5W-3

拡張概念ベースを用いた文書の分析

牧野 俊朗　杉崎 正之　稲垣 博人
NTTサイバーソリューション研究所

1 はじめに

近年のインターネットの発達により、我々が容易にアクセスできる情報の量は飛躍的に増加した。これらの情報を有効に活用する為には、必要とする情報を効率的に選びだすことが必要不可欠であり、情報の検索や分類に対するニーズが高まっている。文書の検索には、文書中の各単語の出現頻度より求めたtf・idf値がよく利用されるが、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。語の意味を扱う方法として、国語辞書の語義文やコーパスを元に、語を意味ベクトルで表現した概念ベース[1]があり、情報検索システムで利用されている[2]。

情報検索システムは、多数の文書から目的とする文書を見つけ出すことは可能であるが、1つの文書中から目的の部分を見つけだすことはできない。けれども、もし1つの文書をその意味内容に応じて、話題毎に分割できれば、従来の情報検索の手法で目的の部分を見つけることが可能となる。本稿では、国語辞書に基づく概念ベースを、インターネットの検索エンジンの検索ログを利用して拡張した拡張概念ベース[3]を用いて得られる文書の部分の意味情報を利用して、文書を意味的に分割する方法について述べる。

2 拡張概念ベース

概念ベースは、各語の意味をベクトル形式で表現したものである。単語W_iは、次のベクトルw_iで表現される。

$$w_i = (w_{i1}, w_{i2}, \cdots, w_{in}) \tag{1}$$

ここで、w_{ij}は単語W_jに対する重みであり、この値が大きい程、単語W_iと単語W_jは関係が深いとする。辞書に基づく概念ベースでは、この値を単語W_iの語義文中の単語W_jの出現頻度を元に決定する。ベクトルの次元数nは、このままだと、辞書中に出現する単語の種類の数となるが、日本語語彙大系[4]の意味属性を用いて、同じカテゴリに属する単語を同一視することによって、約3000に圧縮してある。また、各ベクトルは長さが1になるように正規化してある。

辞書に基づく概念ベースでは、辞書の項目にない語は定義できないため、固有名詞や新しい語で定義されていないものがある。そこで、インターネットの検索エンジンで用いられる検索語には、多数の固有名詞や新語が含まれている点に着目し、ある目的のページを探す為に入力される一連の単語同士は関連が深いと仮定し、その関連度を元に辞書に基づく概念ベースに存在しなかった語を追加したものが、拡張概念ベースである。

概念ベースを用いると、各単語のベクトルの内積を計算することにより単語間の意味の近さを求めることができる。

3 拡張概念ベースを用いた文書内意味区切りの推定

文書を単語の列と見なすと、各語のベクトルより、文書D_iの文書ベクトルd_iを以下のように定義できる。

$$d_i = \sum_{j=1}^{m} w_j \tag{2}$$

2つの文書について、その文書ベクトルの内積を計算することにより、文書間の類似度を求めることができる。これを応用して、1つの文書中に2つの話題が含まれている場合に、その区切りを以下のような手法で推定する。

(1) 文書内の各文について文ベクトルt_iを、文中に出現する名詞の単語ベクトルの和として求める。

$$t_i = \sum_{j=1}^{m} n_j \tag{3}$$

ただし、n_jは各名詞の単語ベクトルである。

(2) 文書を最初の文からk番目の文までの部分と$k+1$番目の文から文書末までの2つの部分に分け、それぞれの部分の部分文書ベクトル$d_{1,k}$,$d_{k+1,n}$を求める。

$$d_{1,k} = \sum_{j=1}^{k} t_j, \qquad d_{k+1,n} = \sum_{j=k+1}^{n} t_j \tag{4}$$

Document analysis by using the extended Concept-Base
Toshiro MAKINO, Masayuki SUGIZAKI,
and Hirohito INAGAKI
NTT Cyber Solutions Laboratories

(3) k を1から $n-1$ まで変化させ $d_{1,k}$ と $d_{k+1,n}$ の内積を計算することにより、文書の前半部と後半部の類似度を求め、その値が最小となる k の値 l を求める。

(4) 1番目から l 番目の文で構成される部分が1つめの話題であり、$l+1$ 番目から文書末までの部分が2つめの話題であると推定する。

本手法は、1つの話題中の文の集合は、同一話題中の他の文の集合と類似しているという仮定に基づいている。例えば、l 番目の文までが1つの話題とすると、$k<l$ の場合は、同一話題中の部分文書ベクトル $d_{k,l}$ と $d_{l+1,n}$ の内積の値が大きくなり、$k>l$ の場合は、同様に $d_{1,l}$ と $d_{l+1,k}$ の内積が大きくなると予想される。よって、$k=l$ の場合の内積の値が最小になると考えられる。

4 実験と考察

新聞記事データから、2つの記事をランダムに選び、それぞれの本文を続けたファイルを100ファイル作成し、記事の切れ目が判定できるかの実験を行った。

図1は、あるファイルを対象とした k の値による文書の前半部と後半部の類似度の変化のグラフである。この例の場合、$k=6$ の時の類似度が最小なので、第1文から第6文までが1つ目の記事で、第7文から第14文までが2つ目の記事と推定する。

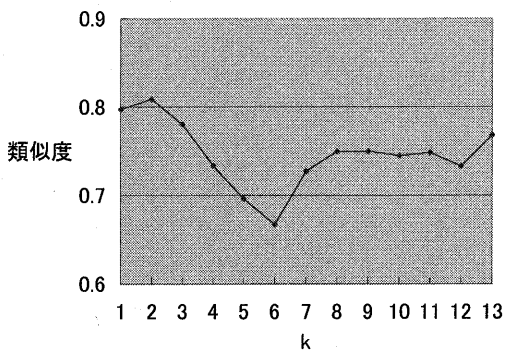

図1 kの値による類似度の変化

このようにして推定した結果と、実際の記事の切れ目の誤差の文の数毎のファイル数を表にしたものが図2である。

100ファイルのうち62ファイルに関しては、正しく記事の区切りを判定することができた。判定が間違っていた38ファイルを分析した結果、原因は以下の3つに大別された。1つ目は、1つの記事中で話題に変化がある場合である。特に記事中に人の談話などが挿入されているものでは、判定の誤りが生じ易い。2つ目は、概念ベース内に含まれていない単語が多い場合である。拡張概念ベースを用いても、まだ記事によっては未知語が多数存在する場合がある。3つ目は、特に誤差が1文の場合であるが、境界の文が極端に短い場合である。この場合、この文をどちらに含めるかの判定に誤りが生じることがある。文が短い場合は、文中に出現する名詞が少なく、またその名詞が未知語となることがある為である。

誤差	ファイル数
0	62
1	15
2	5
3	6
4	3
5	2
6以上	4
判定不能	3

図2 推定結果の誤差

今後、上記の問題点を解決する為に、①より長期の検索ログを用いて拡張概念ベースの単語を増やす。②名詞以外の自立語も対象にする。③同一の未知語が多数回用いられている場合、その使用されている文の範囲を求める。などにより手法の改良を行う。

5 おわりに

拡張概念ベースを用いて、1文書中に2つの話題がある時の話題の区切りを推定する手法を提案した。今後は、手法を改良し推定精度を向上させると共に、1文書中に多数の話題がある時にも、適切な区切りが推定できるように手法の拡張をして行きたい。

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] 熊本、島田、加藤:概念ベースの情報検索への適用－概念ベースを用いた検索の特性評価－, 情処研究報告 99-ICS-115,pp.9-16,1999.

[3] 牧野、杉崎、田中:検索ログを利用した概念辞書の拡張, 第61回情処全大 1P-04, 2000.

[4] 池原他:日本語語彙大系1意味体系, 岩波書店,845P.,1998.