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
{aron@naklab, junkie@naklab, mori@forest, nakagawa@naklab}.dnj.ynu.ac.jp

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。そしてその実現のためには、検索要求や文書の特徴を的確に捉えることが第一であるといえる。

ベクトル空間モデルに代表される従来の検索モデルでは、単語を特徴量の基本単位としている。また、単語の重みには単語の統計情報が一般的に用いられている。これまで、モデルの改良や重みの補正などによって検索精度を向上させようという研究が数多く行われてきたが、単語を基本とするという点では変化は見られなかった。

これに対し本研究では、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マッチング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。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からである。

例えば、単語やその共起情報に基づくモデルでは、2語からなる複合語「システム ファイル」と「ファイル システム」を区別することができない。これは複合語の構造情報を扱わなければ解決できないのである。

2 複合語マッチング

単語を基本量としたモデルでは、異なる語はマッチしないのでマッチング自体が存在しなかった。その語自身が出現しているか否かだけが問題であったのである。一方、本手法では特徴量を単語から複合語へと格上げしたことにより、部分マッチが生ずるようになる。その部分マッチの度合を評価することによって、文書のスコアづけを行なう。

ここでは、まず検索要求文と文書の各1複合語に着目し、そのマッチングの方法について述べる。

2.1 語と語のマッチングによる共通パターンの抽出

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単語(基本語)を表現している。基本語は名詞(形式名詞は除く)あるいは、接頭辞、接尾辞、助数辞である。また、助詞「の」による連体修飾は語の連接と同様に扱っている。これは、「の」によってつくられる名詞句は意味のまとまりとして複合語に近いだけでなく、全く同じ意味構造を持つ場合でも「の」の有無に自由度があることも多いため、このような場合の「表記のゆれ」による検索精度の低下を防ぐという意味もある。

さて、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、語の連接を保存したまま、各複合語の共通部分を抽出することを考える。一つ例を示す。

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

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

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うことにより、検索要求文 Q 内の1複合語 C_j^Q に対する、文書 D_i が含むパターンの集合 $AllP(C_j^Q, C^{D_i})$ が求まる。

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

2.2 $pf \cdot idf$ 法によるパターンの重み

次に、抽出した各パターンの重みについて考える。単語の重みづけには多くの手法が提案されているが、単語の文書内出現頻度 tf と出現文書数 df を用いる $tf \cdot idf$ 法を基盤としたものがほとんどである。本稿では、その $tf \cdot idf$ 法をパターンに対して適用した $pf \cdot idf$ 法を提案する。

まず文書内出現頻度であるが、単語の出現頻度 tf(term frequency) に対し、パターンの出現頻度 pf(pattern frequency) を考える。単語の文書内出現頻度の場合には、文書の「大きさ」を何らかの指標によって表現し、それによって出現頻度を正規化する手法が提案されており、その指標としては、単語数や語彙数、あるいは簡便に文字数が用いられているなどさまざまである。ここでは、文書の内容量を判断する場合にも意味的なまとまりとしての複合語に着目し、複合語レベルの語彙数を指標とすることにした。これによる正規化出現頻度 npf は次のように定義される。

$$npf^{D_i}(P) = \frac{\log_2(pf^{D_i}(P) + 1)}{\log_2 length^{D_i}}$$

ただし、$length^{D_i}$ は文書 D_i における複合語レベルの語彙数である。

次に、パターンの idf を定義する。これは、単語の idf をパターンに拡張したものである。

$$idf(P) = \left(\log_2 \frac{\#doc}{df(P)}\right) + 1$$

ここで、$\#doc$ はコレクション内の全文書数、$df(P)$ はコレクション内におけるパターン P が出現する文書数である。

一般的な傾向としては、構成単語数の多いパターンほど出現文書が限られてくるため、大きな idf の値が与えられることになる。

以上を用いて、文書 D_i におけるパターン P の重み $pw^{D_i}(P)$ を次のように定義する。

$$pw^{D_i}(P) = npf^{D_i}(P) \times idf(P)$$

2.3 文書のランキング

部分マッチのモデルとしては、単語の重みをベクトルの要素とするベクトル空間モデル (VSM) が有名である。しかし、複合語を単位とする場合には部分マッチを考慮するため、文書をベクトルとして表現し演算することはできない。そこで、まず検索要求文側の各複合語に対する文書のスコアを求め、検索要求文全体に対する文書のスコアはそれらの総和とすることにした。

前節で定義したパターンの重み pw 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 $CScore^{D_i}(C_j^Q)$ 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。

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

検索要求文 Q 全体に対するスコアは、Q 中の各複合語に対するスコアの総和と定義する。

$$DScore^{D_i}(C^Q) = \sum_{C_j^Q \in C^Q} CScore^{D_i}(C_j^Q)$$

3 評価

本稿で提案した、$pf \cdot idf$ 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。

3.1 比較対象

比較対象は、$tf \cdot idf$ の重みによるベクトル空間モデルである。出現頻度 tf の文書の大きさによる正規化には、(単語レベルの) 語彙数を用いた。単語を基本量としたモデルでは、複合語レベルの語彙数で正規化することは現実的ではないからである。なお、文字数や単語数よりは単語レベルの語彙数を用いたほうがよい結果が出ると予測したのであるが、実際に調査したところ有意な差は確認できなかった。

正規化出現頻度 ntf は次のように定義した。

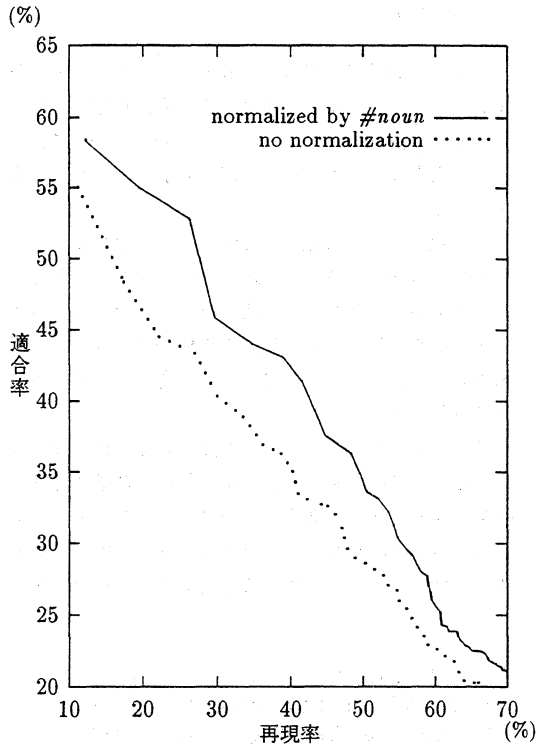

図 1: *tf·idf* における正規化の効果

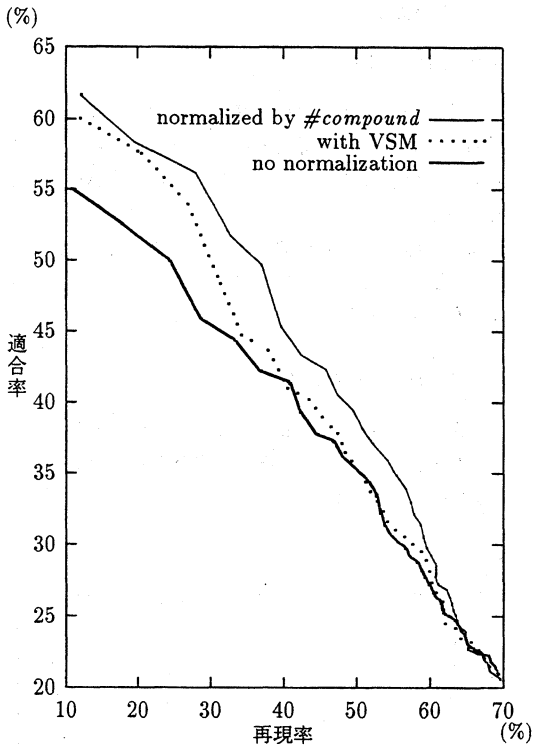

図 2: *pf·idf* における正規化の効果

$$ntf^{D_i}(N) = \frac{\log_2(tf^{D_i}(N)+1)}{\log_2 length^{D_i}}$$

$length^{D_i}$ は文書 D_i における単語の語彙数である。

また、*idf* には以下の定義を用いている。

$$idf(N) = \left(\log_2 \frac{\#doc}{df(N)}\right) + 1$$

ただし、$\#doc$ はコレクション内の総文書数、$df(N)$ はコレクション内で名詞 N が出現する文書数である。

3.2 評価条件

いずれのシステムも、検索要求文、記事本文とも茶筌 (ver.1.0b5) を用いて形態素解析をしている。また、茶筌における未定義語は名詞として扱っている。なお、数字、アルファベット等の連続は一つの形態素として扱うようにフィルタを通している[1]。

[1]この機能を持つ JUMAN を使用していないのは、茶筌 (の辞書) でなければカタカナの複合語を扱えないためである。

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[2]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。

[2]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

3.3 評価結果

出現頻度の正規化

まず、*tf·idf*, *pf·idf*の各重みづけ手法において、出現頻度の正規化の効果を確認した(図1、図2)。どちらも正規化することにより、明らかに検索精度が向上している。

なお、図2の "with VSM" は、以前提案した、ベクトルの要素を複合語とした文書ベクトルにより文書スコアを正規化する手法での値である。

pf·idf と tf·idf との比較

提案手法を用いることにより、*tf·idf*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3)。

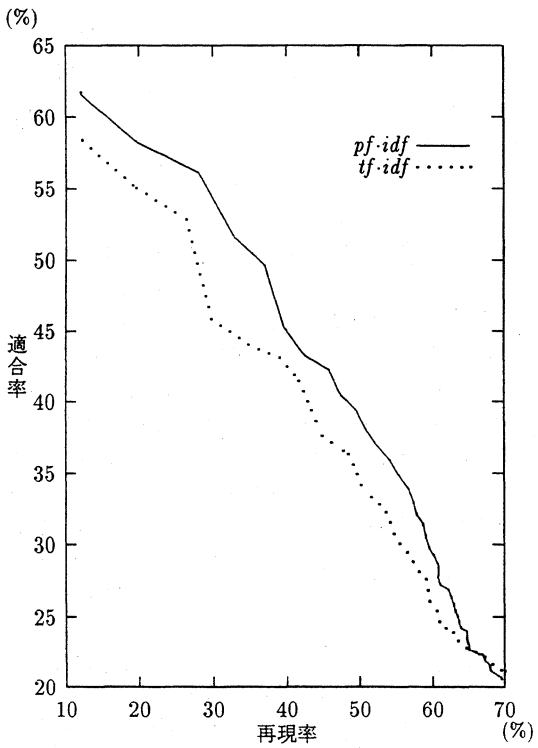

図 3: *pf·idf* と *tf·idf* との比較

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。今回提案した手法はベクトル空間モデルのように文書のベクトルを計算する必要がないため、速度の面でも劣らないものと考えられる。

本手法は文書を特徴づける量として複合語に着目したもので、従来の単語の共起を用いる手法とは異なるものである。この2つの手法を統合し、複合語レベルの共起情報を扱うアルゴリズムを構築すれば、双方の利点を取り込んだ、より検索精度の高いシステムが実現すると考える。

また、シソーラスを用いる手法もいくつか提案されているが、やはり意味のまとまりとしての複合語は無視すべきではなく、単語レベルで単純にシソーラスを利用することは望ましくないと考える。やはり、複合語を意識したシソーラスの利用法を考える必要があるといえる。

謝辞 BMIR-J1を提供してくださった方々、特にリコー小川さん、富士通 松井さんに感謝いたします。また、JUMAN, 茶筌を公開、発展させ続けている方々に感謝いたします。

参考文献

[1] 高木徹, 木谷強. 単語出現共起関係を用いた文書重要度付与の検討. 情報処理学会研究報告 96-FI-41-8, 情報学基礎研究会, 情報処理学会, April 1996.

[2] David A. Evans and Chengxiang Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, June 1996.

[3] 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.

[4] 亀田雅之. 擬似キーワード相関法による重要キーワードと重要文の抽出. 言語処理学会第2回年次大会発表論文集, pp. 97–100, March 1996.

[5] 野口直彦, 稲葉光昭, 野本昌子, 菅野祐司. 単語統計情報と言語情報とを併用した新しい文書検索のモデル. 情報処理学会研究報告 96-FI-44-5, 情報学基礎研究会, 情報処理学会, Dec 1996.

[6] 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.

[7] 山田剛一, 斉藤公一, 森辰則, 中川裕志. 複合語マッチングによる情報検索. 情報処理学会第54回全国大会発表論文集, March 1997.