A System for Collecting Similar Terms from WWW

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

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：

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[†] 廣川 佐千男[††]

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、キーワードを入力と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[2] 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えることにより、我々の細粒度知識獲得方式[2] を従来の Web 検索などと比較することができる (図 1)。通常の検索エンジンは、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。

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

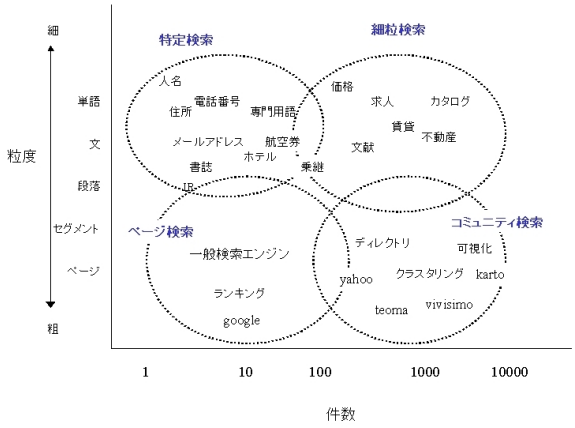

図 1 細粒検索の位置付け

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[3]。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 の左上の部分は、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、書誌情報、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。

3. EIP 構築ツールへの利用

我々の細粒情報収集法は、検索機能と切り離して、EIP(Enterprise Information Portal) 構築のための基礎的技術として十分実用的なものと考えられる。EIP において提供される情報がその組織だけの閉じたものならば、関連する外部情報を利用するために別のアクセス手段が必要となる。例えば、味の素株式会社のサイトでは、6000 メニューからシチュエーション、食材、調理法などで料理のレシピを検索ができるページ「レシピ大百科」のページを設けている。キューピー株式会社でも同様のレシピのページを備えポータル・サイトとしての価値を高めている。このように、外部の情報であっても、同種の情報をまとめることにより EIP としての価値を高めることができる。

4. 表構造に着目した同系統単語抽出と索引付け

特定の対象について体系的にページ群を収集し、その中から必要とする情報のみを抽出、統合し、用語辞典や百科辞典を作る研究がある[1),4),7)]。本研究では、収集するのは Web ページではなく、Web ページに含まれる同系統の単語群である。本手法では、まず、探したい単語の例を 3〜5 個検索エンジンに与え、それを含むページを収集し、得られたページが表構造を含むかどうかを判定する。表構造を含めばその単語が、同一列あるいは同一欄に現れていれば同一列あるいは同一欄に現れる他の単語を同系統とみなす。こうして得られた新たな単語について同じ操作を繰り返すことにより、効率良く同系統の単語知識を増やすことができる。一連の操作で使うのは、表構造の情報だけであって、単語についての知識や自然言語処理など一切必要としない。表の抽出についてはいくつかの研究[5),6),8)] があるが、我々の手法は、表の記述として TABLE タグや OL、UL などのリストを表すタグに限らず繰り返し現れるタグのパターンを発見、抽出する。

一般的の検索エンジンではキーワードによる検索を実現するためには、多数の Web ページを収集し、ページ単位で索引付けしなければならない。また逆にそれぞれのキーワードがどのページに含まれていたかを瞬時に求めるための転置索引を作成しなければならない。ページ単位の検索でなくより詳細な検索のためには、ページがどのような構造をしていてその中のどの部分にキーワードが現れていたかという詳細なインデックス化が必要となる。我々の方式は、同系統の細粒情報を収集する時点ですでに構造解析まで行うので、収集と同時に粒度の高い索引付けが実現できる。

5. 単語のスコア

九州の温泉を探す場合について、本システムの使い方と得られる単語群のスコアの解釈を説明する。利用者は、“二日市温泉”、“原鶴温泉”、“嬉野温泉” のように探す温泉の具体例を 3 個程度入力する。そのような具体例を知らない場合には、例えば「九州の温泉」というキーワードを用いて一般の検索エンジンで検索を行なうことで得

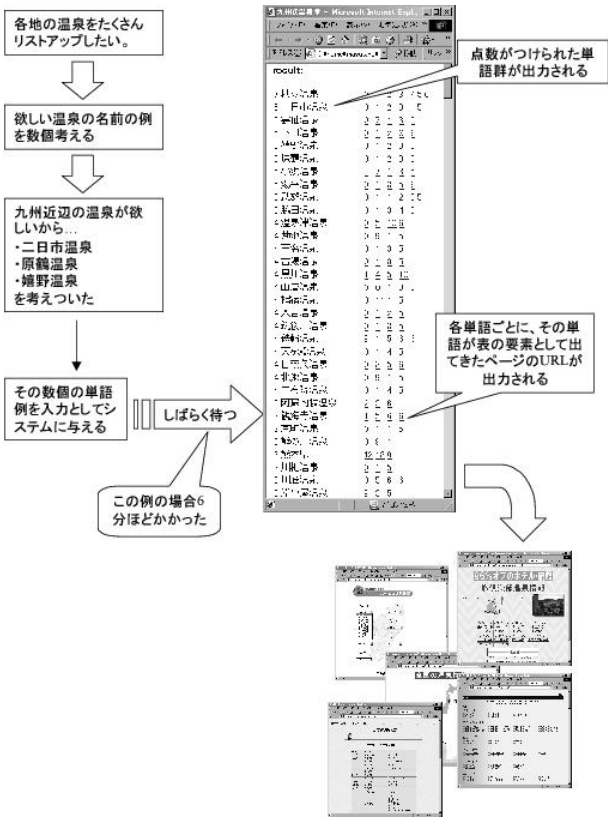

図 2 温泉リスト

られる例を使う。出力として、スコア (点数) が付けられた単語のリストが得られる (図 2)。図 2 において、温泉の名前の横に並んでいるのは、その温泉を含んでいた表の HTML ページの番号である。単語（この場合、温泉名）のスコアは、その単語を含んでいたページの個数である。スコアが高いということは、その温泉名を含んでいる表が沢山あったことを意味し、人気があるあるいは有名であると考えられる。この例についていえば、例として入力した温泉が福岡県、佐賀県の温泉であったため、表に含まれるものも九州のものがほとんどであった。

6. 実験と評価

これまで述べたシステムを実装し、実験を行った。システムは、Windows および Linux 上で動作する Perl5 用のスクリプトとして実装した。

このシステムは 1 回の動作に対して、入力として同系統単語を数語必要とする。今回の実験では、人手でその入力単語を作成し、その結果を人間の感覚で判断し評価をおこなった。

入力例は 14 人で分担して、合計 80 種類作成した。それぞれの入力例は各々3〜5 語ずつから成り、その入力に対してどのようなリストが生成される事を期待したかを、“期待したカテゴリ” として同時に記録した。

それぞれの入力例から得られた単語群が、(1) 十分な大きさのリストになっているかどうか、(2) 本当に同系統の単語になっているかどうか、(3) 期待したカテゴリ

にあっている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

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ると、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思わ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プラコッテ”, “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” の場合は、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「○や×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, 鞄などのブランド, 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, ポーカーの役, 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列から単語が抽出される現症が起き、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事になる。

“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).