単語の係受け構造を利用した WWW上での日本語テキスト検索システム

麻生 和昭, 峯 恒憲, 雨宮 真人
九州大学大学院システム情報科学研究科知能システム学専攻
E-mail:{aso,mine,amamiya}@al.is.kyushu-u.ac.jp

1 はじめに

近年インターネットなどの情報ネットワークの普及に伴い、個人がアクセスできる情報リソースが急激に肥大化してきており、これら膨大な情報の中から求める情報を効率良く得るための有効な情報検索の手段が求められている。

現在WWW上では計算機による情報検索サービスを行うサイトが幾つか存在する。これらのサイトでは一般に幾つかの単語をキーワードとして入力し、それらの単語を全て含むかもしくは一部を含む(AND/OR検索)ページのリストを出力する形式をとっている。

しかしこの方法ではある入力に対しての検索結果が数十件から数百件、ときには数千件に及ぶこともあり、検索結果を更に絞り込むための手段が求められている。

そこで本稿では「単語の係り受け構造を利用した検索手法」を提案し、この手法を適用したプロトタイプ版を例に挙げその有効性を示す。

2 既存の検索システムの問題点

現在WWW上で利用可能な日本語を扱う情報検索システムとしては、TITAN(NTT)・ODIN(東京大学)・千里眼(早稲田大学)等がある。

これらの動作は、多少の差はあるものの概略すると以下の通りである。

1. ユーザから検索のためのキーワードを入力してもらう。(TIATNでは文も入力できるが、その際は入力文から抽出された名詞のみがキーワードとして扱われる。)
2. そのキーワードを含むWWW上のページ(情報)を検索する。この際入力されたキーワードのいくつかを含むページのみを検索するAND検索、キーワードをどれか一つでも含むページを検索するOR検索などが可能である。
3. 何らかの評価式に従い検索結果を順位付け(スコアリング)する。
4. 検索結果の順位が高いものからユーザに提示する。

しかしこれらの手法では検索結果を充分に絞り込めず、その検索結果は一般に数十件、多いときには数百件から数千件に達することもある。そのため既存のシステムでは検索結果の順位付けを行い、その優先順位に従ってユーザに提示するという方法を用いているが、優先度を求める際に用いられる計算式はヒューリスティックなものであり、特に「こうすればよい」といった明確な判断基準は確立されていない。

3 単語間の係り受け関係を利用した検索手法

3.1 基本手順

本手法では検索の際ある単語が出現しているかだけでなく、それらの単語がどのように係っているかも検査する。以下にその手順を示す。

1. 検索文を形態素解析し、単語に分解する。
2. 検索文を係り受け解析し、全ての2単語間の係り受け関係を抽出する。
3. 検索対象文を同様に形態素・係り受け解析し、全ての2単語間の係り受け関係を抽出する。
4. 検索文の全ての単語を検索対象文が含んでいるか検査する。

5. 検索文の全ての2単語間の係り受け関係を検索対象文間が含んでいるか検査する。

例えばユーザが「メイルを読む」ことについて調べたいとする。

このときユーザが「メイルを読む」とそのまま入力すると、システムはその文を形態素・係り受け解析し、「メイル」が「読む」に係っていることを抽出する。

次に検索対象テキスト中から一文ずつ取り出しながら「メイル」と「読む」の両方の単語を含んでいるかを調べ、もし含んでいるなら「メイル」が「読む」に係っているかどうかを調べる。係っている場合にはその検索対象テキスト中の文を検索結果に加える。

3.2 単純なキーワードマッチングを用いた検索手法との比較

以下に「メイルを読む」について係り受けを考慮して検索を行った場合と、単語「メイル」と「読む」のAND検索を行った場合のそれぞれについて検索結果に挙げられる可能性のある文を示す。

テキスト	A	B
以下にメイルを読む方法について述べる。	○	○
受けとったメイルを読むには…	○	○
メイルを出した。しかし申込方法を読むと…	○	×
本を読んだので、メイルを書いています。	○	×

A…AND検索、B…係り受け検索

上記の表から、純粋に単語「メイル」と「読む」のAND検索を行った場合と係り受け関係を考慮して検索を行った場合を比べると、後者が「メイルを読む」という文にそぐわない検索結果を排除できていることがわかる。

このことによって、従来のキーワードマッチング主体の情報検索システムよりも検索結果を絞り込むことが可能であると予想できる。

4 プロトタイプシステムの試作

本検索システムは、九州大学六本松地区で開講されている情報処理入門教育を補助するために開発されたシステム [1] の一機能として、現在試験的に運用している。図1にシステムの概要を、図2,3に実行例を示す。

なお今回のプロトタイプ版では日本語の形態素・係り受け解析に (株) リコーで開発された簡易日本語解析系QJP[2] を用いた。QJPは約50KB程度の小規模な辞書しか必要としない、高速な日本語解析系である。

また今回は本手法の典型的な振舞いを調べるため、抽出する単語間の係り受け関係を最も基本的な「名詞と動詞間の係り受け関係」のみに限定した。ただし例外的に、「ソフトのインストール」のような「(名詞) の (サ変名詞)」という形式は「ソフトをインストールする」というように書き換えることができるため、その係り受け関係も抽出した。

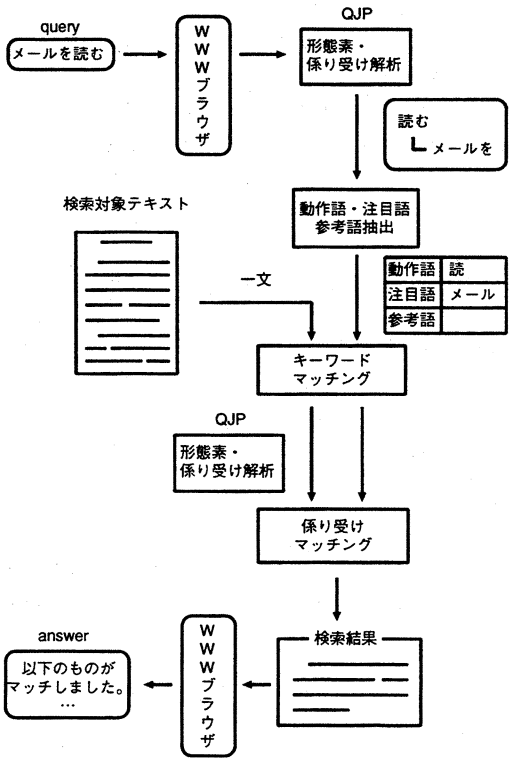

図1:概要図

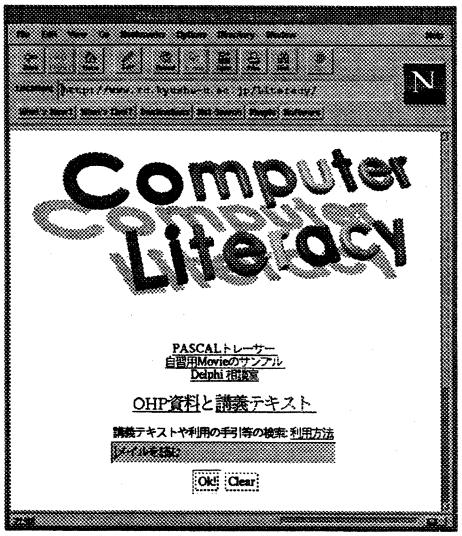

図 2:検索項目の入力

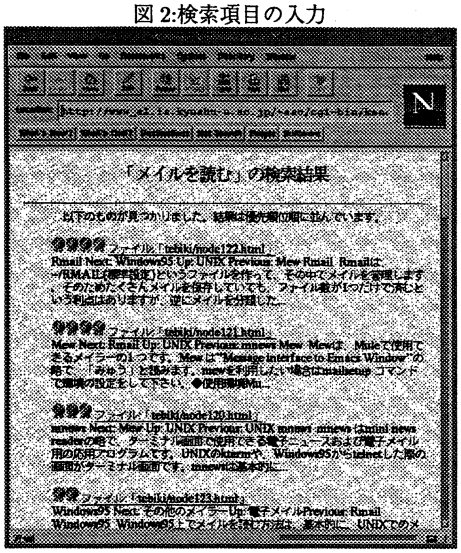

図 3:検索結果の表示

システムはまず検索項目として入力された自然言語の文を、QJP を用いて形態素・係り受け解析する。その結果入力された文から動詞とその動詞に係っている最も重要な名詞の組を抽出する。これらを「動作語」と「注目語」と呼ぶ。実際はもっと詳しく場合分けするのだが、今回のシステムでは簡単のために「注目語」を「動作語」に係る「が・は・を・の格の名詞」として定義し、その他の格の名詞はその文に対する「参考語」として定義し扱っている。次にそれら動作語と注目語の組を用いて検索対象テキスト中の対象文とのマッチングを計る。動作語・注目語が共に対象文中に含まれていれば、対象文を形態素・係り受け解析し、更に係り受け構造が等しいかを調べる。その結果注目語を含む対象文をそのマッチングの度合により以下の 5 つのグループに分類し、ユーザに提示する。

1. ◎： 注目語・動作語・係り受け構造がそれぞれ等しい。
2. ○： 注目語・動作語がそれぞれ部分マッチし、その係り受け構造が等しい。
3. △+: 注目語・動作語が検索対象文に含まれ、注目語が検索対象文中で「が・を・は・の格」である。
4. △： 注目語・動作語が検索対象文に含まれる。
5. ？： 注目語が検索対象文に含まれる。

例えば「メイルを読む」という文を「九州大学情報処理教育センター利用の手引 [3]」の中から検索した場合上記のグループにはそれぞれ以下の表記がマッチする。

◎： (A)メイルを読む
○： 受け取った電子メイルを良く読んでみましょうね。
△+: 電子メイルを送るには、ここを読んでください。
△： 電子ニュースを読むと、メイルでリプライするように書いていることがあります。
？： むやみに電子メイルを送りますと迷惑になりかねないので、注意して使用しましょう。

本システムでは本来は◎と○のグループのみが検索結果となるが、現在はその他 3 つのグループも「参考項目」としてユーザに提示している。

5 実験

本システムを単純なキーワードマッチングと比較するため、以下の実験を行なった。

1. 検索文を決める。
2. 検索文から名詞と動詞の語幹を取りだし、それらを用いてキーワードの AND 検索を行なう。この時の検索結果の件数を a とする。
3. 検索文を本システムで検索する。この時の検索結果の件数 (グループ◎と○のみ) を b とする。
4. $\frac{b}{a}$を計算し、キーワードの AND 検索と係り受け検索との検索結果の比を求める。

横軸に a・縦軸に$\frac{b}{a}$を表したグラフを図 4 に示す。

検索対象ファイルとしては情報処理の講義資料 [4] と情報処理教育センターの利用の手引 (合計約 1Mbytes) を用い、検索文としては実際に学生がシステムに入力した

文 (43 文) と検索対象ファイルから抜き出した単文 (107 文) の計 150 文を用いた。

これらの文は無作為に抽出したため、キーワードの AND 検索で検索した場合、得られる検索結果の数が同数の文が存在する。その際はそれら複数の文の係り受け検索の結果の平均件数を求め、$\frac{b}{a}$を求めた。

図 5 に AND 検索で検索した場合、得られた検索結果が同数であった文の個数を示す。

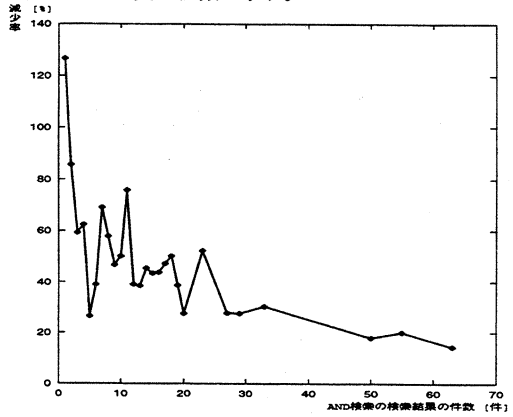

図 4:検索結果の減少率

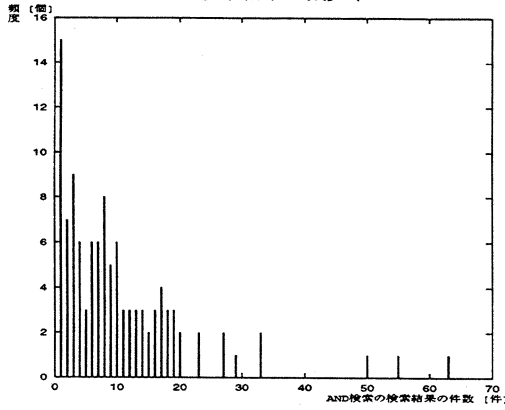

図 5:AND 検索で得られた検索結果の件数の度数分布

この結果、比較的データ数の多い「AND 検索の検索結果が 20 件以下」の場合でもほぼ 50 %程度、最も良い場合は 15 % (AND 検索時 63 件、係り受け検索時 9 件) まで検索結果が減少していることがわかる。

また AND 検索の検索結果が 1 件のデータに対しては減少率が 100 %を越えている。これは AND 検索では検索文から抽出した全ての単語を用いて検索するのに対して、本システムでは入力文から抽出した「動作語」と「注目語」のみを用いて検索しているためで、三語以上の単語を含む文の場合にはこのようなことが起こり得る。しかし「参考語」も検索結果を順位付けする際には用いてるため、検索結果の順位付け自体は単純な AND 検索の結果よりも悪くなることはない。

6 おわりに

本稿では、検索結果の絞り込みに単語間の係り受け情報を利用する方法を提案し、単純なキーワードマッチングに比べその優位性を示した。また、本手法を用いて情報検索システムのプロトタイプ版を作成し、その有効性を確認した。

今後は以下の事を行なう予定である。

- 抽出する係り受け関係の種類とその順位付けの決定
  (形容詞-名詞間・副詞-動詞間等の係り受け関係も順位付けに利用する)
- システムの高速化
  (検索対象テキストを予め日本語解析し、ハッシュ表に蓄える)
- WWW から自動的にデータを収集するモジュールの作成・リンク・実験
  (WWW 上のデータを対象に適切に実時間で検索できるか)

謝辞

本システムを作成するに当たりお世話になりました九州大学大学院システム情報科学研究科の Edutainment Project の皆様に感謝します。また QJP の利用を許可して頂きました (株) リコーと開発者の亀田雅之氏に感謝します。

参考文献

[1] Sachio Hirokawa, Tetsuhiro Miyahara, Tsunenori Mine, Takayoshi Shoudai, Masao Mori, Hiroyuki Sato, Ayumi Shinohara, Masayuki Takeda, "Teaching 2300 Students with WWW – Practice and Experience at Kyushu University", Proc. of ERI'96, pp59–63, 1996

[2] Masayuki KAMEDA, "A Portable and Quick Japanese Parser - QJP", Proc. of COLING'96, pp 616–621, 1996

[3] 九州大学情報処理教育センター 編," 九州大学情報処理教育センター利用の手引 (1996 年版)", 九州大学出版会,1996

[4] 廣川 佐千男, 宮原 哲浩, 峯 恒憲, 正代 隆義, "12 回で学ぶ情報処理", 九州大学生協, 1996