4U-5

WEB 検索におけるキーワード関連語提案システムの検索性能精錬とその応用について

梅永 明寛
九州大学大学院システム情報科学府*

竹下 日出男, 久本 学
九州大学工学部電気情報工学科†

長谷川 隆三, 藤田 博, 越村 三幸
九州大学大学院システム情報科学研究院‡

1 はじめに

近年，事物を調べる際にインターネットを利用することは一般的であり，ユーザは主にサーチエンジンを用いて情報を得ているが，現在世の中に存在している WEB ページはすべてを把握できないほど多くなっており，検索を行っても大量の検索結果を目の前にして，途方にくれることもしばしばある．この問題を解決するために，ユーザーが調べたいキーワードに関連する語（関連語:Relevant word）を提案するシステム [1] が作成されている．このシステムは，ページを HTML のタグ構造と単語の生起頻度に着目して解析し，ユーザが調べたい事柄に関連する語をリアルタイムに提案する．現システムでは確かに関連語は提案するが，形態素解析ソフトの性質も関わって，キーワードに関係のない語や抽象的すぎる語まで拾ってしまう傾向がある．

そこで本研究では，無視語辞書を作成し，これに含まれる単語を関連語群から削除することによって関連語の精錬を試みた．また，検索結果ページの内容を要約して表示する機能も実装したので，報告する．

2 キーワード関連語提案システム

キーワード関連語システムは，ユーザーが与えたキーワードを検索エンジンに投げ，検索エンジンが所得したページを多角的に解析することで関連語を提案することを主目的としており，関連語の評価法は TF/IDF 法と共起確率を用いている．

TF/IDF 法とは文書に含まれる単語の重み付けをする手法である．TF(Term Frequency) とは，ある文書 d における単語 t の生起頻度であり, IDF(Inverse Document Frequency) は「文書中に頻発する語の評価を下げる」というものであり，単語の評価値として，これらの積を用いるのが TF/IDF 法である．TF/IDF 法は単に文書中に現れる単語の数に注目した重み付けをする手法であり，一般的によく使われる手法である．

次に共起確率であるが，ユーザが与えたキーワードに対して得られた HTML 文書内でキーワードの近傍に現れる単語は，キーワードに関連している単語であると考えられる．そのような単語のことを共起語と呼ぶことにする．キーワード群 k に対する検索結果のページ群 $P(k)$ における k の共起語 t の共起頻度を $co(t, P(k))$ とする．共起語 t の生起頻度 $tf(t, P(k))$ を用いて，共起語 t の共起確率 (Co-occur Probability)$cp(t, P(k))$ を定義する．

$$cp(t, P(k)) = \frac{co(t, P(k))}{tf(t, P(k))} \tag{1}$$

次に単語の分類を行うのだが，各単語を各単語のタグ付けのされ方を表すタグベクトルと，ページにおける単語の生起の表すページベクトルという 2 つのベクトルで表現し，ベクトル空間モデルを利用して分類する．ベクトル空間モデルとは，ある情報を多次元空間上のベクトルとして表現し，2 つのベクトルを比較することにより類似度を調べるものである．このモデルでは，2 つのベクトルが同じ方向を指すときに類似度が高いとする．

3 検索性能の精錬とその応用について

3.1 無視語の登録

形態素解析ソフトの性質もあいまって，関連語に相応しくない語を抽出してしまうことが頻繁に起こる．例えば，WEB ページに記載されたプログラムソースの欠片，WEB ページに記載されていた URL の欠片，単独では検索語として意味をなさない語，検索結果順位の高いページに現れたであろう検索語との関連が希薄と断定される固有名詞などである．

検索語として無意味な語がユーザーに提起されることは，ユーザーにとって情報の取捨選択をする手間を増やすことになる．加えて，これらの語がランクインした陰で，関連語として重要な語がランク外に弾き出されている可能性も高い．そこで，これらの単語を無視語として登録し，形態素解析を行う際に無視語は関連語から削除する．無視語を選定するに当たっては，人気検索語ランキング 1 年分より，各月ごとに 200 件ほど無作為に選んだキーワード群 2400 件からランダムに選んだ 50 ワードについて検索を行い，その結果得られた関連語群から追加検索語として不適切と思われる単語を著者の主観で選

* **Akihiro Umenaga**, Graduate School of Information Science and Electrical Engineering,Kyushu University
† **Hideo Takeshita, Manabu Hisamoto**, Department of Electrical Engineering and Computer Science,Kyushu University
‡ **Ryuzo Hasegawa, Hiroshi Fujita, Miyuki Koshimura**, Graduate School of Information Science and Electrical Engineering,Kyushu University

表 1　共起行列 (検索語:トヨタ, ホンダ)

rank	関連語	査定	パッケージ	ダイハツ	スズキ	新車
1	査定	0	2	2	2	4
2	パッケージ	2	0	2	2	2
3	ダイハツ	2	2	0	3	3
4	スズキ	2	2	3	0	4
5	新車	4	2	3	4	0

表 2　確率分布

関連語	査定	パッケージ	ダイハツ	スズキ	新車
出現回数	38	28	35	39	56
出現確率	6.21%	4.58%	5.72%	6.37%	9.15%

ぶことで無視語を選定している.

3.2　観点要約文の作成

次に，関連語の評価値と，単語間の共起の関係に着目し，観点要約文を作成する技術について述べる．適切なキーワードを与えることで検索結果ページ数を減らしても，やはり全部のページ見て回ることは大変時間がかかる．そこで文章の内容を反映する要約文があれば，文章全体を読み通すことなく内容の大枠を認識することができ，情報の取捨にかかる時間の短縮を図ることができる．評価値の高い関連語 20 件に対応する列だけを抜き出し，20×20 行列とし，対角成分に関しては 0 とする．**表 1** は「トヨタ」と「ホンダ」をキーワードとして AND 検索をかけた場合 (05.1.6 現在) の共起行列を 5×5 に縮小したものである．ここでは「ダイハツ」と「スズキ」と同文書中で 3 回同時に出現していることを意味している．また，**表 2** は関連語の上位 10 語の出現頻度と出現確率 (全体が 1 になるように正規化したもの) の小数点第 3 位以下を四捨五入したものの上位 5 語分を抜粋したものを表している．本手法では関連語の評価値とこの出現確率を評価値として用いる.

3.3　要約文抽出の実験例

要約文抽出システムをキーワード「トヨタ」「ホンダ」の検索結果に対して適応した結果を**表 3** に示す.

検索結果 1 位となったページは『トヨタとホンダ』(光文社新書　塚本 潔 著) という本を紹介している通信販売サイトである．この本の論旨はまさしく要約文の通りであり，商品の内容を的確に述べた文章を抽出している.

次に検索結果 2 位のページだが，『週刊朝日』2002 年 7 月 19 日号の「ベール脱いだ国産小型飛行機開発――トヨタ対ホンダ空中戦」という記事についての紹介をしているロータークラフトファンのページである．要約文として抽出された文は，「この両社が何の目的で何をしているのか，本当のところはよく分からない．そこで，いくつかの英米誌を探してみた．その伝えるところを総合すると以下のようになる.」という文の直後に登場する，英米誌をページ作者が要約した文であった.

検索結果 4 位のページは車の最新ニュースに特化したページの中の，「リースするならトヨタ・ホンダ 」(2004.2.25) なる記事で，リース終了時の車両の市場評価と，月々のリース料を比較した結果，トヨタ/レクサスを合わせて 12 のモデルが最高の 5 つ星評価，またホンダも「業界で最もリースの残余評価が高いメーカー」に 5 年連続で選ばれたことを報告しているニュース記事である．その記事の中の「最もリースがおトクなモデル」について述べた箇所を抜き出しているので，これも的確であると言ってよいだろう.

表 3　要約文抽出結果 (keyword:トヨタ, ホンダ)

rank1:http://www.amazon.co.jp/exec/obidos/ASIN/43340311611
トヨタ式とホンダ流どこが違うか― “ 大衆・規模・道具のトヨタ ” vs “ 個性・効率・趣味のホンダ 二〇〇一年九月の中間連結決算で，過去最高益を達成したトヨタとホンダ．ドコモ，ソニーの’ 失速’ という状況の下，日本を真の意味で牽引する企業は，もはやこの二社をおいて存在しない. 本書はトヨタ・張，ホンダ・吉野の両社長のインタビューをはじめ，製造，販売の現場，そして米国，欧州の現地工場への徹底した取材を通して，両社の強さの秘密，知られざる苦悩，そしてライバルに対する思いなどを浮き彫りにする.
rank2:http://helicopt.hp.infoseek.co.jp/honda.html
トヨタ自動車の試験機は「トヨタ・アドバンスド・エアクラフト」(TAA) と呼ばれる概念実証のための実験機である．初飛行は去る 5 月 31 日カリフォルニア州モハービ飛行場でおこなわれた．これをトヨタは単なる「フィジビリティ・スタディ」のためとしている.
rank4:http://response.jp/issue/2004/0225/article58100_1.html
各セグメントでの「最もリースがおトクなモデル」は，ミッドサイズカーではホンダ『アコード』，ミニバンでは『オデッセイ』，小型 SUV では『CRV』，クロスオーバーでは『パイロット』，小型トラックではトヨタ『タコマ』，大型 SUV では『セコイア』，フルサイズトラックでは『タンドラ』，ミッドサイズ SUV では『4 ランナー』，準ラグジュアリーではレクサス『ES330』，とトヨタとホンダが独占.

4　今後の展望

本稿では WEB 検索によって得られたページを解析して関連語を導き出し，それを利用して各所得ページの要約文の抽出する技術を紹介した.

最後に今後の展望であるが，現在は要約文の抽出をするのを単純に関連語の出現頻度で行っているが，これに語と語のつながりを考慮したものを搭載したいと考えている [2].

謝辞

本研究の一部は，日本学術振興会科学研究費補助金・基盤研究 (A)(2)(課題番号：15200002) の補助を受けた.

参考文献

[1] 大石 哲也, 長谷川 隆三, 藤田 博, 越村 三幸:“WEB 検索におけるキーワード関連語提案システム”, システム情報科学紀要第九巻第一号,19-24,2004

[2] 松尾 豊, 石塚 満：“語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム”, 人工知能学会論文誌 17 巻 3 号 D,2002.

[3] 砂山 渡, 谷内田 正彦:“観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装”, 人工知能学会論文誌 17 巻 1 号 B,2001.