2P-01　WWW検索ログを用いた次検索候補単語の提示方式の検討

杉崎 正之　　牧野 俊朗　　田中 一男 *
NTTサイバーソリューション研究所

1　はじめに

インターネットなどに代表されるコンピュータネットワークの普及により、テキスト情報のやりとりが頻繁に行なわれている。その代表格がHTMLファイルであり、複数のコンピュータ上に分散し存在する多量のHTMLファイルの中から欲しい情報を取り出すために、それらを収集し検索できるようにするサービスが提供されている。しかし、そのような検索サービスでは一回の利用で欲しい情報を得ることは難しく、何度も入力単語を変えて利用するのが現状である。そこで、検索入力された検索語に対して再検索を支援するための次検索候補単語の抽出および提示方法について検討を行った。

2　検索入力の分析

単語入力による検索サービスにおいて、利用者が欲しい情報を得ることは一般的には容易ではない。特に、分野を問わず様々な情報が自由な表現で書かれたWebを検索対象とした場合はさらに困難になる。その原因の一つとして、入力単語が探したい情報を見つけるための適切な単語でない場合があり、そのため、利用者側は単語を変えたり追加して入力し検索結果の適合度をあげるという手段を取っている。そこで、検索サービス側がそれを支援することが切望されている。

次検索に利用される候補となる単語 (次検索候補単語) を得るのに以下の3種類が考えられる。それは (1) 検索対象とまったく別に独自に作成したもの (2) 検索対象の文書内から抽出したもの (3) 検索サービス利用者の実際の入力から抽出したものである。(1) は、検索対象が大きく変化しない場合は表記の揺れを吸収するような辞書を用意しておく価値はあるが、Webページのように検索対象が日々変化する場合には不向きである。(2) は、検索対象内に存在する単語の文内共起情報などから提示する手法であるが、Webページが検索対象の場合、大量である上に、かつ、雑多な情報が含まれており適切な次検索候補を抽出するのは難しい。(3) は、人間が検索するために入力した単語であり、利用者にとっては理解しやすい検索語が埋まっている可能性が高い。そこで、今回は次検索候補単語の獲得先として検索入力を用いることにした。

検索時に欲しい情報が入手出来なかった場合の利用者の行動パターンとして (i) 入力単語の変更と (ii) 単語の追加がある。(i) は「プロバイダ」「プロバイダー」などの表記の揺れや「オートバイ」「バイク」といった表現の違いによる再検索であり (ii) は「ウィルス + パソコン」といった詳細な情報に絞り込むような検索である。次検索候補として単語を提示する場合も、単語の羅列ではその単語が絞り込みなのか置換なのかが理解しにくくなる。そこで、使用用途に応じて明示的に分けて提示するための抽出方法を検討した。

3　単語間の関連度

使用用途に応じて単語を提示するために、単語間の関連度としていくつかの関数を定義する。

検索入力された時間情報を利用した単語間の関連度を用いる [1]。単語間の関連度は、利用者 i の検索語 x, y の使用時間差の最小値を tm^i_{xy} として、単語 x,y の間隔関連度 T_{xy} を

$$\begin{aligned} T_{xy} &= \sum assoc(tm^i_{xy}) \\ assoc(tm^i_{xy}) &= a && (tm^i_{xy} = 0) \\ &= 1 && (0 < tm^i_{xy} \leq t_1) \\ &= \frac{t_2 - tm^i_{xy}}{t_2 - t_1} && (t_1 < tm^i_{xy} \leq t_2) \\ &= 0 && (t_2 < tm^i_{xy}) \end{aligned}$$

とした。なお a, t_1, t_2 は定数であり、その値はそれぞれ $2, 60, 300$ とした。これは、同一人物が短時間に入力した単語間の関連度が大きくなるような関数である。

さらに、間隔関連度 T_{xy} の値を用いて、単語 x の特徴ベクトル W_x を

$$W_x = (T_{x1}, \cdots, T_{xj}, \cdots, T_{xn}) \quad (1)$$

とし、特徴ベクトルを用いた単語間の距離 Dis_{xy} を、三角関数の $cos\theta$ で定義する。

A method of suggesting terms for query refinement using query analysis
Masayuki SUGIZAKI, Toshiro MAKINO, and Kazuo TANAKA
NTT Cyber Solutions Laboratories
* 現在、株式会社NTTデータ

間隔関連度のみ	$\cos\theta$ を用いた場合
1. mp3	1. プレイヤー
2. cd	2. プレーヤ
3. dvd	3. mp#
4. ソフト	4. mp3プレーヤー
5. ダウンロード	5. j-pop

図 1: 「プレーヤー」の関連単語の抽出結果

間隔関連度のみ	$\cos\theta$ を用いた場合
1. 設定	1. freebsd
2. インストール	2. solaris
3. vine	3. fvwm95
4. redhat	4. wu-ftpd
5. turbo	5. majordomo

図 2: 「linux」の関連単語の抽出結果

$cos\theta$ によって関連度を定義した場合、当然同じような分布の特徴ベクトルを有する単語間の関連度が高くなると考えられる。言い換えると、2つの単語 x, y(例えば「オートバイ」「バイク」) と一緒に入力される単語が同じ (例えば「中古」「パーツ」「販売」) であれば、x, y の $cos\theta$ の値は大きくなることが予想される。これにより前記 (i) にあるような表記の揺れや表現の違いによる単語を抽出できると考えた。

4 実験と考察

WWW 検索で利用された検索ログを用いた。期間は 2000/05/08～2000/05/14 までの 1 週間分である。

図 1, 2, 3 は、間隔関連度のみを用いた場合の関連単語と、$cos\theta$ を用いた場合の関連単語の抽出結果を関連度の値で降順に並べた例である。それぞれ「プレーヤー」「linux」「本」に関して抽出した結果である。

図 1 では「プレイヤー」「プレーヤ」といった表記の揺れによる単語が上位を占めており、$cos\theta$ を用いた効果が確認できた。さらに図 2 ではいわゆる UNIX 系の OS やアプリケーションの名前が上位を占めている。間隔関連度のみによる関連語 (「インストール」「設定」) は、$cos\theta$ で関連のある単語 (「freebsd」「solaris」) のいづれとも同じように入力されるような単語であった。いわゆる表記の揺れとは異なるが「UNIX 系 OS」という意味で互いに置き換えて利用されるであろう単語が上位になっていた。以上から、$cos\theta$ による関連語は前述の (i) のように検索入力単語を置換するための候補として利用可能であると思われる。

間隔関連度のみ	$\cos\theta$ を用いた場合
1. 検索	1. 書籍
2. 書籍	2. 洋書
3. セブンイレブン	3. isbn
4. 通販	4. 書籍販売
5. 通信販売	5. 六法

図 3: 「本」の関連単語の抽出結果

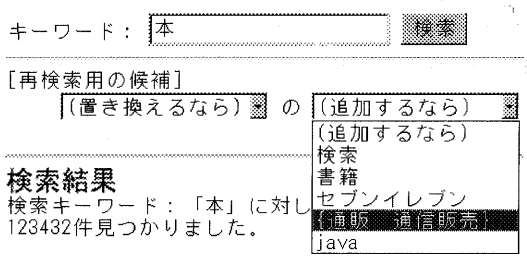

図 4: 次検索候補の提示方式の例

間隔関連度のみの場合、図 1 では「mp3」「cd」「dvd」などのプレーヤーや「ソフト」「ダウンロード」といった、より検索の目的を絞り込むような単語が上位を占めた。また、図 3 では本の「検索」や「通販」に関する単語が上位を占めており、これらもより詳細な情報に絞り込むような単語であった。

以上から、入力した単語を置換した次検索や詳細な情報に絞り込むための次検索のための候補を抽出する見通しはついた。提示方法は、(i)(ii) をそれぞれ明示的に分けて提示し、さらに図 3 のような「通販」「通信販売」など同じように利用される単語は $cos\theta$ などを利用して一つにまとめて提示するなどが考えられる (図 4 が提示方式の一例)。

5 今後の課題

提示すべき適切な単語数や複数の単語の検索入力に対する提示手法などを検討し、フィールドでの評価実験を行いたい。

参考文献

[1] 大久保, 杉崎, 井上, 田中: WWW 検索ログに基づく情報ニーズの抽出情報処理学会論文誌 Vol.39, No.7, pp.2250-2258, 1998