4K－3

複合語マッチングによる情報検索

山田 剛一　斉藤 公一　森 辰則　中川 裕志
横浜国立大学 工学部

1 はじめに

ネットワークの発展により、一般ユーザが大規模データベースに対して検索を行う機会が増えている。多くの場合ユーザが望む出力数は限られているので、文書に対し綿密な重要度付与を行ってランクづけすることが必要である。本発表では、語が複合して意味のまとまりをつくることに着目し、複合語を単位とした類似度計算を行うことによって柔軟なスコアリングを行う手法を提案する。

日本語は、複合語が多く現れる言語である。複合語は全体で一つの概念を表現しているので、文書の特徴量を考える際には、複合語を構成する個々の単語ではなく複合語自身を用いることが望ましいと考えられる。しかし、文書をランキングするために広く用いられているベクトル空間モデルは、ベクトルの要素として単語の重みを用いている。我々はこれを拡張し、複合語の部分マッチに対するスコアを定義することにより、擬似的に基本量を複合語に格上げし、より文書の特徴を的確に捉えることを試みた。

2 複合語マッチング

複合語は検索要求文にも文書にも現れる。ここではまず、検索要求文と文書の各１複合語に着目し、そのマッチングの評価のしかたについて述べる。

語と語のマッチング

検索要求文 Q 内の (複合) 語の集合を C^Q、ある文書 D_i 内の (複合) 語の集合を C^{D_i} とする。そのそれぞれの要素の $C_j^Q, C_k^{D_i}$ に注目し、次のように表現する。

$$C_j^Q = /W_1^Q/W_2^Q/\ \cdots/W_n^Q/$$
$$C_k^{D_i} = /W_1^{D_i}/W_2^{D_i}/\cdots/W_m^{D_i}/$$

ただし、/ は語の区切り、W は複合語を構成する基本語を表現している。基本語は名詞あるいは、接頭語、接尾辞である。また、「の」による連体修飾は語の連接と同様に扱っている。

複合語内では、語が単に共起しているのではなくて連接しているということが重要な意味を持っている。そこで連接を保存したまま、共通部分を抽出することを考える。一つ例を示す。

$$C_j^Q = /A/B/C/D/E/$$
$$C_k^{D_i} = /B/C/E/$$

この場合、$C_j^Q, C_k^{D_i}$ の共通部分である語の列 (パターン) は、$P(C_j^Q, C_k^{D_i}) = \{/B/C/, /E/\}$ となる。/B/ や /C/ はより大きいパターンに含まれているので抽出しない。

このようなパターン抽出を文書 D_i 内の全複合語に対して行い、C_j^Q に対する、文書 D_i が含むパターンの集合 $AllP(C_j^Q, C^{D_i})$ を求める。

$$AllP(C_j^Q, C^{D_i}) = \bigcup_{C_k^{D_i} \in C^{D_i}} P(C_j^Q, C_k^{D_i})$$

さて、各パターンに対する重みであるが、ここではパターン P の文書 D_i における出現頻度 $PatternFreq^{D_i}(P)$ と、パターン P の $IDF(P)$ を用いて次のように定義する (これを pfidf と呼ぶことにする)。

$$PatternWeight^{D_i}(P) = PatternFreq^{D_i}(P) \times IDF(P)$$

ただし、パターン P の IDF の定義は次のようにしている。

$$IDF(P) = \left(\log_2 \frac{DBsize(DB)}{freq(P, DB)}\right) + 1$$

ここで、$DBsize(DB)$ はデータベース DB の全文書数、$freq(P, DB)$ はデータベース DB 内におけるパターン P が出現する文書数である。

文書のスコア

先ほど定義したパターンの重みを用いて、まず検索要求文中の１複合語 C_j^Q に対しての文書 D_i のスコア

Information Retrieval Based On Compound Matching
Koichi Yamada, Koichi Saitoh, Tatsunori Mori and Hiroshi Nakagawa
Division of Electrical and Computer Engineering,
Faculty of Engineering, Yokohama National University

$CScore^{D_i}(C_j^Q)$ を求める。文書内で何通りもの部分マッチが起こる可能性があるので、それぞれのマッチングの度合いに応じたスコアの総和として定義している。

$$CScore^{D_i}(C_j^Q) = \sum_{P_k \in AllP(C_j^Q, C^{D_i})} PatternWeight^{D_i}(P_k)$$

検索要求文 Q 全体に対するスコアは次のように定義する。

$$DScore^{D_i}(C^Q) = \frac{\sum_{C_j^Q \in C^Q} CScore^{D_i}(C_j^Q)}{\sqrt{\sum_{C_k^{D_i} \in C^{D_i}} PatternWeight^{D_i}(C_k^{D_i})^2}}$$

文書中の複合語全体についてパターンの重みを計算し、それをベクトルの重みとみなしてベクトルの大きさで正規化することにより、ベクトル空間モデルのような効果を出している。

3 評価

本稿で提案した、pfidf の重みによる複合語マッチングの有効性を検証するため比較実験を行った。比較対象は、tfidf の重みによるベクトル空間モデルである。

ただし、IDF には以下の定義を用いている。

$$IDF(N) = \left(\log_2 \frac{DBsize(DB)}{freq(N, DB)}\right) + 1$$

$DBsize$: DB 内の総文書数

$freq$: DB 内で名詞 N が出現する文書数

また参考に、以前我々が報告した手法の改良版とも比較した (cnidf)。

いずれのシステムも、検索要求文、記事本文とも茶筌 (ver.1.0b4) を用いて形態素解析をしている。また、未定義語は名詞として扱っている。

評価には、情報検索評価用データベースである BMIR-J1 を利用[1]した。これは文書 600 記事と検索要求文 60 文、およびその正解からなるものである。この正解には A,B の 2 ランクがあるが、今回の評価では同一に扱った。また、検索要求文 60 文の中には複合語を含まないものも多いが、一般的な傾向を知るため総ての検索要求文を利用して評価した。なお、各方式の純粋な比較を行うため、記事の情報は本文のみを利用し、タイトルや付与されているキーワード、記事の重要度等は使用していない。

[1]株式会社 日本経済新聞の協力によって、社団法人 情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが、1993 年 9 月 1 日から 12 月 31 日の日本経済新聞記事を基に構築した情報検索評価用データベース (テスト版) を利用

評価結果

提案手法を用いることにより、tfidf によるベクトル空間モデルに比べ、再現率、適合率とも向上することが確認された (図 1)。

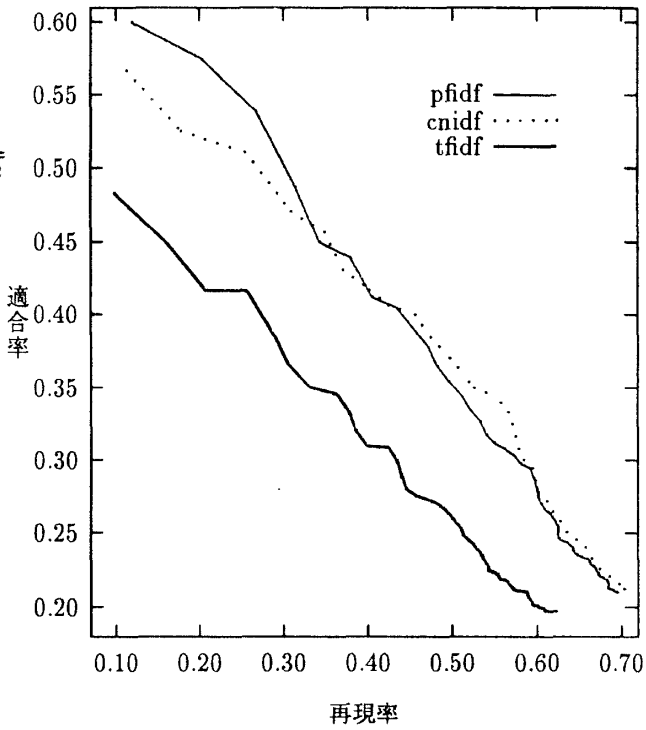

図 1: BMIR-J1 による評価結果

4 おわりに

複合語を検索における基本量とすることにより、検索精度が向上することが示された。本手法と、従来研究されている共起情報やシソーラスを使用する手法とを統合することにより、よりよいシステムが構築できるものと考える。

参考文献

[1] 山田剛一, 森辰則, 中川裕志. 情報検索のための複合語マッチング. 情報処理学会研究報告 96-NL-115-13, 自然言語処理研究会, 情報処理学会, Sept 1996.

[2] Yasushi Ogawa, Ayako Bessho, and Masako Hirose. Simple word strings as compound keywords: An indexing and ranking method for Japanese texts. In *ACM-SIGIR'93*, pp. 227–236, June 1993.