1P-04

検索ログを利用した概念辞書の拡張

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

1　はじめに

インターネットの発達により、電子化された文書が多数存在するようになり、それらを検索、分類したいという要望が高まっている。文書を検索、分類する手段としては、文書中の各単語の出現頻度より求めたtf・idf値を利用する方法があるが、語の表記のみを扱うので、語の意味を扱うことができない点が問題である。語の意味を扱う方法としては、語の意味を属性ベクトルで表現した概念辞書を使う方法がある。概念辞書には、国語辞書の語義文を利用して作成した概念ベース[1]があるが、辞書の見出し語以外の語は定義できないため、固有名詞や新しい語に対応することが難しい。本稿では、インターネット上の検索サーバの検索ログと国語辞書より作成した既存の概念ベースを利用することにより、検索サーバに入力された検索語の属性ベクトルを算出し、概念ベースに容易に語を追加する方法について述べる。

2　概念ベース

概念ベースは、概念を意味特徴を表す属性aと、概念と属性の関連の深さを表す重みwの組(a,w)の集合で表現したものである。つまり、属性を次元とし、重みを値とすると概念は属性ベクトルで表現されることになる。辞書の見出し語を概念とし、その語の語義文中に出現する自立語を属性とし、語の出現頻度を用いて重みを定義するというのが基本的な考え方である。属性に関しては、シソーラスを利用して語義文中の自立語が属するカテゴリを属性とすることによって、表記が異なっても意味的に近い語を同じ属性と見なしている。また、重みに関しては、多くの概念に出現する属性の重みを下げて調整を行っている。概念ベースを用いると各概念の属性ベクトルの内積を計算することにより、概念の意味の近さを求めることができる。

概念ベースは辞書を元に作成されているので、WWWページの検索や分類に用いる場合、文書中に出現する語のなかで概念ベースに登録されていない語が多く存在するという問題がある。WWWページの検索の際に入力された単語を、検索サーバのログから抽出し、概念ベースに登録されているかを調べたところ、登録されている語は3割にも満たない。そこで、検索ログを用いて概念ベースに新たな概念を加える手法を提案する。

Expantion of the Concept Base utilizing a Retrieval Log
Toshiro MAKINO, Masayuki SUGIZAKI
and Kazuo TANAKA
NTT Cyber Solutions Laboratories
* 現在、株式会社ＮＴＴデータ

3　検索ログを利用した概念ベースの拡張

3.1　検索語間の関連度

インターネット上の検索エンジンを利用して、WWWページの検索を行う場合、適切なページが見つかるように、何度かキーワードを入れることがある。例えば、映画の「マトリックス」についての情報が欲しい場合、単に「マトリックス」と入力すると、数学の行列の意味のマトリックスのページも検索される。そこで、「マトリックス　映画」と入力する。また、Matrixのように異なる表記で書かれてるページが検索もれするので、「Matrix　映画」と入力することもある。ここで、同じ情報を求めて入力される一連の単語には、何らかの関連があると考えられる。検索ログを解析すると、入力された単語やその単語が入力された時間の情報が得られるので、これに基づいて検索エンジンに入力された単語間の関連度を求めることができる。検索語x、yの関連度T_{xy}を次の式で定義する[2]。ここで、t_iは、利用者iが検索語x、yを利用した時間の最小時間間隔である。

$$T_{xy} = \sum_i assoc(t_i)$$

また、関数$assoc$は、

$$assoc(t_i) = \begin{cases} 2 & (t_i = 0) \\ 1 & (0 < t_i \leq t_2) \\ (t_2 - t_i)/(t_2 - t_1) & (t_1 < t_i \leq t_2) \\ 0 & (t_2 < t_i) \end{cases}$$

ただし,$0 < t_1 < t_2$
と定義する。また、$t_1 = 60$秒、$t_2 = 300$秒とした。

3.2　関連度を利用した概念の定義

関連度の高い語同士は、意味的な関連も深いと仮定し、語tの属性ベクトル$\vec{a}_t$を語tの関連語iの属性ベクトル$\vec{a}_i$と関連度v_iを用いて以下のように定義する。

$$\vec{a}_t = \frac{\sum_i v_i \vec{a}_i}{|\sum_i v_i \vec{a}_i|} \quad (1)$$

n個の語の属性ベクトルを計算するためには、上記の式を各語毎に作り、n個の連立方程式を解くことになるが、概念ベースに既に存在する語の属性ベクトルを定数とし、他の語の属性ベクトルを変数にすると解析的に解けるとは限らない。そこで、概念ベースに既に存在する語に関しては、その属性ベクトルの値を初期値 $\vec{a}_{i,0}$ として、それ以外の語に関しては、初期値 $\vec{a}_{i,0}$ を $\vec{0}$ として、

$$\vec{a}_{t,k+1} = \frac{\sum_i v_i \vec{a}_{i,k}}{|\sum_i v_i \vec{a}_{i,k}|} \tag{2}$$

により、$\vec{a}_{t,1}$ を求めることにする。同様にして、$\vec{a}_{t,k}$ を用いて、$\vec{a}_{t,k+1}$ を求めることを繰り返し、$\vec{a}_{t,k}$ と $\vec{a}_{t,k+1}$ がほぼ等しくなった時、つまり以下の式で与えられる各語の属性ベクトルの差分の総和 D が十分小さくなった時点で、近似的に式(1)が成り立つと見なす。

$$D = \sum_t |\vec{a}_{t,k+1} - \vec{a}_{t,k}| \tag{3}$$

4 実験と考察

ある1日の検索ログを用いて、実験を行った。検索ログにおいて一定数以上の人が利用した検索語28647語のうち概念ベースにない単語22160語の属性ベクトルの計算を試みた。各語に関する関連語のうち、関連度の高いものから10語を利用して、式(2)により各語の属性ベクトルを計算した。その結果、14065語については、属性ベクトルが求められた。

4.1 属性ベクトル値の収束性

式(2)で、式(1)を近似できるかを調べるために、各語の属性ベクトルの差分の総和 D の k による変化を求めた。結果を図1に示す。図より k の増加に伴い D が小さくなること、すわなち属性ベクトルの値が収束していくことがわかる。よって、式(2)により、順次kの値を増やして計算することによって、式(1)を近似できると言える。

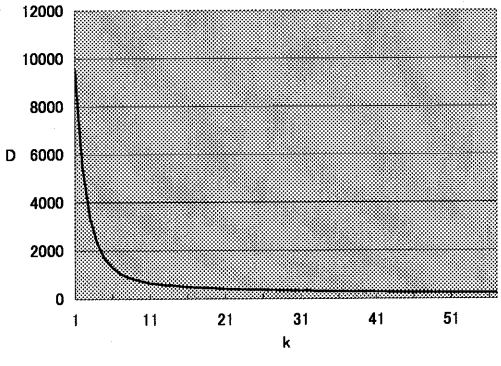

図1 属性ベクトルの差の総和の変化

4.2 語の類似性の評価

概念ベースは、属性ベクトルの内積の値を類似度とすることにより、語の類似性を判断することができる。属性ベクトルの値が十分に収束していると思われる $\vec{a}_{t,50}$ の値を用いて得られた類似語の例を示す。図2は「テレホーダイ」の、図3は「コンピュータ」の類似語とその類似度である。「テレホーダイ」については、「テレ放題」「テレホウダイ」などの別表記、「テレチョイス」「タイムプラス」などの類似サービスなどの語が得られている。類似度の値も高く適切な類似関係が得られていると考えられる。一方、「コンピュータ」に関しては、「テレホーダイ」の場合ほど類似度の高い語は存在しない。「コンピュータ」は今回の手法で追加した概念であり、もとの概念ベースに存在する表記違いの「コンピューター」が類似語として得られていない。これは、インターネットの検索要求では、コンピュータによる辞書やコンピュータに関する用語説明のニーズが高かったために、求めた属性ベクトルが辞書や用語に近づいてしまい、もともとのコンピュータの辞書的な意味から離れてしまったことが原因と思われる。

テレホーダイ	1.00000	コンピュータ	1.00000
テレチョイス	0.99893	用語辞典	0.89343
テレ放題	0.99311	用語集	0.86676
テレホウダイ	0.99268	用語解説	0.82022
NTT 東日本	0.99222	パソコン用語	0.75876
タイムプラス	0.99203	コンピュータ用語	0.75129
エリアプラス	0.99203	オンライン辞書	0.73381
テレほーだい	0.99112	イタリア語	0.73230
IP 接続サービス	0.99072	専門用語	0.72586
てれほーだい	0.98976	辞彙	0.71955

図2 「テレホーダイ」の類似語　　図3 「コンピュータ」の類似語

5 おわりに

検索ログを利用することにより、新語や固有名詞を概念ベースに追加する方法を提案した。今後、長期間のログの利用や、利用する関連語の増加などにより、精度の向上を行い、検索や分類に適用していきたい。

参考文献

[1] 笠原、松澤、石川:“国語辞書を利用した日常語の類似性判別”,情処学論 Vol.38,No.7,pp.1272-1282,1997.

[2] 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”,情処学論 Vol.39,No.7,pp.2250-2258,1998.