# アンカーテキストとリンク構造を用いた同義語抽出手法

黒木 さやか[†1]　山名 早人[†2]　立石 健二[†3]　細見 格[†4]

†1 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

†2 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

†3,†4 日本電気株式会社 〒630-0101 奈良県生駒市高山町 8916-47

E-mail: [†1, †2]{kuroki,yamana}@yama.info.waseda.ac.jp

[†3]k-tateishi@bq.jp.nec.com [†4]i-hosomi@ay.jp.nec.com

**あらまし** Web2.0 に代表される新しい情報発信の仕組みにより，企業や商品に対する一般ユーザの評価は，他の一般ユーザだけではなく，企業にとっても貴重な情報源となっている．しかし，企業や商品の評価に関する Web ページは，それらの略称や俗称を用いて書かれていることが多く，検索クエリに正式名称を入力しただけでは取得することができない．そこで本論文では，アンカーテキストとリンク構造を用いることで，略称や俗称などにも対応した同義語抽出の手法を提案する．関連研究としてクエリの翻訳語を発見する研究が存在するが，同手法により作成される翻訳語ランキングは，翻訳語をトップにすることを目的としており，頻出語が上位にランキングされるようになっている．従って，頻出ではない略称や俗称などの同義語を効率的に抽出することは難しい．提案手法では，既存手法よりも多くの同義語を抽出すると同時に，新しい同義語候補ランキングの指標を提案し，同義語抽出の効率化を試みる．実験では既存手法に比べ，精度を保った上で，網羅性を約 15%向上させることができた．

**キーワード** 同義語抽出，クエリ拡張，アンカーテキスト，リンク構造

# Extracting Synonyms using Anchor Texts and Link Structures

Sayaka KUROKI[†1]　Hayato YAMANA[†2]　Kenji Tateishi[†3]　Itaru Hosomi[†4]

†1 Graduate School of Fundamental Science and Engineering, Waseda University 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555, Japan

†2 Science and Engineering, Waseda University 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555, Japan

†3, †4 NEC Corporation 8916-47 Takayama-Cho, Ikoma, Nara, 630-0101, Japan

E-mail: [†1, †2]{kuroki,yamana}@yama.info.waseda.ac.jp

[†3]k-tateishi@bq.jp.nec.com [†4]i-hosomi@ay.jp.nec.com

**Abstract** Due to the new mechanism of information transmissions, such as Web2.0, general users' evaluations for companies and products have become a valuable information source for companies as well as for the other users. However, Web pages containing companies' evaluations are written using either abbreviated names or common slang so that we cannot obtain those pages by inputting official names as the search engines' query terms. In this paper, we propose the method to extract synonyms including abbreviated names or slang using anchor texts and link structures. There is related research which finds the translations of Web query terms, but this method aims to rank the query's translated term as the Top-1 and frequent terms rank high in the ranking. Therefore efficient extraction of the synonyms which are not frequent, like abbreviated names or slang, is difficult. In our way to make synonyms' rankings, we try to improve the effectiveness of extracting synonyms than the existing research, as well as trying to keep the recall rates at the same time. In our experiments, we can estimate Top-200 ranking of synonyms, the result is a 15% increase in the recall while we are keeping the accuracy.

**Keyword** Synonym Extraction, Query Expansion, Anchor Text, Link Structure

## 1. はじめに

近年インターネットが大幅に普及したことにより，企業や商品に対する評価が Web 上で多く見られるようになっている．これまでの一般ユーザは，自らの評価を公に示す機会に恵まれていなかったが，インターネットを用いることで自由に発言することが可能となった．Web2.0 の概念で表わされるように，ユーザの評価はそれらを閲覧する他のユーザに影響を与え，企業や商品のイメージを決定付けることにつながっている．企業側から見ても Web の情報は，自社に関する忌

憚なき意見を抽出できる，貴重な情報源である．

自社に関する情報を抽出するためには，特定の口コミ掲示板を参照するか，検索エンジンを用いる方法が一般的である．商用検索エンジンは，クエリの表記ゆれを解消する技術などを組み込んでおり，目的の Web ページを効率的に取得することが可能である．表記ゆれ解消の技術とは，漢字とひらがなの違いを吸収する機能，多くのユーザが間違えるスペルを補正する機能などを指す．しかし，ユーザによる評価記事，特にマイナスの評価記事には，企業の略称や俗称しか現れない場合が多く，自然言語処理をベースとした技術だけではこれらの Web ページを抽出することができない．

上記の問題を解決する試みとして，クエリ拡張に関する研究が行われている．クエリと同じ意味を持つ語を利用することで，クエリに関連する Web ページをより多く集めることが目的である．シソーラスを用いた研究[1]では精度の高い同義語を抽出できるが，シソーラスには新語や俗語は含まれていない．クエリログを用いた研究[2]では，新語やマイナーな語は抽出することが可能だが，俗語で検索を行うユーザは少ない．一般的な情報を知りたい場合には，正式名称や略称で検索をすれば十分だからである．

新語や俗称に強い同義語抽出の手法としては，アンカーテキストとリンク構造を用いる手法が効果的であると考えられる．図 1 に表すように，同じ URL を指すアンカーテキストは同義語である可能性が高い．企業のページなどではアンカーテキストに正式名称を用いる半面，個人のページや掲示板などでは略称や俗称を用いる傾向があり，多様な同義語を抽出することができる．クローリングの頻度を上げることで，新語に対応することも容易である．この手法を用いた既存研究[7]では，クエリの翻訳語をアンカーテキストの中から抽出しており，実験では高い精度を出している．

しかし，[7]の手法による翻訳語ランキングは翻訳語をトップにランキングさせることが目的であり，頻出なアンカーテキストが上位にランキングされやすい．従って，頻出ではない略称や俗称などの同義語を効率的に抽出することは難しいという問題がある．同義語抽出の網羅性を高めるためには，頻出ではない同義語ほど抽出できることが望ましい．そこで本論文では，アンカーテキストの類似度指標を新たに提案することで，同義語抽出の網羅性を保ちつつ，精度の高い同義語ランキングを作成する手法について提案する．人手による評価をランキングに反映させる Relevance-Feedback の技術を利用することにより，同義語抽出の網羅性とランキング精度の向上を試みる．

本稿の構成は，以下の通りである．まず 2 節で提案手法に関連した研究をまとめ，3 節で既存研究の問題点について述べる．4 節で提案手法の詳細について述べ，5 節で評価実験を行う．


図 1　　同一 URL にリンクするアンカーテキスト

## 2. 関連研究

### 2.1. クエリ拡張

大量のデータから，検索クエリに関連する文書を探す時，検索クエリと同様の概念を持つ語についても，文字列検索を行うことが効果的であると考えられる．1990 年代までのクエリ拡張分野では，自然言語処理に基づく研究が一般的であったが[3]，インターネットの普及により自然言語処理以外の技術が注目されている．

シソーラスを利用したクエリ拡張技術では，特に Wikipedia を利用した研究がさかんである[1]．クエリと同じ名前を持つ Wikipedia のページに着目し，そのページへのリダイレクトを同義語とする．また，そのページと似たようなサイト内リンクを張るページの項目名も，クエリの同義語と定義している．Wikipedia は人手が作成したシソーラスであるため，精度の高い同義語が抽出できるが，新語やマイナーな語については網羅率が下がる欠点がある．

検索エンジンのクエリログを利用し，クエリ拡張を行う研究もされている[2]．同じセッション内に入力されたクエリは，最初に入力したクエリをユーザが言い換えたものであるとして，クエリログから同義語抽出を行っている．流行の語やマイナーな語を抽出しやすい特徴があるが，俗称などはログに含まれにくい．また，検索ログの多くは公開されておらず，一般で実用化するのは難しいという欠点がある．

### 2.2. コミュニティ抽出

Web から特定の事柄に関するページ群を取り出す手法として，コミュニティ抽出の研究が挙げられる．[4][5]の研究では，「同じ事柄を述べたページ群は相互リンクを張りやすい」という考え方に基づき，Web のリンク構造から完全，または密な 2 部グラフを抽出している．また，コミュニティ内のリンク数が，コミュニティ外のリンク数よりも多いという定義に基づき，Web のリンク構造に s-t 最大フロー問題を適用した研究もある[6]．

我々の提案手法は，URL 間のリンクではなく，アンカーテキストと URL 間のリンクに着目している．しかし，リンクが密になっている部分を抽出する点においては，同じ手法を利用できると考えられる．

### 2.3. リンク構造を用いた研究

提案手法と同様に，アンカーテキストとリンク構造を用いた研究として，クエリ翻訳[7]が挙げられる．[7]では，以下の条件を全て満たすアンカーテキストを，ユーザによって入力されたクエリに対する翻訳語として抽出している．

- 翻訳したい言語のアンカーテキスト
- クエリと同じ文字列のアンカーテキストがリンクする URL 群に対し，最もリンクしているアンカーテキスト

2 つ目の条件は，クエリと同じ文字列のアンカーテキストが持つリンク構造について，類似するリンク構造を持つアンカーテキストを抽出している．本稿では，2 つのアンカーテキストが持つリンク構造の類似度を，アンカーテキストの類似度と呼ぶことにする．

[7]によるアンカーテキストの類似度は，式(1)で表される．翻訳語ランキングを作成する際には，クエリを Ts とし，翻訳語候補 Tt を P(Ts<->Tt)によりランキングする．

$$P(T_s \leftrightarrow T_t) = \frac{\sum_{i=1}^{n} P(T_s|U_i)P(T_t|U_i)P(U_i)}{\sum_{i=1}^{n}\left[P(T_s|U_i)+P(T_t|U_i)-P(T_s|U_i)P(T_t|U_i)\right]P(U_i)} \quad (1)$$

- P(Ts|Ui), P(Tt|Ui):アンカーテキスト Ts, Tt から Ui へのリンク数／URL Ui の in-link 数
- P(Ui): URL Ui の in-link 数／Web 上の全リンク数（HITS[8]による値）
- n: 実験データに含まれる全 URL 数

[7]の実験では，英語のクエリに対し，その翻訳語である中国語をアンカーテキスト群から抽出している．データセットは，検索ログで頻出な 9,709 個の語をアンカーテキスト群として用意している．英語のクエリは，中国語の翻訳語がアンカーテキスト群に存在する語のみを利用し，622 個の英語クエリについて実験を行っている．(1)式を用いた翻訳語ランキングで評価した場合，Top-1 が翻訳語となったクエリが 53%，Top-10 に翻訳語が含まれるクエリは 85%となった．

## 3. 既存研究の問題点と解決策

### 3.1. 提案手法で抽出する同義語

既存研究の問題点を述べる前に，提案手法により抽出する同義語について述べておく．まず，ユーザが特定の企業や人に関する同義語を抽出する際，この企業や人を「対象物」と呼ぶことにする．提案手法で抽出する同義語とは，この対象物を連想できる全ての語である．以下に例を挙げる．

- 対象物の正式名称，正式な略称
- 対象物の翻訳語
- 対象物の一般的な俗称
- 一般的な呼び方ではないが，明らかに対象物であると分かる語

既存研究[7]は，対象物の翻訳語を抽出することに特化した手法であるといえる．3.2 では，翻訳語以外の同義語を抽出する際に障害となる既存研究の問題点について述べる．また 3.3 で，精度と網羅性の高い同義語抽出の妨げとなる Web のリンク構造の問題点について述べる．

### 3.2. 既存研究[7]の問題点

既存研究[7]により定義された(1)式を，全てのアンカーテキストに適用することで，クエリの同義語についてもランキング作成することができると考えられる．すなわち，クエリと似たようなリンク構造を持つアンカーテキストを，クエリの同義語として抽出することが可能である．

一方，既存研究では翻訳語がランキングトップになれば良く，ランキング全体の評価については述べられていない．本研究では同義語抽出の網羅性を高めることを目的としており，頻出ではない略語や俗語などの同義語も上位にランキングする必要がある．

図 2 は，アンカーテキスト A と B のリンク構造を表している．クエリと同じ文字列のアンカーテキストは URL1 のみにリンクするものとする．アンカーテキスト A も，回数は少ないが URL1 のみにリンクしている．一方アンカーテキスト B は，URL1 に対するリンク数がアンカーテキスト A よりも多いものの，URL2 にも多くリンクを持つ．図 2 において，頻出ではない略称や俗語はアンカーテキスト A のようなリンク構造を持ち，頻出だが多くの URL にリンクを持つ汎用語はアンカーテキスト B のように表すことができると考えられる．既存研究[7]により定義された類似度計算では，頻出ではないアンカーテキスト A は，頻出なアンカーテキスト B よりも低く計算されてしまう．これは，URL 側から見たリンク確率を類似度計算に用いているため，アンカーテキストが他の URL へリンクしている情報を全く活用できないからだと考えられる．

4.1 では，頻出ではない同義語も上位にランキングすることができる，新しい類似度指標を提案する．提案手法では，URL 側から見たリンク確率を用いるのではなく，アンカーテキスト側から見たリンク確率を用いて，類似度の計算を行う．

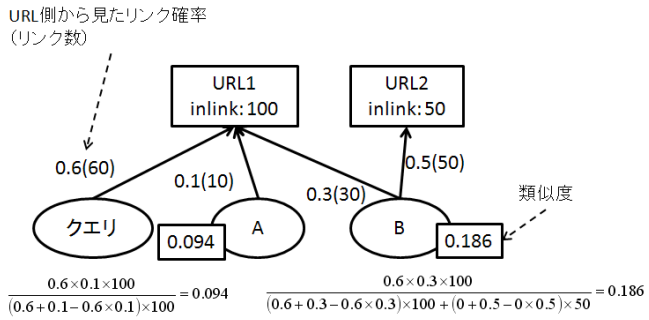


図 2　　頻出度によるリンク確率の変化

### 3.3. Web のリンク構造に関する問題点

更に精度と網羅性を向上させるためには，Web のリンク構造が持つ問題について解決する必要がある．本節では，Web のリンク構造に関する問題点を 3 つに分けて説明し，それぞれの解決策について述べる．

➢ **全ての関連 URL を抽出できていない**

クエリを対象物の正式名称とした場合でも，関連する全ての URL に，クエリと同じ文字列のアンカーテキストがリンクしているとは限らない．図 3 は，クエリを「早稲田大学」にした場合の例である．アンカーテキスト「早稲田大学」から，早稲田大学の英語版トップページである URL「www.waseda.ac.jp/index-e.html」にはリンクがないことが分かる．このため，URL「www.waseda.ac.jp/index-e.html」のみをリンクしているアンカーテキスト「มหาวิทยาลัยวาเซดะ　」（タイ語で早稲田大学）は同義語候補ランキングに出現せず，同義語抽出の網羅性が下がってしまう．URL「www.waseda.ac.jp/index-e.html」には同義語「Waseda University」が最も多くリンクしていることから，図 3(右欄)のように，クエリと同義語のリンク情報をマージすれば良いと考えられる．

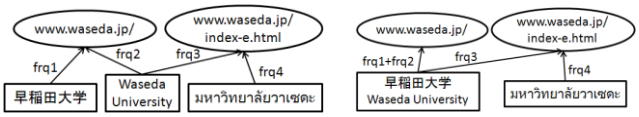


図 3　　同義語アンカーのマージ

➢ **URL の分散により，類似度が低下する**

企業のホームページなどでは，トップページを複数の言語で用意している場合がある．例えば表 1 は，「早稲田大学」のトップページ一覧を表したものである．日本語版や英語版以外にも，ドメインの異なるトップページが存在している．

2.3 で示した既存研究の類似度や，4.1 で定義する提案手法では，クエリが多くリンクする URL に重みがついている．従って，クエリからのリンク数が少ないトップページにリンクする同義語は，類似度が低く計算されてしまう．図 4（左欄）のアンカーテキスト「Waseda Univ.」は，トップページ「www.waseda.jp/top/index-j.html」にリンクしているが，クエリが最もリンクしているトップページ「www.waseda.jp/」にはリンクしていない．「Waseda Univ.」の類似度は低く計算されてしまい，同義語ランキングでは下位に位置することになる．図 4（右欄）のように，トップページのバリエーションを 1 つの URL にまとめることで，同義語の類似度を上げることが望まれる．

表 1　　早稲田大学のトップページ一覧

| |
|---|
| www.waseda.jp/ |
| www.waseda.jp/index-j.html |
| www.waseda.jp/top/ |
| www.waseda.jp/top/index-j.html |
| www.waseda.jp/top/index-e.html |
| www.waseda.ac.jp/ |
| www.waseda.ac.jp/index.html |
| www.waseda.ac.jp/index-j.html |
| www.waseda.ac.jp/index-e.html |
| www.waseda.ac.jp/index-gb.html |
| waseda.ac.jp/ |

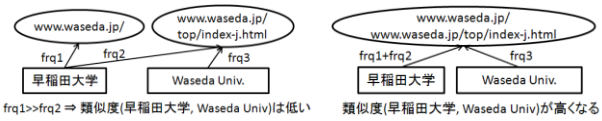


図 4　　関連 URL のマージ

➢ **誤ったリンク情報により同義語候補が増大する**

図 5 のように，対象物とは関係のない URL に，クエリからのリンクが存在する場合がある．これらの URL にリンクするアンカーテキスト群 Ax は，全て同義語候補として抽出されてしまい，同義語候補ランキングの項目数を増やすことにつながる．クエリから見たとき，対象物とは関係のない URL へのリンク確率は小さく，アンカーテキスト群 Ax の類似度は低く計算される．従って誤ったリンク情報は，同義語候補ランキング Top-n の精度には影響しないといえる．

しかし，同義語候補ランキングからより多くの同義語を抽出する場合には，同義語候補数は少ない方が良い．図 5 のように，誤ったリンク情報を削除することで，同義語候補数を削減することができる．

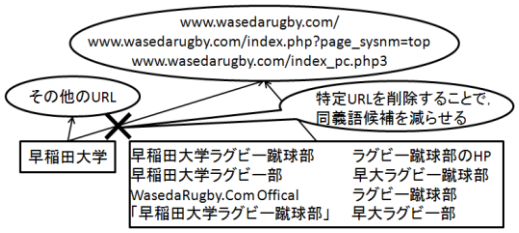


図 5　　特定 URL 削除による同義語候補数の削減

## 4. 提案手法

本節では，クエリの同義語をアンカーテキストとリンク構造から抽出し，それらをクエリとの類似度でランキングする手法について提案する．3 節で述べた通り，精度と網羅性の高い同義語抽出を行うためには，以下の問題を解決する必要がある．

- 既存研究の問題
  - 頻出ではない同義語の類似度が低い
- Web のリンク構造に関する問題
  - 全ての関連する URL が抽出できていない
  - URL の分散により類似度が低下する
  - 誤ったリンク情報により同義語候補が増大する

提案手法では，アンカーテキスト側から見たリンク情報を利用する，新しい類似度指標を用いることで，既存研究[7]の問題について解決する．この新しい類似度指標については，4.1 で詳細に述べる．

Web のリンク構造に関する問題については，Relevance-Feedback の技術を利用して解決する．すなわち，新しい類似度指標によりランキングされた同義語 Top-n に，ユーザが○×を付与することにより，リンク情報の補正を試みる．新しいリンク情報を利用して，同義語候補のリランキングを行い，精度と網羅性の高い同義語ランキングを作成する．Relevance-Feedback を用いたリランキングについては，4.2 で詳しく述べる．

### 4.1. 共起強度による同義語候補ランキング

3.2 で述べたように，既存研究は URL 側から見たリンク確率しか用いておらず，頻出ではない同義語をランキング上位にすることができなかった．本節では，アンカーテキスト側から見たリンク確率を利用することで，頻出ではない同義語も上位にランキングできる，新しい類似度指標を提案する．新しい類似度の指標は共起強度と呼び，以下の式で表される．

$$\text{共起強度}\quad co(a,b)=\frac{2}{\frac{1}{P(b|a)}+\frac{1}{P(a|b)}} \qquad (2)$$

$$\text{条件付き確率}\quad P(y|x)=\frac{\sum_{u\in c(x,y)} frq(x|u)}{frq(x)} \qquad (3)$$

- ✧ frq(x):アンカーテキスト x の総リンク数
- ✧ frq(x|u):アンカーテキスト x から URLu へのリンク回数
- ✧ c(x, y):アンカーテキスト x と y が共通してリンクする URL 群

アンカーテキスト a と b の共起強度は，a と b それぞれの条件付き確率を調和平均したものである．相加平均ではなく調和平均を用いることで，a と b の条件付き確率に差がある場合，最終的な共起強度の値を低く計算することができる．

条件付き確率 P(y|x)は，アンカーテキスト x のリンクについて，x と y が共通してリンクする URL へのリンク確率を示している．共通する URL 数ではなく，URL へのリンク確率を用いて共起強度計算を行うため，クエリと同じ文字列のアンカーテキストから多くリンクされる URL に，重みがついた式になっている．

### 4.2. Relevance-Feedback を用いたリランキング

3.3 でまとめたように，精度と網羅性の高い同義語抽出を行うためには，Web のリンク構造に関する問題点を解決する必要がある．提案手法では，Relevance-Feedback の技術を利用することでリンク情報の補正を行い，新しいリンク情報を用いて同義語ランキングをリランキングする．リランキングのプロセスを，以下に述べる．

#### ① 対象物の同義語候補に対し人手で○×を付与

共起強度による同義語ランキングの Top-n に対し，対象物の同義語と思う場合には○を，異なる語と思う場合には×をつける．どちらか判断できない場合には，○×をつけないことにする．以後のプロセスでは，○をつけた語を「○アンカーテキスト」，×をつけた語を「×アンカーテキスト」と表現する．なお，クエリと同じ文字列のアンカーテキストも「○アンカーテキスト」として扱う．

#### ② ○アンカーテキストのマージ

対象物の同義語と判断されたアンカーテキストについて，リンク情報をマージする．○アンカーテキストのみがリンクしていた URL を，クエリがリンクする URL 群に追加することで，新しい同義語候補を抽出することができる．

#### ③ ○アンカーテキストがリンクする URL のマージ

複数 URL へのリンク分散を解消するため，対象物に関連する URL をマージする．この処理により，クエリがリンクする URL 群の 1 部にしかリンクしていない同義語について，共起強度の値を高く計算することができる．マージする URL は，以下の条件を満たすものである．

- ➢ ○アンカーテキストからのリンク確率の合計が，一定以上となる URL
- ➢ （実験では，1URL に対するクエリからの最大リンク数×0.8 以上）

**④ ×アンカーテキストがリンクする URL について，クエリからのリンク情報を削除**

対象物とは関係のないアンカーテキストを×アンカーテキストとして指定することで，誤ったリンク情報を削除する．クエリから対象物とは関係のない URL へのリンク情報を削除することにより，その URL にリンクするアンカーテキストを，同義語候補から取り除くことが可能である．リンク情報を削除する URL は，以下の条件を全て満たすものである．

- ×アンカーテキストとクエリが共通してリンクする URL
- URL 側から見たクエリのリンク確率の合計が，一定以下の URL（実験では 0.2 未満）

①～④までのプロセスを繰り返すことにより，対象物の同義語ランキングの網羅性と精度を上げていく．②アンカーテキストのマージはランキングの網羅性向上に有効であり，③④URL のマージ・削除はランキングの精度向上に有効である．

プロセスサイクルを終了するタイミングとしては，プロセス①でユーザに示すランキングに，同義語が含まれなくなった時が考えられる．

## 5. 評価実験

本節では，提案手法による同義語ランキングの精度と網羅性を確かめるための実験と評価を行う．

### 5.1. 実験概要

- **実験データ**

実験データは，文部科学省の e-Society プロジェクト[9]において収集した，2006 年 1 月時点の日本語 Web ページである[10]．データの内容を表 2 にまとめる．

実験に用いるアンカーテキストとリンク情報は，ホスト外リンクのみを用いて抽出した．ホスト内リンクには，「前へ」「トップへ」などのナビゲーションを目的に使われているアンカーテキストが多く，同義語抽出の目的には利用できないと判断したためである．また，1 つのアンカーテキストからしかリンクされていない URL は，アンカーテキストを用いた同義語抽出では扱われない．1 つのアンカーテキストからしかリンクされていない URL と，これらの URL にリンクするアンカーテキストは，予めデータセットから削除した．実験で利用したアンカーテキストとリンク情報について，表 3 にまとめる．

表 2　　実験で用いた Web データ

| | |
|---|---|
| 対象ページ | 1,324,268,374 |
| ホスト外リンク | 3,235,910,945 |
| レコード<br>（アンカーテキスト→URL のペア数） | 358,011,591 |

表 3　実験で用いたアンカーテキストとリンク情報

| | |
|---|---|
| アンカーテキスト | 51,822,702 |
| URL | 22,873,005 |
| レコード<br>（アンカーテキスト→URL のペア数） | 82,652,395 |

- **実験に用いたクエリ**

同義語抽出の精度と再現率がジャンルにより異なるかどうかを確かめるため，実験で用いるクエリを複数ジャンルから選択した．ジャンル名と各クエリ数を表 4 に示す．なお，会社名，人名，漫画・アニメ，ゲームのジャンルに属するクエリは，Yahoo! JAPAN 2005 年検索キーワードランキング[11]から抽出した．漫画・アニメ名ランキングに含まれていた「魔法先生ネギま！」は，一致するアンカーテキストが存在しないため，クエリからは除外してある．

表 4　　ジャンル別クエリ一覧

| ジャンル名 | クエリ抽出元 | 数 |
|---|---|---|
| 会社名<br>/サービス名 | 総合ランキング 2005　Top-10 | 10 |
| 人名 | 著名人ランキング 2005　Top-10 | 10 |
| 漫画・アニメ | 漫画・アニメランキング 2005 Top-10 | 9 |
| ゲーム | ゲーム名ランキング 2005 Top-10 | 10 |
| 大学名 | 東京六大学 | 6 |
| | 合計 | 45 |

- **正解セット**

各クエリの正解セットは，Relevance-Feedback によるリランキングから人手で作成した．リランキングを 5 回行って得た同義語候補，もしくは共起強度が 0.01 以上の同義語候補のうち，3 ユーザ中 2 人が同義語と判断したものを正解としている．

- **評価ユーザ**

Relevance-Feedback によるリランキング時の人手による評価は，著者を入れた大学院生 3 ユーザで行った．リランキングは 5 回，または共起強度が 0.01 未満になるまで行い，最終的な同義語ランキングを取得した．5.2～0 における Relevance-Feedback によるリランキングの実験結果は，3 ユーザの実験結果を平均した値である．5.4～5.6 の実験データは，著者によるリランキング結果を用いている．

### 5.2. 各手法の比較実験

既存研究[7]と，共起強度による同義語ランキング，Relevance-Feedback によるリランキングの比較について，精度を表 5 に，再現率を表 6 に示す．クエリは 5.1 で述べた 45 個の語を用い，精度と再現率は 45 個の結果を平均したものである．

既存研究に比べ，共起強度を用いたランキングは精

度と再現率がともに向上していることが確かめられた. また，Relevance-Feedback を用いたリランキングを行うことで，Top-200 までのランキング精度は向上していることが分かる. 全体のランキングを見た場合には，Relevance-Feedback を用いたリランキングの精度が最も低いが，リランキング時に同義語候補が増大するためである．再現率を確認すると，Relevance-Feedback を用いたリランキングと比べ，既存研究では抽出できていない同義語が存在していることが分かる.

クエリにより同義語候補数が異なることを考えると，Top-n のランキングではなく，共起強度による閾値を設ける方が扱いやすい．Relevance-Feedback を用いたリランキングの場合，共起強度を 0.1 以上にすれば再現率が 80%程度となり，精度も Top-100 と変わらないことが確認できた.

表 5　　各手法のランキング精度

| 手法 | Top-10 | Top -100 | Top -200 | 全て | 共起強度 0.1 以上 |
|---|---|---|---|---|---|
| 既存研究[7] | 24.2% | 8.1% | 5.6% | 2.1% | — |
| **共起強度** | 28.7% | 9.9% | 7.2% | 2.1% | 13.5% |
| **リランキング** | 43.9% | 11.9% | 8.1% | 1.4% | 12.2% |

表 6　　各手法のランキング再現率

| 手法 | Top-100 | Top-200 | 共起強度 0.1 以上 | 全同義語候補 |
|---|---|---|---|---|
| 既存研究[7] | 53.1% | 69.0% | — | 95.2% |
| **共起強度** | 63.5% | 82.8% | 69.7% | 95.2% |
| **リランキング** | 70.7% | 87.8% | 79.8% | 99.5% |

※表 5, 表 6 で，太字になっている手法が提案手法

### 5.3. クエリのジャンルによる比較実験

同義語抽出の精度と網羅率について，ジャンルによる違いがあるかどうかを確かめる. Relevance-Feedback によるリランキングについて，精度を表 7 に，再現率を表 8 に示す．精度，再現率とも，ジャンルにより違いはあまり見られなかった．どのジャンルの同義語でも，提案手法で抽出できることが分かる.

個々の特徴を見ていく．会社名は同義語候補の数が多く，ランキング全体の精度は低くなりがちである．正解セット抽出の際，リランキングを 5 回行っても同義語候補の共起強度が 0.1 以上となったため，共起強度 0.1 以上では再現率が 100%に近い値となってしまっている．大学名などはホームページがはっきりしており，流行などの影響を受けないため，精度の高いランキングになりやすいことが分かった．人名，ゲーム，漫画・アニメに関しては，ジャンルの違いよりも，クエリの違いにより同義語候補数に違いが出た．話題の対象物に関しては，関連するホームページやリンクが多く，同義語候補数が多くなることが確かめられた.

表 7　　各手法のランキング精度

| | Top-10 | Top -100 | Top -200 | 全て | 共起強度 0.1 以上 |
|---|---|---|---|---|---|
| 会社名 | 44.7% | 12.6% | 9.7% | 0.4% | 2.9% |
| 人名 | 29.7% | 7.9% | 4.9% | 1.3% | 13.4% |
| ゲーム | 43.3% | 12.7% | 8.4% | 1.9% | 20.8% |
| 漫画・アニメ | 40.0% | 7.5% | 4.5% | 1.2% | 9.6% |
| 大学名 | 72.7% | 22.7% | 15.9% | 2.8% | 15.5% |

表 8　　各手法のランキング再現率

| | Top-10 | Top-100 | Top-200 | 共起強度 0.1 以上 |
|---|---|---|---|---|
| 会社名 | 26.6% | 59.5% | 84.7% | 99.7% |
| 人名 | 40.1% | 77.2% | 85.7% | 68.2% |
| ゲーム | 25.3% | 68.5% | 86.6% | 63.7% |
| 漫画・アニメ | 40.4% | 75.1% | 85.8% | 79.0% |
| 大学名 | 26.5% | 73.1% | 95.4% | 94.1% |

### 5.4. 同義語数と精度, 網羅率

リランキングのサイクルにより，同義語数がどのように変化するかについて実験を行った．変化が分かりやすい例として，クエリ「早大」の実験データを表 9 に示す．閾値は共起強度 0.1 以上としている．同義語数増加率は，サイクル 0 からの増分である.

サイクルを増やすごとに，精度を保ったまま，より多くの同義語が抽出できることが分かった．また，閾値を共起強度 0.01 以上にした場合には, 再現率が 100%になることが確かめられた．同義語候補数は 286 個と増えるが，目視で確認できる量であると考えられる.

本節では，対象物の略称をクエリに選んだが，抽出した同義語数はジャンル 東京六大学に含まれる「早稲田大学」と同じである．すなわち，正式名称と略称のどちらをクエリにしても，同じ同義語数を抽出できることが確認できた.

表 9　　各サイクル時の同義語数と精度
（共起強度 0.1 以上）

| サイクル | 再現率 | 同義語数／同義語候補数 | 同義語数増加率 | 精度 |
|---|---|---|---|---|
| 0 | 79.1% | 34/191 | — | 17.8% |
| 1 | 86.1% | 37/210 | 8.8% | 17.7% |
| 2 | 93.0% | 40/229 | 17.7% | 17.5% |
| 3 | 93.0% | 40/227 | 17.7% | 17.6% |

### 5.5. ○アンカーテキストによるマージ効果

「早稲田大学」のトップページを用いて，○アンカーテキストのマージ, 及び URL マージがどのように機能したかを確かめた. クエリは 5.4 と同様に「早大」で実験を行った．実験結果を表 10 に示す.

サイクル 0 の結果から，アンカーテキスト「早大」は 3 つのトップページへしかリンクしていないことが

分かる．アンカーテキスト「早稲田大学」と「Waseda University」を○アンカーテキストとすることで，クエリがリンクするトップページが 9 つに増えたことが確認できた．

トップページのマージでは，サイクル 3 で 6 つのトップページがマージされた．マージされなかったトップページの特徴としては，URL の形をしたアンカーテキストや，正式名称に記号がついたアンカーテキストから多くリンクされている点が挙げられる．予めこれらのアンカーテキストを削除しておくことで，ランキングの精度向上が望めることが分かった．

表 10　○アンカーテキストと URL のマージ

| サイクル数 | 共起強度計算に用いられるトップページ数 | マージされたトップページ数 |
|---|---|---|
| 0 | 3 | — |
| 1 | 9 | 1 |
| 2 | 10 | 5 |
| 3, 4 | 11 | 6 |
| 5 | 11 | 7 |

### 5.6. ×アンカーテキストによる同義語候補の減少数

×アンカーテキストを指定することにより，同義語候補数がどのように変化するかについて確認する．クエリは「早大」で行った．表 11 の左欄が×アンカーテキストであり，中欄が×アンカーテキストの指定により，クエリからのリンク情報が削除された URL である．右欄は同義語候補の減少数を表している．対象物とは関係のない同義語候補を削除することにより，ランキングの精度を向上させることができた．

表 11　同義語候補の減少数

| ×アンカーテキスト | 削除 URL | 同義語候補減少数 |
|---|---|---|
| 早稲田大学 所沢キャンパス | www.human.waseda.ac.jp/ | 15 |
| 早稲田大学 理工学部 | www.sci.waseda.ac.jp/ | 86 |
| 早稲田大学 法学部 | www.waseda.ac.jp/ hougakubu/index-j.html | 11 |

## 6. おわりに

本稿では，対象物の略称や俗称を対象とした同義語抽出の手法について提案を行った．アンカーテキストとリンク構造を用いることで，シソーラスには存在しない同義語を抽出することができる．既存研究による類似度計算では頻出語ではない同義語を上位にランキングできないという問題があったが，提案手法ではアンカーテキストから見たリンク構造を用いることで，頻出ではない同義語も抽出できるようになった．また，ランキングの精度と網羅性の低下原因となっている Web の誤ったリンク情報を補正するため，Relevance-Feedback の技術を利用した．同義語ランキング Top-n の同義語候補に○×を付与することにより，Web のリンク情報を更新し，同義語ランキングのリランキングを行う．実験では，精度を保った上で，網羅性を既存研究よりも約 15%向上させることができた．

今後の課題としては，より精度の高いランキングを行うことである．同義語候補ランキングの中には，同義語に記号がついたアンカーテキスト，または「ホームページ」や「トップページ」などの定型語がついたアンカーテキストが現れている．自然言語処理の技術を取り入れることで，これらの語句を取り除くことが可能であると考える．また，コミュニティ抽出の手法を取り入れることで，誤ったリンク情報の除去を自動化できると考えられる．

## 文　献


[1] D.Milne, I.H.Witten and D.M.Nichols: “A Knowledge-Based Search Engine Powered by Wikipedia”, CIKM'07, pp.445-454, 2007.

[2] B.M.Fonseca, P.Golgher and B.Possas: “Concept-Based Interactive Query Expansion”, CIKM'05, pp.696-703, 2005.

[3] Y.Qiu and H.P.Frei: “Concept Based Query Expansion”, SIGIR'93, pp-160-169, 1993.

[4] S. R. Kumar, P. Raphavan, S. Rajagopalan and A. Tomkins: “Trawling the Web for emerging cyber communities”, The International Journal of Computer and Telecommunications Networking, Vol.31, pp.1481-1493, 1999.

[5] P. K. Reddy and M. Kitsuregawa: “An approach to relate the Web communities through bipartite graphs”, WISE'01, Vol.1, pp301-310, 2001.

[6] G. Flake, S. Lawrence and C. Giles: “Efficient Identification of Web Communities”, Proceedings of the sixth ACM SIGKDD, pp.150-160, 2000.

[7] W.H.Lu, L.F.Chien and H.J.Lee: “Translation of Web Queries Using Anchor Text Mining”, ACM Transactions on Asian Language Information Processing, Vol.1, No. 2, pp.159-172, June 2002.

[8] J.M.Kleinberg: “Authoritative Sources in a Hyperlinked Environment”, Journal of the ACM, Vol.46, Issue.5, pp.604-632, 1998.

[9] 文部科学省リーディングプロジェクト e-Society: “http://cif.iis.u-tokyo.ac.jp/e-society/”

[10] 早稲田大学山名研究室 e-Society プロジェクト: “http://www.yama.info.waseda.ac.jp/e-society/”

[11] Yahoo! JAPAN　2005年検索キーワードランキング: “http://picks.dir.yahoo.co.jp/new/review2005/”