

サーチエンジンのクエリログを利用した同位語の発見

Discovering Coordinate Terms Using Search Engine Query Logs

山口 雅史[▼] 大島 裕明[♦]
小山 聡[▲] 田中 克己[▲]

Masashi YAMAGUCHI Hiroaki OHSHIMA
Satoshi OYAMA Katsumi TANAKA

サーチエンジンのクエリログを用いて、ある語の同位語を発見する手法を提案する。同位語とは共通の上位語を持つような語のことである。従来研究にも、同位語や、上位語、下位語を求めるようなものは数多くある。しかし、それらは大量のコーパスを用いるもの、または、膨大な計算処理が必要なものがほとんどであった。それに対して我々は、サーチエンジンのもつメタデータの種類であるクエリログのみを用いて、効率よく同位語を発見する手法を提案する。本手法はユーザが入力する検索クエリにおいて、ある語とその同位語となる語とはしばしば同様の語を用いて絞り込みをされるという点に着目している。

We propose a method for searching coordinate terms by using search engine query logs. "Coordinate terms" are terms which have the identical hypernym. There are several researches that acquire coordinate terms, but they need parsed corpora or much time for computation. In contrast, our proposing method only needs query log data, which is a kind of a metadata owned by Web search engine. It needs relatively few time for computation to obtain coordinate terms. Our approach is based on the idea that coordinate terms have common cooccurrent terms in search engine query logs.

1. はじめに

全ての語は他のいくつかの語とそれらとの関係を用いて説明することが出来る。語と語の関係を表す用語で有用なものに、上位語や下位語がある。上位語とはある語の上位概念を表す語であり逆に下位語はある語の下位概念を表す語である。また他にも同位語と呼ばれるものがあり、一般的に「ある語と共通の上位語を持つ語」とされる。本研究では、「同位語同士は共通の話題語を持つ」という考えのもと同位語の発見を行う。ただし、話題語とはある語の話題を表す語である。例えば「京都」の話題語としては「観光」、「地図」などがあり、「巨人」と「阪神」の共通の話題語としては「応援歌」などが挙げられる。

[▼] 学生会員 京都大学情報学研究科博士前期課程
yamaguti@dl.kuis.kyoto-u.ac.jp

[♦] 学生会員 京都大学情報学研究科博士後期課程
ohshima@dl.kuis.kyoto-u.ac.jp

[▲] 正会員 京都大学情報学研究科
foyama_tanaka@dl.kuis.kyoto-u.ac.jp

我々が提案する同位語の発見手法は、サーチエンジンのクエリログのみを用いたものである。クエリログとはユーザが問い合わせに用いた検索語の履歴であり、サーチエンジンは日々この履歴を蓄積している。クエリログの例を表1に示す。クエリログから、ユーザがある語に対してどのような語を追加する事が多いのかを統計的に把握する事が可能である。

なお本研究に用いたクエリログは Overture[1]が提供する「キーワードアドバイスツール」[2]と呼ばれる Web ベースのツールを利用して取得した。ある語に関して問い合わせを行うと、その語を含んだ様々な組み合わせのクエリを前月の検索数と共に取得することが可能である。

2. 関連研究

同位語を発見する従来研究はいくつか存在している。Google Sets[3]というサービスは、複数の語を入力すると、それらの同位語と考えられる語を出力するものである。Google Sets のアルゴリズムは現時点で非公開であるが、Google が収集した大量の Web ページを解析し、同位語のクラスタを発見していると思われる。

同様に同位語発見に関する研究を述べる Churchら[4]は、相互情報量を用いて、意味的に関連があるような語を発見する手法を提案した。正確には同位語発見を目的とした研究ではないが、発見される語には同位語も含まれている。他の研究のいくつかにおいても、この研究の成果である、相互情報量が高い語どうしは、同位語である可能性が高いことを利用している。

Zoubinら[5]による Bayesian Sets はベイズ推定を利用した同位語のクラスタを発見するものである。用いられているアルゴリズムは非常にシンプルで高速であるが、何らかの大規模データを用意することが前提となっている。Linら[6]は類似の語に関するクラスタを生成する手法について提案した。係り受け関係を利用して語同士の類似度を計算することによって、語のクラスタを生成するものである。そのため、大規模コーパスが必要となる。Shinzatoら[7]は HTML 文書から同位語を発見する手法を提案した。

また、サーチエンジンクエリログを対象とした従来研究としては、Web ページの重要度判定や、クエリ補完に関するものがある。Cuiら[8]は、クエリと Web ページの確率的な相関性を基に、適切な補完クエリを求める手法を提案した。Fonsecaら[9]は、クエリログから関連ルールを用いて、関連語を取得し、クエリ補完する手法を提案した。また Wenら[10]はサーチエンジンによって得られたページのうちユーザが閲覧した文書を記録し、クエリと文書との関係性からクエリをクラスタリングする手法について提案した。Silversteinら[11]はクエリログのマイニングにより、カイ2乗値を基に語の出現数の偏りから、共起語を求める手法を提案した。いずれもクエリログを対象とした研究であるが、本研究の目的とは異なるものである。

3. 同位語の発見

3.1 同位語候補集合の取得

検索クエリはしばしば and 検索を行う目的で複数の語がスペースで区切られ作成される。本節では、このような絞り込みに使われる語に着目して同位語の候補を取得する手順を述べる。「トヨタ」を例にあげると「トヨタ カローラ」「トヨタ ディーラー」「兵庫 トヨタ」などのクエリが存在している。このとき、「X カローラ」「X ディーラー」「兵庫 X」などといったような、「絞り込みに用いられる語」と「絞り込みの方向」を表す型を絞込型と呼ぶことにすると、トヨタ

はこれらの絞込型にあてはまる語であると言える。

提案手法における候補語とはこれらの絞込型に当てはまる語を指す。実験においては、絞込型に含まれている語でクエリログを検索し得られた 100 件のうち、型に適合する物を候補語としている。

以下に「トヨタ」の同位語候補を得る手順を示す。

表 1 「トヨタ」,「自動車」のクエリログ¹
Table 1 Query Logs that contain “Toyota” and “automobile”

検索数	キーワード	検索数	キーワード
660026	トヨタ	2081735	自動車 趣味
476901	トヨタ 自動車	513803	自動車 メーカー
143873	トヨタ レンタカー	476901	トヨタ 自動車
75200	トヨタ 中古車	259138	三菱 自動車
49292	トヨタ レンタリース	252657	日産 自動車
31375	ネット トヨタ	193977	ホンダ 自動車
26890	トヨタ ホーム	160289	自動車 保険
25241	トヨタ ネット	119458	自動車
22739	トヨタ カローラ	108154	trend 自動車 保険
20751	トヨタ bb	99555	自動車 試乗 レポート
19092	トヨタ ディーラー	95075	自動車 税
18912	トヨタ レクサス	90130	スズキ 自動車
16402	トヨタ 自動車 ホームページ	84869	マツダ 自動車
16014	トヨタ 紡織	82947	自動車 教習所
15113	トヨタ クラウン	74659	スバル 自動車
13597	トヨタ ファイナンス	63784	自動車ドレスアップ
11835	トヨタ 車体	63134	メルセデス ベンツ 自動車
11668	トヨタ カード	62431	bmw 自動車
11334	トヨタ 博物館	59495	自動車 免許
11154	トヨタ アルファード	53516	フォルクスワーゲン 自動車

まず、「トヨタ」のクエリログを取得する(表 1 左)。検索回数 1 位の「トヨタ」は単独クエリであるため、候補語は取得できない。そこで、2 位以下のクエリに注目すると、2 位は「トヨタ 自動車」であるから、絞込型は「X 自動車」である。そこで、「自動車」でクエリログを検索し「自動車」を含むクエリログを取得する(表 1 右)。そのうち、「X 自動車」の絞込型に適合する物は「トヨタ」以外に「三菱」「日産」「ホンダ」「trend」などが挙げられる。同様に 3 位以下の「X レンタカー」「X 中古車」についても候補語を取得する。「トヨタ」の場合、絞込型は 99 個得ることができ、それらに合致する候補語は重複を除いて、2335 個得ることが出来た。

すなわち、一般化すると以下ようになる。

1. ユーザがクエリ p を与える。
2. サーチエンジンクエリログから p を含むクエリ集合 Q を取得する。
3. Q のそれぞれの要素において、絞込型を取得する
4. それぞれの絞込型に含まれている語を全て含むクエリ集合 R を取得し、絞込型に適合する語を候補とする。

次節以降では、得られた候補集合を評価する手法として、「cos 類似度」および「出現順位加重平均」を用いる方法と、「HITS アルゴリズム」を用いる方法を述べる。

3.2 cos 類似度

3.1 節で述べたように、提案手法は絞込に用いられている語に着目して同位語発見を行うものである。以下、絞込に用

いられている語を並列語²と呼ぶこととする。例えば「トヨタ ディーラー」というクエリにおいて、「トヨタ」の並列語は「ディーラー」であり、逆も然りである。

本節においては、語の特徴量として並列語のベクトルを作成し、それらの cos 類似度を元に同位語判定を行う手法を述べる。まず、ある語 p を含むクエリログを上位 100 件取得し、その集合を $L(p)$ とする。このとき p における語 t の検索数を加味した出現度 $cf_p(t)$ を以下のように定義する。

$$cf_p(t) = \sum_l N(l, t)$$

ただし、 l は $L(p)$ の要素となるクエリログであり、

$$N(l, t) = \begin{cases} st(l) & (\text{if } l \text{ contains } t) \\ 0 & \end{cases}$$

$$st(l) = (l \text{ の月間検索回数})$$

である。これを用いて、語 p の特徴ベクトル $V(p)$ を以下のように定義する。

$$V(p) = (cf_p(t_1), cf_p(t_2), \dots, cf_p(t_n))$$

このとき、語 p と q の類似度を以下のように求める。

$$sim(p, q) = \frac{V(p) \cdot V(q)}{\|V(p)\| \|V(q)\|}$$

このようにして求めた類似度に基づき候補語の評価を行った。

3.3 クエリにおける語の出現順位

本節では、語の出現順位を語の特徴量として利用する事について述べる。ユーザはサーチエンジンにクエリを与える際、絞込の目的で複数の語を組み合わせたことが多いが、その順位にはユーザの意図が反映されているものである。例えば、「トヨタ ディーラー」というクエリの場合、トヨタを 1 番目に指定し、次にディーラーを指定することで、トヨタのディーラーに関する情報を求めようとするユーザの意図が汲み取れる。しかし逆に「ディーラー トヨタ」の順でクエリを与えることは希である。実際 2006 年 4 月のデータによると、「トヨタ ディーラー」の検索数が 19092 件であるのに対し、「ディーラー トヨタ」の検索数はわずか 40 件しか無い。

ここで、出現順位の傾向を表す値 $wao(p)$ を以下のように定義する。

$$wao(p) = \frac{\sum_{l \in L(p)} (st(l) \cdot order(l, p))}{\sum_{l \in L(p)} st(l)}$$

ただし

$$order(l, p) = (l \text{ における } p \text{ の出現順位})$$

である。

上記の「トヨタ」「ディーラー」において、この値はそれぞれ

$$wao(\text{"トヨタ"}) = 1.05104$$

$$wao(\text{"ディーラー"}) = 2.36592$$

となり、明らかに出現順位に偏りがあるといえる。

本研究においては、「同位語関係にある語はそれらの $wao(p)$ 値は近い」という仮説のもとに、同位語の発見を行う。4.1 節ではこの手法の実験結果を示す。

¹ Overture キーワードアドバンスツールより取得

² 絞込型に含まれる語であるとも言える。

3.4 HITS アルゴリズム

候補語の評価に用いる別の方法として、HITS アルゴリズムに基づく手法を提案する。HITS アルゴリズムは Kleinberg[12]が提案したアルゴリズムで、“authority”と“hub”という、Web ページの有用性を表す二つの尺度を用いていることが特徴的である。Web ページ p から Web ページ q へリンクされていることを、 $p \rightarrow q$ と表すとすると、ページ p の“authority”と“hub”は以下のように定義される。

$$auth(p) = \sum_{q:q \rightarrow p} hub(q)$$

$$hub(p) = \sum_{q:p \rightarrow q} auth(q)$$

“authority”値は高い“hub”値を持つページからから多く参照されているとき高くなり、“hub”値は高い“authority”値を持つページを多く参照しているとき高くなるという様に、再帰的な定義がなされている。

先に述べたとおり、我々の手法はクエリログにおける共起関係に基づき同位語を発見するものである。例えば、「トヨタ」と「ホンダ」は共に「ディーラー」、「中古車」と言ったような語と共起するが、この共起関係をリンクとみなすと、これらの語群をコミュニティとみなす事ができる。このようなコミュニティを HITS アルゴリズムにより発見するため、まず、絞込型と候補語の二部グラフを作成する。

「トヨタ」の場合の例を挙げる。

- (1) 絞込型は 99 個得られ、それぞれから得られた候補語は重複を除いて、2335 個である。
- (2) それぞれの絞込型からそれに合致する全ての候補語にリンクを張り二部グラフを作成する。
- (3) 作成した二部グラフに対し、HITS アルゴリズムを適用する。

グラフを図 1 に示す。例えば候補語である「ホンダ」は絞込型「X ディーラー」や「X 中古車」に合致するため、それらからリンクされている。また、「X カローラ」からは「新型」、「70」へリンクしている。

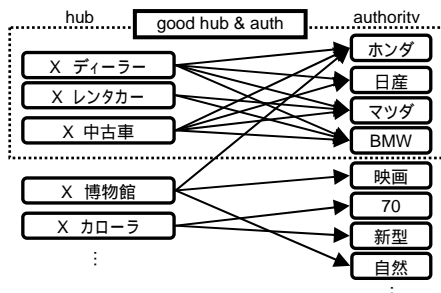


図 1 絞込語と候補語の二部グラフ

Fig 1 Bipartite Graph among narrowing pattern and candidates

HITS アルゴリズムを使う利点としては、候補語を含むクエリログを取得する必要が無いことが挙げられる。先に述べた仮説に基づくと、同位語はコミュニティを形成し他の候補語と比較して、“authority”値が高くなると考えられる。4.2 節ではこの手法の実験結果を示す。

4. 実験結果

4.1 「cos 類似度」および「出現順位加重平均」を用いる手法

得られた候補語それぞれに対し、cos 類似度を求め、出現

順位加重平均 $wao(p)$ の差 δ が一定値以内の語を同位語と判定した。なお今回の実験では $\delta = 0.3$ を用いている。表 2 に「トヨタ」と「ルイヴィトン」の結果を示す。なお、表中の*が付与された語は $wao(p)$ の値がそれぞれ

$wao(\text{トヨタ}) = 1.05$ と $wao(\text{ルイヴィトン}) = 1.01$ との差が $\delta = 0.3$ 以上あるため、除外したものである。cos 類似度を計算する際には、対象となる語の並列語ベクトルを作成するために、今回の実験においては Web アクセスが必要となる。実験一回あたり数千語の並列語ベクトルを作成する必要があるため、十分な量の実験は行えていない。cos 類似度と出現順位加重平均について、ある程度の可能性は示すことが出来たと考えるが、有効性については今後議論していきたい。

表 2 「トヨタ」、「ルイヴィトン」の結果 (cos 類似度)

Table 2 Results for “Toyota” and “Louis Vuitton”

候補語 p	類似度	wao(p)	候補語 p	類似度	wao(p)
日産	0.9630	1.03	gucci	0.9151	1.02
*ドレスアップ	0.9155	1.95	miumiu	0.8697	1.01
光岡	0.9072	1.05	グッチ	0.8361	1.08
マツダ	0.9071	1.06	フェンディ	0.7787	1.03
いすゞ	0.9024	1.09	miu	0.7778	1.14
スバル	0.8763	1.11	ロエベ	0.7709	1.03
フォルクスワーゲン	0.8745	1.02	アナスイ	0.7654	1.01
ヒュンダイ	0.8503	1.01	シャネル	0.7533	1.03
amg	0.8435	1.15	プラダ	0.7509	1.02
*メーカー	0.8395	2.19	クレージュ	0.7307	1.13
suv	0.8347	1.17	コーチ	0.7292	1.19
マセラッティ	0.8336	1.02	*クロコ	0.7284	1.54
bmw	0.8252	1.06	フェリージ	0.7075	1.04
アルピナ	0.8230	1.15	クレイサス	0.7000	1.03
*メンテナンス	0.8160	2.05	dakota	0.6982	1.11
アウディ	0.8067	1.02	*コードバン	0.6889	1.39
ケーターハム	0.8014	1.01	*がま口	0.6808	1.45
*趣味	0.8002	2.00	furla	0.6804	1.02
gm	0.7939	1.32	フェラガモ	0.6733	1.08
*整備士	0.7931	2.01	オーストリッチ	0.6658	1.12

4.2 HITS アルゴリズムを用いる手法

表 3, 表 4 にそれぞれトヨタ、ルイヴィトンの実験結果を示す。なお表には“hub”値上位 20 件の絞込型と“authority”値上位 20 件の候補語を挙げている。高い“authority”値を示した語に、それぞれ自動車メーカー、高級ブランドの名前があり正解であるが、いくつかの不正解も存在する。不正解の中には、同位語ではなく元の語の上位語にあたる語が含まれていることもわかる。「トヨタ」には「車」、「ルイヴィトン」においては、「ブランド」がそれにあたる。また本稿には掲載していないが、「サッカー」において「スポーツ」、「ギター」において「楽器」などが上位に挙がった。これらの例から、上位語も同位語同様の絞込をされることが分かる。

5. おわりに

サーチエンジンクエリログに有益な情報が眠っていることを示した。本稿では、同位語の発見に焦点を当てたが、その他にも様々な有益な情報を得る事が出来ると思われる。今後は同位語だけでなく、上位語、下位語についての発見手法も模索していく所存である。

【謝辞】

本研究の一部は、文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」(リーダ

ー：田中克己，平成 14～18 年度），文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，異メディア・アーカイブの横断的検索・統合ソフトウェア開発（研究代表者：田中克己）および，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者：田中克己，A01-00-02，課題番号：18049041)によるものです．ここに記して謝意を表します．

表 3 「トヨタ」の実験結果(HITS アルゴリズム)
Table 3 Results for "Toyota" (HITS)

絞込型 p	hub(p)	候補語 p	auth(p)
X ディーラー	0.06488	日産	0.01345
X 専門店	0.06162	ホンダ	0.01262
X 純正 部品	0.04719	toyota	0.01187
X リコール	0.04261	マツダ	0.01118
X 自動車	0.03917	スズキ	0.01063
X 中古車	0.03194	bmw	0.00963
X 部品	0.03018	ダイハツ	0.00944
X 純正 ナビ	0.02962	車	0.00913
X 中古車 販売店	0.02714	スバル	0.00886
X レンタカー	0.02251	中古 車	0.00864
X ファイナンス	0.02056	三菱	0.00854
X 自動車 ホームページ	0.02002	ボルボ	0.00751
X 部品 共販	0.01993	自動車	0.00675
X 自動車 株価	0.01608	アウディ	0.00649
X 株価	0.01576	バイク	0.00641
X 中古車 販売	0.01536	フォルクスワーゲン	0.00582
X レンタ	0.01280	ボルシェ	0.00579
中古車 X	0.01239	三菱 自動車	0.00574
X 生産方式	0.01230	ベンツ	0.00539
X ハイブリッド	0.01217	メルセデス ベンツ	0.00497

表 4 「ルイヴィトン」の実験結果(HITS アルゴリズム)
Table 4 Results for "Louis Vuitton" (HITS)

絞込型 p	hub(p)	候補語 p	auth(p)
X 長財布	0.05052	グッチ	0.01410
X 財布	0.04980	ブランド	0.01348
X バスケース	0.04432	シャネル	0.01272
X 名刺入れ	0.04366	コーチ	0.01262
X 直営店	0.04149	エルメス	0.01196
X カードケース	0.04134	gucci	0.01147
X 小銭入れ	0.03978	ブルガリ	0.01125
X バッグ	0.03861	ブラダ	0.01104
X 携帯 ストラップ	0.03721	パーバリー	0.01102
X キーホルダー	0.03379	ポール スミス	0.00993
X 財布 新作	0.03331	coach	0.00989
X 偽物	0.03133	ディオール	0.00906
X 偽者	0.02815	フェラガモ	0.00872
X 新作 財布	0.02814	カルティエ	0.00860
X ネックレス	0.02661	アナスイ	0.00787
X 財布 メンズ	0.02436	ピアノ	0.00772
X ストラップ	0.02377	クロエ	0.00733
X バック	0.02263	サマンサタバサ	0.00721
X バッグ 新作	0.02203	dior	0.00633
X サンGLラス	0.02190	革	0.00624

【文献】

[1] "Overture". <http://inventory.jp.overture.com/>.

[2] "キーワードアドバイスツール".
<http://inventory.jp.overture.com/>.

[3] Google Sets <http://labs.google.com/sets>.

[4] Kenneth Ward Church and Patrick Hanks: "Word association norms, mutual information, and lexicography", Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pp. 76-83 (1998).

[5] Z. Ghahramani and K. Heller: "Bayesian sets", Proceedings of the Nineteenth Annual Conference on Neural Information Processing Systems (NIPS2005) (2005).

[6] Dekang Lin: "Automatic retrieval and clustering of similar words", Proceedings of the 36th annual meeting on Association for Computational Linguistics, pp. 768-774 (1998).

[7] Keiji Shinzato, Kentaro Torisawa: "A simple www-based method for semantic word class acquisition", Proceedings of the Recent Advances in Natural Language Processing (RANLP05), pp. 493-500 (2005).

[8] H. Cui, J. Wen and W. Ma: "Probabilistic query expansion using query logs", Proceedings of the eleventh international conference on World Wide Web, ACM Press, pp. 325-332(2002).

[9] B. F. Federal: "Using association rules to discover search engines related queries".

[10] J.-R. Wen, J.-Y. Nie and H.-J. Zhang: "Clustering user queries of a search engine", World Wide Web, pp. 162-168(2001).

[11] C. Silverstein, M. Henzinger, H. Marais and M. Moricz: "Analysis of a very large altavista query log", Technical Report 1998-014, Digital SRC (1998).

[12] J. M. Kleinberg: "Authoritative sources in a hyperlinked environment", Journal of the ACM, 46, 5, pp. 604-632 (1999).

山口 雅史 Masashi YAMAGUCHI

京都大学大学院情報学研究科博士前期課程在学中．2005 年 京都大学工学部情報学科卒業．Web 環境におけるパーソナライゼーション，クエリログ活用の研究に従事．日本データベース学会学生会員．

大島 裕明 Hiroaki OHSHIMA

京都大学大学院情報学研究科博士後期課程在学中．2004 年 神戸大学大学院自然科学研究科博士前期課程修了．Web 環境におけるパーソナライゼーションの研究に従事．情報処理学会，日本データベース学会，ACM 各学生会員．

小山 聡 Satoshi OYAMA

京都大学大学院情報学研究科社会情報学専攻助手．2002 年 京都大学大学院情報学研究科博士後期課程修了．博士(情報学)．主に機械学習，データマイニング，情報検索の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI 各会員．

田中 克己 Katsumi TANAKA

京都大学大学院情報学研究科社会情報学専攻教授．1976 年 京都大学大学院修士課程修了．京大工博．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society, ACM, 人工知能学会, 日本ソフトウェア科学会, 情報処理学会, 日本データベース学会等各会員．