Webからの同系統単語知識獲得方式

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：情報学シンポジウム講演論文集. 2003, pp.21-24, 2003-01
バージョン：
権利関係：

Web からの同系統単語知識獲得方式

野口正人[†]　廣川佐千男[††]

少数の単語を例として与え同系統の単語を収集する方式を提案し、それを実装したシステムを使った実験結果について述べる。たとえば、二日市温泉、原鶴温泉、嬉野温泉という３つの温泉名から、82 個の温泉名を得る事ができた。本方式では、まず収集したい単語の２、３個の例を検索エンジンに与え、得られるページから表と見なせる部分を抽出する。それらの表で与えたキーワードが同一欄あるいは同一行に現れる場合、その欄、あるいは行に現れるの他の単語を同系統の単語と見なす。この一連の操作を繰り返し、単語を増やすことができる。

1. はじめに

爆発的に成長を遂げた Web 上にある大量の情報の中から目的の情報を探すことは困難な作業である。探索目的に応じて一つのページが見つかれば十分なこともあるが、逆に複数のページ群が必要なこともある。ある特定企業の住所を調べたいというように探す目的が限定され、一つの特定のページが想定される場合には、そのページが見つかるか、あるいは、ランク付けされたページのリストが得られればよい。[2] 一方、ある分野に関連してどのような企業があるか調べたい場合のように、求めるものが一つのページではなく、複数のページあるいはそれらへのリンク集のこともあり、そのためには、検索結果のクラスタリングや、リンク構造を利用し Web コミュニティとしてまとめることがより重要となる。更に、求めるものが Web ページでなく、そこに含まれる部分的情報であれば、さらに高度な後処理が必要となる。そのために、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構築する試みもある。[1),3),6)]

このように、WWW における探索といっても、探す対象が Web ページであるか、あるいは用語や企業名などの単語であるかという、探索対象の単位によって異なるアプローチが必要である。また、探索結果として対象とする単位が一つ得られれば良いのか、一群の結果を必要とするかに応じて、異なるアプローチが必要である。

本研究では、ページではなく、単語を求める情報の単位とし、同系統の単語群を求める方式を提案する。また、考案した手法を用いたシステムを実装して実験を行い、その評価を行った。本手法では、まず、探したい単語の例を２〜３個検索エンジンに与え、それを含むページを収集する。得られたページにおいて、表やリストの形式を含み、しかも同一列あるいは同一欄に与えた単語を含むものを抽出する。同一列あるいは同一欄に現れる単語を同系統とみなすことにより、単語を増やすことができる。

与えられた単語を含む表の発見が本研究の核であり、そのような表と見なせるページが十分沢山存在し、その表構造を自動的に抽出できなければ、有効とはいえない。表の抽出についてはいくつかの研究[4),5),7)] があるが提案する手法は、Web ページを記述する HTML ファイルにおいて繰り返し現れるタグのパターンに着目し、表構造を抽出するものであり、TABLE タグや OL、UL などのリストを表すタグに限ったものではない。

† noguchi@matu.cc.kyushu-u.ac.jp 九州大学システム情報科学府
†† hirokawa@cc.kyushu-u.ac.jp 九州大学情報基盤センター

2. 表の発見と表からの同系統単語抽出 – 温泉リスト

「温泉旅行の計画」についての情報を検索する場合について、本手法の考えを説明する。目的の温泉を何処にするか考える際に、参加者の旅行経験や性格などに応じて多様な条件が考えられるが、その条件は曖昧であり、最終的には人間の判断を必要とする。そこで、全国の温泉地に対して温泉地の名称と住所、泉質や効能などの情報がまとめられた表があれば、その表を見て何処に旅行に行くかを決める事が可能になる。

Web 上には、各地の温泉をリストアップしたページが多数あり、それらのページを見比べて欲しい情報を揃える事が出来る。

本論文で提案する手法では、まず、求めるリストの要素の具体例を３個程度検索エンジンに与え、その結果得られるページからリスト構造となるものだけを選択する。ただし、リストの要素として、与えた具体例を同じフィールドに含むものだけを選び出す。これにより、同系等の単語のリストが得られる。これを繰り返すことにより、少数の単語のリストから始めて、多くの単語のリストへと展開できる。繰り返し行う検索の過程で、ある単語を含むリストが多数あれば、その単語は求めているものである可能性が高い。

このシステムを利用して温泉リストを取得する例を図 1 に示す。例えば、利用者が九州の温泉を列挙したいと考えた場合、“二日市温泉”, “原鶴温泉”,“嬉野温泉” のように欲しい単語の具体例を 3 個程度考え、システムに入力することで、その他の九州の温泉名 (例えば、“杖立温泉”,“雲仙温泉” など) のリストを受け取ることが出来る。

出力されるリストは、温泉の名前とその温泉名が出現する Web ページ数の対になっている。温泉名が出現する Web ページ数は、その温泉と最初に入力した温泉との関連が深ければ深いほど多くなる。今、入力として与えた温泉は全て九州の温泉なので、九州の温泉の Web

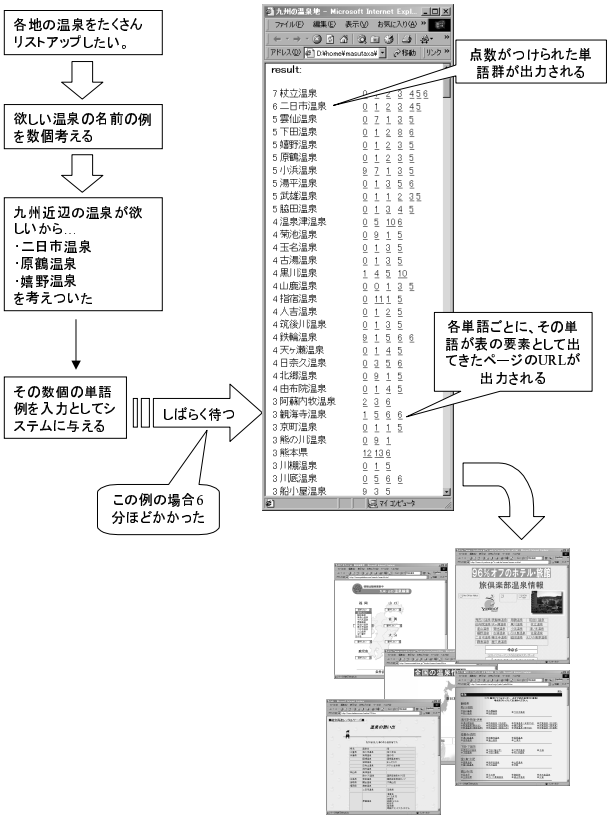

図 1　温泉リスト

ページ数は多くなる。また、一般的な温泉のほうが Web ページ数は多くなるので、“杖立温泉” や “雲仙温泉” などの九州の有名な温泉地は Web ページの数が多くなる。つまり、出現する Web ページの数は、その単語が利用者のほしかった単語に近いものかどうかの指標となっている。本手法ではこの数字を単語のスコアと表現する。

このシステムでは、温泉名だけでなく、その温泉名が出現する Web ページの URL も同時に得られ、そのページから詳細情報を得ることができる。あるいは、新たに得られた温泉名を一般の検索厭人で検索することでも高品質な詳細情報を得ることができる。

以上は温泉の例であったが、このシステムではその単語がどのような種類のものであるかという事は一切関知しない。これは「“二日市温泉”,“原鶴温泉”, “嬉野温泉” といえば、他には？」という質問を人間に質問することに似ている。人間は、自分の判断で 3 つの単語の共通点をみつけ類似例を回答する。それと同様に、このシステムは Web 上に「同系統の単語のリスト」として蓄えられた情報を総合し、類似例を検索する。

3. リスト展開アルゴリズム

本論文で提案する手法は、まず検索エンジンに２～３個で検索を行ない、得られるページから表を検出し、新規キーワードを抽出するという一連の手順を繰り返し実行する。(図 2)

ここに現れる「初期キーワード」や「キーワード集合」、

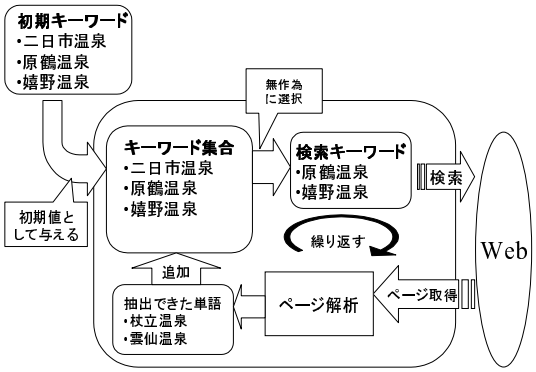

図 2　検索・表の検出・追加の反復手順

「検索キーワード」は、本論文では以下のような意味で用いる。

初期キーワード　この手法に入力として与える数個の単語。出力として得られる単語リストは、この初期キーワードと同系統の単語のリストである。

キーワード集合　手順の途中の時点で、初期キーワードと同系統の単語として抽出された単語の集合。手順が進むにつれて要素が増える。最初は初期キーワードだけであるが、手順を進めることで単語が追加される。

検索キーワード　この手法では、検索と解析の処理を数回繰り返すが、その繰り返し 1 回ごとに決められる、検索の時に用いられるキーワード。これは、キーワード集合から無作為に数個選出される。

3.1 反 復 手 順

この手法は、まず、利用者が欲しい単語群に対して、適切な初期キーワードを考えることから始まる。初期キーワードとしては、目的としている単語集合に含まれている代表的な単語を数個与えればよい。温泉を探すなら自分の知っている有名な温泉名を数個与える。

次に、検索エンジンを用いて初期キーワードを含むページを検索する。そして、その結果見つかったページを解析し、ページ内に初期キーワードが並んだ表が含まれていた場合、その表に並べられたほかの単語も同系統の単語としてキーワード集合に加え、記録する。

こうして得られたキーワード集合のなかから無作為に数個の単語を選び、その単語を含むページを検索エンジンで探し、その単語が並んだ表を探し、新しい同系統の単語を得る。この一連の処理を繰り返すことで同種の単語に関する知識を探索するのがこの手法である。

3.2 ページ解析

ページ解析では、検索によって得られた各ページから検索キーワードと同系統の単語の抽出を行なう。この処理はおおきく別けると、表構造の抽出と同系統単語の抽出の 2 つの手順にわけられる。

表構造の抽出は、まず、得られたページの HTML を木構造に解析する。そして、その木の中でブロックレベル要素の構造だけを見て同じ構造のくり返しとなっている部分を探し、その部分を表とみなす。次に、ブロックレベル要素以外の部分の構造に差異があった場合は、その部分の表現を統合することにより、完全な二次元の表

に変換する。

同系統単語の抽出は、表構造の抽出で得られた表の各列の中で、検索キーワードを最も多く、そのままの形で含んでいる列を探し、その列に並んでいる単語を同系統単語と見なして抽出を行なう。

4. 実験と考察

任意に選んだ数個の例に対して実際にどの様なリストが取得できるかの実験を行なった。実験のシステムは、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。

4.1 九州の温泉

まず、最初に次の様な入力で実験を行なった。

- 二日市温泉
- 原鶴温泉
- 嬉野温泉

これらは、全て温泉地の名前であり、二日市温泉と原鶴温泉は福岡県、嬉野温泉は佐賀県の温泉である。

全部で 73 ページを解析して、そのうち 16 ページから 42 個のテーブル、638 種類の単語を抽出出来た。ただし、638 個の単語のうち 544 個はスコアが 1 であり、スコア 2 以上の単語は 94 個であった。

スコアが低くなる単語は

- そもそも同系統の単語ではないがページの構成の関係で見かけ上並列に並んでいた単語
- 広い目で見れば同系統の語ではあるが、深い関連があるわけではない単語
- 漢字の違いやスペルミス、もしくは通称などの正式な名前とは異なる単語

などであると考えられる。特にスコア 1 の単語はその傾向が顕著であり、無視すべき語群であるといえる。

スコア 2 以上の単語に対して、単語の種類をスコア別に集計したものを表 1 に示す。

スコアが 2 以上の単語群には温泉地名以外の単語は 9 個含まれていた。それらのうち、2 つは “温泉名” と “大分” であり、それ以外の 7 つは九州の各県名を “○○県” の形で表記した語であった。

この表では、スコアが小さくなるにつれて、九州以外の温泉名や温泉名以外の単語が混じっていく様子が見られる。また、もともと温泉地の数が多く、有名な温泉も多い熊本県や大分県を除くと、スコアが高い部分に福岡、佐賀、長崎の温泉が並んでいる。これは、入力として与えた温泉地が、2 つは福岡県、1 つは佐賀県の温泉なので、九州北部の県の方がスコアが高くなりやすかったためだと思われる。

テーブルが見付かった 16 ページ中 9 ページからはテーブルが 1 つしか見つからず、複数のテーブルが見つかるページは多くが県別や地域別に並んでいた。

また、殆ど全てのテーブルが、所在地などの情報や詳細なページへのリンクを持っていた。

4.2 スタジオジブリ作品

次に入力として与えた単語は

- “ラピュタ”
- “ナウシカ”

表 2 スタジオジブリ作品の出力結果

23	もののけ姫	10	おもひでぽろぽろ
21	となりのトトロ	8	火垂るの墓
19	風の谷のナウシカ	8	平成狸合戦ぽんぽこ
18	天空の城ラピュタ	6	もののけ山
15	耳をすませば	5	海がきこえる
15	魔女の宅急便	4	となりの山田くん
14	紅の豚	4	ルパン三世
12	千と千尋の神隠し	4	熱風

- “千と千尋の神隠し”

である。これらは、すべてスタジオジブリ製作の映画の名前である。ただし、“ラピュタ” と “ナウシカ” は、正しくはそれぞれ “ 天空の城ラピュタ ”、“風の谷のナウシカ” という名称である。

一般に “ジブリ作品” と呼ばれている映画は、実験を行なった 2001 年 12 月の時点では 1983 年公開の “風の谷のナウシカ” から 2001 年公開の “千と千尋の神隠し” までの 13 作品であり人手でも容易に列挙が可能な量ではあるが、不完全な名称を入力として与えた場合の例として実験を行った。

出力された結果でスコア 4 以上のものを表 2 に示す。

入力として与えた通称の単語 2 件は “ナウシカ” がスコア 1、“ラピュタ” がスコア 2 の単語として出力されたが、“風の谷のナウシカ” はスコア 19、“天空の城ラピュタ” はスコア 18 と正式名称に補完された形で上位に示されている。また、目的としていた 13 作品の正式名称はスコア 4 以上の部分に全て含まれていた。

スコア 4 以上の部分に目的の単語以外のものが 3 つ混じっていた。それら 3 つの単語は全て Web サイトの名前であり、単語が抽出されたテーブルはリンク集であった。これは、最もスコアの高い “もののけ姫” や入力として与えた “千と千尋の神隠し” などとまったく同名の Web サイトがあったため、Web サイト名のリストであるリンク集も入力と同系統の単語とみなされてしまったためである。

スコア 3 以下の単語は多くがリンク集から抽出された Web サイト名であったが、そのほかには

(1) 映画中に使われた曲の名前
(2) (映画に限らず) ビデオや DVD で販売されている作品名
(3) 宮崎駿 (入力に与えた 3 作品の原作・脚本、兼監督) がかかわってはいるが、スタジオジブリ制作ではない映画やテレビアニメーション作品の名前

なども少量ずつではあるが含まれていた。

4.3 ハーブの種類

次に入力として与えた単語は

- “ジャスミン”
- “レモングラス”
- “ラベンダー”

である。

これらは一般に “ハーブ” と呼ばれる香りを持った植物の名前である。ハーブは料理用に、ハーブティー用に、薬用に、染色用にと利用法が多々あり、同じ植物でも利用する部分や形態によって名称が異なる場合もある。ま

表 1 出力された単語の種類の分布

単語の	九州内の温泉								九州外	温泉地名
スコア	福岡	佐賀	長崎	熊本	大分	宮崎	鹿児島	合計	の温泉	以外
7	0	0	0	1	0	0	0	1	0	0
6	1	0	0	0	0	0	0	1	0	0
5	2	2	2	1	1	0	0	8	0	0
4	1	1	0	6	3	1	1	13	1	0
3	1	1	1	1	4	1	1	10	5	2
2	3	1	1	7	7	3	6	28	18	7

表 3 ハーブの種類の出力結果

22	レモングラス	8	フランキンセンス
18	ローズマリー	8	ローズ
16	ベルガモット	8	ローズウッド
16	ペパーミント	7	クラリセージ
16	ラベンダー	7	シトロネラ
14	ジャスミン	7	ライム
13	サイプレス	7	レモンバーム
13	バジル	6	ジュニパーベリー
13	レモン	6	ネロリ
11	サンダルウッド	6	バニラ
11	ユーカリ	6	パチュリー
10	グレープフルーツ	6	レモンバーベナ
10	シダーウッド	5	クローブ
10	ミルラ	5	シナモン
9	イランイラン	5	ジャーマンカモミール
9	ジンジャー	5	スペアミント
9	ゼラニウム	5	ティートゥリー
8	ジュニパー	5	ニアウリ
8	セージ	5	ハニーサックル
8	パイン	5	ヒソップ
8	フェンネル	5	ブチグレン

た、ハーブという言葉の定義自体が曖昧なこともあるため、ハーブの完全なリストの作成は困難である。

出力された結果うち、スコアが 5 以上になった単語のリストを表 3 に示す。

出力されたページは多くがアロマセラピー用の香油のリストであったが、ガーデニング用の植物としてのハーブのリストも少なからず含まれていた。中には植物としての性質から、お茶として飲用したときの効能まで幅広く情報を持っていたページもあった。また、一部のページは通信販売のページであった。

出力されたリストの中には “レモン” や “パイン” などの単語も上位に来ているが、それらの単語が抽出されたページは全てハーブに関するページであるので、アロマセラピーでの香りの効能など「ハーブとしてのレモン」の情報が記述されたページを容易にみつける事が出来る。これは検索エンジンには無い本手法の利点の一つである。

5. 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) 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`

3) 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.

4) Yalin Wang, Jianying Hu, “A Machine Learning Based Approach for Table Detection on The Web”, The Eleventh International World Wide Web Conference (WWW2002).

5) 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, August 2001.

6) 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 年.

7) 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).