Web 上の表検索

野口, 正人
九州大学大学院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：人工知能学会全国大会論文集. 17, pp.25-28, 2003-06. 人工知能学会
バージョン：
権利関係：

1C4-02

Web 上の表検索

Table Search on the Web

野口 正人[*1]　廣川 佐千男[*2]
Masato Noguchi　Sachio Hirokawa

[*1]九州大学大学院システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University

[*2]九州大学情報基盤センター
Computer and Communications Center, Kyushu University

1. はじめに

膨大な Web から目的の情報を探すことは困難な作業であり、検索エンジンは必須のツールといえる。しかし、検索エンジンで探し出すことができるものは、一つの Web ページであったり、いくつかまとまった Web ページ群でしかない。求めるものが Web ページでなく、商品名とその価格のように部分的情報である場合、検索結果として得られたページの中から必要とする情報を取り出さなければならない。検索結果が 500 件得られたとしても、それぞれのページに含まれる粒度は均質でなく、1 ページ中に求める情報を沢山含むものもあれば、関連するキーワードが一つ出現するだけのものもある。また、複数のページ群において探す情報が含まれる場合には、必要な情報は個別ページでそれぞれ異なるパターンとして現れるので、全体をまとめて見るためには、一度それらのファイルを収集し、その後に必要な部分を個別に抽出する処理を行なわなければならない。我々は、このような問題を、検索結果の件数と検索対象粒度という二つの観点から捉える分類法（図 1）を提案している [6]。

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、図では左下の部分に位置付けられる。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 [7] が重要となっている。一方、図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[*1] 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。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, 8, 9]。

本稿では、表のように同一ページ中に多数の同系統情報を含むページを効率よく収集する方式と、収集したページに含まれるレコードについての粒度の細かい検索方式について提案する。

本システムでは HTML 中の表とみなせる構造に着目し、その表のデータを索引付けして蓄え、検索を行うものである。

Web ページに現れる表の抽出については我々以外にも、いくつかの研究が発表されている。例えば、

- HTML の TABLE タグに限定したもの、

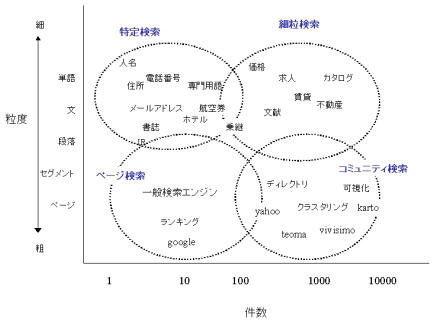

図 1: 検索システムの分類

- 数値や地名、人名などの文字の特徴を用いるもの、
- 共通的なテンプレートを人手で記述するもの、
- 予め準備した多数の例を与えることにより表のパターンを学習するもの [10]

などがある。本稿で述べるシステムで用いられている解析手法は、HTML の木構造中に現れるくり返し構造に着目することで、TABLE タグに限定されず、自然言語の知識や単語の知識を一切必要としないため、ページは一意的に解析される。これにより、データベースはページの内容が変化しない限り、後から再解析する必要もない。また、本システムでは、予め索引付けされたデータベースを用いることで、高速に多様な検索を行うことが出来る。

2. 細粒検索システムの実装

2.1 システムの全体像

図 2 はシステム構成の模式図である。システム内には 1 つの大きな表情報データベースがある。このデータベースにはそれまでに収集してきた各種のページやそのページ中に存在していた表の情報が蓄えられている。システムの利用者は、表検索 UI 部を用いることで、このデータベースから必要な情報を検索することができる。また、システム管理者は、ページ取得ワーカー部やページ解析部に働きかけることで、表情報データベースに新たな情報を蓄えることができる。

表検索 UI 部 システム利用者に表検索部へのインターフェースを提供する。表検索 UI 部では、システム利用者からの

連絡先: 九州大学情報基盤センター 廣川研究室, 〒 812-8581
福岡市東区箱崎 6-10-1, hirokawa@cc.kyushu-u.ac.jp

*1 http://citeseer.nj.nec.com/

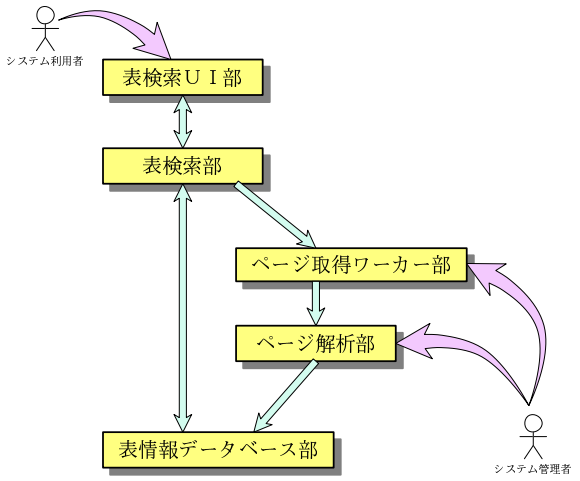

図 2: システム構成図

検索要求を、表検索部が受け付ける事ができる表検索クエリーへ変換し、返された検索結果を利用者に分かりやすい形に変換して表示する。

表検索部 表検索クエリーを受け取り、クエリーに応じた検索を行い、結果を返す。検索の対象となる情報は表情報データベース部から取得する。ただし、表情報データベースに蓄えられた情報で不足していると感じられた場合は、必要に応じて、ページ取得ワーカー部に必要な情報の収集を要求し、収集が終わって新しい情報が表情報データベースに蓄えられた後、再度表情報データベース部にアクセスして情報を取得する。

表情報データベース部 表検索部からの検索要求や情報問い合わせに応じる。この問い合わせはデータベース内の情報を直接参照するような、極めて単純なもののみである。また、ページ解析部から新規の表情報を受けとり、データベースへの格納を行う。

ページ取得ワーカー部 表検索部からの要求、または、システム運用者からの要求に応じて、Web ページを取得し、ページ解析部にデータを渡す。ページ取得要求としては、特定の URL リストを渡す方法と、情報のキーワード等を与える方法の 2 種類がある。

ページ解析部 ページ取得ワーカー部、もしくは、システム運用者から渡された Web ページから表構造を抽出し、抽出された情報を表情報データベースに格納する。

2.2 ページの収集と解析

ページの解析方法については、以前発表を行った参考文献 [5] のページ解析手法を用いる。この手法は、HTML の木構造上におけるくり返し構造を手がかりに、表とみなせる構造の抽出を行うもので、自然言語の知識や単語の知識などの予備知識を一切必要としない手法である。そのため、ページの情報のみからの一意的な情報の抽出が可能であり、予めページを解析し、解析結果を整理して蓄えておくことが可能である。手法の詳細については参考文献 [5] を参照して頂きたい。

本解析手法では、そのページがどのようにして収集されてきたか、といったこともページ解析の結果に影響をあたえない。そのため、必要なページデータさえ集まればどのような収集方法であってもかまわない。ページを収集するもっとも一般的な方法は、ページ収集用ロボットを走らせる方法であろう。この方法は、とくに収集するページが定まっておらず、内容を問わずとにかく大量のページを収集したい場合に適している。

しかし、一方で、本システムの場合、特定の分野に限定したページ群の収集を行う必要が生じる事があるので、そのようなページ収集を行うエンジンを補助的に利用する。このページ収集のための技法に関する研究も多数ある [3, 4, 2] が、システム全体はページ収集エンジン (前述のページ取得ワーカー部) の収集方法とは無関係にページの検索を行うことができるので、その実装を特に限定することはしない。最も簡単な実装の例をあげるならば、例えば、一般の検索エンジンを用いて必要なページ群に関するキーワードを含むページを検索し、そのページを取得すればよい。(これは参考文献 [5] で用いた手法である)

2.3 データベースの内容

解析された情報はデータベースに格納される訳だが、何をどのように格納するか、という点については詳細な考察が必要である。この点については、次節で述べる。

2.4 検索手順と検索クエリー

データベースの内容が定まった後は、そのデータベースからの検索の方法に付いて定める必要がある。検索には目的によって数種類の利用法が考えられるので、各々について詳しく後述する。

3. 表情報データベースに格納する情報

ページのデータはページ取得ワーカーによりなんらかの基準で収集され、解析される。ページの解析では、まず、そのページに関する基礎的な情報が解析された後、ページ内に含まれる表の抽出が行われる。その結果、そのページに、どのような表がふくまれており、それぞれの表にはどのような単語が含まれていたかが解析きる。

本システムでは、Web ページ、表、単語を扱う際、各々に固有の ID を割り振って識別し、取り扱う。以降では、ページ、表、単語のそれぞれの情報を扱うインデックスについて個別に述べる。

3.1 ページ情報

Web ページに関しては、ページ ID からそのページの情報を検索するための” ページ情報インデックス” と、URL からページ ID を検索するための” 逆ページ情報インデックス” がある。“ページ情報インデックス” には、そのページの URL やタイトルなどと共に、そのページに含まれていた表の表 ID のリストが格納されている。“逆ページ情報インデックス” は、検索の利便のためのインデックスであり、ページ URL によってインデックス付けがされ、URL からページ ID を検索することができる。

3.2 表情報

取得した Web ページは HTML 木構造を解析され、中に含まれている表が抽出される。抽出された表は、それぞれに表 ID がふられ、表 ID はページ情報インデックスに格納される。また、表中の各セルに現れた単語で、未知の単語は単語 ID が割り振られる。

表情報に関してもページ情報と同様に、” 表情報インデックス” と、“逆表情報インデックス” がある。“表情報インデックス” は、表 ID をキーとして、その表の大きさやページ内での

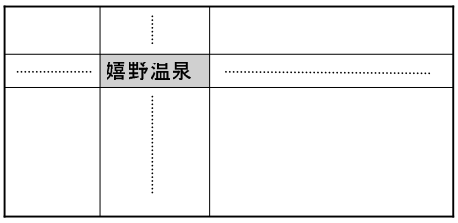

図 3: 単純単語検索

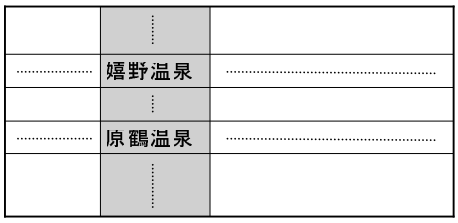

図 4: 同型統語検索

存在位置、各セルの中の文字列を表す単語 ID などの情報が格納されている。一方、“逆表情報インデックス” では、単語 ID をキーとして、その単語が含まれる表の場所を表す「表 ID、行位置、列位置」の組のリストが格納されている。

3.3 単語情報

ページ解析の際に見付けられた未知の単語は、単語 ID が割り振られる。実際の文字列と単語 ID の対応は” 単語 ID インデックス” と” 逆単語 ID インデックス” に格納される。

“単語 ID インデックス” には、キーに単語 ID が、内容として実際の文字列が格納される。“逆単語 ID インデックス” は、任意の文字列から、それに対応する単語 ID を検索するための物である。文字列検索の場合、部分文字列にマッチを取る場合と、厳密に全文一致で取る場合の 2 種類がある。そのため、実装ではハッシュ関数による高速な厳密一致検索用インデックスと、bigram による部分文字列検索用インデックスの 2 種類を用意した。

4. 細粒検索システムを用いた各種検索

利用者がどのような目的をもって、このシステムを利用するか、という事に関する考察も必要である。

前節で述べたように、表情報データベースにはいろいろな観点からのインデックス付けがおこなわれているので、表の情報に関するいろいろな検索を行うことが出来る。

基本的な検索の手順としては、まず、検索質問中の単語から単語 ID を逆単語 ID インデックスを用いて検索し、その単語 ID を逆表情報インデックスで検索することで、その単語の出現位置を調べることが出来る。調べた位置情報を元にどのような処理を行うか、結果としてどのような物を受け取るか、といったことは何通りか利用法が考えられるので、以降ではその利用方法ごとに述べる。

4.1 単純単語検索

最もシンプルな検索質問としては例えば次のようなものがある。

> 表の中に「嬉野温泉」という語を含んでいる表は?

この検索質問で得られる表は、例えば図 3 のようなものである。この形式での検索を行った利用者の意図は正確には不明だが、おそらく「嬉野温泉」に関係したなんらかの情報 (例えば、嬉野温泉の所在地や連絡先など、または、嬉野温泉となんらかの共通点がある温泉名など) を検索しようとしていると考えられる。

この場合、前述のようにして単語の現れる表を特定し、その表の中身を返してやればよい。

4.2 同系統語検索

単純単語検索ではその意図が不明であるが故に、多彩で雑多な表が結果として出力され、利用者の意図した物を返す事は難しい。

そこで、もう少し利用者の意図が表現された形での検索質問として、同型統語検索が考えられる。この同型統語検索の例を挙げると、例えば、次のようになる。

> 同じ表の同じ列に「嬉野温泉」と「原鶴温泉」が含まれているような表は? また、その列に含まれている他の単語は?

この検索質問によって得られるのは図 4 のような表 (もしくは、表の色を付けた列) である。この形式の検索質問を用いることで、利用者は、指定した複数の単語と同系統の単語のリストを得ることができる。また、検索によって得られた表は指定した単語が並べられており、それらの単語の差異を比較した表であることも多い。

この検索を行うには、まず、前述の方法で各々の単語の出現位置を特定した後、それぞれの単語が同じ表で同じ列に出現しているかどうかを調べ、目的の箇所が見付かった場合は、その列か、その表全体を返せばよい。単語の出現位置のリストが予め表 ID と列番号でソートされていた場合、同じ列に出現している箇所を探す走査はリスト長の線形時間で行うことが出来る。

4.3 レコード検索

さらに複雑な検索質問も考えられる。たとえば、「嬉野温泉」の所在地が知りたい場合、予め所在地の判っている温泉名を使うことで、次のような検索質問をつくることができる。

> 「原鶴温泉」と「福岡」が同行に、「原鶴温泉」と「嬉野温泉」が同列に現れる表において、「嬉野温泉」のある行の、「福岡」あるの列に書かれた単語は?

文章にすると複雑だが、意図している表は図 5 のような表である。直感的な表現を用いれば、次のように言うことも出来る。

> 「原鶴温泉」が「福岡」なら、「嬉野温泉」は何?

この場合、利用者の意図がかなり明確にあらわれ、表に要求される条件もかなり厳しいものになっているため、この結果得られた単語は利用者の期待どおりの単語となる可能性はかなり高い。

検索の実装としては、先に同系統語検索によって対象となる表を絞り、「福岡」の単語位置情報を用いて不適合の表の振り

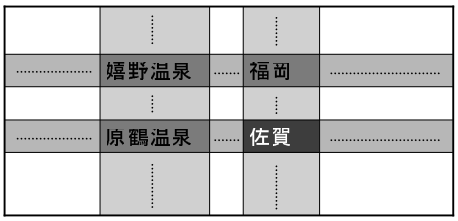

図 5: レコード検索

落しや “県名” の列の特定を行う。この場合も、単語位置情報のリストが予めソートされて格納されていた場合は線形時間で走査ができる。最終的に 1 つの表だけが残れば、該当箇所の単語を結果として返せばよいし、もし、複数の表がマッチした場合は、それぞれの表の該当箇所の単語を集計し、数の多い順で表示すればよい。

4.4 検索クエリーに関する考察

前述の 3 種類の検索は、条件の指定と出力の指定の仕方を規定することで、統一的な検索クエリーの書式をつくることが出来る。

例えば、次のようなルールで検索式を書くこととする。

- 検索単語はクォート (‘) でくくる
- 同列に並ぶべき要素はカンマ (,) で区切って並べる
- 同行に並ぶべき要素はコロン (:) で区切って並べる
- 同行に並ぶべき要素と同列に並ぶべき要素が両方とも在る場合は、コロンで繋がれたかたまりをカンマで区切って並べる
- 出力して欲しい要素に相当する部分はクエスチョンマーク (?) を書く
- 検索質問中に?マークが含まれなかった場合は、検索質問の条件を満たした表のリストを出力するものとする

この場合、前述の 3 つの検索の例は次のように表すことが出来る。

```
‘嬉野温泉’
‘嬉野温泉’ , ‘原鶴温泉’ , ?
‘嬉野温泉’:‘福岡’ , ‘原鶴温泉’:?
```

このように、統一的なクエリーを定めると、システムにとっては解析や検索が行いやすく、利用者にとってもある程度直感的に分かりやすくなる。

本システムでは、このような検索クエリーを表検索部の入力、つまり表検索 UI 部とのインターフェースとして採用する。表検索 UI 部では、この検索クエリーをユーザーから直接受け取ったり、このクエリーよりも分かりやすい別の形式 (例えば、$n \times n$ 個の格子状に並んだテキストボックスに検索したい語を書き込む、といった形式) のインターフェースをユーザーに公開し、内部で検索クエリーに変換して表検索部に渡すような処理をおこなったり出来る。必要ならば、もっと複雑で用途を限定した検索インターフェースも作成可能だろう。

5. まとめと今後の課題

本稿では、表に特化することで粒度の細かい情報を検索することができる細粒検索システムに対して、その構造や技術的な面での考察、および、従来の検索エンジンとは異なる特殊な検索方式についての考察を行った。

我々は今後、この細粒検索システムを実装し、さらに多彩で詳細な検索の考案、実装を行う予定である。

参考文献

[1] S.Brin, “Extracting patterns and relations from the world wide web”, WebDB Workshop at EDBT ’98, 1998

[2] S.Chakrabarti, M.van den Berg, B.Dom, “Focused Crawling: A New Approach to Topic-specific Web Resource Discovery”, WWW8 Conference, 1999.

[3] S.Chakrabarti, K.Punera, M.subramanyam , “Accelerated Focused Crawling through Outline Relevance Feedback”, Proc. WWW2002, 2002.

[4] M. Diligenti, F.M. Coetzee, S. Lawrence, C. L. Giles, M. Gori, “Focused Crawling using Context Graphs”, In Proceedings of the International Conference on Very Large Databases (VLDB-00), pp.527-534, 2000.

[5] 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.

[6] 野口正人, 廣川佐千男, “Web からの同系統単語知識獲得方法”, 情報学シンポジウム講演論文集,pp.21-24, 2003.

[7] 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`

[8] S.Sato, M.Sato, “Toward Automatic Generation of Web Directories.” Proc. of International Symposium on Digital Libraries 1999 (ISDL’99), pp127-134, 1999.

[9] 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.

[10] Yalin Wang, Jianying Hu, “A Machine Learning Based Approach for Table Detection on The Web”, The Eleventh International World Wide Web Conference, 2002.