社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－NL－180 (20)
2007／7／25

検索語の共起情報を利用した単語クラスタリングとWeb検索への応用

有田 一平† 菊池 英明‡ 白井 克彦†

†早稲田大学理工学部理工学研究科 〒169-8555 東京都新宿区大久保3-4-1
‡早稲田大学人間科学学術院 〒359-1192 埼玉県所沢市三ヶ島2-579-15

E-mail: †{i_arita,shirai}@shirai.cs.waseda.ac.jp, ‡kikuchi@waseda.jp

あらまし 本研究の目的はWeb検索への応用を用途とした単語の類義関係を表すシソーラスを自動構築することである。Web上のドキュメントへのメタデータの付加や、ユーザの検索クエリ作成補助などの用途を想定している。このために、検索語の共起頻度というユーザの検索行動を属性として単語間の類義関係を解析することによって単語クラスタリングを行い、分類語彙表のような階層型の分類番号をもったシソーラスを構築した。被験者へのアンケート調査との比較による精度の評価を行い提案法の有効性を確認した。さらに、このシソーラスの応用例として、索引語への分類番号付加を行い、類義語ナビゲーション機能と、分類番号の階層構造を利用した前方一致検索による同一カテゴリ語の一括検索機能を備えた検索アプリケーションの開発を行った。

キーワード 単語クラスタリング，検索クエリ，Web検索，共起

Word Clustering Using Concurrent Search Queries

ARITA Ippei† KIKUCHI Hideaki‡ and SHIRAI Katsuhiko‡

†Faculty of Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
‡Faculty of Human Sciences, Waseda University 2-579-15 Mikajima, Tokorozawa-shi, Tokyo, 359-1192 Japan

E-mail: †{i_arita,shirai}@shirai.cs.waseda.ac.jp, ‡kikuchi@waseda.jp

Abstract The purpose of the research is to construct thesaurus applied to web search automatically. We expect this thesaurus to be used for adding semantics to documents on the web as metadata and navigating users' search query creation. For this purpose, We conducted word clustering using co-occur search queries that represent users' search action, and constructed semantic information that has hierarchical semantic number like Bunruigoihyo. We also measured precision as conducted questionnaire survey. Additionally, we developed search application as an application of this semantic information. We attached semantic number to each index term, therefore this application provides similar word navigation, and category search using wildcard match based on the hierarchical numbers.

Keyword Word clustering, Search query, Web retrieval, Co-occurrence

1. はじめに

1.1. 研究の背景と目的

本研究は、Web検索において頻繁に検索語として入力される固有名詞や次々と現れる新語に対して、単語クラスタリングを行うことによって類義関係を解析し、Web検索に幅広く活用できるシソーラスを構築することを目的としている。

既存の主なWeb検索エンジンは、Web上のドキュメントを解析することによって単語単位の索引を作成し、ユーザが入力したクエリと索引語の一致に基づいて結果を返す全文検索を基本としているものが多い。この方法を利用すると、クエリとして指定した語を含む文書を確実に抽出することができる。しかし、語の表現は多様性に富むものであり、ユーザの検索要求が必ずしも特定の単語によって十分に表現できるとは限らない。その結果、検索語と索引語の間に不一致が起こり、再現率が低下してしまうという欠点も含んでいる。このため、満足できる検索結果を得るためには、検索クエリを複数回指定しなければならないことも多く、検索エンジンの特性を考慮した検索語を入力することに慣れる必要もある。また、検索要求に合った検索語を思いつかない場合も多々ある。そういった問題を解決するために、シソーラスを利用した検索クエリ拡張やセマンティック・ウェブなどの意味を考慮した

検索システムへの要求が高まっている。

意味を考慮した検索システムの実現には、その基となるシソーラスが必要となる。既存のシソーラスには分類語彙表（国立国語研究所）[1]、日本語語彙体系（NTT）[2]、EDR（日本電子化辞書研究所）[3]などが挙げられるが、Web検索では、検索語として固有名詞や新語などが指定されることが多いため、既存のデータでは十分に対応することができない。例えば、あるWeb検索エンジンの二ヶ月間における検索語トップ40,000における分類語彙表内単語のカバー率を調べた結果、トップ40,000のうち一語のクエリである36,061語において、分類語彙表に存在する単語が5308語、カバー率は僅か14.7%である。

また、用途を特定しない汎用的な意味データを用いたシソーラス検索では逆に検索精度が低下するという報告もされている[4]。このために、精度の高い検索を実現するためには、Web検索に適したシソーラスを構築する仕組みが必要となる。

1.2. 先行研究

単語クラスタリングによってシソーラスを構築する研究は様々な取組みがされている。単語のクラスタリングは、類似の観点、利用する属性と距離尺度の表現、そしてどのクラスタリング・アルゴリズムを利用するかによってさまざまな方法が考えられる。これらの要素をクラスタリングの目的に合わせて選択していくことが重要となる。

先行研究では、主に語の同一文書内における共起関係を用いた取組みが多くなされている。日本語意味マップの作成を目的として連体修飾要素の係り受け関係に基づいて自己組織化マップでクラスタリングを行ったもの[5]、閲覧済みWebページ中の語の共起情報を利用したもの[6][7]などが挙げられる。また、文書中の言語的特徴以外を利用した例としては、Webサイトのハイパーリンクを利用したもの[8]もある。

Web検索サービスの大規模なログから語の共起関係を抽出する取り組みはそのデータ入手の難しさから多くは見受けられない。前例としては、Microsoftの行った検索語と閲覧したドキュメントの関係に基づくものがある[9]。

1.3. 本研究の位置づけ

本研究の特徴は、単語クラスタリングに際して、大規模検索エンジンのログにおける検索語の共起頻度に焦点を当て、この情報を行列化した上で類義関係の解析を行う点である。同一文書内の共起情報を利用した場合には、形態素解析による構文解析やTF･IDFなどによる重み付けなどが必要とされるため、それらの機械処理の精度に依存する部分が大きい。一方、検索語の共起情報を利用すれば、ユーザによって作成されたキーワードを直接利用するため、より高精度のクラスタリングが実行できると考えられる。

このシソーラスは各単語間の距離に基づいて階層型クラスタリングを行い、類義関係を木構造にまとめたものである。階層構造の分類番号を持たせており、分類語彙表とほぼ同様の形式となっている。

用途としては、索引語へメタデータとしてシソーラスが持つ情報の付加、UIによるユーザの検索クエリ作成のナビゲーションなどを想定しており、応用例として実際にアプリケーションの開発を行った。これにより、特定のカテゴリに属する複数の語の高速検索や、UIでユーザに類義語を提案する仕組みを整えた。

本論文は全6章からなり、2章では提案するクラスタリング手法の説明、3章ではクラスタリングの具体的な手順の説明、4章ではクラスタリングの評価、5章では構築したシソーラスを利用した検索アプリケーションの紹介、6章ではまとめと今後の課題について述べる。

2. 提案する単語クラスタリング手法

本章では提案する単語クラスタリング手法について述べる。なお、本研究はWeb検索への応用を指向しているため、検索サイトでの検索頻度ランキング上位の語に対する有効性を重視する。

2.1. 類似の観点

単語クラスタリングにおける類似の観点について、本研究では共起するクエリの類似性に着目する。Yahoo! JAPANの提供する関連検索ワードWebサービス[1]を利用して予備的に調べたところ、例えば2つの航空会社の名称に対して100語の頻出共起語のうち66語までが共通するなど、共起語の類似性が語の類似性につながる傾向が見られた。

2.2. 利用する属性と距離尺度の表現

先述した類似の観点に従って、単語間の距離を表現する。本手法では複数語クエリの共起情報を利用する。まず、その複数語クエリの傾向を調べるために、検索クエリ全体に対する検索語数の集計を行った。ある検索サイトの4ヶ月間におけるクエリ（異なり5,871,428、のべ21,657,148）を集計した結果、複数語クエリの割合は異なりで25%程度、のべで15%程度であった。割合としては多くはないが、検索クエリの総数が多けれ

1）関連検索ワードWebサービスは、実際にYahoo!検索で使用されたキーワード情報をもとに、指定されたキーワードとよく組み合わせて検索されるキーワード情報が得られるサービスである。詳細は次のURLから確認できる。(http://developer.yahoo.co.jp/search/webunit/V1/webunitSearch.html)

ば、クラスタリングに利用するには十分な量の共起情報が取得できる。

この複数語クエリを基に、各検索語の共起頻度を属性として利用し、クラスタリング対象語をベクトル空間上に表現する。共起クエリを基底とした検索語・共起クエリ行列を作り、それぞれの成分を共起頻度情報とする。こうして表された検索語ベクトル間のユークリッド距離を単語間の距離表現とする。

2.3. クラスタリング・アルゴリズム

クラスタリング・アルゴリズムとしては、階層型クラスタリングを利用する。この手法を用いる理由は、前述したような索引語に階層構造の意味番号を付加して前方一致検索を行うことを可能にするような意味データの作成に適しているからである。階層型クラスタリングの結果はクラスター形成順序を辿ることで樹形図として二分木を形成することができる。つまり、木構造を辿るだけで階層構造を取得することができる。これによって、木構造の上位階層から下位階層へと辿るときに上位の桁から番号を付加し、階層構造の分類番号によって構造化することができるようになる。

3. 本手法によるクラスタリングの手順

本章では、クラスタリング手順の詳細について述べる。

3.1. 共起クエリの取得

まず、検索語ランキング上位の各語の共起クエリを取得する。図 3-1 に検索語ランキングトップ 1000、10000、40000 における共起クエリの異なり取得数、図 3-2 にのべ取得数をまとめた。予備実験からは、共起クエリの異なり取得数が少ないとクラスタリング精度に劣化が生じるという結果を観察している。特に、異なり取得数が 10 未満の語においてこの傾向は顕著であった。この集計から、トップ 1000 では約 9 割、トップ 10000 では約 7 割、トップ 40000 では約 3 割の語がクラスタリング可能ということになる。ただし、今回扱ったデータが 4 ヶ月間と限られたものであったため、この規模を増やせばランキング下位の語に対する取得数増加も見込める。

一方、共起クエリの取得数が多すぎる場合、行列化を行う際に次元数が膨大になるという問題が生じる。このため、大規模データを扱う際には、出現頻度が 1 度のみの複数語クエリを排除するなどの処理が必要となる。

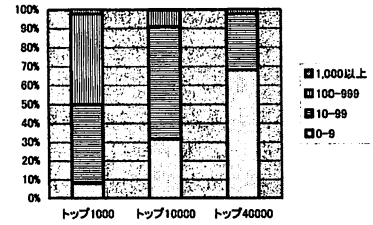

図 3-1. 共起クエリ異なり取得数

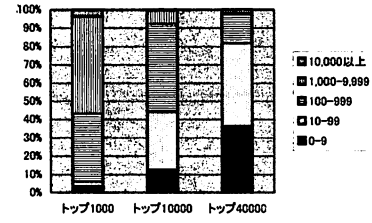

図 3-2. 共起クエリのべ取得数

3.2. 検索語・共起クエリ行列の生成と距離表現

前節で示した共起クエリを利用して、その出現頻度を基底とした行列を作成し、頻度分析、主成分分析や潜在意味解析による基底変換などのベクトル最適化処理を行い、各検索語間の距離をユークリッド距離で表現する。

まず、共起クエリを元に、行に検索語、列に共起クエリの異なり列をとり、検索語と共起クエリが共起する頻度を成分とするような検索語・共起クエリ行列を作成する。

また、検索語によって取得する共起クエリ数が異なるので、各検索語ベクトルの分散を 1 とするように正規化を行った。行列の各成分をその検索語ベクトルの標準偏差で除算することによって求める。構築するベクトル空間は疎なものとなるため、この処理を行わないと共起クエリ数が極端に低い語が他の殆どの語の類義語として出現してしまう。

しかし、検索語数が増えることにより検索語・共起クエリ行列の次元数が膨大になるという問題からこれを効率的に扱う方法が必要となった。

この対応策として、主成分分析や潜在意味解析[10]を利用した次元数の圧縮と、非ゼロ要素以外の次元のみを計算するという二通りの方法を試みた。どちらも、検索語・共起クエリ行列がスパースな行列になりやすいという特性を利用したものである。

主成分分析や潜在意味解析では、効率的な固有値・特異値計算手法[11][12]を利用すれば少ないメモリ空間で高速に処理を行うことができ、基底変換により相関の高い共起クエリをまとめることでクラスタリング精度の向上が期待できる。しかし、実験を行ったところ、主成分分析や潜在意味解析後のクラスタリング結果と非解析時の結果に殆ど差がなく、固有値の累積寄与率も、ほぼ線形に増加するという傾向が見られた。図 3-1 は 100 語の検索語を表現した検索語・共起クエリ行列に対して行った主成分分析結果である。この結果から、相関の高い共起クエリの数が少なく、精度向上を狙った次元数の縮約が上手く機能しなかったと捉えることができる。

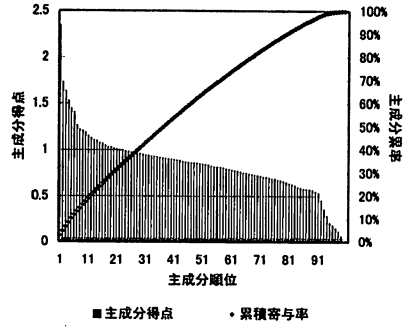

図 3-3: 主成分順位と寄与率

そのため、より単純なロジックで高速に距離が求めることができる非ゼロ要素以外の次元のみを計算するという方法を主に用いた。これは、各検索語ベクトルの非ゼロ要素のみを配列に格納し、その成分をその標準偏差で除算した後に、成分間の距離の二乗和だけを求めることにより、高速にユークリッド距離の計算を行ったものである。これによって、共起クエリリストから行列を生成せずに、直接検索語ベクトル間のユークリッド距離を求められるため、大幅なメモリ領域と計算量の削減ができた。

最後に、各検索語ベクトル間の距離をユークリッド距離によって表現する。

3.3. 階層型クラスタリング

次に、前節で求めたユークリッド距離に基づいて、階層型クラスタリングを行う。最近隣法、最遠隣法、群平均法、重心法、メディアン法、Ward 法、McQuitty 法の 7 手法で実験を行った結果、Ward 法、最遠隣法、McQuitty 法で良好な結果が得られた。最遠隣法によるクラスタリングが有効だったことから、生成した検索語ベクトルもはっきりとしたクラスターを形成していたと考えることができる。特に、Ward 法によって特に高精度の結果が得られたため、この手法を採用することにした。

3.4. 階層型分類番号の出力

階層型クラスタリングの出力は、通常では樹形図でクラスター間の距離と形成順序を表すことが多い。しかし大規模データの表示が難しく、本検索システムでは分類語彙表のような階層構造を持った分類番号によるシソーラスの構築を目的としている。このため、クラスター形成順序を辿り、階層構造の番号をもたせた形式での出力を行った。

クラスター形成順序は二分木によって表すことができるが、図 3-2 のように二分木をルートから辿って枝分かれするごとに上位の桁から 0 と 1 を割り振ることによって階層構造をもった番号を与える。クラスター形成時に下位の桁から 0 と 1 を割り振るという逆の捉え方もできる。

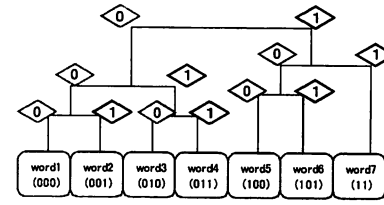

図 3-4: 各単語への分類番号付加

この例では 0 と 1 のみの二進数表記によって階層構造を表しているが、これを 4 進法、8 進法、16 進法へと基数変換することによって 2 階層、3 階層、4 階層へと複数階層のデータを一まとまりのものとすることもできる。

4. クラスタリングの評価と考察

前章で述べたクラスタリング手法の精度を、被験者によって選択された類似語をクラスタリング時に利用する検索語ベクトル間のユークリッド距離の近い語がどの程度カバーできるかという検証によって行った。本章ではその評価方法と結果について述べる。

4.1. 評価方法

まず、検索語ランキング上位の語から順に 300 語と 1000 語の二種類の評価用データを用意した。尚、アダルト語などの不適切語は除外してある。次に、300 語の評価用データからランダムに選択した 10 語を評価対象とした。

評価は被験者 15 名に対するアンケート調査によっ

て行った。被験者は、上記の 10 語に対して類似していると思われる語を、評価用データから類似度の高い順に 5 語選択した。そして被験者によって行われた類似度評価を以下のように集計し類義語を定義した。

① 各被験者の行った類似度判定に対して、以下のように得点を与える
- 1 位の単語の得点を 5 点とする
- 2 位の単語の得点を 4 点とする
- 3 位の単語の得点を 3 点とする
- 4 位の単語の得点を 2 点とする
- 5 位の単語の得点を 1 点とする

② 全被験者による得点を集計し、各評価対象語につき得点の高い語 5 語を類義語と定義する

その後、クラスタリング過程の行列計算で求められる検索語ベクトル間のユークリッド距離の近い語と、被験者の類似度判定から集計した類義語 5 語の比較を行い、ユークリッド距離の近い語によってどの程度の語をカバーできているかの検証を行う。評価はユークリッド距離の近い順から 5 語、10 語、20 語、30 語の各単語群で、被験者判断による類義語 5 語のうち何語をカバーできるかの割合を求めることによって行った。

4.2. 評価結果

はじめに、300 語での評価について述べる。(i)共起頻度から生成した検索語・共起クエリ行列をそのまま利用した場合、(ii)共起頻度を対数値に変換した行列を利用した場合、(iii) 共起クエリのうち上位 200 語のみを用いて共起の有無を 1/0 で表した行列を利用した場合の評価結果を表 4-1 に示す。

表 4-1. 300 語の評価結果

行列表現方法	次元数	評価範囲(語数)			
		5	10	20	30
(i)	149979	28%	34%	50%	62%
(ii)	149979	40%	50%	62%	70%
(iii)	25693	40%	50%	64%	72%

成分を対数化することによって精度は向上し、さらに上位 200 語のみに絞りさらに共起頻度を用いないことにより計算量を減らしても精度は下がらなかった。最終的に上位 30 語までのカバー率は 72%となった。

次に、クラスタリング語数を増加させることによる精度の変化を調べるために、1000 語での評価を同様に行った。その際、共起クエリの上位 200 語のみ(前述の(iii)に相当)を利用したデータのみで行った。評価結果は表 4-5 の通りで、クラスタリング語数を増やしても精度の大幅な劣化は見られなかった。

表 4-2. 1000 語（共起クエリ上位 200 語）の評価結果

	評価範囲(語数)			
	5	10	20	30
平均	36%	42%	58%	70%

クラスタリング語数が増えることによってベクトル空間がより密なものになり、階層型クラスタリング後の結果は主観的にはより高精度になったように見受けられた。対象とする語数をこれ以上増やす場合、本評価方法では実現困難なため評価を行っていないが、本手法の有効性が維持できる可能性がうかがえる。

5. 意味情報に基づく検索アプリケーション

構築したシソーラスを、索引語への意味情報付加と UI での類義語ナビゲーションに応用した検索アプリケーションについて述べる。

5.1. 索引語への分類番号付加

このアプリケーションは村田らが「階層構造データ列の簡易な高速アルゴリズム[13]」で提案した分類語彙表の分類番号を文書内単語にメタデータとして付加するというアイデアを基に、本研究で構築したシソーラスの Web データへの応用を試みたものである。

具体的な処理の流れについて説明する。まず Web から収集した HTML データからテキストを取りだし、形態素解析を行い単語単位に分割する。通常の検索エンジンではこの形態素解析結果の代表表記を索引語として用いるが、本システムでは分割した形態素に対して、シソーラスの対応する分類番号を付加し、それを単位とした意味情報付きのインデックスを作成した。シソーラスに分類番号の存在する単語であれば分類番号に変換、そうでなければ単語のままインデックス化している。

このように分類番号を意味情報として索引に付加することによって、階層構造を利用した同一カテゴリ語の一括検索などを行うことができる。「101*」というようなクエリで分類番号による前方一致検索をかけることで、一つのクエリで特定カテゴリに属するすべての語を表すことができるため、複数の語を高速に検索することができる。

5.2. 類義語ナビゲーションと検索処理

前節で述べた意味情報付きのインデックスに対しての検索を実現したシステム「意味索」を構築した。本システムのインタフェースを図 5-3 に示す。図の例は「英語」と「学習」という検索語を入力した例であるが、英語に対しては「スペイン語」、「中国語」、「韓国語」といった語が類義語として提示されている。また、類義語リストの下にある「more」という文字列をクリ

ックすることによって、シソーラスをさらに1階層上がって類義語の表示を行う。これにより、さらに多くの語を類義語リストとして取得することができるようになっている。そのうえで「searchAll」をクリックして表示されている類義語リストのすべての語を論理和とする検索が分類番号の前方一致により実行される。一般に検索クエリが想起できない利用場面が考えられるが、本アプリケーションの類義語リスト表示は、検索クエリ作成の補助としても役立つものである。

図 5-1 意味索の検索画面

6. おわりに

本研究では、Web検索への応用を目的とした階層型分類番号を持つシソーラスの自動構築と、それを応用した検索アプリケーションの開発を行った。

シソーラスの自動構築では、検索クエリの共起頻度の行列化を行い、行列最適化処理を行った後に、ユークリッド距離によって各単語間の距離を表現し、階層型クラスタリングを行った。これにより、Webドキュメントへの意味情報付加や検索クエリ作成の補助などの用途で実用になる精度のシソーラスを構築することができた。このシソーラスは音声認識における言語モデル構築への応用や、シソーラス自体としての利用という用途も考えられる。

検索アプリケーションの開発では、上記の技術を利用して、シソーラスの階層構造に基づく類義語の探索や、類義語の一括検索が行えるアプリケーションの開発と公開を行った。

本手法ではシソーラスの構築をユークリッド距離による非類似度表現を利用してクラスタリングすることによって行ったため、得られたものは各単語間の距離のみであり、単語間の関係の抽出までは行えていない。これについては、各共起クエリの品詞情報の分析や、アソシエーション分析による共起関係の特性を分析、または既存辞書との統合により実現を目指す。

また、大規模データを扱う際に大量のメモリ空間や計算量が必要となるため、新語のアップデート処理などの機能を加えて、定期的に全体のクラスタリングを行い、新語を発見する度にそれを追加する機構も重要である。

参考文献

[1] 国立国語研究所, "分類語彙表 －増補改訂版－", 大日本図書刊, 2004.

[2] NTTコミュニケーション科学研究所, 日本語意味体系 CD-ROM版, 岩波書店, 1999.

[3] 日本電子化辞書研究所, "電子化辞書仕様説明書第2版", 2001.

[4] 徳永, "言語と計算-5 情報検索と言語処理", 東京大学出版会, pp.143-148, 1999.

[5] 馬青,神崎享子,村田真樹,内元清貴,井佐原均," 日本語名詞の意味マップの自己組織化",情報処理学会論文誌 Vol.42, No.10, 2001

[6] 安川美智子, 山田篤, "Web閲覧履歴に基づくシソーラス自動構築", DEWS2004 I-2-04 2004.

[7] 安川美智子, 山田篤, "Web検索エンジンを用いた用語検索履歴からのシソーラス自動構築手法の評価と改良", DEWS2005 5C-i8, 2005.

[8] Zheng Chen, Shengping Liu, Liu Wenyin, Geguang Pu, Wei-Ying Ma, "Building a Web Thesaurus from Web Link Structure", SIGIR 2003, pp.58-55, 2003.

[9] J. Wen, J. Nie, and H. Zhang. Query clustering using userlogs. ACM Transactions on Information Systems (ACMTOIS), Vol. 20, No. 1, pp. 59.81, January 2002.

[10] Thomas K Landauer, Peter W.Foltz, Darrell Laham, "Introduction to Latent Semantic Analysis", Discourse Processes, 25, pp259-284, 1998.

[11] Iain Duff, Roger Grimes, John Lewis, "User's Guide for the Harwell-Boeing Sparse Matrix Collection", October 1992.

[12] MICHAEL W.BERRY, LARGE SCALE SPARSE SINGULAR VALUE COMPUTATIONS, International Journal of Supercomputer Applications, Vol.6, No1, pp.13-49, 1992.

[13] 村田, 内山, 金丸, 井佐原, "階層構造データ列の簡易な高速検索アルゴリズム", 情報処理学会研究報告, 2005.