

Webを用いた反意関係抽出手法の提案

渡邊 達也[†] 灘本 明代[‡]

[†] 甲南大学理工学部 〒658-8501 兵庫県神戸市東灘区岡本 8-9-1

[‡] 甲南大学知能情報学部 〒658-8501 兵庫県神戸市東灘区岡本 8-9-1

E-mail: [†] na08_festival@yahoo.co.jp, [‡] nadamoto@konan-u.ac.jp

あらまし 現在, 多くの人々は自分の関心事やわからない事を調べる為に, インターネットの検索を利用している. これに対し, 自分の興味の無いことや小さな疑問に関してインターネットから情報を取得しようとするのはまれである. そのため, 豊富な Web の情報を十分に活かし切れていない上に情報に偏りができ, 視野を狭めている場合が多い. そこで, 本研究ではユーザの興味のある語句と, その反意語に当たる語句を用いて, その反意関係の要因となっている単語を抽出し提示する手法を提案する.

キーワード 情報検索 関係抽出

Extracting Antimony Relationship between Two Objects using Web Content

TATSUYA WATANABE[†] AKIYO NADAMOTO[‡]

[†] Faculty of Science and Engineering 8-9-1 Okamoto, Higashinada-ku, Kobe, Hyogo, 658-8501 Japan

[‡] Dept. of Intelligence and Informatics 8-9-1 Okamoto, Higashinada-ku, Kobe, Hyogo, 658-8501 Japan

E-mail: [†] na08_festival@yahoo.co.jp, [‡] nadamoto@konan-u.ac.jp

Abstract Nowadays we usually use search engine on the Internet when we search for something which we are interesting in. In contrast, it might be rare case to search something that we are not interesting by using the Internet. We don't exploit the wealth of information on the Internet. Then, our viewpoint becomes narrow and our knowledge becomes one-side. In this paper, we propose how to extract the factor of the antimony of two words by using Web content. Concretely,

Keyword Information retrieval Relation Extraction

1. はじめに

現在, 自分の関心事やわからない事を調べる為に, インターネットの検索をよく利用している. これに対し, 自分の興味の無いことや小さな疑問に関してインターネットから情報を取得しようとするのはまれである. そのため, 豊富な Web の情報を十分に活かし切れていない上に情報に偏りができ, 視野を狭めている場合が多い. そこで, 本研究ではユーザの興味のある語句と, その反意語に当たる語句を用いて, その反意関係の要因となっている単語を抽出し提示する手法を提案する.

人は何かの象徴を提示するとき, 「A と言えば X, B といえば Y」という言い方を良く使う. 例えば, 「東京と言えばもんじゃ焼き, 大阪と言えばお好み焼き」などである. ここでは「東京」と「大阪」から「粉物」

を連想し, その「粉物」から各々「もんじゃ焼き」と「お好み焼き」を導き出している. つまりは, この関係には A と B に共通するカテゴリ C から A と B が反意語となる要因 X と Y を連想している (図 1 参照). そこで本論文ではこの関係に着目し, このカテゴリ C を「要因カテゴリ」と呼び, X, Y を反意語の要因語と呼び, これらを提示するシステムを提案する.

システムの流れは以下の通りである.

(1) ユーザは反意関係であると思われる A と B の 2 単語を入力する.

(2) システムは図 2 に示すクエリ構造を用いてユーザの入力した単語 A に対して検索エンジンからサマリ・タイトルを取得する. しかし, 検索エンジンから取得した名詞の多くが略語や表記揺れが多いため, 上位概念をうまく取得することが出来ない. それらの問題を解消するために, Wikipedia から A と B のペー

ジ内の名詞を取得する．これは，ページ内に存在する名詞の多くはクエリと何かしらの関係があるため，反意語を取得しやすい．また，表記揺れを起こさないというメリットがある．

(3) (2) で取得した検索結果の文に対し形態素解析を行う．ここでは，Mecab[2]を使用する．

(4) (3) の結果から名詞のみを対象とし，その出現頻度 tf 値を求める．ここで「サ変接続・代名詞・形容動詞語幹・非自立・助詞」を属性とする名詞をストップワードとする．

(5) (4) の結果から tf 値の閾値 α 以上の単語を要因語の候補とする．

(6) ユーザの入力した単語 B についても同様 (2) から (5) の処理を行う

(7) (5) と (6) の結果から
単語 A の要因候補 $\{X \mid X \in x_1, x_2, \dots, x_m\}$ と，
単語 B の要因候補 $\{Y \mid Y \in y_1, y_2, \dots, y_n\}$ の各々の要素の上位概念を上位語階層データ[3]を用いて取得し，その上位概念 C が同じ要素のペアを要因語 (x_i, y_j) とする．この上位概念 C が要因カテゴリとする．

(8) ユーザに要因カテゴリ C，要因語 (x_i, y_j) を提示する．

以下，2 章では関連研究を，3 章では予備実験，4 章では上位概念抽出手法，5 章ではカテゴリ生成，6 章では実験・考察について述べ，7 章でまとめについて述べる．

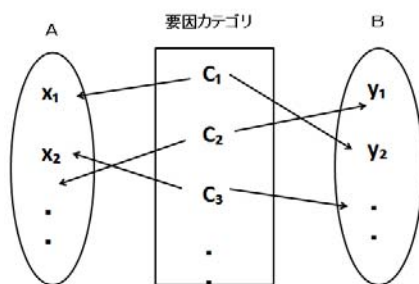


図 1. クエリの構造

Fig1. The Structure of Query

A	B
A と言えば X	B と言えば Y
X と言えば A	Y と言えば B
A といえば X	B といえば Y
X といえば A	Y といえば B

図 2. [～ と言えば・・・] の構造

Fig2. Sample of Query

2. 関連研究

同位語類を抽出する手法については大島ら [4, 5, 6, 7, 8] の研究が挙げられる．この研究ではユーザが与えた一語に対して，Web 検索エンジンが持つ情報から同位語とそのコンテキストを発見する手法を提案している．この手法では同位語が並列助詞「や」によって接続されることに着目し，「～や～」の直前後に出現する語を取得し，両方に出現する語を同意語と見なしている．発見された同位語に対しては同位語が持つコンテキストを取得することで同位語に成りえた要因提示としている．本稿で提案する手法はこの大島らの提案に基づき Web 検索エンジンの検索結果のみから知識抽出を行い，「や」の並列関係を「たとえば」に置き換え反意関係抽出法に応用した．

蓬萊ら [9] はユーザの与えた 2 語間の関係が現在の検索エンジンでは機能しないことに着目し，2 語の概念間のつながりの証拠となる文をユーザに提示する手法を提案している．手法としては検索エンジンから二つのクエリに関する文を抽出し，形態素解析を行い名詞の出現頻度を調べている．また，それらの周辺語・上位語に対しても同様の作業を行う手法も提案されているため，2 つのクエリの密粗に関係なく結果を得ることが出来る．この研究は Simpson 係数や上位概念を用いる点で非常に似ているが，我々は上位の結果だけではなく，反意関係に当たる全てのものを提示することを試みる点が異なる．

3. 予備実験

本章では，反意関係抽出の予備実験として被験者 4 人から対立する 2 つの単語（クエリ）と，要因カテゴリ・要因語を収集した結果について述べる．

実験結果よりユーザが反意語と連想する単語・要因語は同じ上位概念であることが判明した．“任天堂ーソニー”（上位概念は企業）を反意語とした場合，連想された要因語は“WiiーPS3”（上位概念は家庭用ゲーム機），“ニンテンドーDSーPSP”（上位概念は携帯型ゲーム機）のように同一の上位概念を持っている．そこで，本研究ではこの関係に着目し，反意語抽出を行うことにした．また，クエリ同士が同じ特徴を持つことからそれらの要因候補 A, B は固有名詞・企業名群，一般語群，地理語句に分類できる．（図 3 参照）．

4. 上位概念抽出手法

本章では，カテゴリ生成の前段階である名詞の上位概念抽出手法を上位語階層データ[3]を用いて提案する．ある単語の上位概念を抽出する際に問題になるの

一般語		固有名詞・企業名		地理語句	
テニス	サッカー	日本放送協会	フジテレビジョン	東京	大阪
数学	国語	ニッカウエスキー	サントリー	大阪環状線	山手線
バドミントン	テニス	任天堂	ソニー	日本	アメリカ合衆国
相撲	柔道	読売ジャイアンツ	阪神タイガース	神戸市	横浜市
アナログ放送	デジタル放送	キャノン	ニコン	ニューヨーク	ロサンゼルス
英語	フランス語	本田技研工業	トヨタ自動車	兵庫県	神奈川県
テレビゲーム	カードゲーム	麒麟麦酒	サントリー	神戸市	函館市
美術館	博物館	全日本空輸	日本航空	北海道	沖縄県
イヌ	ネコ	Google	Yahoo!	日本	中国
ゴルフ	野球	YouTube	ニコニコ動画	南極	北極

図 3 クエリの分類分け

Fig3. Type of Queries

は、ある単語に対して上位概念というものがいくつも存在することである。例えば「阪神タイガース」の上位概念は「日本のプロ野球チーム」が一般的ではあるが、「阪急阪神ホールディングス」も「阪神タイガース」の親会社ということでは間違いではない。また、「杉山愛」と「小椋久美子（オグシオ）」を比較した場合、ひとつ上の上位概念だと「テニス選手」、「バドミントン選手」となりカテゴリとしてユーザに提示するには不十分である。しかし、もうひとつ上の上位概念ではともに「スポーツ選手」となり、同一のものとして判断出来る。このように一つ上の上位概念は様々なものがあるとともにもその一階層下に属する名詞が少ないので、もうひとつ上の上位概念を抽出することにする。上位概念の上位概念の有能性は内村ら[10]の「Wikipediaを用いた概念構造の抽出」によって明らかとされている。実際には図 5 に示す精度の結果を得ることができ、ユーザに要因カテゴリとして提示する。要因カテゴリを付加してユーザに提示することで、反意関係であることが明確に示すことが出来る。そして、あまり知られていない語句同士の場合においても語弊を防ぐことが可能になると考えた。

5. カテゴリ生成法

4 章により抽出した上位概念にはクエリ A, B に関係性のないものが含まれているため、以下の手法を組み合わせ精度の高いカテゴリを生成する。

5.1 Simpson 係数

概念間の関係の強さを測るため手法としては、Web 検索ヒット件数を利用して類似度を求める Simpson 係数を用いた。以下に Simpson 係数の式を表す。

$$\text{Simpson係数}(A,B) = \frac{H(|A| \wedge |B|)}{\text{Min}(H(|A|), H(|B|))}$$

ここで、 $H(|A|)$, $H(|B|)$ はクエリ A, B に対する検索ヒット数を表す。

クエリのカテゴリを生成する過程として、 $H(|A|)$ or $H(|B|) > H(|C|)$ の関係が成り立つ。ここで、C は 4 章で抽出した要因カテゴリであり、 $H(|C|)$ は上位概念に対する検索ヒット数を表す。

そのため、結果が密集してしまうので、結果をある程度粗にする必要がある。また、予備実験の結果より Simpson 係数には以下の問題があることが判明した。

- (1) クエリ A, B に関係ない上位概念であっても係数が密になる。
- (2) 珍しいカテゴリ C に対しては分母が小さくなりすぎる。

精度をよくするため分母に $H(|A|) \wedge H(|B|) \wedge H(|C|)$ [式 1] を Simpson 係数に加える (図 4 参照)。これはクエリ A, B, カテゴリが共に記載されている文章が多いということは何かしらの関係性があると仮定することが出来る。

予備実験では“東京ー大阪”をクエリとして反意語の抽出を行った際、野球と物理単位というカテゴリが検出された。双方の [式 1] での検索ヒット数は野球が約 67,500,000 件、物理単位では約 1,360 件であった。これは、読売ジャイアンツと阪神タイガースというプロ野球球団と“東京ー大阪”の関係性から結果が密であるため検索ヒット数が多いと考えられる。物理単位については関係性がないことから粗となる。これらの結果から [式 1] を加えることで関係性の差異を結果に反映することが出来る。

$$\text{Simpson係数}(A,B,C) = \frac{H(|A| \wedge |B| \wedge |C|)}{\text{Min}(H(|A|), H(|B|), H(|C|)) + \underbrace{H(|A| \wedge |B| \wedge |C|)}_{\text{式 1}}}$$

図 4 [式 1] を加えた Simpson 係数式

Fig4. Expression's one is added to the Simpson's Coefficient

	名詞	上位概念	上位概念の上位概念
1	アナウサギ	ウサギ科	ウサギ
2	原辰徳	読売ジャイアンツの選手	読売ジャイアンツ及び東京巨人軍の選手
3	真弓明信	阪神タイガースの選手	阪神タイガース及びその前身球団の選手
4	紅茶	茶	農産物
5	オレンジ	果実	食品
6	東京大学	日本の国立大学	日本の大学
7	通天閣	大阪市の観光地	大阪市の観光地
8	東京タワー	東京都の観光地	東京都の観光地
9	柴犬	犬の品種	犬
10	イエネコ	猫の品種	猫

図 5 上位概念の上位概念

Fig5. Superordinate concept above

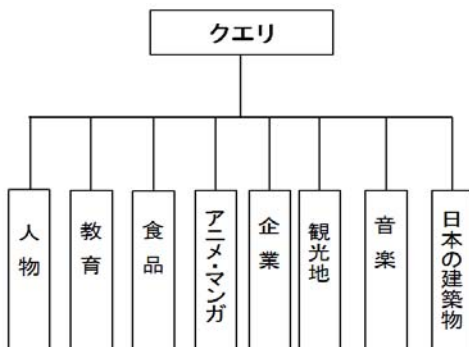


図 6 カテゴリ構造

Fig6. The Structure of Category

5.2 比率・関係性

Simpson 係数の値はクエリ A, B の組み合わせによって大きく違うため、閾値を固定化することは難しい。そのため本稿では、表 1 と 5.1 の結果を元に検索ヒット数と Simpson 係数の値に着目して、どのような組み合わせにでも対応出来るよう、閾値を変更する方式にした。検索ヒット数は 100,000,000 件を、Simpson 係数は 0.3, 0.5 の二つを閾値としている。また、検出されるカテゴリ数が非常に多いため、A-B の検索ヒット数の下位 1/2 件は関係性がないと判断出来るので、結果から取り除くことを行う。ここまでの段階でシステムを“東京-大阪”をクエリとして実装した場合、「大阪府の歴史」や「東京都の観光地」というカテゴリが検出された。これらは各々のクエリに引っ張られた結果であるのが明白である上に、このような上位概念を持つ名詞には 2 つ上の上位概念が存在しない。(図 5 参照)。そのため、これらについては以下の例外処理を行う。

5.3 例外処理

上位概念の上位概念にクエリが含まれているもの(例:クエリが大阪, 上位概念が大阪の歴史)はクエリとの関係性が密であるが、2 語の要因カテゴリとして

は不適切である。しかし、このような上位概念を持つ名詞を省いてしまってもいけないため、例外として以下の処理を行い、結果に反映する(図 7 参照)。

(1) Simpson 係数が 0 でない限り、カテゴリとして認定する。

(2) 5.2 の処理を行わない。

(3) (1) によって認定された上位概念については各々のクエリ部分を X に置き換え、結果の処理を行う。

表 1 Simpson 係数・A-B の検索ヒット数

Table1.Simpson's Coefficient and number of search results for A-B

クエリA-B	A-Bの検索ヒット数	Simpson係数
大阪-パン	83400000	0.199045346
ご飯-パン	60000000	0.212014134
日本-中国	979000000	0.252319588
デジタル-アナログ	48300000	0.478217
テレビ-ラジオ	152000000	0.425770308
東京-大阪	749000000	0.56742424

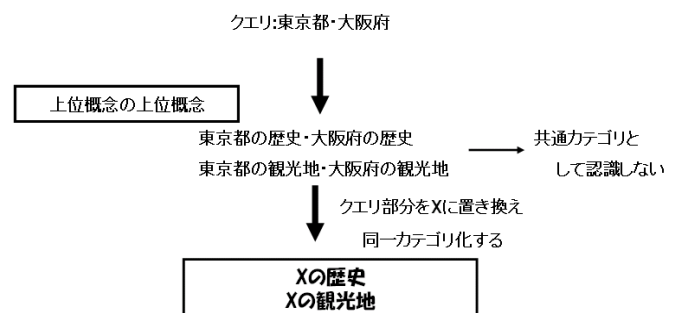


図 7 例外処理

Fig7. Exception handling

6. 実験

6.1 実験

本章では 5 章で得られたクエリをもとに提案システムの実験を行い、システムの有用性を測った。結果の“X”は例外処理によって置き換えた反意語 A, B 部分で

ある．また 5 章で得られた結果以外ものを“その他”の分類として実験を行った．実験結果の一部を表 2 に示す．

表 2 実験結果
Table2.Result of Experiments

		地 理 語 句	
		要因カテゴリ	
		VS	
Good Result	クエリA 東京都	クエリB 大阪府	
	東京ドーム	阪神甲子園球場	
	奥多摩	上町台地	
	山手線	大阪環状線	
	東京タワー	通天閣	
Bad Result	明治神宮	住吉大社	
	牛井	たこ焼き	
	牛井	鰻	
	手羽先	大阪寿司	
	プール	図書館	
		要因カテゴリ	
		VS	
Good Result	クエリA 沖縄県	クエリB 北海道	
	西表石垣国立公園	利尻礼文サロベツ国立公園	
	那覇空港	新千歳空港	
	中城湾	津軽海峡	
	ガレッジセール	タカアンドトシ	
Bad Result	首里城	五稜郭	
	沖縄美ら海水族館	タンチョウ	
	だし	ジンギスカン	
	沖縄本島	樺太	
	ウィルコム沖縄	コンサドーレ札幌	

		固 有 名 詞 ・ 企 業 名	
		要因カテゴリ	
		VS	
Good Result	クエリA 阪神タイガース	クエリB 読売ジャイアンツ	
	藤村富美男	野球殿堂	
	阪神甲子園球場	競技場	
	真弓明信	xの選手	
	サンテレビジョン	テレビ局	
Bad Result	バット	スポーツ器具	
	ボール	スポーツのルール	
	外野手	xの選手	
	麒麟麦酒	サントリー	
	キリンホールディングス	酒類メーカー	
Good Result	ハードランドビール	アジアのビール	
	モルツ		
	玄米茶	お茶	
	ウォッカ	カクテル	
	梅酒	リキュール	

		一 般 語	
		要因カテゴリ	
		VS	
Good Result	クエリA バドミントン	クエリB テニス	
	オグシオ	日本のオリンピック選手	
	小椋久美子	日本のx選手	
	新田玲子	日本のx選手	
	川前直樹	各国のx選手	
Bad Result	日本バドミントン協会	日本の国内競技連盟	
	世界バドミントン選手権大会	x	
	女子	日本のスポーツ競技大会	
	杉山愛	オリンピック日本選手団	
	イヌ	ネコ	
Good Result	x科	イヌ亜科	
	xの品種	ブルドッグ	
	フードル	x	
	種族	ゲーム	
		猫草	

		そ の 他	
		要因カテゴリ	
		VS	
Good Result	クエリA ドラえもん	クエリB クレヨンしんちゃん	
	小学校	学校を舞台とした作品	
	ドラえもん	xの登場人物	
	野比玉子	xの登場人物	
	源静香	xの登場人物	
Bad Result	アニメーション映画	日本の映画制作会社	
	クレヨンしんちゃん	日本のアニメ作品	
	ロボット	xの登場人物	
	テレビ朝日	日本の放送事業者	
	テレビ朝日	テレビ朝日	
Bad Result	大阪	飯	
	串カツ	カツ丼	
	大阪寿司	あなご飯	
	鰻丼	牛丼	
	焼肉	牛丼	

6.2 考察

適合率は東京ー大阪は 40%，阪神タイガースー読売ジャイアンツは 51%，バドミントンーテニスは 30%，ドラえもんークレヨンしんちゃんは 52%と全体的に低い結果であったが，普段の検索では発見しにくい関係の抽出に成功した．しかし，よい結果の大半が Wikipedia から取得した名詞であり，検索エンジンから取得した名詞のほとんどが表記揺れや略語などが多く含まれているため，うまく上位語階層データ[3]とマッチングすることが出来ずにゴミとして処理されてしまう現状がある．それに加え，クエリの関係が親密であるほど A には B の，B には A の要因語候補が含まれているため，結果の重複や真逆の要因語を提示してしまった．また，“牛井”や“女子”のように様々なものごとに関連付けることが出来る名詞への対応策を示唆しなければならない．

① 地理語句

Wikipedia に記述されている目次構造が酷似しているため，例外処理により市町村などの反意語を多く抽出することが出来た．しかし，上記で上げた“牛井”や“女子”が多く含まれていることが問題となっている．

② 固有名詞・企業名

クエリが企業名の場合，平仮名・英数字が混在している独特な商品や同一間のある商品のものも多く，上位語階層データ[3]とマッチングすることが出来ない．また，検索エンジンで取得した名詞は略語で書かれているものが多く存在することが適合率を下げている問題として挙げられる．

③ 一般語

一般語は下位概念（スポーツ類では選手や企業名，動物類では種類など）についての記述が多いため，様々な要因語を取得することが出来る反面，反意語の要因が同じものばかりである．

④ その他

有名なアニメ・キャラクターなどであると，一般語と同じような結果が得られた．上位概念が違うクエリ同士であると当然と言うべきではあるが，反意語を抽出することが出来ずにいるため，反意関係ではないことが明らかになった．

7. まとめ

本研究では，ユーザの興味のある語句と，その反意語に当たる語句を用いて，その反意関係の要因となっている単語ペアの抽出手法を上位概念抽出手法とカテゴリの自動生成手法を用いて提案した．これにより，検索することが楽しくなり，視野の拡大に繋がれば重量である．今後の課題を以下に示す．

・ユーザの入力した一語に対して，反意語に当たる語句と要因を提示する手法への改善．

・入力ワードと要因候補との関係の制約を強くし，適合率を上げる．

・表記揺れ・略語への対応策．

謝 辞

本研究の一部は，平成 21 年度科研費特定領域研究域「コミュニティ型コンテンツのコンテンツホール検索に関する研究」（課題番号：21013044，代表：灘本明代）及び甲南大学平生太郎基金科学研究奨励助成金による．ここに記して謝意を表します．

文 献

- [1] <http://ja.wikipedia.org/wiki/%E9%99%B0%E9%99%BD>
- [2] <http://mecab.sourceforge.net/>
- [3] 高度言語情報融合フォーラム（ALAGIN）上位語階層データ <http://nlpwww.nict.go.jp/corpus/>
- [4] 大島 裕明，小山 聡氏，田中 克己，” web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見”，情報処理学会論文誌（トランザクション）データベース，Vol. 47, No. SIG19, TOD 32, pp. 98-112, 2006 年 12 月
- [5] 大島 裕明，小山 聡氏，田中 克己，”Web からの兄弟カテゴリ文書の検索”，情報処理学会研究報告，Vol.2006, No.77, 2006-DBS-140(II), pp.153-160, 2006 年 7 月
- [6] 大島 裕明，田中 克己，”正解語ペア漸増による関連語取得のための両方向構文パターン発見”，電子情報通信学会，第 1 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2009），B9-1, 2009 年 3 月
- [7] 大島 裕明，小山 聡氏，田中 克己，”サーチエンジンのインデックスを利用した同位語検索と同位語コンテキストの発見”，情報処理学会研究報告，Vol.2006, No.77, 2006-DBS-140(II), pp.161-168, 2006 年 7 月
- [8] 大島 裕明，山口 雅史，小山 聡氏，田中 克己，”Web 検索とクエリログを併用した同位語発見手法”，日本データベース学会論文誌 DBSJ Letters, Vol.5, No.4, pp.37-40, 2007 年 3 月
- [9] 蓬萊 博哉，AdamJatowt, 田中 克己，”Web からの文抽出と概念辞書を用いた概念間の関係支援”，iDB フォーラム 2008, 情報処理学会 研究報告「データベースシステム」, Vol.2008, No.88, 2008 年 9 月
- [10] 内村 圭佑，灘本 明代，”Wikipedia からの概念抽出に基づくなぞかけ文の自動生成”，第 1 回データ工学と情報マネジメントに関するフォーラム（DEIM2009）