4W-02　WWW検索ログを用いた次検索候補単語の抽出方式の改良

杉崎 正之　　牧野 俊朗　　稲垣 博人
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。検索入力された単語に対して再検索を支援するための次検索候補単語抽出の改良方法について検討した。

2　従来手法

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く、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る。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おり、本研究の目的はシステム側で次検索に使う単語の候補を提示することである。

従来行なった抽出・提示方法[1]は、単語間の関連度を用いて検索ログから検索語の置き換え候補と追加候補を抽出し表示する方法である。検索ログに対し、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

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

とする。さらに単語xの特徴ベクトルW_xを

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

とし、特徴ベクトルを用いた単語間の距離を三角関数の$cos\theta$で定義し、この値が大きい単語を置き換え候補単語、それらを除くT_{xy}の値が大きい単語を追加候補単語とした。この手法での抽出例を図1, 図2に示す。

今回、置き換え候補に関し、置き換える理由を考えると(1)表記のゆれによる置換('リナックス', 'linux')、(2)粒度を換えて置換('debian', 'redhat')(3)別の単語と置換('solaris')があると考えられ、これらの単語を区別して抽出できないか更なる検討を行った。

3　改良方法

いくつかの方法を検討した。

[手法1] 間隔関連度T_{xy}の計算で用いている関数$assoc(tm^i_{xy})$は、同時に入力された$(tm^i_{xy} = 0)$単語に関して関連がある$(a > 0)$と評価する関数である。しかし、複数の検索語によるAND検索の場合、同時に入力された単語は置き換え単語として考えにくい。そこで、置き換え候補単語の抽出を関数$assoc(tm^i_{xy})$において$tm^i_{xy} = 0$のとき$a = 0$として単語間の関連度を計算し、その値が大きい単語を置き換え候補として抽出する。同様に追加候補単語の抽出は$tm^i_{xy} \neq 0$のとき$assoc(tm^i_{xy}) = 0$として単語を抽出する。

[手法2] たとえば複数の解釈ができる単語を検索語として入力した場合、それぞれの解釈にあった単語が混合して置き換え候補として抽出される場合がある。そこで抽出された単語集合をクラスター分析手法を用いてクラスタリングすることでそれぞれの解釈毎の単語集合の生成し、これによって上記(1)(2)(3)を区別を試みる。

置き換え候補となる単語の各特徴ベクトルの値は類似する可能性があり、適切に単語集合が分かれない。そこで、特徴ベクトルの要素を形成している単語でその他の単語の特徴ベクトルにおいて非ゼロとなる要素数を計算し(これをmとする)、その値を使って各特徴ベクトルの値を更新するようにした。すなわち、新しい特徴ベクトルの各要素の値をT'_{xy}とすると、

$$T'_{xy} = T_{xy} \times log(\frac{n}{m})$$

A method of suggesting terms for query refinement using WWW access log
Masayuki SUGIZAKI, Toshiro MAKINO, and Hirohito INAGAKI
NTT Cyber Solutions Laboratories

$\cos\theta$ を用いた場合	間隔関連度のみ
1. ロミオとジュリエット	1. 映画
2. バスター	2. matrix
3. mtx	3. スクリーンセーバー
4. レオン	4. 画像
5. チャーリーズ エンジェル	5. 壁紙
:	:

図 1: 従来手法による「マトリックス」の抽出結果

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

図 2: 従来手法による「linux」の抽出結果

となる (n は全単語数)。

4 実験と考察

WWW 検索で利用された検索ログを用いて抽出を行なった。期間は 2000/10/30～2000/11/26 の一カ月分である。間隔関連度の計算で使用した t_1, t_2 の値は、それぞれ 60, 300 とした。

手法 1 についての抽出例を図 3, 4 に示す。「マトリックス」に関する抽出例を見ると、従来手法での抽出ではいくつかの映画の他いとるが置き換え単語として抽出されているが、手法 1 ではマトリックスの表記のゆれと思われる単語が集中的に抽出されている。また「linux」に関する抽出例では、手法 1 のほうがより linux に関連する単語や表記のゆれとなる単語が上位を占めている。このように、手法 1 のほうが表記のゆれや入力に対してより関連があるような単語が上位を占めることが分かった。しかし、関連語として抽出される単語数を計算すると従来手法では平均 24.53 個に対し手法 1 では平均 2.34 個であり、手法 1 により置き換え単語として抽出される単語数は非常に少なかった。よって、多くの単語あるいはより広範囲の検索を行なうための単語を提示するということでは、従来手法と併用することが効果的であると思われる。

手法 2 について「マック」の置き換え候補単語をクラスタリングした結果の図が 5 である。「マック」に対し、表記のゆれと考えられる「macintosh」「imac」「マクドナルド」が置き換え候補として抽出されており、クラスター分析を行うことでパソコン関連の単語とそれ

$assoc(0)=0$ の場合	$assoc(t)=0(t>0)$ の場合
1. matrix	1. 映画
2. マトリクス	2. スクリーンセーバー
3. 映画	3. 画像
	4. 壁紙
	5. サングラス
	:

図 3: 手法 1 による「マトリックス」の抽出結果

$assoc(0)=0$ の場合	$assoc(t)=0(t>0)$ の場合
1. unix	1. インストール
2. turbolinux	2. 設定
3. redhat	3. turbo
4. リナックス	4. windows
5. vine	5. rpm
:	:

図 4: 手法 1 による「linux」の抽出結果

以外の単語をある程度グルーピングできていることが分かる。しかし、この手法がすべての置き換え候補に対して有効に働くわけではなく、また、図内の点線で囲んだ部分は主観的に付けたのだが、このようなグルーピングをどのようにして行うか更なる検討が必要である。

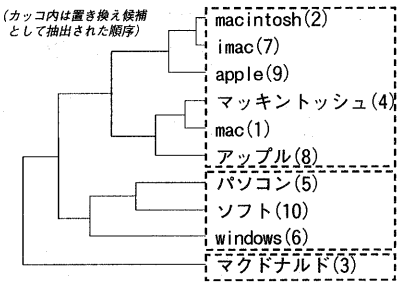

図 5: 手法 2 による「マック」の抽出結果

5 今後の課題

従来手法と手法 1 の組み合わせ方法とその評価、および、クラスタリング手法について検討を行なっていきたい。

参考文献

[1] 杉崎, 牧野, 田中: WWW 検索ログを用いた次検索候補単語の提示方式の検討, 情処第 61 回全大 (3), pp.113-114, 2000.10