日本語の表現特性を利用したテキスト情報検索システムの検討

1U－4

杉崎 正之　森 大二郎　大久保 雅且　田中 一男

NTTヒューマンインタフェース研究所

1 はじめに

インターネット上ではWWWのページを対象にしたさまざまな全文検索サービスが行われている。しかし、多くのサービスにおいて、検索結果を利用者の用途に応じて絞り込むことが困難になっている。この原因の一つとして、検索結果を利用者の目的に合わせて提供できていないことにあると考え、サービス提供側で検索の目的を入力する手段を用意し、それに合わせて検索結果を提示するシステムを検討している。

本報告では、助詞の役割を利用し、利用者の目的に応じて検索結果自体やその提示方法を変更する手法を検討した内容について述べる。

2 現在の検索サービス

インターネット上ではWebを検索対象としたgoo (http://www.goo.ne.jp) やAltavista (http://www. altavista. com) や、各新聞社が自前で発信している新聞記事の検索などが提供されている。

新聞記事の検索などは、検索サービス利用者がある程度明確な目的を持って利用することができる。

しかし、分野を問わずさまざまな情報が自由な表現で書かれたWebを検索対象としたサービスの場合、その利用者が欲しい情報を引き出すのは困難である。今回、その理由として、利用者が獲得したいと思っている目的に応じて、検索システムが結果の出力を行っていないことにあると考えた。すなわち、Webの検索システムは、雑多なWebの中から様々な情報を引き出せるがために目的を絞った検索ができていない。

この問題の一解決策として、あらかじめ検索対象を分類しておく方法も考えられる。しかし、この場合、分類結果あるいは分類体系を維持していくのが困難であり、また、検索したい対象と分類体系が一致するかどうかの問題も残る。

雑多なWebを対象にして、検索単語と利用目的を入力するとそれに合った情報を検索結果の中でより上位に提供することが重要であると考えた。

A study on information retrieval system using japanese expression characteristics
Masayuki SUGIZAKI, Daijiro MORI, Masaaki OHKUBO, and Kazuo TANAKA
NTT Human Interface Laboratories

3 検索目的と検索結果

次に検索の目的と検索結果の関係について考える。Webの検索システムの一般的な利用目的は、(i) ファイルの在る場所を得る、(ii) 製品カタログを調べ検索結果を比較する、(iii) 単語の意味を調べる、など多岐に渡って存在する。(i)(ii)(iii) とも、単語のマッチングによる検索アルゴリズムだけでは的確に獲得できない。なぜなら、単語が存在している場所に書かれている情報が、必ずしも目的と合った情報であると限らないからである。また、検索結果の順序付けも、TF/IDF[1] のような単語の出現頻度やファイルのサイズ情報を利用しているために、それぞれの目的に合っていない。

前記 (i) のとき、ファイルの拡張子やリンク情報から情報を収集することで検索対象を絞り込むことができるし、その機能を提供する検索システムも存在する。(ii) のような価格などの定型的な情報は、保存データのメンテナンスの必要性もあり一般にはデータベースなどがより使いやすい。(iii) のような目的のシステムは、辞書の検索などがある。しかし、辞書の更新は非常に大変であり、日々増加する新しい単語や言葉を追従し続けることは困難である。それに比べ、Webページの集合は常に内容が更新されており、なんらかの結果が得られることが期待できる。特に、辞書に掲載されていない最新の単語の情報が欲しい場合に、Webページから情報を得ようとすることは非常に有効な手段である。

(iii) のような情報を、それが埋まっているWebページからあらかじめ抽出できれば、そこから絞り込みや順序付けする情報を得ることができる。しかし、検索対象となるWebデータは統一したフォーマットで情報を記述されておらず、また、各ページで共通に使われているHTMLタグも利用方法が自由なため、タグ情報を頼りに有効な情報を抽出することは容易ではない。そこで、我々は、タグ情報などではなく日本語文の中から順序付けに利用できる情報を探した。

4 システムの検討

4.1 検索システムの概要

日本語の文章に注目すると、単語に関して詳しい説明をする文章も存在するが、端的な説明を行う場合、助

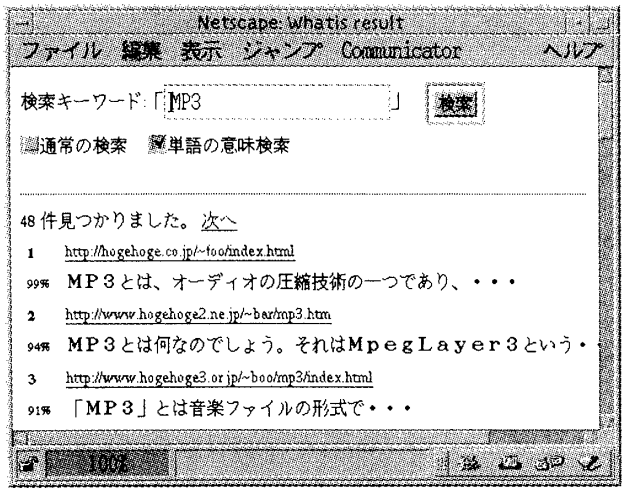

図 1: 検索システムのイメージ

詞の「とは」や「って」などを用いた文が使われる場合がある。これらの助詞は「命題、定義の主語を表す」役割を果たしており [2]、「A とは B である」という文に対し、A という単語 (あるいは文) の説明に B という単語 (あるいは文) が利用されている。この助詞を含む文内に存在する単語間の関係は、他の文内に存在する単語間の関係に比べ、意味的な関わりが強いのではないかと考えた。そこで、共通の役割の助詞に注目し「A とは B である」や「A って B です」などの A と B に存在する単語の共起関係を抽出し、A の説明によく使われる単語のグループを生成する。検索入力時に単語に加えてチェックボックスなどで「単語の説明が欲しい」という指示があった場合に、検索単語のグループの単語を利用して検索結果の順序付けを変更することにした。図 1 がその検索システムのイメージである。

今回、助詞の「とは」を含む文を抽出し、その特徴を調べた。

4.2 助詞を含む文の抽出

単語の共起関係を調べるべく、Web ページを対象に「とは」を含む文の抽出を行う。その際、「ちょっとは」「あとは」などの表記上の「とは」を含む文は共起関係の計算に不要であり、このような文は形態素解析処理を行って抽出しないようにした。形態素解析エンジンには InfoBee[3] を用いた。抽出部分は、HTML タグを除いた「とは」を含む文とそれ以降 2 文までとした。抽出する文の数を複数としたのは、「A とは何でしょう？」のように A の定義を問いかけるだけの文もあるためであり、逆に多くの文を選ぶことを避けたのは、説明の

	抽出処理前	抽出処理後
URL 数 (文の数)	約 226 万 URL	約 24 万 URL (約 49 万)
データ量	約 13GB	約 102MB

表 1: 「とは」を含む文の分析結果

文から話が発展している可能性があり、A と B の共起関係の抽出には不向きであると考えたからである。文の抽出処理を約 226 万 URL を対象に行い、その結果を表 1 に示す。

4.3 考察

「とは」を含む文の抽出結果を分析すると、「A とは関係なく」「A とは違って」など助詞の役割が比較対象を指すのに使われている文が存在する。また、Web の特徴であるリンクによる情報の表現により「A とは」に続く説明が別のページに書かれていたり、改行によって複数の文に分断されている場合があった。前者は、助詞の役割が異なることから、共起関係の抽出には省く必要がある。現在、「とは」を含む文の特徴的な「言い回し」を分析し辞書化することで削除する方法を検討している。後者に関しては Web 上の表現の自由さからくる問題であり、こちらは現在検討中である。

5 今後の課題

今回、単語の共起関係を抽出するために、助詞を含む文の抽出を行った。今後は、単語のグルーピング方法と、検索結果の順序付けへの利用方法を検討し、評価を行いたい。また、ここで獲得した共起関係は Web から獲得できる最新の単語を含んでいる可能性があり、Web の検索結果出力時の順序付けだけではなく、共起関係から計算される語のグループを用いた Web の自動分類などにも有効に活用したい。

参考文献

[1] G. Salton: Automatic Text Processing, Addison Wesley, 1989

[2] 現代語の助詞・助動詞, 国立国語研究所, 1951

[3] 井上, 大久保, 杉崎: InfoBee テキスト情報検索技術, NTT R&D 10 月号, pp.1103-1108, 1997