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

大島裕明[†] 小山聡[†] 田中克己[†]

本研究では，ユーザが与えた1語のクエリに対して，Web検索エンジンが持つ情報のみから同位語とそのコンテキストを発見する手法について提案する．同位語とは，共通の上位語を持つような語のことである．従来研究として，同位語や，上位語，下位語などを求めるような研究は数多くあるが，それらはWeb上の文書を利用するものも含めて，巨大なコーパスを解析して大量の結果を求めるというものであった．我々の提案する手法では，Web文書のタイトルやスニペットといったWeb検索エンジンが持つ情報のみを，少ない回数のWeb検索によって取得し，それらを解析して同位語を発見する．提案手法では，ある語に対する同位語は並列助詞「や」で接続されることを利用してWeb検索エンジンに対するクエリを作成して，その検索結果のみから同位語を得る．そこでは何の事前準備も必要なく，また，あらゆる分野の語に対して同位語を発見することができる．さらに，発見された同位語とクエリの語の背後にあるコンテキストも同時に取得する．このような同位語発見は，Web検索におけるクエリ拡張や想起支援や，何かを調べるにあたって他のものと比較したいときの比較対象の発見など，幅広い分野で利用することができると考えられる．

Discovering Coordinate Terms with Their Contexts Using Web Search Engine Index

Hiroaki Ohshima,[†] Satoshi Oyama[†] and Katsumi Tanaka[†]

We propose a method of using only a Web search engine index to discover coordinate terms, i.e., terms that have the same hypernym. Several research methods acquire coordinate terms, but they require huge corpora or many Web pages. Our proposed method uses only the information in a Web search engine index such as titles and snippets of Web pages. These are obtained by a few Web searches, and then they are parsed to discover coordinate terms. We focus attention on coordinate terms that are connected by the coordinating particle "ya," and use those to make queries for a Web search engine. Our method does not require any preprocessing, and can find coordinate terms for terms in any field. At the same time, we find the background context between a query term and each discovered coordinate term. Such a service for discovering coordinate terms can be used in any field for such purposes as query expansion, word remembrance support system, or finding comparable objects.

1. はじめに

あらゆる語は様々な他の語を使って説明される．我々がある語を知っているというためには，その語に関連する他の語をいくつも知っていなくてはならない．語と語の関係性というものは重要であり，そのため，多くの辞書ではそのような関係性を説明しており，また，ある語に対して関連する語を発見するような研究も数多く行われている．本稿で我々が目的としているのもそのような研究の1つであり，同位語を発見するというものである．

† 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

語の関係を表す語には様々なものがあるため，それらについて説明しておく．なお，ここでは，「語」は単語もしくは成句を指すものとする．たとえば，“白鳥” も “白鳥の湖” も「語」である．「上位語」と「下位語」はそれぞれある語に対する上位概念や下位概念を表す語である．“トマト” の上位語は “野菜” であり，逆に，“トマト” は “野菜” の下位語である．「同意語」や「同義語」はまったく同じ意味を持つ語である．たとえば，“贈り物” と “贈答品” と “プレゼント” はすべて「同意語」「同義語」であり，文脈にかかわらずそれらを入れ替えても意味は変わらない．「類義語」は “装置” と “設備” や，“愛情” と “仁愛” といった，意味が似ている語のことである．本稿で対象とする「同位語」は，ある語に対して共通の上位語を持つような

別の意味の語のことである．すなわち，ある語 X がある語 Y の同位語であるとき，X と Y には共通の上位語 Z が存在する．たとえば，“トマト” と “ジャガイモ” は，共通する上位語として “野菜” が存在するため「同位語」である．共通の上位語を持つ語には「同位語」以外に「同意語」や「類義語」があるため，共通の上位語を持つ語のうち特に意味の異なる語のみを「同位語」とする．

我々が提案するのは，Web 検索エンジンが持つ情報のみを用いた同位語発見の手法である．ユーザは，クエリとして 1 語を入力する．その与えられたユーザのクエリの語の前後に並列助詞「や」を付加して，Web 検索におけるクエリを 2 つ作成する．それらのクエリによって Web 検索エンジンで検索を行うことで，検索結果としてタイトルとスニペットが得られる．それらのみを利用して，クエリの語と並列助詞「や」で接続されるような語を発見し，見つかった回数に応じてランキングを行い結果を返す．

同位語を発見すると同時に，見つかった同位語のコンテキストの発見も行う．クエリに対して発見された同位語のそれぞれには，異なる背景がある場合がある．そこで，複数の異なる形式でコンテキストを表現し，それらを提示することによって，ある語がなぜ同位語と判定されたのかをユーザが理解するのを助ける．本稿の同位語発見において，発見されたそれぞれの同位語に対して提示するコンテキストは，

- クエリの語と発見された同位語がともに使われている文
- 発見された同位語を特徴づける語とその重み
- クエリと発見された同位語との力関係

の 3 つである．また，上記にあげた，発見された同位語を特徴づける語とその重みをベクトルと見なして，発見された同位語のクラスタリングを行い，同位語の中でどのような語が似たようなコンテキストで使われているかを提示する．

また，クエリが多義語などであった場合，多様な意味や側面からの同位語が一度に得られてしまうが，特定の意味や特定の側面に対する同位語を求めるために，背景となる語をユーザが明示的に指定することによって，クエリの語の特定の側面における同位語を発見するような手法も提案する．

以降，2 章で本研究の位置づけについて，3 章で関連研究について，4 章で同位語発見手法の詳細について，5 章で発見された同位語のコンテキスト発見について，6 章でまとめについて述べる．

2. Web 検索エンジンインデックスのクイックマイニング

我々の研究の目的は，Web 検索エンジンが持っているインデックスデータのみから知識を抽出することである．これまで，Web 検索エンジンが保有するインデックスデータは Web ページを検索するためだけに用いられてきたが，少し工夫してインデックスデータを利用することによって，多くの役立つ知識を取得することができる．つまり，これまで我々は多くの情報を利用せずに放置してきたのである．本稿ではそのような知識取得の 1 つとして，同位語の取得を行った．

同位語などの，ある語に対して関連のある別の語を発見するための研究は，これまでにも数多く行われてきている．次章でそれらの研究について述べるが，それらのほとんどは，巨大なテキストコーパスや大量に収集した Web ページを解析したり，語の共起性や相互情報量の計算などの処理に多くの時間を使ったりする必要があった．それに対して我々の手法は，Web 検索エンジンのインデックスデータが利用可能であれば，それ以外に何の情報も事前処理も必要なく，必要なときに必要な知識を即座に得ることができる．新たな追加コストをほどんどかけることなく知識を得ることができれば，様々なアプリケーションにおいてその知識を利用することができる．ある語に対して別の関連する語を取得することができるサービスは，適応範囲が広いものであり，たとえば，Web 検索におけるクエリ拡張やキーワードの想起支援，何かを調べているときに同類の別のものとの比較をしたいときなどの比較対象の発見などでの応用が可能である．

本稿の同位語発見において，我々が解析するのは，Web 検索エンジンを利用したときに検索結果として得られるタイトルとスニペットのみである．提案手法では，Web 検索エンジンに対して 2 種類のクエリを与え，そこで得られた検索結果を解析する．たとえば，それぞれのクエリに対して 100 件ずつの検索結果を利用するとすれば，処理のために必要な時間は Web 検索の結果を取得する時間がほとんどであり，テキストの解析自体は量が大きくないためほとんど時間がかからない．

このような，Web 検索エンジンのインデックスデータのみを解析して知識を抽出するのを，我々は，“Web 検索エンジンインデックスのクイックマイニング” と呼ぶ．Web 検索エンジンのインデックスには，検索結果として出力されるような Web ページの URL，タイトル，スニペット，推定総検索件数などが含まれる．

Web 検索エンジンインデックスのクイックマイニングには，いくつかのメリットが存在する．まず，知識の取得のために必要な追加コストが低いことがあげられる．Web 検索エンジンのインデックスデータは検索サービス提供のためにコストをかけて作成されたものである．しかし，それを知識取得のために流用する際には特に事前準備などは必要なく，追加コストは低いといえる．また，適応分野が広いこともメリットとしてあげられる．Web には日々様々な分野における文書が追加されており，Web 検索エンジンはそのような文書をつねにインデックス化している．そのため，ありとあらゆる分野において適応可能となる．

従来研究のように，大規模コーパスなどに対して多くの処理を行って網羅的に知識を取得すること自体は今後も必要であると考えているが，我々が提案する Web 検索エンジンインデックスのクイックマイニングによる知識取得も今後広い分野で求められるものと考えている．今回行ったのは同位語発見だが，これ以外にも多くの知識取得を行える可能性があると考えている．

3. 関 連 研 究

同位語を取得するシステムとして Google Sets☆というサービスが Web 上に存在する．いくつか同位語と考えられる語を与えると，それらが属するような同位語の一群を見つけて結果として返す．Google Sets のアルゴリズムは公開されていないが，Google が収集した Web ページに含まれる語に対して大規模なクラスタリングを行い，それによって同位語のクラスタを大量に生成しているようである．また，現時点では英語のみに対応しており，日本語では利用することができない．

同位語を求めるためには，電子化された辞書を用いることも考えられる．そのような辞書としては，WordNet [1),☆☆]や EDR 電子化辞書[2)]，言語工学研究所デジタル類語辞典（シソーラス）[☆☆☆]などが存在している．これらを用いれば，上位語や下位語，また，同位語も取得することができる．人手による辞書の必要性がなくなることはないが，あらゆる語を網羅することは不可能であり，時の流れとともに更新していくことを考えれば，自動的にある語に対して関連のある別の語を発見するような手法は重要である．

同位語の発見に関する研究はいくつか存在している．Church ら[3)]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について提案した．厳密には同位語の発見を目的としたものではないが，発見される語には同位語が多く含まれることになる．この研究の，相互情報量が高くなるような語どうしは同位語である可能性が高いという知見は，他の研究のいくつかにおいても利用されている．Ghahramani ら[4)]による Bayesian Sets は Google Sets と同様のシステムの作成を目指したものである．語の共起テーブルのような大規模なデータに対して，ベイズ推定を用いて同位語のクラスタを発見する．アルゴリズムはシンプルで高速である．ただし，本アルゴリズムを適応するためには EachMovie や Grolier encyclopedia といった大規模データを用意する必要がある．Lin [5)]は類似するような語のクラスタを作成する手法を提案した．係り受け関係を利用して語どうしの類似度を計算することによって，類似する語のクラスタを生成する．そのため，係り受け解析が行われている大規模コーパスが必要となる．Shinzato ら[6)]は HTML 文書から同位語を発見する手法を提案した．HTML の構造に着目し，同レベルに列挙されているような語を，同位語である可能性がある語として取得する．それらの中でも特に語どうしの相互情報量や共起度が高いものが同位語である可能性が高いことを利用し，同位語取得を行う．解析する HTML 文書は，Web 検索エンジンを利用した検索結果をもとにして Web 文書を大量に収集しており，Web 検索エンジンインデックスのクイックマイニングとは異なるものである．彼らはまた，HTML の構造を利用して上位語や下位語を取得する研究[7)]も行っている．

上位語，下位語を求めるような研究もいくつも存在している．上位語と下位語を求めることができれば共通の上位語を持つ下位語の集合を求めることによって，同位語を発見することも可能である．Hearst [8)]は “such as” といったような上位語と下位語が現れるようないくつかのパターンにあてはまる語を発見することで，上位語と下位語を取得する手法を提案した．Sanderson ら[9)]は概念階層の抽出を行う手法を提案した．大量の文書群において，ある 2 語の出現の仕方に包含関係が見られるときに，その 2 語の関係性に上下関係を見るというものである．Glover ら[10)]は，ある語に対して親の概念を表す語，自分自身を指す語，子の概念を表す語，を取得する手法を提案した．ユーザがいくつかの文書例を与えると，それらと全体文書における語彙の出現の違いから，親，自身，子，という 3 つの関係性を持つ語を発見する．

共著者である Oyama ら[11)]は，ある語に対する詳

☆ http://labs.google.com/sets
☆☆ http://wordnet.princeton.edu/
☆☆☆ http://www.gengokk.co.jp/ruigo.htm

細語を発見する手法を提案した．HTML 文書でタイトルに出現するか本文に出現するかという簡単な構造情報を利用して詳細語かどうかを判定する．その際には Web 検索エンジンの結果で得られる Web 上での推定検索件数を利用しており，Web 検索エンジンインデックスのクイックマイニングであるといえる．Turney[12] や，Baroni ら[13] は，類義語の取得において Web 検索エンジンを利用した推定検索件数を用いて共起度や相互情報量を計算する手法を提案している．そこでは，そのような検索エンジンの持つ情報が従来的なコーパスの解析で得られる情報に代わるものとして利用できることを示唆しており，今後，Web 検索エンジンインデックスのクイックマイニングがより重要になってくると考えられる．

また，研究の目的は異なるが，相澤ら[14] は，コーパスからの知識抽出を行う手法で，特にコーパスが大規模になっていったときにも利用できるような手法の研究を行っている．その評価のために，大規模コーパスから並列助詞「や」に着目してテストセットにおける同位語の正解集合を求めている．

4. 同位語の発見

4.1 並列助詞「や」に着目した同位語の発見

本節では同位語の発見手法について述べる．我々のアプローチは，下記の 2 つの仮定に基づいている．

(1) 並列助詞「や」は同位語を接続することができる．
(2) 語 X と語 Y が並列助詞「や」によって接続され，「X や Y」「Y や X」という両方のパターンが存在するとき，語 X と語 Y は同位語である可能性が高い．

1 つ目の仮定は，あたりまえであると考えられる．並列助詞「や」によって接続されるのは，単語でも複合語でもかまわない．2 つ目の仮定も特別なものではなく，従来研究においても利用されているものである．相澤ら[14] は大規模コーパスから「A や B」「B や A」というパターンがともに 100 回以上出てくるような語のペアを発見し，それらを同位語の正解ペアとして提案手法の評価のために用いている．この，両側に出現する場合に同位語である可能性が高いという仮定が成り立つと，「や」で接続されている語が同位語であることの精度を上げるのに役立つばかりでなく，複合語を正確に切り出すことにも役立つことになる．これについては後で述べる．

並列助詞には「や」以外にも「と」「も」「とか」などがあるが，それらよりも「や」を利用した場合が良い結果を出すことは 4.3 節で述べる．

我々のアプローチは下記の 4 つのステップからなる．

(1) ユーザがクエリの語を与える．
(2) Web 検索エンジンに対するクエリを 2 つ作成して検索を行い，結果を取得する．
(3) 検索結果のタイトルとスニペットを解析する．
(4) 同位語の候補となる語をランキングして提示する．

まず，ユーザはクエリの語を与える．クエリの語は単語でも複合語でもかまわないが，1 語である必要がある．

次に，Web 検索のためのクエリを 2 つ作成する．並列助詞「や」をクエリの語の前後に付加したものである．たとえば，ユーザのクエリが「白鳥の湖」であるとき，Web 検索に対する 2 つのクエリは「”白鳥の湖や”」と「”や白鳥の湖”」となる．引用符で括っているのは，多くの Web 検索エンジンが実装しているフレーズ検索に対応するものである．フレーズ検索では，引用符で括られた検索語がそのまま出現するようなページを検索することができる．Web 検索の結果は，リスト形式で提示されることが多く，各アイテムはタイトル，URL，スニペットからなるのが一般的である．スニペットは，検索された Web ページの中に含まれいくつかの文で，検索語が出現するような文が含まれる．作成したクエリを基にして Web 検索を行い，得られたタイトルやスニペットが解析対象となるテキストである．

解析するテキストの中から，「白鳥の湖や」の直後と，「や白鳥の湖」の直前に現れるような語を取得する．そして，両側に出現するような語が見つかれば，それが本手法で同位語と見なす語となる．たとえば，「白鳥の湖」の例では，以下のような文が検索結果のスニペットで発見される．

(S1) まぁ，この曲自体，白鳥の湖や**くるみ割り人形**と比較して，こういうシンフォニックな演奏でも聞き映えするように書かれてるから...
(S2) 選ばれた曲は**くるみ割り人形**や白鳥の湖，カルメン，惑星‥のようなポピュラーなものから，...

この場合，「くるみ割り人形」という語が「白鳥の湖」と並列助詞「や」と接続されて前後両方において出現しているのが分かる．そのため，「くるみ割り人形」という語を「白鳥の湖」に対する同位語であると判定する．

このような「や」で接続されて出現するのがより多く検出される語の方が，あまり検出されない語よりも同位語として適していると考えられる．そのようなラ

ンキングを行うために，実際には並列助詞「や」の前後での出現回数を数えている．並列助詞「や」のどちらか一方でしかで出現しないような語は最終的には同位語とは見なさず取り除かれることになる．

Web 検索によって取得したタイトルとスニペットの解析において，まず，各文の分かち書きを行う．たとえば，先ほどのスニペットの「白鳥の湖」の付近を分かち書きすると，下記のようになる．

(S1') まぁ/，/この/曲/自体/，/白鳥/の/湖/や/くるみ/割り/人形/と/比較/して/，/

(S2') 選ば/れ/た/曲/は/くるみ/割り/人形/や/白鳥/の/湖/，/カルメン/，/惑星/・/・/

このとき，発見すべき同位語も単語や形態素に分割されてしまうため，単純に並列助詞「や」の隣の語を取得するというのではなく，語を連結しながら複合語でも発見できるようにしている．すなわち，1 つ目のスニペットの場合，

- 「くるみ」
- 「くるみ割り」
- 「くるみ割り人形」
- 「くるみ割り人形と」
- 「くるみ割り人形と比較」

のそれぞれを 1 度ずつ出現したものとして数え上げる．

表 1 が実際に「白鳥の湖」というクエリが与えられたときに，「"白鳥の湖や"」と「"や白鳥の湖"」というクエリで Web 検索を行い，それぞれにおいて 100 件ずつの検索結果を取得して，それらのタイトルとスニペットから並列助詞「や」の前後で出現した語とその出現回数の結果の一部である．複合語として正しい「くるみ割り人形」のみが並列助詞「や」の両側で出現しており，その他の複合語として誤った区切りのものはどちらか一方でしか出現していないことが分かる．このように，「や」の両側にあるという条件を課すことによって，複合語を正しく取り出すことが可能となる．

解析の結果，両側に出現する語は出現数に応じてランキングが行われる．その際に用いる指標としては，「や」の前後に出現した数の相乗平均，調和平均，相加平均などが考えられる．同じ回数出現したとしてもどちらか一方に偏って出現するよりも，両側に同様に出現していた方のスコアを高くするように，相乗平均を用いてランキングを行っている．

以上が，同位語発見の手法である．

4.2 実　　験

4.2.1 実験における諸条件

提案手法について，Web 検索エンジンを利用して同位語を取得する実験を行った．

表 1 問合せ「白鳥の湖」において「や」の前後で発見される語の頻度

Table 1 Frequencies of phrases preceding and following "ya" for the query "Swan Lake."

同位語の候補となる語	「や」前の出現数	「や」後の出現数
人形	6	0
割り人形	5	0
はくるみ割り人形	1	0
曲はくるみ割り人形	1	0
くるみ割り人形	4	17
くるみ割り人形と比較	0	1
くるみ割り人形と	0	2
くるみ割り	0	17
くるみ	0	18

まず，Web 検索エンジンの結果を取得する必要があるが，これらはプログラム上から Google API☆1や Yahoo!ウェブ検索 Web サービス☆2 などによって利用することができる．今回作成したシステムでは Google API を利用した．また，システムが作成する 2 つのクエリに対して Web 検索エンジンから取得する検索結果の最大数は，それぞれのクエリに対して 100 件とした．

先ほど，複合語の取得のために形態素をいくつか連結したものを 1 語として扱うと述べた．本システムでは，日本語の分かち書きのために形態素解析器「茶筌」☆3を用いており，たとえば，「くるみ割り人形」の場合は 3 つの形態素に分けられる．この場合，「くるみ割り人形」を 1 語と見なすためには，3 つの形態素を連結しなくてはならない．実験時には最大 6 つの形態素を結合したものまで 1 語と見なすようにした．

4.2.2 評価に用いる検索語

いくつかの語から同位語を取得して手法の評価を行った．我々が検索語として用いたのは，集英社の imidas 2006 から取得した 143 語である．imidas 2006 は，現在話題になっている事柄に関する用語辞典の最新版☆4である．一般の辞書には含まれないような語も取り上げられ，広い分野について網羅している．本研究では Web 検索エンジンのインデックスデータを用いる利点として，シソーラスのような辞書では取り上げられない語，新しい語，広い分野の語，などに対しても同位語が得られることを利点としてあげており，imidas 2006 には評価に用いるのにふさわしい語が含まれていると考えられる．imidas 2006 には分野区分

☆1 Google API http://www.google.com/apis/

☆2 Yahoo!ウェブ検索 Web サービス http://developer.yahoo.co.jp/search/web/V1/webSearch.html

☆3 http://chasen.naist.jp/hiki/ChaSen/

☆4 2006 年 6 月現在.

表 2 評価に用いた検索語のすべて
Table 2 Terms for evaluation.

景気判断，アジェンダ 21，全人代，骨太の方針，決済用預金，コーポレートガバナンス，雇用調整，M&A，貿易摩擦，人民元，経済学，日本 21 世紀ビジョン，E-JAPAN 戦略，知的財産権，エネルギー需要，建設市場，直取引，世代，広告媒体，ニュービジネス，農産物，平成の大合併，小泉政権，臨時国会，内閣，郵政民営化関連法案，ローカルマニフェスト，憲政，イラク復興支援，インド洋津波，55 年体制，胡錦涛，6 者協議，東アジア首脳会議，オーストラリア，インド洋大津波，ウズベキスタン，ロシア憲法，GCC 諸国，アフリカ，西バルカン，ヨーロッパ憲法，ラテンアメリカ，ブッシュ政権，国連改革，国際法，ナショナリズム，NPT，アメリカ陸軍，スマトラ沖地震，京都議定書，3R，マクロビオティック，リノベーション，都市再生，公益法人制度改革，ジェンダー，教育，7・5・3 問題，晩婚化，加齢，国民年金，社会保障，司法制度改革，海老沢 NHK 会長，アスベスト健康被害，給与所得控除，ライフデザイン，エタネルセプト，腹腔鏡手術，代謝症候群，ウイルス，悪性新生物，細胞，運動能力，母子健康手帳，自己決定権，組織健康度，星間物質，成層構造，津波災害，GPS 気象学，鉱床，誤差，物理学，元素，生物界，遺伝子，ファンデルワールス力，量子力学，ホログラム光ディスク，デスクトップ検索，情報セキュリティー，コンピューターウイルス，愛知万博，スペースデブリ，航空機，エネルギー資源，放射性同位体，化審法，科学リテラシー，ロハス，老化，プレイステーション 3，モーツアルト，レゲトン，純愛映画，中村勘三郎，コンバージョン，ヴェネツィアビエンナーレ，ローマ法王，ロボット，国家，入口遺跡，国語，年少者，和スイーツ，ロケ地ツアー，トレッキング，リコール，デジタル放送，コンパニオンアニマル，ガーデニング，クラシックレース，機動戦士 Z ガンダム，失踪日記，陸上競技，世界水泳選手権，オリックスバファローズ，大リーグ，ワールドカップ，ゴルフ場経営，グランドスラム，ラグビーワールドカップ，NFL，卓球，ノルディックスキー，体操競技，ボート競技，F1，総合格闘技，幕内，オフィシャルスポンサー制度

として 143 の区分が設けられており，評価に用いる語としては基本的にはそれらの区分において最初の見出し語を用いることとした．ただし，見出し語が複数の語の併記であったり，説明的であったりするなど，1 語と見なすにはふさわしくない場合があったため，場合に応じて，下記のように評価に用いる語を取得した．

(1) 語が併記されている場合は，最初の語を取得した．「スマトラ沖地震／インド洋大津波」からは「スマトラ沖地震」を，「愛知万博とロボット」からは「愛知万博」を用いた．
(2) 動向や行為を示すなど，1 語と見なせないような場合は主題と考えられる語を取得した．「給与所得控除の見直し」からは「給与所得控除」を，「55 年体制の終わり」からは「55 年体制」を用いた．
(3) (2) のように主題を取得できない場合は，次の見出し語を用いた．「2 強の時代」の次の「直取引」を用いた場合などがこれにあてはまる．

表 2 が，このようにして取得された 143 語のすべてである．様々な分野における様々な粒度の語が存在していることが分かる．

4.2.3 正誤の判定

各語に対してシステムから結果を取得し，得られた語に対して人手で正誤の判定を行った．クエリの語の同位語である語が正解であるが，特に正誤の判定は以下のような注意点に基づいて行った．

(1) 正解の語はクエリの語と共通の上位語を持たなくてはならない．
(2) 共通の上位語を持っていてもほぼ同様の意味を持つ語である同意語や類義語の場合は不正解とする．

表 3 同位語発見の実験結果のまとめ
Table 3 Summary of the experimental results of the coordinate term discovering.

	143 語すべて	平均
同位語として出力された語の数	1,086 語	7.59 語
うち正解数	847 語	5.92 語
うち不正解数	239 語	1.67 語
適合率	78.0%	69.1%

(3) クエリの語の上位語や下位語の関係（is-a）にある語は不正解とする．
(4) クエリの語と集約関係（part-of）にある語は不正解とする．
(5) クエリの語と正解の語との共通の上位語は，結果のそれぞれの語によって異なっていてもかまわない．たとえば，「ジャガー」がクエリの語である場合，共通の上位語として「車」を持つ「メルセデスベンツ」と，共通の上位語として「動物」を持つ「ヒョウ」は，それぞれを正解として扱う．
(6) 共通の上位語は，通常考えられる程度での上位語でなければならない．たとえば，「京都大学」と「Google」は共通の上位語として「組織」を持つと考えることも可能であるが，通常考えられる上位語よりもより上位の語であるため，このような場合は不正解とする．

4.2.4 評　　価

表 3 は，実験結果の集計である．まず，143 語に対して結果として出力された同位語の総数は 1,086 語であり，クエリ 1 語あたり 7.59 語の結果が出力された．そのうち，正解と判定されたものは総数で 847 語あり，クエリ 1 語あたりでは 5.92 語となった．不正解

表 4 同位語が発見できなかった語
Table 4 Terms for which no coordinate term was discovered.

M&A，日本 21 世紀ビジョン，直取引，郵政民営化関連法案，憲政，55 年体制，ロシア憲法，西バルカン，ヨーロッパ憲法，7・5・3 問題，海老沢 NHK 会長，アスベスト健康被害，組織健康度，成層構造，GPS 気象学，ホログラム光ディスク，ヴェネツィアビエンナーレ，入口遺跡，世界水泳選手権，ラグビーワールドカップ，オフィシャルスポンサー制度

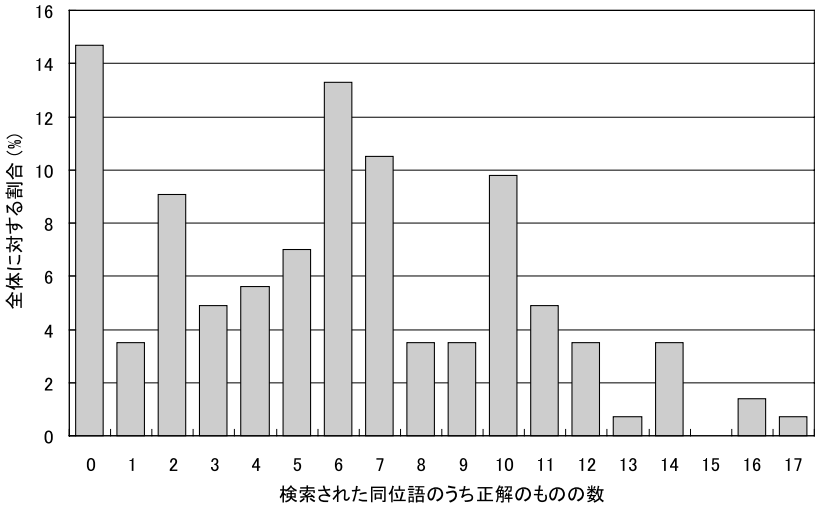

図 1 各語に対して発見された同位語数の分布
Fig. 1 The distribution of the number of discovering coordinate terms for each term.

と判定されたものは総数で 239 語あり，クエリ 1 語あたりにすると 1.67 語であった．得られた同位語の総数 1,086 語に対して正解と判定された語が 847 語あったため，全体における適合率は 78.0%となる．また，143 語のそれぞれの結果において適合率を求め，その適合率の平均をとった場合は，69.1%となった．

143 語のうち，同位語として正解がまったく得られなかったクエリは 21 語あり，全体の 14.7%において同位語が得られなかったことになる．**表 4** であげた語が正解がまったく得られなかった語のリストである．基本的には，Web 検索で同位語「や」を付加したらあまり検索結果が得られなかった場合に，結果が得られなくなる．そのような語の多くは，語としてまだあまり使われていないようなものである．たとえば，「組織健康度」「7・5・3 問題」「オフィシャルスポンサー制度」などは，Web 全体においても検索で 30 件未満しか見つからず，「や」を付加するとほとんど，もしくは，まったく Web 検索の結果が得られない．また，「日本 21 世紀ビジョン」のように，語としてはある程度の知名度があっても同位語がそもそも少ない場合もある．いくつかの語は，分割すれば同位語が発見できるようになるものがある．たとえば，「世界水泳選手権」では，「世界水泳」とすれば，「世界陸上」「オリンピック」「世界柔道」といった同位語を取得することができるし，「郵政民営化関連法案」では「郵政民営化」とすれば多くの同位語を取得することができる．このように，少し工夫することによって同位語を取得できるようになる場合は多い．

図 1 は，各語に対して見つかった正解同位語数の分布である．143 語中，正解数が 0 であったのは先ほど述べた 21 語，全体の 14.7%である．最も多い場合は正しい同位語を 17 語取得することができた．表 3 で，平均の正解数は 5.92 語となっているが，このグラフでも正解数が 6 語や 7 語の分布が多いことが読み取れる．

図 2 は，同位語として発見された語における適合率の分布である．正解数が 0 のときは適合率も 0 としているため，適合率が 0 の語の割合が全体の 14.7%となっている．正解が見つかるクエリの場合は，結果の適合率は良く，ほとんどが 70%以上の適合率である．また，全体においても，半数以上のクエリにおいて適合率が 80%以上になっている．

出力された 1,086 語の 22.0%にあたる 239 語は不正解と判定された．図 2 からは，適合率が 50%を下回る場合があることが読み取れる．不正解であった場合を分類すると，以下のようなものがあった．

- 複合語の共通部分の省略による間違い

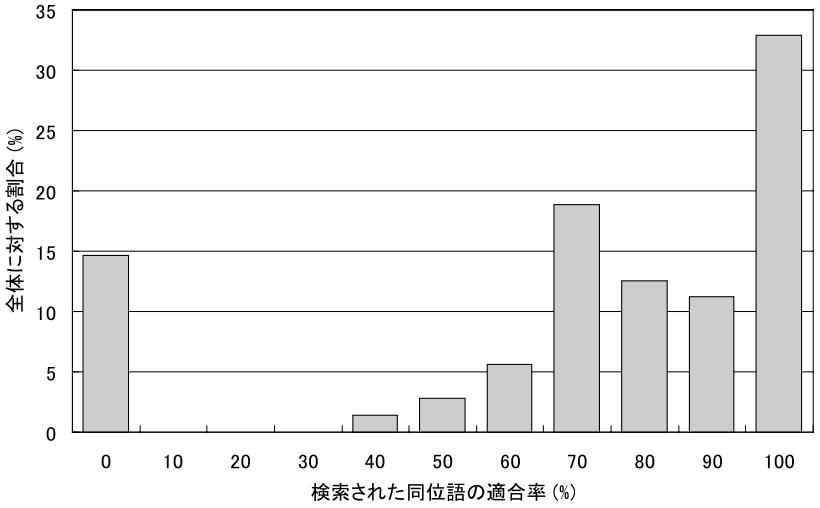

図 2 同位語発見の適合率の分布
Fig. 2 The distribution of the precision of the coordinate term discovering.

- 関連はあるが同位語ではない語
- 関連がない語や一般的な語
- 上位語
- 同意語

まず，不正解のいくつかは，クエリの語とその同位語がともに複合語であり，その後半部分が共通である場合に，その部分が省略されることによるものであった．たとえば，「スマトラ沖地震」という語に対して，「新潟県中越」という不正解の語が出力された．利用したスニペットには「新潟県中越やスマトラ沖地震」が 1 回，「スマトラ沖地震や新潟県中越」が 10 回出現していた．このとき，「新潟県中越やスマトラ沖地震」は，意味としては「新潟中越地震やスマトラ沖地震」であると考えられるが，「地震」が両方の語に共通であるため，省略されている．もう一方の「スマトラ沖地震や新潟県中越」は，「スマトラ沖地震や新潟県中越地震」という記述の一部であった．前後の出現回数が大きく異なることや，「新潟県中越」を完全に含んだ「新潟県中越地震」も結果の出力には含まれていることなどを利用して，このような間違いを少なくできる可能性がある．

関連はあるが同位語とはいえない語というのは，たとえば，「イラク復興支援」に対する「年金改革」のような語である．「小泉政権における政策」という上位語を考えたときには，共通の上位語を持つ語ととらえることも可能であるが，上位語として一般的とはいえないため不正解とした．

関連がない語が出力されることもある．これは特に，クエリの語が「や」で接続されている記述の一部を構成するような場合に起こることが多い．たとえば，「遺伝子」に対して「治療」という不正解の語が出力されたが，これは「遺伝子治療や遺伝子診断」，「関連遺伝子や治療薬の病態に対する作用」といった記述が存在したためである．これら 2 つの記述においては「遺伝子」は「や」で接続されている記述の一部である．同様の間違いによって出力されるものには，「情報」「利用」「方法」といった，複合語の末尾においてよく使われ，かつ，その語単体でもよく使われるような語もある．このような間違いを少なくするためには，分かち書き以上の自然言語処理を行って「や」で接続されている部分がどこであるかを判断したり，「情報」といった語はストップワードとして無条件に取り除いたりすることが考えられる．

上位語が出力されるのは，同位語において上位語を含む複合語が存在している場合や，単純に下位語と上位語が「や」で接続されている場合などがあった．たとえば，「国語」に対して「教育」が出力されたが，これは「英語教育や国語教育」や「国語や教育指導」という記述が存在したためである．また，「量子力学」に対して「物理」が出力されたが，これは「物理や量子力学」や「量子力学や物理化学」という記述が存在したためである．

同意語である間違いとしては，「ロハス」に対する「ローハス」があげられる．スニペットを見てみると，「ロハスやローハスと呼ばれ」，「ローハスやロハスとも表記する」といったような文章で使われていることが分かった．このような間違いを少なくするためには，「呼ぶ」「読む」「表す」などの日本語の表記のゆれを説明

表 5 様々な並列助詞による結果例
Table 5 Example results by various coordinating particles.

並列助詞	クエリ「京都大学」で発見された語 (並列助詞の前に出てきた回数:並列助詞の後に出てきた回数)
や	東京大学 (26:3)，大阪大学 (5:12)，神戸大学 (1:6)，同志社大学 (1:3)，九州大学 (1:2)， 東京工業大学 (2:1)，名古屋大学 (1:2)，京都府立医科大学 (2:1)， 奈良先端科学技術大学院大学 (1:2)，広島大学 (1:1)
と	京都市立芸術大学 (4:7)，NTT (4:6)，東京大学 (11:2)，早稲田大学 (6:2)，神戸大学 (2:2)，技術 (1:1)
か	東京大学 (8:2)，大阪大学 (2:4)，同志社大学 (1:2)
とか	大阪大学 (1:7)
も	今年 (2:3)，今 (4:1)

並列助詞	クエリ「フェラーリ」で発見された語 (並列助詞の前に出てきた回数:並列助詞の後に出てきた回数)
や	ポルシェ(23:22)，ランボルギーニ (5:15)，マセラティ(4:10)，ランボ (6:5)，ルノー (7:2)， アルファ(1:4)，アルファロメオ (1:2)，ロータス (1:1)，マクラーレン (1:1)，ベンツ (1:1)
と	ブリヂストン (7:2)，ルノー (13:1)，マクラーレン (1:8)，馬 (5:1)，ポルシェ(4:1)，シューマッハ (4:1) ランボルギーニ (1:3)，言う (1:2)，オリンパス (2:1)，ミハエル (1:1)
か	ポルシェ(7:4)，マクラーレン (2:4)，ピニンファリーナ (2:2)，ルノー (2:2)
とか	ポルシェ(1:3)
も	それ (4:1)，これ (1:2)，ルノー (1:1)，今 (1:1)，塗装 (1:1)

並列助詞	クエリ「情報処理学会」で発見された語 (並列助詞の前に出てきた回数:並列助詞の後に出てきた回数)
や	電子情報通信学会 (12:20)，通信学会 (14:2)，情報通信学会 (12:1)，人工知能学会 (4:3)， 音楽情報科学研究会 (1:1)，電気通信学会 (1:1)，電気学会 (1:1)，ACM (1:1)，言語処理学会 (1:1)
と	電子情報通信学会 (13:17)，情報ネットワーク法学会 (2:14)，KBS (4:1)，教育システム情報学会 (3:1)， IEEE (1:3)，物理学会 (1:2)，本学会 (2:1)，日本ソフトウェア科学会 (1:1)，ACM (1:1)

するような文か否かの判定が必要になると考えられる．

出力結果の適合率が 50%を切る語は，出力される語数が少なく，かつ，上記の間違いうちの複合語の共通部分の省略による間違いが起こりやすいような語であった．たとえば，「ゴルフ場経営」では，「リゾート開発」「不動産業」「ホテル経営」の 3 語を正解としたのに対し，「リゾート」「不動産」「ホテル」「理事会」の 4 語を不正解としたため，適合率が 50%を切ってしまった．

以上，同位語発見の手法について評価を行った．様々な分野における様々な粒度の語に対して，平均して 78%の適合率で 5.92 語の同位語を取得することができた．本システムの有効性が示された．

4.3 「や」以外の並列助詞との比較

並列助詞には「や」以外にも「と」「か」「とか」「も」などがある．これらを用いた場合について考察を行う．**表 5** は，それぞれの並列助詞を使った場合における結果の比較を示している．

クエリに「京都大学」と「フェラーリ」を用いた場合には，発見された語の数も適合率も「や」が最も良い結果である．「や」は列挙以外の用法があまり見受けられず，Web 検索エンジンから得られた文書より効率良く同位語候補を取得できる．そのため，発見できる語の数が多くなり，精度も良くなっていると考えられる．逆に，「か」「とか」「も」を用いた場合は，少数の語しか発見されなかった．これは，比較的列挙以外の用法が多いことや，使用頻度が低いことが原因であると考えられる．

「と」も，多少は他の用途で使われることがあり，「や」と比べると発見される語の数は少ない．また，発見される語は「や」で発見される語とはいくらか異なる性質のものとなる．たとえば「フェラーリ」に対して「ブリヂストン」が発見されているが，これは，F1 においてパートナーシップである関係から取得されたものと考えられる．「や」では，スーパーカーのメーカや，F1 におけるライバルチームなどが発見されている．「京都大学」に対しては「と」を使った場合に「NTT」が発見されているが，これも何らかの研究において協力関係にあったという文脈から発見されたものである．このような違いは，それぞれの並列助詞の用法の違いから生じるものと考えられる．

日本語における列挙の方法としては，限定列挙と非限定列挙がある．限定列挙は列挙されるものの数があらかじめ決まっている場合であり，「と」「か」などが用いられる．それに対して，非限定列挙は列挙されるものの数が不定であり，「など」「等」といった語を付加することで，列挙されているもの以外にもさらに列挙される可能性が残る場合であり，「や」「とか」「も」などが用いられる．たとえば，「私は英語と日本語を話すことができる」というのは限定列挙であり，「私

は英語や日本語を話すことができる」というのは非限定列挙である．

同位語の中にも，協力関係にあるような語と競合関係にあるような語がある．非限定列挙ではその両方をあまり区別することなく列挙される可能性があるが，限定列挙では頻度としては協力関係にあるような語の方がよく列挙されるものと考えられる．また，非限定列挙では例示する目的で多くの同位語が並べられる可能性が高いのに対して，限定列挙では例示目的の用途はあまりなく，具体的に何らかの関わりがあるようなものを列挙する場合が多い．そのような違いが「や」と「と」で得られる語の違いを生み出していると考えられる．

実際に検索結果のタイトルとスニペットの解析においてもそのような違いを見ることができる．クエリが「情報処理学会」の場合に「や」でも「と」でも「電子情報通信学会」が結果として得られるが，「や」を用いて Web 検索を行ったときに得られる文では，「情報処理学会や電子情報通信学会の論文誌などで…」「電子情報通信学会や情報処理学会にもインターネット関連の分科会が発足しており…」というようにある程度漠然とした文脈で並べられている．それに対して，「と」を用いて Web 検索を行ったときに得られる文では，「電子情報通信学会と情報処理学会が主催の FIT2004…」といったように具体的な事柄において協力関係にあったような文脈で並べられている．

語によっては「や」よりも「と」を使った方が良い結果を出力するような場合もあるのだが，日本語における用法の違いに着目すると，同位語発見という目的においては，協力関係，競合関係を問わずに様々な同位語が取得できた「や」が適していると考えられる．

5. 同位語のコンテキスト発見

5.1 発見された同位語のコンテキスト

発見されたそれぞれの同位語は，それぞれ異なった背景，コンテキストを持つ場合がある．たとえば，クエリの語が多義語である場合は，それぞれの意味において異なる同位語が存在していると考えられる．多義語ではない語であっても，語が表す事柄が様々な側面を持っており，各側面における同位語が異なる場合も十分に考えられる．

発見された同位語を提示するときに，それらがどのようなコンテキストを持っているかということも同時に提示することによって，ユーザがなぜそれぞれの語が同位語と判定されたのかを理解するのを助けることができると考えられる．

特に，一元的なコンテキスト表現ではなく，様々な表現によるコンテキストを同時に見せることにより，その効果は高くなると考えられる．そこでまず，発見されたそれぞれの同位語に対するコンテキストとして，以下の 3 種類による表現を行った．

（例文） クエリの語と発見された同位語がともに使われている文

（特徴語） 発見された同位語を特徴づける語とその重み

（力関係） クエリと発見された同位語との力関係（どちらがより有名か）

さらに，発見された同位語すべてをクラスタリングしたときにどのようなクラスタが生成されるかを表すことによって，発見された同位語どうしの類似性も理解できるようにした．

これらはすべて同位語を発見する際に用いるデータのみを利用して得られる情報であり，コンテキスト発見のために Web などからの新たな情報取得を行う必要はない．以後の各節において，それぞれの表現によるコンテキストをどのように取得するか述べる．

5.2 コンテキスト：例文

例文は，解析したテキストの中から，同位語とクエリの語が並列助詞で接続されている文を抜き出したものである．たとえば，クエリが「樋口一葉」のとき，「夏目漱石」「野口英世」「森鴎外」「泉鏡花」などが同位語として発見される．「夏目漱石」が同位語として出力されたということは，「夏目漱石や樋口一葉」や，「樋口一葉や夏目漱石」といった部分を含んだ文が存在したことになる．そのような文を例文として示す．たとえば，「夏目漱石」の場合は以下のような例文が提示される．

- 夏目漱石や樋口一葉をはじめとする多くの文豪たちのゆかりの地であり，大名屋敷の一部が今も残り，江戸の歴史に触れられる本郷は，
- どうして夏目漱石や樋口一葉が紙幣になって，私はならないのだね．

これにより，夏目漱石と樋口一葉のゆかりの地が本郷であることや，両者が紙幣になったことなどを読み取ることができる．また，同様に「野口英世」に対する例文としては以下のようなものが提示される．

- 今回のお札は野口英世や樋口一葉が肖像になりましたが，
- 野口英世や樋口一葉と聞くと薄幸の人というイメージで見てしまう．

これらからも，両者がお札の肖像であることや，ともに薄幸であったらしいことなどを理解することがで

表 6 クエリ「自由民主党」の同位語に対して抽出された特徴語の例
Table 6 Examples of feature terms for some coordinate terms of "Liberal Democratic Party."

労働党		緑の党	
労働党	32.7	ドイツ	23.9
Respect	22.8	緑の党	22.4
スコットランド	20.0	議席	19.1
議席	19.1	政党	17.0
松下	18.0	CSU	16.8
政経	18.0	首相	15.6
塾	15.8	Frankfurt	13.2
補選	15.2	Dusseldorf	13.2

きる．

コンテキストの表現としては特に工夫があるものではないが，実際にどのような文で使われているかが分かるため，直接的な理解に役立つと考えられる．

5.3 コンテキスト：特徴語

特徴語とは，発見された同位語とクエリの語とがともに現れるような文や周辺文章のコンテキストを特徴づける語であり，その重みは TF-IDF を用いる．たとえば，クエリが「樋口一葉」のときに「夏目漱石」が同位語と判定されるが，例文のところで取得した「夏目漱石や樋口一葉」や「樋口一葉や夏目漱石」といった部分を含んだ文そのものだけでなく，その文を含んだスニペット全体やそのスニペットを持つ Web 文書のタイトルも，「夏目漱石」という語が「樋口一葉」という語の同位語として使われるときのコンテキストを特徴づけるものと考えられる．

以下は，特徴語を求める処理である．各同位語に対して，「や」でクエリの語と連結したような部分が見つかった文を含むスニペットとタイトルをすべて収集する．それらから，すべての語の出現回数，すなわち Term Frequency（TF）を求める．この出現回数が多いような語が，発見された同位語とクエリの語のコンテキストを特徴づけるような語であると考えられる．さらに，一般的な語などの重みを下げるため，Inverse Document Frequency（IDF）を掛け合わせる．IDF は，Web 検索によって取得されたすべてのタイトルとスニペットから求める．クエリの語の前後に「や」を付加した 2 種の Web 検索に対するクエリのそれぞれにおいて検索結果を 100 件求めた場合，合計 200 件のタイトルやスニペットにおいてある語 $term$ が出現した回数を $DF(term)$ とすると，その語に対する IDF 値は $\log(\frac{200}{DF(term)})$ で計算される．

表 6 は「自由民主党」というクエリにおいて得られた同位語に対して，どのような特徴語が抽出されたかを示している．「自由民主党」というクエリに対しては，「公明党」「労働党」「緑の党」「共産党」などが同位語として発見される．「公明党」や「共産党」は日本の政党であり，すぐに理解することができるが，「労働党」や「緑の党」は日本の政党ではなく，日本人には理解しがたい．そこで，特徴語によるコンテキスト表現を見ることによって，理解を助けるのである．

表 6 は「自由民主党」というクエリに対して得られた「労働党」「緑の党」の 2 語についての特徴語によるコンテキスト表現である．

「労働党」が「自由民主党」と同位語であると見なされるようなコンテキストでは，「スコットランド」という語が特徴語として得られていることが分かる．実際には，労働党はイギリスの政党であり，ユーザがそれを理解するのを手助けする情報であるといえる☆．

「緑の党」では，「ドイツ」が特徴語であることが分かる．これにより，ユーザはこの場合の「自由民主党」や「緑の党」がドイツのものであること，「CSU」☆☆というものも関わりがあることが理解できる．

特徴語によるコンテキスト表現では，発見された語がクエリの語と共起するような文や，その周辺の文において頻出するような語を抽出している．例文によるコンテキスト表現では発見された語がクエリの語が共起する個々の場合が表されているため，より具体的で詳細な背景を知ることができると考えられる．それに対して，特徴語では関連するタイトルやスニペットのすべてを利用してコンテキストが作成されており，より総合的な背景を知ることができると考えられる．

また，特徴語と重みはベクトルと見なすこともできるため，コンピュータにも扱いやすい表現である．後述するクラスタリングにおいては，発見されたそれぞれの語の特徴を表すベクトルとして利用する．

5.4 コンテキスト：力関係

クエリと発見された語の力関係，またはどちらが有名であるかということも，取得することが可能である．それは，「や」の前後のどちらで多く出現したかということによって表される．

たとえば，六甲山系には，「六甲山」と「摩耶山」という山があるが，「六甲山」の方が有名である．「六甲山」というクエリで同位語の発見を行うと，「摩耶山」が得られるのだが，そのとき，「摩耶山や六甲山」が 2 回検出されるのに対して，「六甲山や摩耶山」は

☆ 実データを精査すると，2005 年のイギリスの総選挙において保守党がスコットランドで 1 議席しか獲得することができず，自由民主党や労働党が支持を得たとのことである．また，松下政経塾関係者が情報発信しているものがあったため，「松下」「政経」といった語が特徴語になっている．

☆☆ CSU はキリスト教社会同盟というドイツにおける地域政党の 1 つ．

表 7 「や」の前後の出現回数と Web における総ページ数の関係
Table 7 Correlation between the number of appearance of terms before and after a coordinating particle “ya” and the number of Web pages.

同位語	クエリ	「や」前	「や」後	同位語 Web 数	クエリ Web 数	「や」比率	Web 比率
羅臼岳	利尻岳	1	2	113,000	24,700	0.33	0.82
十勝岳	トムラウシ	3	14	245,000	186,000	0.18	0.57
十和田湖	八甲田山	15	20	693,000	330,000	0.43	0.68
鳥海山	月山	20	16	482,000	1,550,000	0.56	0.24
磐梯山	安達太良山	30	13	428,000	152,000	0.70	0.74
至仏山	平ヶ岳	3	2	113,000	24,500	0.60	0.82
大渚山	雨飾山	4	7	970	67,100	0.36	0.01
中禅寺湖	男体山	18	13	305,000	223,000	0.58	0.58
浅間山	草津白根山	20	7	689,000	67,400	0.74	0.91
鹿島槍ヶ岳	五竜岳	12	20	71,600	61,800	0.38	0.54
薬師岳	黒部五郎岳	11	10	124,000	44,600	0.52	0.74
蝶ヶ岳	常念岳	13	22	54,100	136,000	0.37	0.28
霧ヶ峰	美ヶ原	8	17	417,000	410,000	0.32	0.50
武甲山	両神山	10	9	97,100	76,600	0.53	0.56
蛭ヶ岳	丹沢山	6	7	21,000	79,600	0.46	0.21
御岳	恵那山	15	6	506,000	159,000	0.71	0.76
北岳	間ノ岳	33	6	641,000	70,300	0.85	0.90
聖岳	光岳	11	9	84,600	119,000	0.55	0.42
高野山	大峰山	11	4	1,220,000	177,000	0.73	0.87
阿蘇山	祖母山	11	5	398,000	78,400	0.69	0.84

16 回検出される．明らかに「六甲山」の方を先にする場合の方が多いことが分かる．これは，人が「六甲山」の方が有名であるために先にしていると考えられる．実際に，Google 検索で「六甲山」,「摩耶山」というクエリで Web 検索を行ったとき，「六甲山」では推定総ページ数が 890,000 件であるのに対して，「摩耶山」では推定総ページ数が 123,000 件となっている．大雑把ではあるが，推定総ページ数が有名度を代弁していると考えると，この場合は，「や」の前後の出現回数の違いが有名度を表しているといえる．

表 7 は，「や」の前後での出現回数の違いと，Web 検索で得られる推定総ページ数の相関を表す表である．まず，左から 2 つめの列はクエリの語である．これは，日本百名山から 20 個を無作為抽出した．一番左の列は，そのクエリの語から発見された同位語のうち，最もランキングが高い語である．左から 3 列目，4 列目は発見された同位語の「や」の前後における出現回数である．「同位語 Web 数」の列は発見された同位語で Web 検索を行ったときに得られる推定総ページ数，「クエリ Web 数」の列はクエリの語で Web 検索を行ったときに得られる推定総ページ数である．「「や」比率」の列は，同位語発見における総検出回数に対して，「発見された同位語 + “や” + クエリの語」という形で検出された回数の比率である．たとえば，一番上の「利尻岳」というクエリに対して「羅臼岳」という語が同位語として発見されているが，そのとき「羅臼岳や利尻岳」は 1 回,「利尻岳や羅臼岳」は 2 回検出されており，「「や」比率」の項目は 1/(1 + 2) で，0.33 となっている．一番右の列の「Web 比率」は同様の比率を Web における推定総ページ数から求めたものである．つまり，「羅臼岳」の推定総ページ数は 113,000 であり，「利尻岳」の推定総ページ数は 24,700 であるため，「Web 比率」の項目は 113,000/(113,000+24,700) で，0.82 となる．Web における総ページ数がある語の有名度を表していると仮定したとき，「「や」比率」と「Web 比率」に正の相関が見られれば，我々の同位語発見手法において「や」の前後に現れる回数からある程度の力関係，すなわち相対的な有名度を推定することが可能であることになる．

相関係数を計算すると，0.55 となり，実際に正の相関が見られた．よって，「「や」比率」で表されている値が 1 に近いほどクエリよりも発見された同位語がより有名であることを表し，0 に近いほどクエリのほうがより有名であることが表されていることが明らかになった．

5.5 発見された同位語のクラスタリング

これまで，発見された個々の同位語に対するコンテキスト発見について述べたが，本節では発見された同位語のうち，どの同位語が類似したコンテキストを持っているかを，クラスタリングを行うことによって提示する手法について述べる．

先述のとおり，発見された同位語のコンテキストの 1 つとして，TF-IDF によって重み付けされた特徴語が求められる．それらの特徴語と重みを特徴ベクトル

表 8 クラスタリングの結果例（クラスタリングを停止させるコサイン類似度の閾値は 0.025）
Table 8 Example results of clustering (The threshold cosine value for stopping clustering is 0.025).

クエリの語	クラスタの特徴を表す語	分類された同位語
ダヴィンチ	エロ，ファ，イタリア，街，絵	ミケランジェロ，キリスト教，ラファエロ，ニュートン，モナリザ
	投資，株，不動産，日経，新聞	アセット，ケネディ，パシフィック
法隆寺	時代，ヒノキ，倉，材，院	東大寺，薬師寺，正倉院，四天王寺，聖徳太子，奈良，興福寺
	条約，文化財，重要，計画，男性	姫路城，桂離宮
	回答，木造，建築，様式，細工	伊勢神宮，飛鳥寺
	像，光背，釈迦三尊，蝋，練物	大仏，東京国立博物館
スルメ	神，松前，塞，五徳，火鉢	餅，昆布，コンブ，お餅
	ザリガニ，糸，ざりがに，珍味，エサ	イカ，ヤリ，もち，煮干し，サキイカ，煮干，チッポ，食パン
	祖母，キムチ，コラム，足，灰	タコ，タコの足

と見なして，その特徴ベクトルの類似性を基にしてクラスタリングを行う．発見された同位語の中でも特に類似のコンテキストを持つ語は同一のクラスタを形成すると考えられる．距離計算にはコサイン類似度を用い，クラスタリングの手法としては最長距離法を用いた．試作したシステムではクラスタリングを停止させるコサイン類似度の閾値を 0 から 0.05 まで変化させることが可能である．以下では，0.025 を閾値として用いている．

表 8 は同位語のクラスタリングによるコンテキストの表現例である．表中でクラスタの特徴を表す語として提示しているのは，クラスタに属する同位語の特徴語の重みをすべて足しあわせたときに上位にくる語から，クラスタに属する語自身を除くことによって得られる語である．

このクラスタリングの結果により，たとえば，「ダヴィンチ」という語は多義語であり，投資や株に関連する語としての意味を持っていることや，その場合の同位語にどのようなものがあるかを見ることができる．「法隆寺」の例では，1 つめのクラスタが奈良のお寺などのクラスタであると考えられる．2 つめのクラスタには「姫路城」「桂離宮」がある．これらは重要な文化財の 1 つとして法隆寺が扱われているようなコンテキストでの同位語であることが分かる．「スルメ」の例では，1 つめが神事に関するクラスタであり，2 つめが珍味やザリガニのエサのクラスタであることが分かる．「食パン」という一見同位語でないような語が発見されていても，ザリガニのエサとしての側面から見ると同位語と判定しても間違いではない語があることも見てとれる．

以上，本節においてはクラスタリングによるコンテキスト表現について，前節までは発見されたそれぞれの同位語に対する 3 種類コンテキスト表現について述べた．ユーザは発見された同位語に対して，これらのコンテキストを基に様々な観点から理解を深めることができる．そして，さらに，クエリの語そのものに対する理解をも深めることになると考えられる．

5.6 背景語の指定によるクエリのある側面に対する同位語発見

前節まで，発見された語が持つコンテキストの違いを表現することや，クラスタリングによって多義語のそれぞれの意味や背景の違いから見たときの同位語クラスタを作成することについて述べた．しかし，どのような背景，コンテキストを持つ同位語を求めたいか，ユーザがあらかじめ分かっていて指定したいというような場合がある．

たとえば，「ジャガー」という語は多義語であり，「動物」や「車」といった複数の側面からまったく別の意味を持つ．これら「動物」や「車」などの背景やコンテキストを表す語を「背景語」としてユーザが指定することで，指定されたコンテキストを持つような同位語の発見を行う手法についても考えた．非常に単純な手法であり，同位語発見のために Web 検索を行うときに，ユーザが指定する「背景語」を Web 検索エンジンに対するクエリに追加するだけである．ユーザがクエリの語を「ジャガー」として，背景語が「車」であるとすると，Web 検索の際に用いるクエリは，「” やジャガー” ∧ 車」と「” ジャガーや” ∧ 車」となる．得られた結果からは，通常と同じように同位語を発見する．

表 9 でいくつかの例を示している．まず，「ジャガー」というクエリで背景語を指定しない場合は，「ヒョウ」や「ロールスロイス」といった異なる側面に対する同位語が混在していることが分かる．背景語に「動物」を指定すると，動物としてのジャガーの同位語のみが発見できていることが分かる．背景語を「車」とすると今度は欧米の自動車メーカなどが同位語としてあがっている．エレキギターに「ジャガー」というモデルが存在するため，背景語を「ギター」とすると，他のエレキギターのモデルの名前が同位語として現れるようになる．「ウイルス」がクエリの場合にも，パソコン，PC に感染するウイルスに対する同位語と，人体に感染するウイルスに対する同位語でそれぞれ別の

表 9 背景語を利用した同位語発見の結果例
Table 9 Example results of the coordinate term search using context terms.

クエリ	背景語	発見された語（「や」の前に出てきた回数:「や」の後に出てきた回数）
ジャガー		ヒョウ (6:4)，マドンナ (5:2)，ロールスロイス (3:3)，BMW (1:8)，ベンツ (5:1)，メルセデス (2:2)，ムスタング (2:2)，プジョー (1:2)，アストンマーチン (2:1)，ローバー (1:2)，蛇 (1:2)，ロールス (1:1)，バク (1:1)，ローバー (1:1)，オオカミ (1:1)，リンクス (1:1)
	動物	ヒョウ (18:3)，ピューマ (5:9)，ライオン (9:5)，オセロット (5:2)，チーター (2:5)，トラ (3:1)，クーガー (1:2)，毒蛇 (1:2)，鳥 (1:2)，バク (1:2)，オオカミ (1:2)，ピューマといった (1:1)，トラ，ヒョウ (1:1)，オオアリクイ (1:1)
	車	ベンツ (13:5)，BMW (4:9)，メルセデス (5:6)，ロールスロイス (6:5)，ローバー (4:5)，ボルボ (2:4)，アストンマーチン (4:2)，ポルシェ(2:2)，ロールス (2:2)，レンジローバー (1:3)，プジョー (1:3)，アルファロメオ (2:1)，アウディ(2:1)，ベントレー (2:1)，メルセデスベンツ (1:1)，アストンマーティン (1:1)，MG (1:1)
	ギター	ジャズマスター (15:14)，ムスタング (9:8)，ストラト (8:1)，マドンナ (5:1)，ジャズマス (2:2)，モズライト (1:1)
ウイルス		スパイウェア (8:21)，ワーム (6:21)，スパム (3:8)，細菌 (9:2)，ハッカー (3:4)，不正アクセス (3:4)，情報 (3:3)，スパムメール (1:4)，不正侵入 (2:1)，WINNY (2:1)，ファイル (1:1)，サイト (1:1)
	パソコン	スパイウェア (5:36)，不正アクセス (13:11)，ワーム (2:16)，トラブル (6:2)，ハッカー (3:3)，パソコン (4:2)，不正侵入 (2:2)，攻撃 (1:3)，セキュリティ情報 (1:1)
	PC	スパイウェア (9:19)，ワーム (3:30)，不正アクセス (4:2)，ファイル (4:1)，スパム (2:2)，不正侵入 (1:2)，ハッカー (1:2)，侵入 (1:1)，トロイの木馬 (1:1)，スパムメール (1:1)
	人体	細菌 (64:39)，菌 (8:4)，カビ菌 (2:4)，細胞 (4:2)，癌 (1:3)，ガン (1:2)，バイ菌 (1:1)

ものが求められている．

このように，あらかじめ，クエリの語のどのような側面における同位語を発見したいかが分かっている場合は，背景語を指定することによって，精度良く同位語発見を行うことができることが分かった．

6. ま と め

本稿では，Web 検索エンジンのインデックス情報のみを用いて同位語を発見する手法について提案を行った．Web 検索エンジンのインデックス情報には知識として利用できる情報が多く存在しており，少し工夫することによって，それらの役立つ知識を取得することができる．本稿ではそのような知識取得の 1 つとして，ユーザが与えた 1 語に対する同位語の取得を行った．従来研究においては巨大なコーパスを解析して大量の結果を取得する，といったことが行われているが，我々の手法では Web 検索エンジンのインデックスが利用できる状況であれば，事前処理などの大きな追加コストをかけることなく，あらゆる分野における語に対して，必要なときに同位語を即座に取得することができる．

まず，同位語が並列助詞「や」によって接続されることに着目して，ユーザから与えられたクエリの語の前後に「や」を付加して Web 検索に対するクエリを作成する．それらのクエリをもとに Web 検索を行い，得られた検索結果のタイトルやスニペットから，クエリの語と「や」で接続されている語を見つけることで同位語の発見を行う．評価のための検索語を用意し，実験を行ったところ，平均的には 70%から 80%程度の適合度でクエリ 1 語あたり 6 語程度は同位語が求められることが分かった．

ユーザがクエリの語や発見された同位語を理解するために，発見された同位語が持つコンテキストを複数の形式で表現して提示することも行った．発見された個々の同位語に対するコンテキストとして発見するのは，例文，特徴語，クエリの語と発見された同位語の力関係の 3 つであり，別々の観点からユーザの理解に役立つことを示した．また，発見された同位語全体に対してクラスタリングを行うことで，類似したコンテキストで使われる同位語をクラスタで提示することも行った．多義語のように複数の側面を持つような語に対しては，ユーザがその背景となるような語を指定することで，クエリの語の特定の側面における同位語を取得する手法についても提案を行った．

謝辞 本研究の一部は，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」，文部科学省科学技術振興費プロジェクト「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」（代表：田中克己），平成 18 年度科研費特定領域研究「情報爆発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」（課題番号：18049041，代表：田中克己），および，平成 18 年度科研費若手研究（B）「参照の同一性

判定に基づく複数 Web ページの検索閲覧方式の研究」（課題番号：16700097，代表：小山聡）によるものです．ここに記して謝意を表します．

参 考 文 献

1) Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J.: Introduction to WordNet: An on-line lexical database, *International Journal of Lexicography*, Vol.3, No.4, pp.235–312 (1990).
2) 独立行政法人情報通信研究機構：EDR 電子化辞書 2.0 版仕様説明書，株式会社日本電子化辞書研究所 (2001).
3) Church, K.W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Proc.27th Annual Meeting of the Association for Computational Linguistics*, pp.76–83 (1998).
4) Ghahramani, Z. and Heller, K.: Bayesian Sets, *Proc. 19th Annual Conference on Neural Information Processing Systems* (*NIPS2005*) (2005).
5) Lin, D.: Automatic Retrieval and Clustering of Similar Words, *Proc. 36th annual meeting on Association for Computational Linguistics*, pp.768–774 (1998).
6) Shinzato, K. and Torisawa, K.: A Simple WWW-based Method for Semantic Word Class Acquisition, *Proc. Recent Advances in Natural Language Processing* (*RANLP05*), pp.493–500 (2005).
7) Shinzato, K. and Torisawa, K.: Acquiring Hyponymy Relations from Web Documents, *Proc. Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting* (*HLT-NAACL04*), pp.73–80 (2004).
8) Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora, *Proc. 14th International Conference on Computational Linguistics*, pp.539–545 (1992).
9) Sanderson, M. and Croft, B.: Deriving concept hierarchies from text, *Proc. 22nd ACM SIGIR Conference* (*SIGIR'99*), pp.206–213 (1999).
10) Glover, E., Pennock, D.M., Lawrence, S. and Krovetz, R.: Inferring hierarchical descriptions, *Proc. 11th International Conference on Information and Knowledge Management* (*CIKM'02*), pp.507–514 (2002).
11) Oyama, S. and Tanaka, K.: Query Modification by Discovering Topic from Web Page Structures, *Proc. 6th Asia Pacific Web Conference* (*APWEB'04*), pp.553–564 (2004).
12) Turney, P.D.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, *Proc. 12th European Conference on Machine Lerning* (*ECML 2001*), pp.491–502 (2001).
13) Baroni, M. and Bisi, S.: Using cooccurrence statistics and the web to discover synonyms in a technical language, *Proc. 4th International Conference on Language Resources and Evaluation* (*LREC 2004*), pp.1725–1728 (2004).
14) 相澤彰子，中渡瀬秀一：係り受け関係を利用した類語・例文辞書構築法と大規模コーパスへの適用，人工知能学会第 20 回全国大会発表論文集，2E1-5 (2006).

(平成 18 年 6 月 20 日受付)
(平成 18 年 10 月 2 日採録)

(担当編集委員　福島 俊一)

大島　裕明（学生会員）
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．主に Web 検索，パーソナライゼーションの研究に従事．日本データベース学会，ACM 各学生会員．

小山　　聡（正会員）
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，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（正会員）
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．博士（工学）．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，日本データベース学会等各会員．