category network. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, WI '06*, pp. 456–462, Washington, DC, USA, 2006. IEEE Computer Society.
- [5] Pavan Kapanipathi, Prateek Jain, Chitra Venkataramani, and Amit Sheth. Hierarchical interest graph from tweets. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pp. 311–312. International World Wide Web Conferences Steering Committee, 2014.
- [6] 胡寅駿, 谷田泰郎. Wikipedia のカテゴリ情報を用いた twitter ユーザの関心分野の抽出 (評判・評価, 第 3 回テキストマイニング・シンポジウム). 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 113, No. 213, pp. 17–21, sep 2013.
- [7] V Subramaniaswamy and S Chenthur Pandian. Effective tag recommendation system based on topic ontology using wikipedia and wordnet. *International Journal of Intelligent Systems*, Vol. 27, No. 12, pp. 1034–1048, 2012.
- [8] 林幸記, 江口浩二. カテゴリ構造を用いた確率的トピックモデルの効率的推定とその応用. Technical Report 5, 神戸大学大学院工学研究科情報知能学専攻, 神戸大学大学院工学研究科情報知能学専攻, jul 2009.
- [9] Robert West, Ingmar Weber, and Carlos Castillo. Drawing a data-driven portrait of wikipedia editors. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, WikiSym '12*, pp. 3:1–3:10, New York, NY, USA, 2012. ACM.
- [10] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference, iConference '11*, pp. 122–129, New York, NY, USA, 2011. ACM.