D-4-2

複合語を考慮したテキスト自動分類

Text Classification at the Compound Word Level

斎藤 剛 Tsuyoshi SAITO　熊谷 貴禎 Kiyoshi KUMAGAI　小林 学 Manabu KOBAYASHI　平澤 茂一 Shigeichi HIRASAWA

早稲田大学 理工学部 経営システム工学科

Department of Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University

1. はじめに 文書自動分類システム構築のためには，キーワードの選択が重要となる．しかし日本語のような膠着言語では単語の区切りが不明瞭なため形態素解析が必要となり，その処理の過程で多くの複合語は分解され形態素となってしまう．したがって，その形態素を組み合わせた複合語もキーワード候補として用いることで精度が向上すると考えられる．そこで本稿では形態素をカテゴリ内頻度情報に基づいて再結合し，キーワード候補となる語集団の精度向上を行う．その際，データマイニングにおけるアプリオリアルゴリズムを用い，現実的な計算量での複合語キーワード生成を可能にする．さらに提案手法をベンチマークデータに適用し，その有効性を明らかにする．

2. 従来研究 キーワードの候補として形態素解析の結果，得られた名詞をそのまま用いる分類ルール抽出法は，最も基本的な技法として古くから存在する[1]．しかし，対象となるテキストから単語の抽出漏れや語の区切りの誤りがあると，たとえ重要度評価が優れていてもキーワードによる分類精度を向上させることができない．これは形態素解析では辞書に登録されている語にマッチした文字列を語とみなすため，まだ辞書に登録されていない新語を抽出できないことに起因する．例えば「年間-国民-所得」のように本来なら切れるべきではないところで切れてしまう．これを防ぐには人手で辞書を更新することもできるが，常に増加し続ける専門用語などには対応しきれていない．また，単独ではカテゴリを特徴付けないが，組み合わされることによって有効性を持つという複合語も考えられる．その結果，どのような分類のモデルを利用するにしろ，この最小単位の形態素を用いたままでは，キーワードの分類精度を向上させるには限界がある．

一方，形態素解析を用いずに，キーワード抽出している研究[2]もある．しかし，形態素解析を用いないこれら従来手法では，必然的にキーワードの品詞情報を失うこととなる．それにより品詞情報を必要とする分類モデルを適用することが不可能となり汎用性の低下は避けられない．

3. 提案方式 テキストの分類ではないが，長野らは文書を端的に表現する単語の抽出手法[3]を提案した．本研究はこの手法にデータマイニングのアプリオリの手法[4]を応用して計算量を削減し，最小支持度をみたす複合語をキーワードの候補とすることにより分類精度を向上させる方式を提案する．

この利点としてはアプリオリを用いているため，大量の文書の全品詞に関して処理を行うことが可能であることが挙げられる．それによって，文書による表記の違いや同義語の影響は少なくなる[3]ことが期待できる．さらに語尾の表現などもキーワードとして抽出されることも期待できる．以下でこの手法の定式化を行う．

① 文書を要素に持つ第k番目のカテゴリをC_kとする．C_kに属する全ての文書中の全ての単語t_iを抜き出し候補集合$R^m_{Ck}=\{t_1,t_2,\cdots\}(m=1)$とする．また，その要素をフレーズと呼ぶ．

② 候補集合R^m_{Ck}のフレーズt_iに対し，カテゴリ内出現頻度$IP(C_k,t_i)$を，カテゴリC_kに含まれるt_iを持つ文書の数$x(C_k,t_i)$を用いて以下のように計算する．$|C_k|$はカテゴリC_k中の文書数．

$$IP(C_k,t_i)=\frac{1}{|C_k|}x(C_k,t_i) \qquad (1)$$

関数$IP(C_k,t_i)$はカテゴリ内でフレーズを持つ文書の存在確率を示し，これが高いということは，そのフレーズが，このカテゴリでは一般的であることを示す．

③ ユーザによって与えられている閾値(最小支持度) θよりもIP値の小さいフレーズを切り捨てる．これをm段階目のラージアイテム集合L^m_{Ck}とする．すなわち

$$L^m_{Ck}=\left\{t_i\in R^m_{Ck}\mid IP(C_k,t_i)\geq\theta\right\} \qquad (2)$$

とする．

④ カテゴリC_k中の文書を参考にして，L^m_{Ck}に含まれるフレーズの(文書中での)前後の単語を付加し，それを$m+1$段階目の候補集合R^{m+1}_{Ck}とする．なお付加する単語はL^m_{Ck}に含まれていなければならない．また，$\circ$は連接を表す．すなわち

$$R^{m+1}_{Ck}=\left\{t_i\circ t_j\mid t_i\in L^1_{Ck},t_j\in L^m_{Ck}\right\} \qquad (3)$$

とする．

⑤ $m\leftarrow m+1$とし，②～④の作業をラージアイテム集合が得られなくなるまで繰り返す．これはアプリオリ[1]により実現可能である．

⑥ ①～⑤を全てのカテゴリについて繰り返し，$L^m_{Ck}(m\geq 2,\forall k)$の和集合をとり$L_{CALL}$とする．

⑦ L_{CALL}と形態素解析の結果得られた名詞に対し，相互情報量を用いてキーワードの抽出を行う[5]．

4. シミュレーション 分類結果が既知の毎日新聞1994年分[6]に形態素解析を行った結果から9カテゴリ8,243記事を学習データ，4,100記事をテストデータとして抜き出した．それに対し，本手法を用いて抽出したキーワードと，従来手法どおりに形態素解析結果の名詞から相互情報量を用いて抽出したキーワードを用いて，それぞれテストデータの分類を行い正解率の比較を行った．ただし，抽出キーワード数は同一とする．

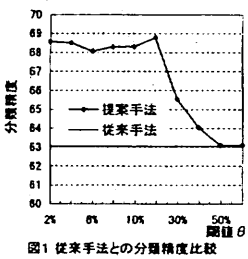

図1 従来手法との分類精度比較

その結果，支持度の閾値θを20%としたとき最適で68.8%となり，それ以上閾値を上げると，生成される複合語が極端に減少し精度が減少した．一方20%より閾値を下げても分類精度にそれほど大きな変化はない．従ってθの選び方は比較的自由にとることが可能である．なお，アプリオリにより削除されたフレーズには相互情報量の高いものは含まれていないことは確認済みである．

5. まとめと今後の課題 アプリオリアルゴリズムを用いることによって形態素解析結果の品詞情報を維持したまま複合語生成が可能となり，分類精度の向上が確認できた．今後は閾値を統計的に有意な値を求める方法の研究を行いたい．

《参考文献》

[1] 徳永健伸，情報検索と言語処理，東京大学出版会，1999.
[2] 湯浅夏樹，外川文雄："概念識別子の頻度分布を利用した文書分類"，情報処理学会研究報告 Vol.95 No.87 (95-FI-39), pp.33-40, 1995.
[3] 長野 徹，那須川哲哉："テキストマイニングのための情報抽出"，日本アイ・ビー・エム(株)東京基礎研究所研究報告.
[4] R.Agrawal, R.Srikant : "Fast Algorithms for Minings Association Rules", *Proceeding of the 20th VLDB Conference* pp. 487 – 499 (1994).
[5]相澤彰子："語と文書の共起に基づく「特徴量」の定義と応用"，情報学基礎自然言語処理 136-4 , pp. 25-32 , 2000.
[6] CD-毎日新聞94，毎日新聞社，日外アソシエーツ.