

Wikipedia のリンク構造を利用した関係性ラベルの抽出手法

瀬戸口 司[†] 浅野 泰仁[†] 吉川 正俊[†]

[†] 京都大学大学院情報学研究科 〒606-8501 京都市左京区吉田本町 36-1

E-mail: [†]setoguchi@db.soc.i.kyoto-u.ac.jp, ^{††}{asano,yoshikawa}@i.kyoto-u.ac.jp

あらまし Wikipedia の好ましい特徴として、現実世界の一つの概念や対象が Wikipedia の一つのページに対応している点と、ページ間の関連がリンクとして表現されている点が挙げられる。これらの特徴から、単語間の関連強度を求める研究などで、Wikipedia は情報源としてたびたび利用されてきた。本研究では Wikipedia のリンク構造およびカテゴリ構造を利用して対象と対象の関係性を表すラベルを取得する手法を提案する。提案手法では、まずページ内のリンクをクラス分けし、次に各クラスを特徴づける単語を抽出した上で、これをページとリンク先との関係性のラベルとする。

キーワード Wikipedia, リンクマイニング, 意味抽出, セマンティックウェブ

Tsukasa SETOGUCHI[†], Yasuhito ASANO[†], and Masatoshi YOSHIKAWA[†]

[†] Graduate School of Informatics, Kyoto University

36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 JAPAN

E-mail: [†]setoguchi@db.soc.i.kyoto-u.ac.jp, ^{††}{asano,yoshikawa}@i.kyoto-u.ac.jp

1. はじめに

現代はインターネットが普及し、人が様々な情報にアクセスできる時代である。インターネット上には膨大な量の情報が存在しているので、自分が求める情報を効率的に探し出す検索の技術が重要である。

現在の Web 検索システムは、クエリ（あるいはその同義語）を含むページを返す。例えば「カレーに牛肉を何グラム入れるべきか答えよ」という課題が出題された時、ユーザは Google などの検索システムに「カレー 牛肉 グラム」と入力すれば、楽に正解が書かれたウェブページを得ることができる。これはクエリによって十分限定されたページのリストが返され、ここから正解を探すのにかかる労力が少なくて済むということである。

しかし調べるべき対象の名称が明確ではない場合（「オムレツにかかっているこの黒いソース」）や、そもそも調べるべき情報が未定の場合（「カレーについて 2000 字のレポートを書きたい」）などは、検索によって正解を得ることが難しい。このような状況でユーザは分かっているクエリだけからページリストを取得し、膨大なページリストの中から次に投入すべきクエリを手動で探すことを余儀なくされる。この時、ページランク [3] などの手法で検索順位が決定するシステムであれば、ページランクが高くなる傾向にあるトピックばかりがユーザの目に触れる。もしユーザの求める情報がマイナーな情報であれば、ユーザは目的の情報を発見するまでに相当の労力を支払う必要があ

るだろう。

ある対象についての様々な側面の知識を得たい時、人はしばしば百科事典を用いる。Web 上で利用可能な百科事典として、Wikipedia [1] が存在している。これは日本語版だけで約 80 万件の記事を掲載する大規模な百科事典である。しかしページの記述自体が膨大であることも少なくなく、例えば【日本】についての記事は約 9 万文字を含むなど、Wikipedia を活用すること自体がユーザにとって負担であることも少なくない。

本研究ではブラウジングによる知識獲得を支援する方法として、概念間の繋がり関係性をラベル付けすればよいと考えた。関係性ラベルとは二者間の関係を表す単語のことで、例えばオムレツに対してのケチャップは「ソース」であり、ケチャップに対してのオムレツは「料理」である。オムレツにかけるソースを一覧提示すれば、「オムレツにかかっているこの黒いソース」の名称をユーザが探すことは容易となるだろう。またカレーと関係を持つ概念の関係性ラベルを一覧提示すれば、カレーという概念の全ての側面が可視化され、「カレーについて 2000 字のレポートを書きたい」と思っているユーザに、カレーの何について書けばいいかのヒントを与えることができる。

関係性ラベルを取得する際の情報源として、本研究では Wikipedia を用いる。現実世界の一つの概念や対象が Wikipedia の一つのページに対応している点や、ページ間の関連がリンクとして表現されている点など、Wikipedia には本研究にとって好ましい特徴を持っているからである。提案手法ではまずある

ページ内のリンク先のページを所属するカテゴリを利用してクラス分けし、次に各クラスを特徴づけるカテゴリの名称を、ページとページの関係性を表すラベルとする。

本研究ではページ集合のカテゴリへの所属をカテゴリハッシュなる形式で表現し、ページ集合をまとめて1つの集合とする操作をカテゴリハッシュの積として表現した。また、ページのクラス分けがどれほど妥当かを表す指標として、カテゴリハッシュの score を定義した。さらにこれらを用いて関係性ラベルを取得する「草むしり法」なるアルゴリズムを検討した。これらを用いて、概念とそれに関係する対象との関係性ラベルを表示するシステムを試作した (図 1)。このシステムは部分的には目的を達成しているものの、より良いアルゴリズムの検討やカテゴリ的な通常ページの利用などを今後の課題とする。なお今回の試作システムを <http://smokycat.info:3000/> で公開しているので、実際に動かしていただければ幸いである。

オムレツ	GO
オムレツと関係のある対象	
ソース	ドミグラスソース クリームソース ケチャップ ウスターソース
卵料理	鶏卵 トルティージャ 卵焼き
洋食	チキンライス オムライス
日本の米料理	
共和国	フランス 中華人民共和国
乳製品	クリーム (食品) パター
スプレッド	
その他	スポンジケーキ タマネギ 日本 食品 ハム 餃子 果実 食文化 塩 焼 きそば 品川工業所 チーズ カレー コショウ フライパン モン・サン＝ミシェル

図 1 本研究の試作システム

以下本稿では、第 2 章では関連研究について述べる。第 3 章では提案手法について述べる。第 4 章では作成した試作システムの挙動を述べる。第 5 章で今後の課題について述べる。

2. 関連研究

Wikipedia は概念間の意味を抽出するための情報源として好ましい特徴を持つ。それは現実世界の一つの概念が Wikipedia の一つのページに対応している点や、ページ間の関連がリンクとして表現されている点などである。これらの特徴ゆえに、Wikipedia のページのリンク構造を解析することで、間接的に概念の繋がりを解析することができる。そのため Wikipedia はしばしば研究の題材として利用されてきた。

Milne [4] は Wikipedia のリンク構造を用いて、単語間の関連強度を推測する研究を行った。Milne の手法ではまずページを、そこに含まれるリンクを次元とするベクトル空間にプロットする。ページとページの間連強度は、ベクトルとベクトルのコサイン相関によって求める。そしてページ間の関連強度を、それに対応する 2 単語間の関連強度として用いる。これは「似たようなページは、同じページにリンクしている可能性が高い」という、共参照を用いた関連強度の算出方法である。ベクトル空間モデルの不要な次元を削除する方法としては LSA [5] などがあるが、Milne の手法も次元を選ぶための手法であると考えられる。Wikipedia にページが存在しているということは、その概念は一定以上の重要性を持つということである。またペー

ジからページにリンクが貼られているということは、それらの概念間に一定以上の関係性があるということである。Milne の手法はこれらの事実を用いてベクトル空間モデルの次元に用いる素性データを選択する手法である。

Milne の研究ではもうひとつの課題に取り組んでいる。それは曖昧な語の解決である。例えば plane という単語があった時、それが飛行機 (air plane) のことなのか、平面 (plane) のことなのか、大工道具のカンナ (plane) のことなのか曖昧さが残り、単語と Wikipedia のページを機械的に対応付けることが困難である。しかし plane と sky の関連強度を問われた時、通常は「飛行機と空」のことであり、「大工道具と空」のことでないと人は考える。このような人間の性質を捉え Milne は、単語と対応しうる全ての Wikipedia のページをベクトル化し、最も相関値が強くなる組み合わせを最終的に選ぶという手法を提案している。

より「意味」に着目した研究としては、Zhang [6] らの研究がある。Zhang は Wikipedia のリンク構造を用いて、二つの概念の繋がりをよく説明する概念を抽出する研究を行った。Wikipedia のページはある概念に相当しており、Wikipedia のリンクは概念間の繋がりを表していると考えられる。ページ A と B の繋がりは、A からリンクをたどって B に至る経路が多ければ多いほど強くなる。しかしリンクを経由する数が多くなると、関係性は弱まっていく。Zhang の研究ではこの特徴を減衰流 (Generalized Maximum Flow) を利用して表現し、あるページからあるページのフローに大きく寄与した中間ページを、関係性を説明するページとして抽出する。

リンク構造を持つデータの繋がりを可視化することはユーザにとって有用である。Wikipedia 以外にもこのようなデータは多々あり、それを可視化するシステムもいくつか存在している。TwitterGraph [7] は twitter アカウントのフォロー関係を利用し、クラスタを形成してユーザに提示する。あの人検索スパイシー [8] は WEB 上から個人の実名に関する情報を収集・蓄積し、これらの情報をもとに人脈を推定し、提示する。

3. 提案手法

この章では提案手法について述べる。提案手法では、まずあるページからリンクされているページの集合を、それらの所属するカテゴリ構造を利用してクラス分けする。次に各クラスのページが共通して所属するカテゴリのカテゴリ名を、リンクの関係性のラベルとする。

まず第 1 節で Wikipedia のリンク構造の概要を述べる。第 2 節でカテゴリ構造をどのように利用するかについて述べる。第 3 節で本研究で用いる概念の定義を行う。第 4 節でクラス分けのアルゴリズムについて述べる。

3.1 Wikipedia の構造

Wikipedia には数多くのページが存在しているが、これには数種類あり、データ上は namespace という属性で区別されている。本研究では通常ページ (namespace 0) とカテゴリページ (namespace 14) を利用する。通常ページとはある対象について説明しているページである。カテゴリページとは通常ページ、

およびカテゴリページを分類するためのページである。以下、本稿では通常ページを単にページと呼び、カテゴリページを単にカテゴリと呼ぶ。また実際のページを表す時は、通常ページを【オムレツ】のように表し、カテゴリページを【Category:卵料理】のように表す。

カテゴリは階層的に作成することができ、また一つのページが複数のカテゴリに属することも可能である。例えば【オムレツ】は【Category:卵料理】および【Category:フランス語の語句】に属し、【Category:卵料理】は【Category:卵】および【Category:料理】のサブカテゴリである。カテゴリは複数カテゴリに所属でき、かつ閉路も稀ながら存在する。カテゴリというと通常は木構造を思い浮かべるが、Wikipedia のカテゴリ構造は一般の有向グラフである。

データ保持のルール上は、ページのリンク構造とカテゴリの所属構造の間に違いはない。また実際に【日本のゴルフ場一覧】のように、極めてカテゴリ的な通常ページも存在している。とはいえほとんどの場合は、通常ページには対象についての説明が記述され、カテゴリページは通常ページを分類するために編集されている。

以下本稿では、通常ページがカテゴリページに属することを、階層的な所属も含めて単に所属と呼ぶ。通常ページからカテゴリページに至るまでにいくつの所属関係を経由する必要があるかをホップ数とよび、 n ホップというように表記する。

本研究ではページ間の関係性を説明するラベルの抽出に、Wikipedia のカテゴリ構造を利用する。

3.2 カテゴリ構造の利用方法

カテゴリとカテゴリの関係性は限定的である。例えば【Category:料理】に属するページから【Category:調味料】に属するページにリンクが存在するとき、両者の関係は「この調味料はこの料理の材料である」という関係に限られる。ゆえにリンク先のページの所属するカテゴリのカテゴリ名を関係性ラベルとすれば、人は両者の関係性を容易に推測することができる。

ただし関係性ラベルとして用いるカテゴリは適切に選ぶ必要がある。例えば【塩】は【Category:調味料】や【Category:塩化ナトリウム】に属している。確かに塩には調味料としての側面と化学物質としての側面がある。しかし料理にとっての【塩】は調味料であり、化学物質としての塩とはあまり関係がない。

Wikipedia は百科事典であり、網羅的な説明的記述が目指されている。この特徴を利用すると、関係性ラベルとして用いるカテゴリを適切に選択することができる。例えば【焼き鳥】のページには【サンショウ】、【唐辛子】、【タレ】など、【Category:日本の調味料】に 1 ホップで所属するページへのリンクが 13 個存在する。化学物質については、【唐辛子】→【Category:香辛料】→【Category:香料】→【Category:化学物質】というように遠まわしに化学物質カテゴリに所属するページへのリンクが 9 個含まれる。【焼き鳥】のページからリンクされているページに対して【Category:日本の調味料】と【Category:化学物質】のどちらがより支配的かというと、これは前者であると考えるのが自然であり、また現実にも即している。ゆえに本研究では、あるページとリンク先のページの関係性のラベルを定

める時、リンク先のページの集合全体に対してより大きな支配力を持つカテゴリの名を優先的に用いる。

以上をまとめると、あるページからリンクされているページを似たもの同士クラス分けし、各クラスを特徴付けるカテゴリ名を関係性ラベルとして抽出することが本研究の具体的なゴールとなる。

3.3 本研究の諸概念の定義

この節では本研究に用いる諸概念の定義を行う。第 1 項でページが所属するカテゴリを表現する形式（カテゴリハッシュ）の具体例を示し、第 2 項でそれを厳密に定義する。第 3 項でカテゴリハッシュ同士の積を定義する。第 4 項でページのクラス分けの妥当性を示す指標として、カテゴリハッシュを指数に取る関数 score を定義する。

3.3.1 カテゴリハッシュの具体例

Wikipedia のカテゴリ構造は単純な木構造ではなく、一般の有向グラフである。あるページがあるカテゴリに属する度合いをどのように算出するかは一考の余地があるが、本研究ではそのページからそのカテゴリに到達するまでの最小のホップ数を指標とする。最小ホップ数が小さければ小さいほど、そのカテゴリとの関係が深いと考える。あるページから、そのページが所属する複数のカテゴリへの最小のホップ数を、本研究ではハッシュとして保持し、これをカテゴリハッシュと呼ぶことにする。カテゴリハッシュはカテゴリ名を指数に取り、最小ホップ数を返す一方向関数である。

具体例としてページ【卵】のカテゴリハッシュを作成する。まず【卵】が 3 ホップ以内で所属するカテゴリは、表 1 のようになっている。【卵】→【Category:卵料理】→【Category:料理】→【Category:食文化】と【卵】→【Category:畜産物】→【Category:食品】→【Category:食文化】など、一つのカテゴリへの複数の経路が存在することが確認できる。

表 1 【卵から 3 ホップのカテゴリ（最小ホップ以外も含む）】

ページ	1 ホップ	2 ホップ	3 ホップ
卵	卵	生殖	生物学
卵	卵	生殖	生理学
卵	卵	発生生物学	動物学
卵	卵	発生生物学	基礎医学
卵	卵	発生生物学	植物学
卵	卵	発生生物学	生物学
卵	卵料理	卵	生殖
卵	卵料理	卵	発生生物学
卵	卵料理	料理	調理
卵	卵料理	料理	食品
卵	卵料理	料理	食文化
卵	畜産物	畜産	農業
卵	畜産物	食品	栄養学
卵	畜産物	食品	資源
卵	畜産物	食品	食文化

このカテゴリ構造をグラフで示すと図 2 のようになる。カテゴリハッシュはページからあるカテゴリへの最小ホップ数を保持するものであるため、不要な経路は無視しなければならない。

この操作は、グラフ上では冗長な枝を開放除去してページを根とする木構造を作成することに等しい（図 2）。

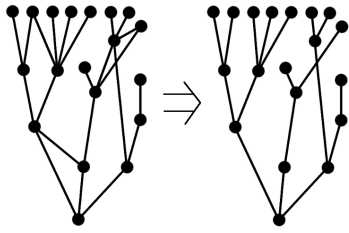


図 2 冗長経路の削除

実装上は SQL の INNER JOIN を用いて数ホップ先までの全ての経路を求め、これを幅優先探索してページからカテゴリへの最小ホップ数を調べる。【卵】のカテゴリハッシュを求める、表 2 のようになった。

表 2 【卵】のカテゴリハッシュ

最小ホップ数	カテゴリ名
1	卵料理, 卵, 畜産物
2	生殖, 畜産, 発生生物学, 食品, 料理
3	基礎医学, 資源, 食文化, 調理, 栄養学, 農業, 生物学, 生理学, 動物学, 植物学
4	農林水産業, 医学, テーマ別の文化, 生命科学, 動物, 生活, 技術, 健康, 環境, 食科学, フードビジネス, 家事, 生化学, 理学, 経済, 植物, 自然科学
5	人間, テーマ別に分類したカテゴリ, 医療, 化学, 生活科学, 自然, 生態学, 社会, 科学, 労働, 産業, 主要カテゴリ, 学際領域, 有機化学, 家庭, 生命, 応用科学, 生物, 人体, 健康科学, 学問の分野, 文化
...	...

カテゴリ構造を上がっていくと、より包括的なカテゴリが現れる。【卵】の例では 4 ホップの時点で【Category:理学】や【Category:経済】などのカテゴリに属している。このような情報をページのクラス分けに利用したとしても、ユーザにとって有益なクラス分けとはならないだろう。また、調べるホップ数を増やせば計算すべき情報が増え、システムがユーザにとって快適な時間内に応答できなくなる。これらの理由から、実用上は 3 ホップ先までのカテゴリハッシュを用いれば十分であると考えられる。

3.3.2 カテゴリハッシュの定義

カテゴリハッシュには、それがどのページ集合に対するものかという属性と、カテゴリの所属関係を何ホップ先まで用いるかという 2 つの属性がある（ページ集合とした理由については次項で述べる）。これをそれぞれ下添字と上添字で表す。

カテゴリハッシュに対応するページ集合の元が単一のページ p である時、ホップ数の上限を l とし、これに対応するカテゴリハッシュ $H_{\{p\}}^l$ がカテゴリ C というキーに対して返す値を次のように定義する。

$$H_{\{p\}}^l(C) = \begin{cases} n & (p \text{ から } C \text{ に } n \text{ ホップで到達可能} \wedge n \leq l) \\ \infty & (\text{otherwise}) \end{cases}$$

例えば【鶏卵】の 3 ホップ先までのカテゴリハッシュは表 3 のようになる。また【トルティージャ】の 3 ホップ先までのカテゴリハッシュは表 4 のようになる。

表 3 $H_{\{\text{鶏卵}\}}^3$

最小ホップ数	カテゴリ名
1	卵, 畜産物, 卵料理, 鶏
2	生殖, キジ科, 畜産, 家禽, 発生生物学, 食品, 料理
3	農業, 鳥類, キジ目, 生物学, 家畜, 生理学, 基礎医学, 走禽類, 資源, 食文化, 植物学, 動物学, 調理, 栄養学

表 4 $H_{\{\text{トルティージャ}\}}^3$

最小ホップ数	カテゴリ名
1	スペインの食文化, キューバの食文化, 卵料理
2	卵, 各国の食文化, 地中海の食文化, スペインの文化, 中南米の食文化, キューバの文化, 料理, ヨーロッパの食文化
3	生殖, スペイン, 中近東の食文化, 地中海, 大陸別の食文化, 地域別の食文化, 発生生物学, キューバ, アフリカの食文化, 食文化, 食品, 調理, 各国の文化, ラテンアメリカ, ヨーロッパの文化

3.3.3 カテゴリハッシュの積

本研究の目的はページとページの関係性を表すラベルを抽出することである。本研究では、似た関係性であろうページをクラスにまとめ、それらに共通したカテゴリのカテゴリ名をラベルとする。ここで、クラス内のページが共通して所属するカテゴリを求める必要がある。

ページ p_1 と p_2 が存在し、 p_1 がカテゴリ C に所属するが p_2 は C に所属しないとき、ページの集合 $\{p_1, p_2\}$ は C には所属しないと考えられる。また p_1 が 1 ホップで C に所属し、 p_2 が 2 ホップで C に所属するとき、 $\{p_1, p_2\}$ が C に所属する度合いはより弱い方、即ち 2 ホップであることが妥当であると考えられる。このような「ページの集合のカテゴリへの所属の度合い」を表現するために、カテゴリハッシュの積を以下のよう定義する。

$$H_{P_1}^{l_1} \cdot H_{P_2}^{l_2} = H_{P_1 \cup P_2}^{\min\{l_1, l_2\}}$$

$$H_{P_1 \cup P_2}^l(C) = \max\{H_{P_1}^l(C), H_{P_2}^l(C)\}$$

具体例として、【鶏卵】（表 3）と【トルティージャ】（表 4）のカテゴリハッシュの積を表 5 に示す。【鶏卵】と【トルティージャ】に共通する属性が抜き出されていることが確認できる。

表 5 $H^3_{\{\text{鶏卵, トルティージャ}\}}$

最小ホップ数	カテゴリ名
1	卵料理
2	料理, 卵
3	食品, 発生生物学, 生殖, 調理, 食文化

3.3.4 カテゴリハッシュの score

前項ではカテゴリハッシュの積を定義した。カテゴリハッシュ同士をかけ合わせるということは、それが担当する 2 つのページ集合を合併して 1 つのページ集合とすることに等しい。本研究では、同じクラスに属するページは、リンク元のページから見て全て同じ関係性にあることになる。適切に関係性を抽出するためには、適切に合併を行なってページをクラス分けする必要がある。ここで、ページの集合がクラスとしてどれだけ妥当かを示す指標が必要となる。

ページ集合 P の元である各ページが似ていれば似ているほど、それらは共通のカテゴリに所属する機会が多くなり、 P のカテゴリハッシュは多くのカテゴリを含むと考えられる。ゆえにカテゴリハッシュの中身の多さをクラスの妥当性として用いることができる。ただしページからのホップ数の大きい、つまり上位のカテゴリであれば、そこにページが所属する機会が増える。ゆえに下位のカテゴリに所属するページはよく似ているが、上位のカテゴリに所属するページはそこまでは似ていない。

これらの特徴を踏まえ、本研究ではクラス P の妥当性をそのカテゴリハッシュ H_P^l を用いて以下のように定義する。ホップ数の上限 l はグローバルな定数として別途定める。以降これをクラスの score 値と呼ぶ。

$$\text{score}(H_P^l) = \sum_C \frac{1}{2^{H_P^l(C)}}$$

3.4 草むしり法

本研究の具体的なゴールは、score 値があまり減少しない組み合わせでカテゴリハッシュをかけ合わせて行き、少数のカテゴリハッシュを得ることである。このような処理は例えば、ハッシュをかけ合わせた時の score 値の減少率をハッシュ間の距離とする空間を考え、その上で最短距離法による階層的クラスタリングを行えば実現できると考えられる。

あるページ p に含まれるリンクの数を δ^+p と表すとすると、階層的クラスタリングを行うには、 $O((\delta^+p)^3)$ の時間計算量がかかる。Wikipedia のページに含まれるリンク数は平均 61 個であり、確かにこの方法で計算できるページも存在する。ところが Wikipedia には数多くのリンクを含むページが存在する。例えば【日本】は 2284 のページへのリンクを含む。後述する試作システムでも階層的クラスタリングを試みてみたが、このように多くのリンクを含むページについてはシステムが実用的な時間で応答しなかった。効率的なクラスタリングアルゴリズムの検討は今後の課題とする。

本稿では近似的であるが高速なアルゴリズムを用いる。イメージとしてはカテゴリハッシュは上に向かって枝が複数伸びた草のような形をしており、このアルゴリズムはあたかもカテ

グリハッシュの草を効率的にむしるような振る舞いをする。ことから、以降このアルゴリズムを草むしり法と呼ぶことにする (アルゴリズム 1)。

草むしり法ではまずリンク先のページ集合 $\mathcal{P}$ について、各 $p \in \mathcal{P}$ が 1 ホップで所属するカテゴリから最も出現頻度が高いカテゴリ C を選択する。そして C に 1 ホップで所属するページを合併し、1 つのクラスとする。この時に、 C のカテゴリ名を関係性ラベルとして用いる。クラス分けされたページを $\mathcal{P}$ から取り除き、この操作を繰り返す。これらの操作はカテゴリ C が $\mathcal{P}$ について最も影響力があり、 C に属するページ同士を合併した時に最も score 値の減少が抑えられるだろう、という推測に基づいている。

アルゴリズム 1 草むしり法

$\mathcal{P} \leftarrow \{p \mid p \text{ は現在のページからリンクされているページ}\}$

while $\mathcal{P} \neq \emptyset$ **do**

$\mathcal{H} \leftarrow \{H_p^1 \mid p \in \mathcal{P}\}$

if $\mathcal{H} = \emptyset$ **then**

break

end if

$C \leftarrow \mathcal{H}$ に最も多く含まれるカテゴリ

$\mathcal{Q} \leftarrow \{p \mid p \in \mathcal{P} \wedge H_p^1 \text{ が } C \text{ を含む}\}$

if $\mathcal{Q}$ の要素数 ≤ 1 **then**

break

end if

$\mathcal{Q}$ の全ての要素が所属するカテゴリを各 $q \in \mathcal{Q}$ のラベルとする

$\mathcal{P} = \mathcal{P} - \mathcal{Q}$

end while

各 $p \in \mathcal{P}$ に「その他」とラベルを付ける

4. 実 験

4.1 利用したデータ

Wikipedia は掲載記事に GNU GPL を適用しており、このライセンスのもとで誰もがデータを再利用することができる。本研究ではこれを利用し、試作システムを作成した。本稿では日本語版 Wikipedia の 2012 年 6 月 30 日のダンプファイル [2] を利用した。

この節では以下、ダンプファイルから本研究に有用なデータを抽出する手順の概要を記す。本研究では公開されている SQL ダンプのうち、redirect, page, pagelinks, categorylinks テーブルを利用する。

redirect とは同義語を解決するため等に用いられるテーブルで、約 51 万件のリダイレクトが登録されている。リダイレクトが設定されたページはユーザーに閲覧されることが無くなるが、データとしては削除されずに残っていることがある。これはノイズであるため、リダイレクトが設定されたページからのリンクは削除する。リダイレクトが設定されたページへのリンクが存在することがあるが、これはリダイレクト先へのリンクに変換する。

page は全種類のページを保持しているテーブルであり、約 220 万件のデータがある。このうち通常ページは namespace=0

なる属性を持っており、約 130 万件のデータがある。ここからリダイレクトを解消すると、約 80 万件となる。

全種類のページ間のリンクは pagelinks テーブルに保持されている。このテーブルには 7400 万件のデータが登録されている。このうち通常ページ間のリンクは約 5600 万件である。ここからリダイレクトを解消し、存在しないページへのリンクを削除すると、約 5000 万件となる。各ページは平均して 61 個のリンクを含む。

カテゴリは page テーブルのデータ上は namespace=14 なる属性を持っているページであり、約 10 万件のデータがある。category には「スタブカテゴリ」と呼ばれる、編集上有用なカテゴリが存在する。しかしこのカテゴリは Wikipedia を通常利用する場合には隠蔽されており、意味抽出の目的の上でもノイズである。スタブカテゴリは「隠しカテゴリ」に属することを利用し、本研究では全てのスタブカテゴリを削除する。ここでは約 1000 件のスタブカテゴリを削除した。

通常ページおよびカテゴリが上位のカテゴリに 1 ホップで所属する関係は categorylinks テーブルに保持されている。

ページがカテゴリに 1 ホップで所属する関係は約 355 万件登録されている。ここからリダイレクトを解消し、存在しないページを含むリンクを削除すると、約 350 万件となる。さらにスタブカテゴリへの所属関係を削除すると約 290 万件となる。各ページは平均 3.5 個のカテゴリに 1 ホップで所属する。

カテゴリ間の 1 ホップでの所属関係は約 23 万件登録されている。ここからリダイレクトや存在しないカテゴリを含む関係を約 2000 件を削除する。さらにスタブカテゴリを含む関係を約 5000 件削除する。各カテゴリは平均 2.2 個のカテゴリに 1 ホップで所属する。

以上の処理結果を、3 つのテーブルにまとめる。ページからページへのリンクを保持するテーブル、ページからカテゴリへの 1 ホップでの所属を保持するテーブル、カテゴリからカテゴリへの 1 ホップでの所属を保持するテーブルである。これらが本研究が必要とするリンク構造の全てである。

4.2 試作システム

前節で作成したデータを草むしり法によって解析し、関係性ラベルを取得するシステムを実装した。このシステムによると、例えば【オムレツ】とリンク先ページとの関係は表 6 のようにラベル付けされる。

表 6 卵と関係する概念とその関係性ラベル

関係性ラベル	リンク先ページ
ソース	ドミグラスソース, ケチャップ, ウスターソース, クリームソース
卵料理	卵焼き, トルティージャ, 鶏卵
洋食日本の米料理	オムライス, チキンライス
共和国	中華人民共和国, フランス
乳製品スプレッド	バター, クリーム (食品)
その他	果実, 日本, ハム, タマネギ, フライパン, スポンジケーキ, チーズ, 食文化, コショウ, 塩, 焼きそば, 餃子, モン・サン＝ミシェル, 品川工業所, カレー, 食品,

この例ではオムレツとソースの関係がうまくラベル付けされている。これら 4 つのソースのページは全て、1 ホップで【Category:ソース】に所属するため、このように抽出できた。

【塩】や【コショウ】は「調味料」とラベル付けされることが理想であるが、今回の実験ではどちらも「その他」とラベル付けされてしまった。これは【コショウ】が 1 ホップで【Category:調味料】に所属せず、【コショウ】→【Category:香辛料】→【Category:調味料】のように 2 ホップで所属するためである。【塩】と【コショウ】を 1 つのクラスとして【Category:調味料】でラベル付けするためにはアルゴリズムを改善する必要がある。

もう一つ実例を挙げる。【カレー】とそのリンク先ページとの関係は表 7 のようにラベル付けされる。

表 7 カレーと関係する概念とその関係性ラベル (一部)

関係性ラベル	リンク先ページ
カレー	レッドカレー, チキンティッカマサラ, カレーライス, コルマ, グリーンカレー, スープカレー, イエローカレー, カレー粉, サーフ, カレー・ヴルスト, ドライカレー
香辛料	香辛料, マサラ, 唐辛子, ショウガ, コリアンダー, ウコン, ニンニク
インドの食文化	ダール, チキンティッカ, サンバル, ナン, インド料理
日本のカレー	カレーパン, カレー丼, カレー鍋, インディアンスパゲッティ, カレーラーメン
肉料理	餃子, サンデーロースト, ローストビーフ, 肉
共和国	ドイツ, パキスタン, インド, ベトナム
アジアの食文化	エシャロット, 香り米, ココナッツミルク, コブミカン
精油	ニンジン, レモングラス, ハーブ
曖昧さ回避	スパイス, てんぷら, カレー (曖昧さ回避)
君主国	日本, イギリス, タイ王国
病理学	アミロイド, 腫瘍, 悪性腫瘍
ベトナムの食文化	フランスパン, ベトナム料理, 魚醤

この例では、カレーの種類や材料となる香辛料などがうまく得られている。ここで注目したいのが、「病理学」ラベル付けされたクラスについてである。Wikipedia によると

クルクミンの生理作用として抗腫瘍作用や抗酸化作用、抗アミロイド作用、抗炎症作用などが知られている。

とのことであるが、実際のテキスト中からこの一文を発見することは労力がかかる。しかしこのシステムの出力のようにクラスごとにまとめて提示されると、ユーザはカレーに病理学的な側面があることを楽に知ることができる。「カレーについて 2000 字のレポートを書きたい」と考えているユーザに、本システムは何を書けばいいかのヒントを提供できるだろう。

5. 今後の課題

本稿では草むしり法なる近似的なアルゴリズムを用いた。しかし理想的には score の減少が最小になることが保証されたアルゴリズムを利用すべきである。本研究は今後、アルゴリズムの検討を深める必要がある。

また、カテゴリ的な通常ページをラベル付けのために用いることができるようにすることも、今後の課題である。例えば国と国の関係性は多岐に渡るが、Wikipedia のカテゴリ構造のみを用いたのであれば「共和国」や「君主国」といったラベルが付けられ、国家間の関係性を表すラベルは得られない。「A 国は連合国で、B 国は枢軸国で、かつ連合国と枢軸国は交戦中」という情報があれば A 国と B 国が交戦中であることも知ることができるが、現在の Wikipedia では【枢軸国】も【連合国】も通常ページであり、カテゴリではない。このようなカテゴリ的な通常ページをカテゴリと同様にラベル付に用いることができれば、本研究のシステムで可能なことが増えるであろうと考えられる。

6. おわりに

Web ページのブラウジングを支援するためには、概念間の関係性をラベル付けして整理することが有用だろうと考えられる。このために本稿では、Wikipedia のリンク構造からページとページの関係性を表す関係性ラベルを取得する手法を提案した。実験では高速だが近似的なアルゴリズムを用いてラベル付けを行った。試作システムは部分的には本研究の目的を達成しているが、今後アルゴリズムをさらに検討する必要がある。なお本研究で作成した試作システムを <http://smokycat.info:3000/> で公開しているので、実際に動かしていただければ幸いである。

謝 辞

Wikipedia のシステムおよび掲載記事の作成・維持に尽力されている方々に、この場を借りてお礼を申し上げます。

文 献

- [1] “Wikipedia”, <http://ja.wikipedia.org/> (2013/1/11 access)
- [2] “jawiki dump progress on 20120630”, <http://dumps.wikimedia.org/jawiki/20120630/> (2013/1/11 access)
- [3] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. “The PageRank Citation Ranking: Bringing Order to the Web.”. <http://ilpubs.stanford.edu:8090/422/> (2013/1/11 access)
- [4] David Milne. “Computing Semantic Relatedness using Wikipedia Link Structure”. *Proceedings of the Fifth New Zealand Computer Science Research Student Conference (NZCSRSC '07)*, pp. 63-70, 2007.
- [5] Scott Deerwester, Susan T Dumais, George W Furnas,

- Thomas K Landauer, Richard Harshman. “Indexing by latent semantic analysis”, *Journal of the American Society for Information Science*, Vol.41, No.6, pp.391-407, 1990.
- [6] Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa. “A Generalized Flow-Based Method for Analysis of Implicit Relationships on Wikipedia”. *IEEE Transactions on Knowledge and Data Engineering*, Vol.25, No.2, pp.246-259, 2013.
 - [7] “TwitterGraph”. <http://twittergraph.wetcradle.com/> (2013/1/11 access)
 - [8] “あのひと検索スパイシー”. <http://spysee.jp/> (2013/1/11 access)