

## Web からの同系統単語知識獲得についての実験

野口, 正人  
九州大学システム情報科学府

廣川, 佐千男  
九州大学情報基盤センター

<https://hdl.handle.net/2324/2978>

---

出版情報：全国大会講演論文集. 65, pp.223-226, 2003-03. 情報処理学会

バージョン：

権利関係：ここに掲載した著作物の利用に関する注意 本著作物の著作権は（社）情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。

# Webからの同系統単語知識獲得についての実験 A System for Collecting Similar Terms from WWW

野口 正人<sup>†</sup> 廣川 佐千男<sup>††</sup>

ある概念についての例を多数集めたいとき、検索エンジンにその概念をキーワードとして与えても、得られるのはそれに関連するページ群であり、個別に単語を抜き出しまとめ直す作業が必要となる。我々は、概念を単語として与えるのではなく、概念のインスタンスとなる単語例を3~5個与え、それらを表の一部として含むWEBページ群を収集することにより効率的に同系統の単語群を抽出する方式を開発している。この手法では、表において同列あるいは同行に現れる単語を同系統とみなすことにより自然言語の知識を用いずに多数の単語を収集することができる。本発表では、企業名、人名、商品名、作品名、地名などの80種類の例についての実験結果を述べる。

## 1. はじめに

Webには膨大な数のホームページがあり、所在(URL)さえ分かれば必要な情報を瞬時に入手できるようになった。その意味で、Webは人類がこれまでもったことがない真の百科辞典ともいえる。百科辞典との大きな違いは、百科辞典には全体の構成を行なった編集者がいるのに対し、Webではそのように全体的な意図で構成されているわけではなく、個別のページが独立に関係なく作られている点にある。従ってWeb上の情報を活用するには検索が必要となる。現在一番広く使われているWeb検索は、キーワードを入力してそのキーワードに関連するWebページを検索結果として返すものである。ある企業のホームページを知りたい場合のように、一つのページが求める対象であればよいが、その企業のある商品、あるいはその商品の市場価格を調べたい場合には、通常、検索結果として得られる複数のページを順番に詳細に見て、メモをとったりして利用者は求める情報を構成しなければならない。我々は、文献<sup>2)</sup>においてこのような詳細情報を効率良く収集する方式を提案した。本稿では、その方法を実装したシステムを用いて行なった評価実験について述べる。企業名、人名、商品名、作品名、地名などの80種類の例について、3~5個の例を与えるだけで多数の同系統単語の収集に成功した。本手法は、従来の検索エンジンと比べ、細粒度の情報を効率良く収集するものといえる。

## 2. 細粒検索

検索結果の件数と検索対象の粒度という二つの観点から捉ることにより、我々の細粒度知識獲得方式<sup>2)</sup>を従来のWeb検索などと比較することができる(図1)。通常、検索エンジンは、キーワードに応じた少数のWebページを返すことがその第一義的目的なので、本稿ではページ検索と呼び、図1では左下の部分に位置付けられる。

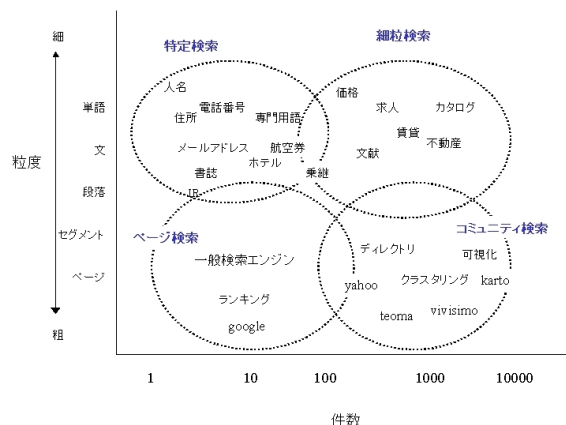


図1 細粒検索の位置付け

そこでは得られたページを個別に閲覧しなければならないので、より適切なページを上位に提示するために、ランキングが重要となっている<sup>3)</sup>。同様にページを対象としていても、検索結果として単独のページではなく、複数のページ群が期待される場合には、yahooのように予め分類されたディレクトリ構造として表示する方法や、teomaやvivisimoのように関連するページ群をクラスタリングして表示する方法、あるいはグラフィカルな可視化を用いたkartooなどがあり、本稿ではコミュニティ検索と呼ぶ。図1の左上の部分、細粒度の単一情報を求める検索サービスであり、本稿では特定検索と呼ぶことにする。人名、住所、電話番号、メールアドレス、書籍情報、専門用語、企業決算公告データなどがある。これらは個別のDBを持っていて、検索結果としてWeb情報を返すわけではないのでWeb検索と呼ぶには広すぎるかもしれないが、そのようなDBはなんらかの形でWeb情報を収集した結果として構築されたものといえる。図1の右上の細粒検索と表した部分では、例えば、関連研究を調査するときに利用する文献検索siteseerのように、理想的には関連するすべての論文の情報が検索結果として要求される。本稿で述べる手法がもっとも有効に利用できる、細かい粒度で同系統の情報を多数求める検索である。

<sup>†</sup> noguchi@matu.cc.kyushu-u.ac.jp, 九州大学システム情報科学府, Graduate School of Information Science and Electrical Engineering, Kyushu University

<sup>††</sup> hirokawa@cc.kyushu-u.ac.jp, 九州大学情報基盤センター, Computer and Communications Center, Kyushu University



にあっていのかどうかという観点から評価した結果、下の4通りに分析できた。

- a 大体期待していた通りのリストが出来た (47)
- b 期待していたものとは異なるが、意味のあるリストが出来た (8)
- c 期待した単語群は得られているが、上位に関係ない単語が多く含まれていた (8)
- d 殆ど有用な出力結果が得られなかった (17)

全80種類の入力例の“期待したカテゴリ”と、それらが実行の結果どのように分類されたかを表1に示す。

全体の約6割が、入力を作成した人間の期待に沿う結果となっていたが、約2割の例はリストの抽出に完全に失敗していた。以下に、失敗した原因についての考察を記す。

#### 6.1 完全にリスト生成を失敗したものの

リストが殆ど抽出出来なかったの例をみると、失敗の原因としては次の様なものが考えられる。

- (i) 初期キーワードの中に検索エンジンでも殆ど見つからない単語が多く含まれていた

本手法では、キーワード集合から選択した数語の検索キーワードを検索エンジンに与えることで、検索キーワードを含むページを探している。初期キーワードの単語が検索エンジンに与えても殆ど検索結果が返ってこないような単語の場合、対象となるページの絶対数が少ないため、単語を全く抽出できないことがあり、そのような単語ばかりを初期キーワードとしてあたえと、リストの生成は失敗する。

- (ii) 初期キーワードが一般的過ぎる単語のため、検索エンジンで関係ないページばかりが見つかった

例えば、“月の名前”として‘1月’、‘2月’、‘3月’、‘4月’、‘5月’を入力したが、このような単語は非常に一般的な単語であり、これらの語を全て含んでいても、1列には並んでいないようなページが非常に多く存在する。単語が一般的であればあるほど、そのようなページが検索結果として大量に返される事になるが、本実験では検索結果の上位50件までを走査の対象としたので、その範囲内に本手法に対して有用なページが含まれていない場合、単語集合の生成は失敗することになる。

- (iii) 初期キーワードが表として並ぶことがあまり無い(比較されたり、列挙されたりすることが無い)ものだった

“チョコレート菓子”や“日本酒の銘柄”のように、インターネット上にそれらを列挙したページが殆ど存在しないとされるような単語の場合、本手法は失敗しやすい。

そのような性質を持ったものでも成功している入力例が幾つか存在するが、それらは多くの場合、その例の入力単語が非常に「個性的」であった。つまり、それらの単語を検索エンジンに与えた結果としてかえって来るページが少なく、その単語が出てくる文脈は同系統単語の列挙の場合が殆どであるような単語であれば、Web上に情報が少なくても成功することがある。

#### 6.2 期待と異なった意味の単語群が集まったもの

8件が“期待していたものとは異なるが意味のあるリストが出来た”となった。

例えば、“首都”や“オリンピック開催都市”として、前者は“東京”、“スリジャヤワルダナブラコッテ”、“キャンベラ”、“ソウル”、“ロンドン”を、後者は“東京”、“アテネ”、“ローマ”、“バルセロナ”、“シドニー”を入力単語として与えたが、両者共、得られた結果はもっと一般的な世界の主要都市のリストとなり、入力作成者の意図には合わない結果となった。

また、“アフリカの国名”、“自動車メーカー”も、同様に(多少スコアにかたよがりがあるものの)期待していたカテゴリより広い意味でのカテゴリと誤認された単語群が生成された。

一方、“甲子園の常連高校”、“ノートパソコン”などは、入力として与えた単語の一部にのみ共通する特徴である‘高校の所在地’や‘パソコンのメーカー’が強調され、それらの観点からのリストになってしまっていた。つまり、前述の国名などの例とは逆に、より狭いカテゴリと誤認されたことになる。

“ギリシャ神話の神”の例の場合、各地の神話をモチーフとして使ったゲームがあり、そのゲーム中にそれらの神々が登場するため、そのゲームの登場キャラクターをまとめた表が抽出の対象となり、結果として各地の神話の神々のなまえが混ざって抽出された。

“色”の場合、ショッピングサイトなどで服の色や模様を選ぶリストボックスから抽出されることが多かったため、色だけでなく模様の種類も結果に含まれていた。また、前述の一般的過ぎる単語の性質も満たしているため、ノイズの多い結果になっていた。

これらの例に共通して言えることは、利用者が期待していたカテゴリとしてのリスト(例えば首都のリスト)よりも、それとは異なったカテゴリとしてのリスト(例えば主要都市のリスト)の方が、インターネット上のWebページに圧倒的に多く存在する場合、そちらのカテゴリに引っ張られて利用者の期待とは異なったリストが生成されたと考えられる。

#### 6.3 ノイズの多いリストが生成されたもの

同じく8件あった“期待した単語は出てきているが上位に関係ない単語が多く出てきていた”に分類されたものの場合、多くの例では単語の抽出の対象となった表に、試合やレースの結果表が含まれていた。例えば、“投げ技”は柔道の試合の対戦表の結果の欄に多く現れていた。また、“棋士の名前”や“メジャーリーグのチーム名”、“F1ドライバー”なども同様にそれぞれのチームや人の対戦表に単語が含まれているため、これらの表が対象となった。“車の会社名”や“バイクのメーカー”も、それらの会社の持っているレーシングチームの試合結果を示す表が走査対象となった。

これらの表には「点数や勝数など数字」や「 $\times$ 」や「 $\times$ 」などの記号が多く含まれる。また、全く関係の無い表であってもそれらの数字や記号を含んでいることは多い。本手法では基本的にキーワード集合に含まれている単語が複数並んでいる列のみを単語抽出の対象とするので、

表 1 実験結果

大体期待していた通りのリストが出来た (47 件)
キノコ, プログラミング言語, SMAP のメンバー, 小説作家, インターネットサービスプロバイダ, 大学, JR 九州の駅, 声優, 自動車メーカー, 鞆などのブランド, プロテニスプレーヤー, 漱石の作品, 有名な映画, 国名, 温泉, あるシリーズ小説のサブタイトル, お笑い芸人, 競争馬, あるアーティストの曲, あるゲームのキャラクター, 麻雀の役, プロレスラー, 学会名 (分野関係なし), コンピュータウイルス, 教科名 (小・中学校), 星, 冬のスポーツ, 寿司ネタ, トランジスタの型番, コンピュータ雑誌, テレビアニメ, CPU の種類, 目薬の名前, あるテレビアニメシリーズのサブタイトル, テレショップ, ボーカーの役, 電器店, 北欧神話の神, 三国志の武将, ある漫画の登場人物, あるテレビアニメの登場人物, 通販の会社, 日付, あるゲームのアイテム, あるアニメシリーズに出てくる型番, 検索エンジン
期待していたものとは異なるが、意味のあるリストが出来た (8 件)
首都, オリンピック開催都市, 自転車メーカー, 色, アフリカの国名, 甲子園の常連高校, ノートパソコン, ギリシャ神話の神
期待した単語は出てきているが、上位に関係ない単語が多く出てきていた (8 件)
投げ技, 棋士の名前, メジャーリーグのチーム名, 車の会社名 F1 ドライバー, 航空機, バイクのメーカー
殆ど有用な出力結果が得られなかった (17 件)
日本酒の銘柄, 空手形, ツール・ド・フランスの選手, 宗教, クワガタ, 月 (1月,2月,...), 山脈, チョコレート菓子, クリスマス・ソング, 日本の神話の神, 弦の名前, ある会社の取締役, 研究会メンバー, 料理用語, 京都名物の食べ物, GREEN DAY の曲, 遺伝子データベースの accession No.

そのような数字や記号が抽出される可能性は少ない。しかし、一部の表は特殊で複雑な形をしており、例えば、ある1つの試合の結果表示として、‘チーム A’, ‘1-2’, ‘チーム B’ というように語が並んでいた場合は ‘1-2’ という誤った単語が抽出されてしまうことがある。一旦そのような語が抽出され、キーワード集合に含まれてしまうと、その語を含む全ての表で間違っただけから単語が抽出される現象が起き、結果としてそれらの不適切な単語が高いスコアを得てしまうことになる。

“航空機” の場合は、飛行機の運航時刻表などが単語抽出の主な対象となったが、時刻や便の番号などの数字がキーワード集合に紛れ込んだため、上記と同様の現象が起きていた。

その他、全ての例にいえることとしては、その単語集合のカテゴリを表す言葉、例えば、温泉のリストの場合は “温泉名”、国のリストの場合は “国名” といった単語が、比較的高いスコアで抽出されることが多かった。これは、表のヘッダの部分にそれらの単語が記述されている場合が多く、そのヘッダと表の本体とに構造的な差異が無い場合、ヘッダに記述された単語も同系統の単語とみなされるために起こったと考えられる。

## 7. まとめと今後の課題

本論文では、数個の単語例を与えることで Web から同系統の単語集合の抽出を行う手法を述べた。この手法は次の様な特徴をもつ。

- 数個の単語を与えるという簡単な操作だけで、それに関するリストを取得することができる。
- 最初の単語を与える部分以外は完全に自動化できる。
- アルゴリズムは単語の意味や、それが何のリストであるか、という情報を一切必要としないので、分野を問わず利用できる。
- HTML の構造を利用して解析を行い、自然言語の知識を利用しない。
- 抽出されるリストは、初期キーワードを与えた人間の意図をくみとったリストになりやすい。

この手法の実装システムを用いた実験の一部について

評価を行なった。本手法は単語の知識や自然言語の知識を使わない単純な手法であるにも関わらず、多様な単語集合の作成が可能であり、有効な手法であることを示した。紙面の都合よりより一般的な評価実験、適用可能分野の検討は詳しく述べられなかったが、手法の改良も含め別稿で発表予定である。

## 参 考 文 献

- 1) S.Brin, “Extracting patterns and relations from the world wide web”, WebDB Workshop at EDBT '98, 1998
- 2) 野口正人, 廣川佐千男, “SoftPath を用いた同系統単語抽出方式”, 人工知能学会研究会資料 SIG-KBS-A203, pp.15-20, 2002.
- 3) L.Page, S.Brin, R.Motwani, T.Winograd, “The PageRank Citation Ranking: Bringing Order to the Web”, 1998, <http://www-db.stanford.edu/~backrub/pageranksub.ps>
- 4) S.Sato, M.Sato, “Toward Automatic Generation of Web Directories.” Proc. of International Symposium on Digital Libraries 1999 (ISDL'99), pp127-134, Tsukuba, September 28-29, 1999.
- 5) Yalin Wang, Jianying Hu, “A Machine Learning Based Approach for Table Detection on The Web”, The Eleventh International World Wide Web Conference (WWW2002).
- 6) Kristina Lerman, Craig A.Knoblock, Steven Minton, “Automatic Data Extraction from Lists and Tables in Web Sources”, Automatic Text Extraction and Mining workshop (ATEM-01), IJCAI-01, 2001.
- 7) 梅原雅之, 岩沼宏治, 永井宏和, “事例に基づく HTML 文書から XML 文書への半自動変換”, 人工知能学会誌, 16 巻 5 号 B, 2001.
- 8) William Cohen, Matthew Hurst, Lee S.Jensen, “A Flexible Learning System for Wrapping Tables and Lists in HTML Documents”, The Eleventh International World Wide Web Conference (WWW2002).