# 上位下位関係に基づく複合クエリの集約手法

松本　拓馬†　　北山　大輔†

† 工学院大学情報学部コンピュータ科学科　〒 163–8677 東京都新宿区西新宿 1-24-2
E-mail: †j111098@ns.kogakuin.ac.jp, kitayama@cc.kogakuin.ac.jp

**あらまし**　Web 検索において，言葉では言い表せるが検索クエリとしては適切に表現できない場合がある．そのような状況の場合，ユーザは，AND および NOT などのブーリアン検索を用いてクエリを拡張し，目的とする情報を試行錯誤しながら見つけなければならない．そこで，本稿では，拡張した複合クエリを下位概念となる語へ置き換えられると仮定し，ユーザが拡張した複合クエリを 1 語に置き換える手法を提案する．これは，Web 検索を行う際に概念的に下位となる語を用いるほど，検索結果が多義性のない具体的なものとなり無関係な Web ページを除外することができることから，複合クエリにおける検索結果と同様の結果を得られると考えたためである．ユーザは，本手法を用いることによって，複合クエリとは異なる観点から検索を行うこと，および，複雑な複合クエリを具体的な語に置き換えることができる．

**キーワード**　クエリ変換，Web ページクラスタリング，Web 検索，上位下位関係

## 1. はじめに

近年，インターネット情報空間には膨大な量の Web ページが蓄積されており，今もなお爆発的に増え続けている．このような状況の中，目的の情報を探す方法として Google や Bing をはじめとする検索エンジンがある．検索エンジンを利用すると，ユーザは検索クエリを入力するだけで Web ページを探すことができる．しかし，情報要求の段階には，言葉では表現できるものの，目的の情報を見つけるための検索クエリが表現できない段階がある．たとえば,「スクリプト言語」について調べる際に，そのキーワードを忘れた場合や知らない場合は検索クエリを適切に表現することは難しい．この場合，検索クエリの生成と検索を繰り返して目的の情報を見つけなければならない．

この解決方法のひとつとして，クエリ拡張があげられる．クエリ拡張は，ユーザが入力した検索クエリに対して，関連する単語の追加や置き換えなどを行う手法である．たとえば,「スクリプト言語」を「スクリプト言語 AND 種類」へ拡張したり，「スクリプト言語 AND 種類」を「スクリプト言語 AND タイプ」へ置換したりすることができる．このように拡張したクエリはユーザに提示され，ユーザは目的に応じて選択し検索を行える．ここで，ユーザが入力する検索クエリの語数は，年々増加傾向にあることが知られている [1]．検索クエリの語数を増やすと，目的の情報を絞り込むことができるが，検索結果は少なくなってしまう．そのため，意味を変えずに単語を減らすアプローチが重要であると考えられるが，クエリ拡張において，複数の単語を対象とした置き換え手法は提案されていない．

そこで，本稿では，2 語以上から構成されるクエリを 1 語へ集約する手法を提案する．以降では，2 語以上から構成されるクエリを複合クエリ，1 語へ集約したクエリを集約クエリと呼ぶ．本手法では，複合クエリの下位概念となる語に着目し，下位語に対して Web 検索結果をクラスタリングすることで集約クエリを決定する．なお，本研究では，クラスタリング手法を用いることにより，複合クエリの検索結果と Web ページの分類は同じものの，異なる観点となるような検索結果をもつ集約クエリを出力することを目標としている．

以降，2 章では本研究における手法の概要および関連研究について述べ，3 章では Web 検索結果のクラスタリング手法について述べる．4 章ではクエリの集約手法について述べ，5 章では実験および考察について説明する．そして，6 章でまとめと今後の課題について述べる．

## 2. 本研究のアプローチ

### 2.1 概　　要

本研究では，以下の手順で集約クエリを決定する．

（1） 複合クエリの Web ページクラスタリングを行う．
（2） 複合クエリの下位語を辞書から複数取得する．
（3） 下位語を絞り込み，集約クエリ候補を複数取得する．
（4） 集約クエリ候補の Web ページクラスタリングを行う．
（5） 手順 4 の結果から集約クエリを選定する．

たとえば，検索クエリを「プログラミング言語 NOT オブジェクト指向」とした場合は，まず，Web ページクラスタリングを行う．次に，辞書から「プログラミング言語」の下位語である「スクリプト言語」や「Java」などからなる単語集合を取得する．そして，単語集合から集約クエリとなりうる語（集約クエリ候補）を絞り込み，それらについて Web ページクラスタリングを行う．最後に,「プログラミング言語 NOT オブジェクト指向」の Web ページクラスタリング結果に最も結果が近い単語を集約クエリとする．この例の場合は,「スクリプト言語」や「アセンブリ言語」,「手続き型言語」などを理想の検索クエリの集約結果といえる．

### 2.2 関 連 研 究

#### 2.2.1　クエリ拡張

クエリ拡張の研究は，文書の検索を対象に古くから盛んに行われてきた．我々が知る限り，概念辞書を用いて検索クエリを

拡張する手法は 90 年代には研究がなされている．一方，近年は，情報爆発により特定の情報を見つけるのが困難となったことから，Web 検索を対象とする手法が盛んに研究されている．

吉田ら [2] は，検索結果を相関グラフによって可視化し，ノード間の距離から検索クエリの生成，および，リランキングを行う手法を提案している．彼らの手法では，ユーザが相関グラフのノードの操作することで，AND 検索，NOT 検索，および，OR 検索を組み合わせた検索クエリを生成できる．大石ら [3] は，検索クエリを構成する単語と関連性の強い単語を，センテンス間の距離から抽出するアルゴリズムを提案し，それを用いたクエリ拡張システムを提案している．大塚ら [4] は，Q&A サイトを用いて，話題の多様性および情報要求の曖昧さの二つの視点から拡張クエリを作成し，ユーザの情報要求支援する Web 検索システムを提案している．

これらの研究は，いずれもクエリを拡張する手法であり，複合クエリの集約は行っていない．

**2.2.2** Web 検索結果のクラスタリング

Web 検索結果をクラスタリングする手法として，HTML 文書構造やリンク構造などから Web ページを分類するものや，タイトルおよびスニペットから分類するものなどが存在する．安川ら [6] はクラスタ型の検索エンジンにおいてユーザにとって理解しやすいクラスタを生成するために，検索エンジンに蓄積された複合クエリのログを用いてユーザが入力した検索クエリの関連語を抽出し，関連語のみに限定したクラスタリング手法を提案している．平尾ら [5] は複合名詞の構成に着目し，クラスタが上位下位関係となる階層的クラスタリング手法を提案している．仁科ら [7] は，話題が類似したクラスタが複数出力されないように Web 検索結果を語句間の意味関係を考慮してクラスタリングする手法を提案している．

本研究では，互いに類似しないクラスタの代表語を生成できることから，仁科らの手法を参考にクラスタリングを行う．

## 3. Web ページクラスタリング

一般に，Web ページのクラスタリングを行うと，ある規則を基にして複数の Web ページをクラスタ単位に分類できる．本研究では，クラスタリングによる分類が似ているクエリどうしは互いに置き換え可能であると仮定し，Web ページクラスタリングを集約クエリ発見に利用する．

本手法における Web ページクラスタリングは，仁科らの手法を参考にしている．我々は彼らの手法を一部変更しており，高速形態素解析システム MeCab(注1) [8] を用いる点，名詞集合からストップワードを削除する点，形態素のパターンから複合名詞を作成する点，および，重要語を抽出できなくなるまで抽出する点で異なる．本研究における Web ページクラスタリングの手法は次の通りである．

（1） 入力クエリを基に Web 検索を行い，Web ページ $n$ 件分の検索結果（タイトル，スニペット）を取得する．

（2） 取得した検索結果のタイトルとスニペットに対して形態素解析を行い，複合名詞を含むすべての名詞を抽出する．

（3） 名詞集合からストップワードを除外する．

（4） tf-idf 法を用いて名詞集合のランキングを作成し，クラスタの話題を表すと考えられる互いに類似していない重要語を抽出できなくなるまで抽出する．

（5） 名詞集合を，コサイン類似度を用いて重要語を核とした単語グループに分類する．

（6） 単語グループを用いて Web ページを分類する．

3.1 節，3.2 節で本研究における Web ページクラスタリング手法を述べ，3.3 節で出力例を示す．

### 3.1 形態素解析による名詞の抽出

手順 2 では，すべての Web ページのタイトルとスニペットについて，Web ページごとに形態素解析し，名詞の抽出を行う．この際に，形態素の並びが「名詞の連続」,「名詞＋接尾名詞」,「接頭詞＋名詞」のパターンのいずれかに該当する場合は，形態素をつなぎ合わせて複合名詞として抽出する．ただし，形態素解析の際に名詞と判定される形態素のうち非自立の名詞および形容動詞語幹は抽出しない．

### 3.2 名詞集合内のストップワードの除外

手順 3 では，抽出した複合名詞を含む名詞集合から，連続した数字列，年月日，およびストップワードの辞書に該当する単語を除外する．ストップワードの辞書には，1 語のみの語をはじめ，スニペット上に多く含まれやすい “株式会社” や “お知らせ” などの語が登録されている．ストップワードをこの過程で除外することで，名詞集合を順位付けした際にストップワードが上位になることを防ぐことができる．

### 3.3 Web ページクラスタリングの出力例

検索クエリを「オブジェクト指向」とした場合を例に，本手法における Web ページクラスタリングの出力結果について説明する．なお，手順 1 における検索結果の取得では，Bing(注2) を用いて 10 件の検索結果を取得した．

表 1 「オブジェクト指向」における単語グループ

| 重要語 | 単語グループ |
|---|---|
| 考え方 | oriented，相互作用，メソッド，システム，オブジェクト同士 |
| オブジェクト指向プログラミング | 注目，プログラム，Wikipedia |
| クラス | |
| 中心 | プログラミング手法 |
| Java | 開発 |

表 2 「オブジェクト指向」における Web ページクラスタリング結果

| クラスタの代表語 | クラスタ内の Web ページタイトル |
|---|---|
| 考え方 | オブジェクト指向 - Wikipedia,<br>オブジェクト指向とは 【 OO 】 【 object oriented 】 - 意味/解説 ... ,<br>5 分で絶対に分かる：5 分で絶対に分かるオブジェクト指向 (1/6 ... ,<br>オブジェクト指向とは - はてなキーワード |
| オブジェクト指向プログラミング | オブジェクト指向プログラミング - Wikipedia,<br>オブジェクト指向 - ようこそ、Dayan のページへ ,<br>Insider.NET >オブジェクト指向プログラミング超入門 - @ IT |
| クラス | オブジェクト指向 - Seiichi Yoshida's Home Page,<br>オブジェクト指向とは - はてなキーワード |
| 中心 | オブジェクト指向とは - PHP 用語 Weblio 辞書,<br>オブジェクト指向 - ようこそ、Dayan のページへ |
| Java | 5 分で絶対に分かる：5 分で絶対に分かるオブジェクト指向 (1/6 ...,<br>Insider.NET >オブジェクト指向プログラミング超入門 - @ IT ,<br>オブジェクト指向とは 【 OO 】 【 object oriented 】 - 意味/解説 ... |

(注1)：http://mecab.sourceforge.net/

(注2)：https://datamarket.azure.com/ dataset/bing/search/

手順 5 における単語グループの生成結果を表 1 に示す．表では，互いに類似しない重要語を核とした単語グループが 5 つ生成され，グループに重要語と類似する単語が分類されている．

手順 6 では，以下のようにして，すべての Web ページを分類する．まず，タイトルおよびスニペットに重要語を含む場合は，重要語を代表語としたクラスタに分類する．この際，複数の重要語を含んでいた場合は，複数のクラスタに属する．次に，タイトルおよびスニペットに一番多くの単語を含む単語グループを見つけ，重要語を代表語としたクラスタに分類する．

このようにして分類した結果を，表 2 に示す．表では，重要語を核とした 5 つのクラスタに Web ページが分類されている．また,「オブジェクト指向 - Wikipedia」など複数のクラスタに属した Web ページがあることがわかる．

## 4. クエリの集約

本研究では，ユーザが入力した検索クエリの下位概念となる語を複数探し，集約を表すクエリにふさわしい単語を決定する．検索エンジンにおいて AND や NOT を用いて検索を行うと，あるクエリによる検索結果をより絞り込むことが出来る．一方，あるクエリの下位語を検索に用いると，より多義性のない具体的な検索結果となり，無関係な Web ページを検索結果から除外することができる．このことより，我々は AND や NOT を用いた複合クエリを，複合クエリを構成する語の下位語へ置き換えられると仮定し，本手法ではそのような単語をクエリの集約結果とする．たとえば，検索クエリを「プログラミング言語 NOT オブジェクト指向」とした場合は,「プログラミング言語」の下位語である「スクリプト言語」や「Java」などからなる単語集合のうち,「スクリプト言語」や「アセンブリ言語」,「手続き型言語」などを理想の検索クエリの集約結果とする．

検索クエリの下位概念を探す方法として，本研究では上位下位概念の辞書を用いた．しかし，本研究においてユーザから入力される検索クエリは必ず 2 語以上で構成されるため，検索クエリを構成する単語のうち，どの単語を用いて辞書から下位語を抽出するかが問題となる．ここで，地名やサイト名をあらわす単語を除く語で構成された 2 語の検索クエリは，1 番目の単語が 2 番目の単語（主要部）を修飾する傾向が高いことが知られている [11]．このことから，我々は，2 番目の単語の下位語を用いるよりも 1 番目の単語の下位語を用いた方がより具体的な事物に関する語が取得でき，精度の向上が期待できると考えた．たとえば,「プログラミング言語 AND 習得」という検索クエリでは，2 番目の単語「習得」は抽象的な語であり，1 番目の単語「プログラミング言語」が「習得」を説明している．そのため,「プログラミング言語」の下位語を用いた方がより具体的な語を取得できるはずである．以上より，本研究では検索クエリの 1 番目の単語における下位語を辞書から抽出する．

4.1 節で上位下位関係データベースについて述べ，4.2 節，4.3 節でユーザの入力した検索クエリを基に集約クエリを決定するまでの流れについて述べる．

### 4.1 上位下位関係データベース

上位下位関係データベースは，上位下位関係抽出ツール Version1.0(注3) [9] [10] を用いて，Wikipedia(注4) の 2014 年 6 月 24 日時点の全ページにおける上位下位関係を抽出し，作成したものである．上位下位関係抽出ツールでは，表 3 に示す Wikipedia における 3 種類の情報源から上位下位関係を抽出できる．各情報源において抽出される情報は，上位語，下位語，および，上位下位関係の信頼度の評価スコアである．たとえば，Category の情報源から抽出したデータでは，上位語「アニメ作品」に対する下位語は 3916 語あり，そのうちの下位語「名探偵コナン」の評価スコアは 1.240835 となっている．上位下位関係のデータは，Category を情報源に抽出したデータが 42 万 8815 件，Definition を情報源に抽出したデータは 224 万 8103 件，Hierarchy を情報源に抽出したデータは 604 万 1107 件である．

表 3　情報源における上位下位関係抽出方法の違い

| 情報源 | 上位下位関係抽出方法 |
|---|---|
| Category | カテゴリタグの単語から上位下位関係を抽出する． |
| Definition | 「～とは」などの定義文から上位下位関係を抽出する． |
| Hierarchy | 箇条書きなどの階層構造から上位下位関係を抽出する． |

本研究では，下位語のデータ数が一番多い，Hierarchy の情報源から抽出したデータを用いる．以降では，Hierarchy の情報源から抽出したデータを格納したデータベースを上位下位関係データベースもしくは単にデータベースと呼ぶ．また，上位下位関係の信頼度の評価スコアが負の値の場合において，上位下位関係とならない語が多く見受けられたため，負の値である下位語は抽出しない．ただし，評価値が正の値の語においても，そのような語が存在することがあるが，本研究では許容する．

### 4.2 集約クエリ候補の取得

上位下位関係データベースから下位語を完全一致検索で取得すると，下位語がとても少ない，あるいは存在しないことが多い．たとえば，上位語として「オブジェクト指向」を持つ語をデータベースで完全一致検索すると，0 件となり下位語が存在しない．これは，データベース上で上位語が「オブジェクト指向プログラミング言語」や「オブジェクト指向開発方法論」などのように登録されているためである．そのため，本研究ではデータベースから下位語を検索する際は，前方一致検索を行うこととした．このようにすることで，多くの下位語を取得できるようになり，先の例では，54 語の「オブジェクト指向」の下位語を取得することが可能となる．しかしながら，これらすべての下位語に対して，4.3 節で述べる集約クエリの判別手法を適用すると，多大な処理時間を要してしまう．そこで，下位語の関連語集合を作成し，検索クエリと関係のある語であるかを評価することで，下位語集合内から候補を絞り込むこととした．以下では，絞り込み手法について述べる．

#### 4.2.1 関連語集合の作成

まず，次の手順で下位語の関連語集合を作成する．

（1） 検索クエリから 1 番目の単語を抽出する．

（2） 抽出した単語と前方一致する単語を上位下位関係データベースから取得する．

(注3)：http://alaginrc.nict.go.jp/hyponymy/

(注4)：http://ja.wikipedia.org/

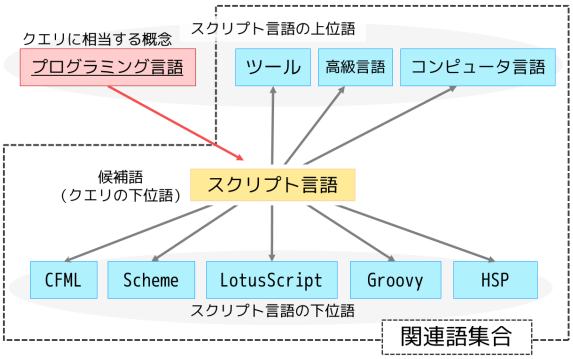


図 1　候補語「スクリプト言語」と関連する語

（3）　手順 2 で取得した単語の下位語を取得する．

（4）　取得した下位語の上位語および下位語を取得する．

（5）　手順 3，4 で取得した単語の和集合を関連語集合とする．

以降では，手順 2 において取得した下位語と手順 3 において取得した下位語の下位語を区別するために，手順 2 において取得した下位語を候補語と呼ぶ．図 1 に候補語「スクリプト言語」と関連する語を示す．図中の上部 4 語が「スクリプト言語」の上位語を表しており，下部 5 語が下位語を表している．ただし，データベースから取得できる「スクリプト言語」の関連語は図示されている語よりも多い．以降では，この図を例に関連語集合の作成手順を説明する．検索クエリを「プログラミング言語 NOT オブジェクト指向」とした場合を考える．

まず，検索クエリの 1 番目の単語は「プログラミング言語」なので,「プログラミング言語」と前方一致する上位語をデータベースから取得する．図中では,「プログラミング言語」と前方一致する語のひとつとして「プログラミング言語」が存在することを，赤色の長方形が表している．なお，この例では「プログラミング言語」と完全一致しているが，前方一致する語を抽出するため,「プログラミング言語の種類」などの語も抽出する．

次に,「プログラミング言語」の下位語をデータベースから取得することで，候補語を得る．図中では，黄色の長方形がこれを表しており,「プログラミング言語」の下位語のひとつとして「スクリプト言語」が存在することがわかる．

そして，候補語「スクリプト言語」の上位語および下位語をすべて取得する．図中の青色の長方形がこれを表しており，「スクリプト言語」の上位語として「ツール」や「高級言語」などの語が,「スクリプト言語」の下位語として「CFML」や「Scheme」などの語が存在することがわかる．

最後に，手順 3 および手順 4 で取得した単語を元に関連語集合を作成する．図中では，点線で囲まれた部分の語集合が関連語集合となる．赤色の長方形の語を関連語集合に含めないのは，検索クエリと前方一致していることから，既に検索クエリと関係性があることが分かっているためである．

### 4.2.2　候補語の絞り込み

候補語の関連語集合作成の次の手順として，候補語の絞り込みを行う．この手順では，前の手順で抽出した候補語を順位付けすることで，10 語に絞り込む．

まず，関連語集合を基にすべての候補語のスコアを算出し，

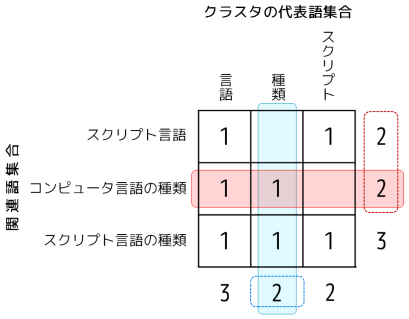


図 2　$pm(R_k, C_q)$ の算出例

順位付けを行う．ある候補語を $k$ としたときのスコアは次式で計算する．なお，$q$ は検索クエリ，$C_q$ は検索クエリ $q$ によって生成されたクラスタの代表語集合，$R_k$ は候補語 $k$ における関連語集合を表す．

$$candidate(k) = \frac{pm(R_k, C_q)}{|C_q| + |R_k| - pm(R_k, C_q)} \tag{1}$$

ここで，$pm(R_k, C_q)$ は次の手順より算出する．

（1）　評価値の初期値を 0 とする．

（2）　関連語集合 $R_k$ とクラスタ代表語集合 $C_q$ において，すべての代表語と部分一致する数が最小の関連語を取得する．部分一致数が同じ関連語が複数あった場合はランダムに取得する．

（3）　取得した関連語と部分一致するクラスタの代表語から，$R_k$ 全体で最も部分一致数が少ない代表語を取得する．最小値の代表語が複数あった場合はランダムに取得する．

（4）　評価値に 1 を加算する．

（5）　取得した関連語および代表語を各集合から除外する．

（6）　$R_k$ に 関連語が存在する場合は，手順 2 に戻る．存在しない場合は評価値を返す．

$pm(R_k, C_q)$ の算出例を図 2 に示す．図は，関連語集合が { スクリプト言語，コンピュータ言語の種類，スクリプト言語の種類 }，クラスタの代表語集合が { 言語，種類，スクリプト } である場合の例である．格子内の数値は，関連語集合とクラスタの代表語集合における部分一致数を表しており，格子外の数値は部分一致数の総和を表している．また，赤点線で囲まれた部分および青点線で囲まれた部分は，それぞれ，関連語集合における最小の部分一致数，その関連語に含まれるクラスタの代表語集合における最小の部分一致数を表している．たとえば，手順 2 において，関連語集合における最小の部分一致数である「スクリプト言語」および「コンピュータ言語の種類」からランダムに取得した結果,「コンピュータ言語の種類」が取得されたとする．この場合，手順 3 では,「コンピュータ言語の種類」に部分一致する代表語である「言語」および「種類」から，全体で最も部分一致数の少ない代表語「種類」を取得する．このように，$pm(R_k, C_q)$ は関連語と代表語を部分一致によって対応付けを行い，対応数を評価値として算出する．したがって，$candidate(k)$ は，候補語 $k$ における関連語集合と，検索クエリによって生成されたクラスタの代表語集合において，要素どうしが部分一致している語数が多いほど高い値となる．

次に，候補語のランキングから上位 10 語を抽出し，絞り込

み結果とする．上位 10 語としたのは，実行時間の短縮に加え，10 語以内に集約クエリの正解になりえる語が含まれることが多かったためである．なお，以降では絞り込み結果 10 語を集約クエリ候補と呼ぶ．

### 4.3 集約クエリの選定

集約クエリ候補から集約クエリを選定するためにスコアを算出する．スコアの算出は，$k$ を集約クエリ候補，$C_k$ を集約クエリ候補 $k$ によって生成されたクラスタの代表語集合，$c$ をクラスタの代表語，$|c|$ をクラスタ $c$ に属する Web ページ数としたとき，次式で行う．

$$equiv(k) = \frac{|C_k \cap C_q|}{|C_k|} \times \sum_{c \in C_q} w(c, C_k) \quad (2)$$

$$w(c, C_k) = \begin{cases} |c| & (c \text{ が } C_k \text{内のいずれかの要素と完全一致する場合}) \\ 0 & (\text{上記以外}) \end{cases} \quad (3)$$

ただし，クラスタの代表語集合 $C_k$ には，集約クエリ候補 $k$ のクラスタリング結果に加えて，集約クエリ候補 $k$ 自体も含める．これは，検索クエリ $q$ の Web ページクラスタリングにおいて，集約クエリ候補 $k$ がクラスタの代表語となる場合があるのに対し，集約クエリ候補 $k$ のクラスタリングでは，集約クエリ候補 $k$ 自体の語はクラスタの代表語とならないためである．

$equiv(k)$ の前半部分では，集約クエリ候補によって作成されるクラスタの代表語集合に，どの程度正解クラスタの代表語が含まれているかを求めている．正解クラスタの代表語とは，検索クエリ $q$ によって作成されるクラスタの代表語集合 $C_q$ の各代表語のことである．つまり，前半部分の式は，検索クエリによるクラスタリング結果に一番近いクラスタリング結果となった単語ほど高い値となる．

一方，後半部分では，重みを求めている．この重みは，検索クエリと集約クエリ候補のクラスタの代表語集合を比較し，完全一致したクラスタに属する Web ページ数を総和したものである．なお，検索クエリのクラスタリング結果と集約クエリ候補のクラスタリング結果では，クラスタの代表語が同じであっても属する Web ページ数は異なる．しかし，集約クエリの選定においては，検索クエリのクラスタリング結果を基準とするため，検索クエリのクラスタリング結果から Web ページ数を求めることとする．たとえば，検索クエリのクラスタ代表語集合が { コンピュータ，言語，解説 }，集約クエリ候補のクラスタ代表語集合が { コンピュータ，スクリプト言語，解説 } の場合，クラスタの代表語「コンピュータ」と「解説」が完全一致しているので，検索クエリにおける「コンピュータ」および「解説」クラスタに属する Web ページの総和を重みとする．

つまり，$equiv(k)$ は，適合率に正解クラスタの重要度の総和を掛け合わせたものであり，検索クエリのクラスタリング結果において多くのページが属しているクラスタを，高い適合率で多く生成できる集約クエリ候補が高い値となる．

## 5. 実験および考察

本手法を実装し，候補語の絞り込みの精度，および，集約クエリが置き換え可能であるかを評価するための実験を行った．

表 4　候補語の絞り込み結果

| 複合クエリ | 候補語数 | 集約クエリ候補を含む割合 | 実行時間 |
|---|---|---|---|
| 紅茶 AND 三大 | 74 | 60% | 4:48 |
| 計測 AND 放射線 | 78 | 40% | 5:30 |
| 公用語 AND フィンランド | 290 | 40% | 19:59 |
| タイ料理 AND レシピ NOT 激辛 | 50 | 40% | 3:08 |
| 水族館 AND 日本 AND ラッコ | 52 | 20% | 3:36 |
| 銀行 AND 中心機関 | 415 | 0% | 29:13 |

表 5 「東京 AND 紅葉」における集約クエリ候補のランキング

| 順位 | 集約クエリ候補 | 正解クラスタ数 | 全クラスタ数 | 適合率 | 重み | スコア |
|---|---|---|---|---|---|---|
| 1 | 小石川後楽園 | 14 | 79 | 0.177 | 117 | 20.734 |
| 2 | 六義園 | 11 | 67 | 0.164 | 95 | 15.597 |
| 3 | 向島百花園 | 10 | 96 | 0.104 | 89 | 9.270 |
| 4 | 薬師池公園 | 11 | 94 | 0.117 | 76 | 8.893 |
| 5 | 洗足池 | 9 | 86 | 0.104 | 68 | 7.116 |
| 6 | 東京大学大学院理学系研究科附属植物園 | 8 | 85 | 0.094 | 63 | 5.929 |
| 7 | 国立科学博物館 | 5 | 63 | 0.079 | 29 | 2.301 |
| 8 | 夢の島熱帯植物館 | 6 | 98 | 0.061 | 35 | 2.142 |
| 9 | 中央公園 | 5 | 74 | 0.067 | 23 | 1.554 |
| 10 | 日本科学未来館 | 0 | 5 | 0 | 0 | 0 |

表 6 「計測 AND 放射線」における集約クエリ候補のランキング

| 順位 | 集約クエリ候補 | 正解クラスタ数 | 全クラスタ数 | 適合率 | 重み | スコア |
|---|---|---|---|---|---|---|
| 1 | シンチレーション検出器 | 15 | 71 | 0.211 | 67 | 14.154 |
| 2 | 半導体検出器 | 12 | 59 | 0.203 | 52 | 10.576 |
| 3 | 直接測定 | 6 | 51 | 0.117 | 27 | 3.176 |
| 4 | ネットワーク・アナライザ | 7 | 90 | 0.077 | 31 | 2.411 |
| 5 | 実車走行評価 | 6 | 78 | 0.076 | 30 | 2.307 |
| 6 | 圧力 | 5 | 58 | 0.086 | 25 | 2.155 |
| 7 | スペクトラムアナライザ | 6 | 78 | 0.076 | 26 | 2.000 |
| 8 | 長さ | 5 | 79 | 0.063 | 26 | 1.645 |
| 9 | 慣性計測装置 | 4 | 69 | 0.057 | 20 | 1.159 |
| 10 | 間接測定 | 1 | 52 | 0.019 | 13 | 0.250 |

Web ページクラスタリングにおける Web 検索は，Bing を用いて 100 件の結果を取得する．

### 5.1 候補語の絞り込みの評価

まず，4.2.2 節で述べた候補語の絞り込み手法によって，集約クエリ結果となりうる候補語を取得できているかどうかを検証する．そのために，すべての候補語を 4.3 節の式 $equiv(k)$ を用いて順位付けし，上位 10 語中に絞り込み後の集約クエリ候補となった 10 語が含まれる割合を算出した．6 つの複合クエリについて算出した結果を表 4 に示す．たとえば，複合クエリ「紅茶 AND 三大」では，74 語の候補語を $equiv(k)$ を用いて順位付けし，その上位 10 語において，候補語の絞り込み手法によって選ばれた集約クエリ候補 10 語が 60%含まれていたことを表している．この結果から，全候補語における $equiv(k)$ の順位付けにおいて，上位 10 語のうち平均 33.3%が集約クエリ候補であることがわかる．

高い精度で絞り込みが出来ない原因として，各候補語における関連語数の少なさがあげられる．表 4 に示した複合クエリにおける関連語数の中央値の平均は 4.3 であり，候補語における関連語数は全体的に少ないことがわかる．このように，候補語の関連語数が少ない場合，$equiv(k)$ 式の $pm(R_k, C_q)$ が低い値となり，結果として複合クエリと関連のある候補語であるかを評価することが困難となる．そのため，絞り込みは，候補語の関連語数によって精度が大きく変化すると考えられる．

次に，候補語の絞り込みを行わない場合にかかる実行時間に関して，計測した結果から絞り込みの必要性を確認する．表 4 の実行時間は，それぞれの複合クエリで候補語の絞り込みを行

表 7　評価実験の結果

| 複合クエリ | 回答者数 | 比較手法 | | 提案手法 | |
|---|---|---|---|---|---|
| | | 集約クエリ（上位 3 語） | 正解率 | 集約クエリ（上位 3 語） | 正解率 |
| 犬 AND アイフル | 5 人 | チワワ，アイフル犬，人気 | 67% | チワワ，十石犬，日本スピッツ | 33% |
| 遊び AND 子供 NOT スポット | 7 人 | 子ども，昔，外遊び | 33% | 福笑い，人間の遊び，言葉遊び | 0% |
| 寺院 AND 舞台 AND 京都 | 7 人 | 清水寺，清水，お寺 | 33% | 清水寺，法隆寺，松尾寺 | 33% |
| 計測 AND 放射線 | 6 人 | 放射線計測，放射線測定器，測定 | 33% | シンチレーション検出器，半導体検出器，直接測定 | 0% |
| 甘味料 AND ノンカロリー | 9 人 | ノンカロリー甘味料，砂糖，人工甘味料 | 33% | 天然甘味料，スクラロース，アスパルテーム | 33% |
| コンビニ AND 品揃え AND 豊富 | 7 人 | 商品，セブン，ネットスーパー | 0% | セブンイレブン，デイリーヤマザキ，ミニストップ | 33% |
| ブラジル NOT サッカー | 3 人 | ブラジル国債，世界，日本 | 33% | セラード保護地域，ゴイアス歴史地区，サン・ルイス歴史地区 | 67% |
| 東京 AND 紅葉 | 3 人 | 名所，紅葉スポット，東京都 | 0% | 小石川後楽園，六義園，向島百花園 | 67% |
| スポーツ AND イギリス AND バット | 7 人 | 野球，クリケット，ボール | 0% | アシックス，クリケット，ミズノ | 0% |
| 植物 AND 観察 AND 小学生 | 6 人 | 研究，夏休み，植物観察 | 33% | 馬場大門のケヤキ並木，バラ園，朝顔 | 0% |
| テーマパーク AND ハリーポッター | 7 人 | ハリー，ポッター，USJ | 33% | ユニバーサル・スタジオ・フロリダ，ウォルト・ディズニー・スタジオ・パーク，エプコット | 0% |
| 菓子 AND 極細 AND グリコ NOT ポッキー | 5 人 | ポスカ，お菓子，プリッツ | 33% | プリッツ，スナック菓子，ようかん | 33% |
| 水族館 AND 日本 AND ラッコ | 7 人 | 飼育，鳥羽水族館，日本水族館立体生物図録 | 0% | 日本の水族館，ベルリン動物園，トレド動物園 | 33% |
| 銀行 AND 中心機関 | 7 人 | 中心，機関，機関銀行 | 0% | 埼玉りそな銀行，北日本銀行，みちのく銀行 | 0% |
| 魚 AND ニモ | 5 人 | カクレクマノミ，映画，ファインディング・ニモ | 67% | ピラニア，ソードフィッシュ，テトラ | 0% |
| マリオ AND 大砲 AND 名前 | 7 人 | 攻略，城，ゲーム | 0% | バタクリボー，ゲッソー，ボムへい | 0% |
| 料理 AND 辛い | 5 人 | タイ料理，韓国料理，レシピ | 33% | 麻婆豆腐，キムチ，おでん | 33% |
| 紅茶 AND 三大 | 6 人 | 世界三，大紅茶，大銘茶 | 33% | 祁門紅茶，ウバ，アッサム | 33% |
| アニメ AND 日常系 AND ほのぼの | 5 人 | 日常，日常系アニメ，漫画 | 67% | めだかボックス，ポヨポヨ観察日記，ドン・ドラキュラ | 0% |
| 小惑星 AND はやぶさ | 5 人 | 小惑星探査機，イトカワ，地球 | 33% | イトカワ，金星横断小惑星，準衛星 | 0% |
| 山 AND 雪化粧 | 5 人 | 雪，浅間山，写真素材 | 33% | 美ヶ原，朝日岳，冠松次郎 | 0% |
| ニュース番組 AND NHK | 7 人 | ニュース，番組，NHK 総合 | 33% | NHK ニュース 10，NNN 朝のニュース，FNN 福井テレビニュース 6 | 33% |
| 公用語 AND フィンランド | 7 人 | フィンランド語，スウェーデン語，英語 | 33% | エストニア語，ハンガリー語，サーミ語 | 0% |
| スノーボード選手 AND 有名人 | 8 人 | スノーボード，芸能人，スポーツ選手 | 0% | 青野令，平野歩夢，ショーン・ホワイト | 100% |
| 温泉 AND 大分 AND 有名 NOT 宿 | 4 人 | 大分県，別府，別府温泉 | 33% | 別府温泉，東山温泉，筋湯温泉 | 33% |
| | | 平均 | 28.0% | 平均 | 22.7% |

わなかった際の集約結果を得るまでの時間を表している．この結果から，候補語数に応じて多くの時間がかかることがわかる．これに対し，候補語の絞り込みを行う場合は 10 語のみを $equiv(k)$ で順位付けすればよいため，表における 6 つの複合クエリの平均実行時間は 1 分 26 秒と比較的短い時間で実行することができる．そのため，本手法によって集約クエリを得る場合は絞り込みが必要であるといえる．

### 5.2　集約クエリの置換えに関する評価

#### 5.2.1　実験方法

複合クエリを集約クエリに置き換え可能かどうかを確認するために，被験者 10 名に対してアンケートを行った．被験者には，複合クエリ，提案手法における上位 3 つの集約クエリ，および，比較手法における上位 3 つの集約クエリを 15 セット提示し，それぞれについて置き換え可能である集約クエリを選択してもらった．ただし，置き換え可能な集約クエリが存在しない場合は，選択しなくてもよいこととした．また，複合クエリは 25 語用意し，その中からランダムに 15 語抽出したものを被験者に提示することとした．これらの複合クエリは，理想とする集約クエリが上位下位関係データベース上に存在するものを用いた．また，複合クエリを構成するすべての単語と関係する単語が複合クエリの検索結果には含まれていると考えられることから，比較手法として，複合クエリの検索結果における名詞を tf-idf 法による順位付けする手法を用いた．具体的には，3 章で述べた手順における手順 1 から手順 3 までを行い，tf-idf 法を用いて順位付けを行う．なお，ある Web ページにおけるタイトルとスニペットのセットを文書とするとき，すべての文書においてある単語が出現する数を tf，すべての文書においてある単語が出現する文書数を df とした．

#### 5.2.2　提案手法における集約クエリ候補の順位付け

実験で用いた 2 語の複合クエリについて，本手法によって集約クエリ候補を順位付けした結果を表 5 および表 6 に示す．表の列は，4.3 節の式 $equiv(k)$ と対応しており，正解クラスタ数は $equiv(k)$ における $|C_k \cap C_q|$，全クラスタ数は $|C_k|$，適合率は $equiv(k)$ の前半部分の式，重みは $equiv(k)$ の後半部分の式，スコアは $equiv(k)$ を表している．

表 8　「銀行 AND 中心機関」における全候補語のランキング

| 順位 | 集約クエリ候補 | 正解クラスタ数 | 全クラスタ数 | 適合率 | 重み | スコア |
|---|---|---|---|---|---|---|
| 1 | 第 2 地方銀行 | 12 | 76 | 0.157 | 86 | 13.578 |
| 2 | 地方銀行 | 11 | 80 | 0.137 | 86 | 11.825 |
| 3 | 商工組合中央金庫 | 6 | 47 | 0.127 | 66 | 8.4255 |
| ⋮ | | | | | | |
| 48 | 中央銀行 | 7 | 87 | 0.080 | 37 | 2.977 |
| ⋮ | | | | | | |
| 230 | 日本銀行 | 3 | 71 | 0.042 | 13 | 0.549 |
| ⋮ | | | | | | |

表 9　正解クエリにおける集約クエリ順位の割合

| 集約クエリの順位 | 比較手法（21 語） | 提案手法（17 語） |
|---|---|---|
| 1 位 | 47.6% | 70.6% |
| 2 位 | 19.0% | 17.6% |
| 3 位 | 33.3% | 11.8% |

表 5 は，複合クエリを「東京 AND 紅葉」とした場合における集約クエリ候補のランキング結果である．2 万 3965 語ある候補語の中から，東京の紅葉名所である「小石川後楽園」や「六義園」などを高いスコアにできている．一方，表 6 は，複合クエリを「計測 AND 放射線」とした場合における集約クエリ候補のランキング結果である．放射線検出器である「シンチレーション検出器」および「半導体検出器」が上位となっているが，クエリ候補には「ネットワーク・アナライザ」や「圧力」など明らかに不正解の集約クエリ候補が含まれている．実験では，ランキング結果から上位 3 位の集約クエリ候補を提案手法の集約クエリとして用いた．

#### 5.2.3　実験結果と考察

全回答者のうち，過半数の被験者が置き換え可能と判断したクエリを正解とするときの評価実験の結果を表 7 に示す．ここ

で，複合クエリ「犬 AND アイフル」を例に表について説明する.「犬 AND アイフル」における集約クエリの列は，両手法の集約結果の上位 3 語として，比較手法においては「チワワ」,「アイフル犬」,「人気」が，提案手法においては「チワワ」,「十石犬」,「日本スピッツ」が順に得られたことを表している．また，下線は回答者 5 人のうち，過半数が置き換え可能として選択された正解の集約クエリを表している．そして，正解率の列は，集約クエリのうち正解のクエリとなった割合を表している．

まず，提案手法の集約結果のうち，正解率が 0%となった複合クエリ「銀行 AND 中心機関」について考察する.「銀行 AND 中心機関」は,「中央銀行」や「日本銀行」などが理想の集約結果として考えられるが，提案手法では上位 3 語に含まれなかった．ここで,「銀行 AND 中心機関」の全候補語について集約した結果を表 8 に示す．その結果,「中央銀行」の順位は 48 位,「日本銀行」は 230 位と低い結果となった．この原因として,「銀行 AND 中心機関」のクラスタリング結果において,「中心」クラスタに含まれる Web ページ数が 43 と多いことがあげられる．このように，多くの Web ページが属すクラスタが存在する場合は，そのクラスタが生成されるだけで集約クエリ候補が高いスコアとなることが問題点といえる．

次に，両手法における平均正解率について考察する．評価実験の結果，比較手法の平均正解率が 28.0%であるのに対し，提案手法は 22.7%となり，比較手法に比べ提案手法の平均正解率は 5.3%劣る結果となった．しかしながら，提案手法では複合クエリと異なる観点のクエリのみが正解となっている中，比較手法では複合クエリから構成されるクエリが正解に複数含まれている．たとえば，複合クエリ「犬 AND アイフル」では「アイフル犬」が，複合クエリ「計測 AND 放射線」では「放射線計測」などが正解となっている．このため，提案手法は，比較手法よりも複合クエリに関して，異なる観点となる集約クエリを多く出力できると考えられる．また，多くの被験者が複合クエリを置き換え可能と判断しており，比較手法と同程度の平均正解率となったことから，複合クエリを下位語へ集約することは比較手法と同程度有効であったと考えられる．

最後に，集約クエリの順位付けについて，表 9 に両手法における集約クエリの順位の割合を示す．たとえば，比較手法についてならば，1 位となった集約クエリが 21 語中の 47.6%，2 位となった集約クエリが 21 語中の 19.0%，3 位となった集約クエリが 21 語中の 33.3%あったことを表している．この結果から，比較手法よりも提案手法の順位付けのほうが，より適切な順序関係となっていると考えられる．

## 6. まとめと今後の課題

本研究では，2 語以上から構成されるクエリを 1 語へ集約する手法を提案した．具体的には，複合クエリの下位概念となる語に着目し，Web ページクラスタリングによる類似する結果を表す語を集約結果とした．実験結果より，集約クエリの置き換えが可能であることや集約クエリの順位付けの有効性を確認した．今後の課題として，集約クエリによって異なる観点となる検索結果を取得することができるかを評価するとともに，候補語の絞り込み精度を向上させるための候補語の絞り込み方法の見直しなどを行う予定である．また，複合クエリの 1 番目の単語から下位語を検索することについて，実験で用いたクエリは作成されにくいと考えられることから，複合クエリの中で広義な単語を取得するなど下位語の取得方法の検討を行う予定である．

## 謝　　辞

本研究の一部は，平成 26 年度科研費若手研究 (B)(課題番号 : 24700098) によるものです．ここに記して謝意を表すものとします．

## 文　　献

[1] Marti A. Hearst 著，角谷和俊，田中克己 監訳 :『情報検索のためのユーザインタフェース』，第 4 章 クエリ指定，pp.102-103，共立出版，2011.

[2] 吉田 大我，小山 聡，中村 聡史，田中 克己 : Web 検索結果におけるキーワード出現相関の可視化と対話的な質問変換，電子情報通信学会第 18 回データ工学ワークショップ第 5 回 DBSJ 年次大会 (DEWS2007)，2007.

[3] 大石哲也，峯恒憲，長谷川隆三，藤田博，越村三幸 : 関連単語抽出アルゴリズムを用いたクエリ拡張，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)，C4-3，2009.

[4] 大塚 淳史，関 洋平，神門 典子，佐藤 哲司 : 情報要求の言語化を支援するクエリ拡張型 Web 検索システム，第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM2011)，F6-3，2011.

[5] 平尾 一樹，竹内 孔一 : 複合名詞に着目した Web 検索結果のクラスタリング，情報処理学会研究報告. 情報学基礎研究会報告，Vol.2006，No.94，pp.35-42，2006.

[6] 安川 美智子，横尾 英俊 : クエリログから獲得した関連語のクラスタリングに基づく Web 検索，電子情報通信学会論文誌. D, 情報・システム，Vol.90，No.2，pp.269-280，2007.

[7] 仁科 朋也，内海 彰 : 単語グループに基づく Web 文書クラスタリング，自然言語処理，Vol. 17，No. 4，pp. 23-41，2010.

[8] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.

[9] 隅田飛鳥，吉永直樹，鳥澤健太郎 : Wikipedia の記事構造からの上位下位関係抽出，自然言語処理，Vol.16，No.3，pp.3-24，2008.

[10] 山田 一郎，橋本 力，呉 鍾勲，鳥澤 健太郎，黒田 航，Stijn De Saeger，土田 正明，風間 淳一 : Wikipedia を利用した上位下位関係の詳細化，自然言語処理，Vol.19，No.1，pp.3-23，2012.

[11] 中渡瀬 秀一，大山 敬三 : 検索クエリにおける修飾構造の調査，電子情報通信学会技術研究報告 (思考と言語)，Vol. 110，No.407，pp.49-52，2011.