キーワード抽出を用いた文書自動分類手法

4Q－4

岡田 真　小山 雅史　獅々堀 正幹　青江 順一
徳島大学工学部知能情報工学科

1．はじめに

近年，インターネット，CD－ROMなどの普及によって電子化された大量の文書が流通するようになった．ここで，文書があらかじめ分類されていると，検索範囲が狭くなり，検索効率は大幅に向上するという理由で，文書分類は重要な研究テーマである．

文書の分類手法は大きく２つに分けることができる．一つはあらかじめ人間が与えた分類に沿って文書を分類させる手法であり，もう一つは似たような文書をグループ化する事によって文書を自動的に分類する手法である[1],[2],[5]．本稿では，前者の場合について，既に分類してある文書から不要語を定め，それを用いて，未知の文書からキーワードを抽出し，人間が定めた特定の分野に分類させる手法を提案する

以下，２．では字種中心のキーワード抽出手法を紹介し，３．では抽出されたキーワードから不要語を決定する処理について述べる．４．では実験により，本手法の有効性を示す．最後に５．で，まとめと今後の課題について触れる．

2．キーワード抽出

キーワード抽出は，文書から重要語を決定する技術であり，文献検索，テキスト編集など幅広い分野で応用されている基本的かつ重要な技術である[3]．キーワード抽出法には統制語辞書（シソーラス）を使用する統制語方式と，シソーラスを使用しない自由語方式の２つの方式がある．統制語方式とは，キーワード候補をあらかじめシソーラスに用意しておき，シソーラスに登録されたキーワードが対象文書内に存在するか否かによって抽出処理を行う方式である．統制語方式は抽出処理は単純だが，管理に多大な工数を要するシソーラスが必要となる．また自由語方式とは，まず対象文書を形態素解析などにより単語に分割し，次に分割された単語列からキーワードパターンとの照合や頻度情報などの重み付け処理を通してキーワードを抽出する方式である．自由語方式は抽出処理は複雑となるが，シソーラスの管理が不要である．このため，一般的には統制語方式よりも自由語方式の方がより多く利用されている．

自由語方式における従来の方法は，キーワード抽出の際に形態素解析や，キーワードパターンとの照合など必要とする．それらの処理は，大量の文書を扱う本研究の場合，大きな負担になる．その処理の軽減のため，本手法では字面手法を用いる．

字面手法は，特定の字種のみで構成された文字列をキーワードとして抽出する．具体的には重要語を構成する品詞（名詞，代名詞等）の大部分が，日本語文書中では漢字列，カタカナ列，数字列及びアルファベット文字列の組み合わせにより表記されることを利用し，文法的な解析を行うことなく字種によってのみ抽出する．

この手法は，他の手法に比べ高速であるため，本研究のように，大量の文書を扱う場合には適しているが，処理の性質上，分類を誤らせるようなキーワード（以後，不要語と呼ぶ）も数多く抽出してしまう．従って，分類の精度を上げるために，不要語を取り除く処理を行う必要がある．

3．不要語についての処理

抽出されたキーワード中にしめる不要語の割合が多いほど分類の精度が低下する．そのため，それらをキーワード抽出時に取り除く処理を行う．

ここで，不要語の評価を行うため，不要語率を定める．不要語率は，全分野で，あるキーワードを含

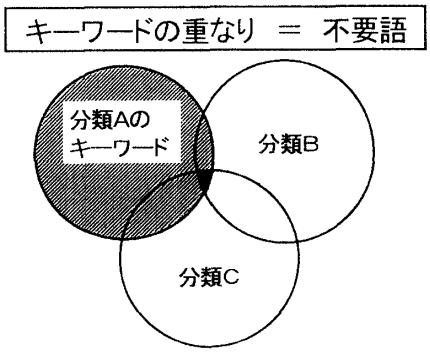

図1 分類分野とキーワードの関係

An Automatic Clustering Method Using
Keyword Extraction
Makoto Okada,Masafumi Koyama,
Masami Shishibori and Jun-ichi Aoe
Department of Information Science and
Intelligent Systems,
University of Tokushima
2-1 Minamijosanjima,Tokushima,Tokushima 770 Japan

んでいる分類分野の割合である．不要語率が高いものほど，不要語として扱われる．

図１にあるように，キーワード中に，いくつかの分類分野に重複して現れるものがある．重複する度合いが大きいほど，それらのキーワードは分野を特定する力が 弱くなり，不要語率が高まる．

不要語の自動抽出は以下の手順で行われる．

（1）　大分類（４．１節参照）ごとにまとめた文書から，頻度の上位からキーワードを一定数抽出し，大分類ごとの辞書を作成する．

（2）　全文書をまとめた文書から，頻度上位のキーワードを一定数抽出し，それに対して，以下を行う．

（3）　カウントを０に初期化する．
（1）で作成したすべての大分類辞書に対して，キーワードを検索し，頻度が閾値より大きい場合，カウントをインクリメントする．

（4）　キーワードのカウントより不要語率を決定し，それが閾値よりも高いものを不要語とする．

４．実験と評価

４．１　不要語の決定

不要語の決定は，まずあらかじめ人間が１１の大分類，１４３の小分類に分類した約５０Mバイト，１５，１０６個のファイルを用意する．そして，各ファイルに対してキーワード抽出を行い，その結果，各大分類の間で頻出しているキーワード中から上位のものを選び出し，それらを不要語と定めた．

集めた文書の中から“ゴルフ”や，“野球”に関する文書をそれぞれ分類したとする．それらの分類はさらに“スポーツ”として分類出来る．このとき，“ゴルフ”，“野球”を小分類，“スポーツ”を大分類と呼ぶ．

“政治”，“科学・技術”などを大分類とした時，“国会”，“行政・内閣”，“選挙”などが“政治”の小分類となり，“エレクトロニクス”，“コンピューター”などが“科学・技術”の小分類となる．

４．２　文書分類

不要語の決定時に用いたデータから１割に当たる１，５３０個のファイルを除外しておき，残りのファイル群にキーワード抽出を行い，その結果得られたキーワードを各小分野ごとに集計して各分野を表す特徴ベクトルを作成した[4]．それらを用いて，除外しておいた文書ファイルについて，分類する実験を行った．その結果を人間の分類と比較し，適合率を用いて評価する．適合率とは，全入力文書に対する，文書数を百分率で表したものである．

その結果を表１に示す．括弧内は不要語除去を行わなかった場合の数値である．小分類単位では１，００４個のファイルが正しい分野を第一候補に挙げている．適合率は約６５．６％である．第三候補までに正しい分野を候補に挙げたのは，１，２７８個で適合率は約８３．５％である．大分類単位では１，３１１個が第一候補に正しい分野を挙げており，適合率は約８５．７％である．不要語削除を行わない場合，第一候補で大分類が正しく分類されたファイルの適合率は約７０％であり，本手法を用いることにより，約１５％の精度の向上が得られ，本手法が有効であることが確認できた．

表１　分類結果

第一候補で正しく分類されたファイル	1,004(個) 65.6(%) (51.3(%))
第三候補までで正しく分類されたファイル	1,278(個) 83.5(%) (67.4(%))
大分類を第一候補で正しく分類されたファイル	1,311(個) 85.7(%) (70.6(%))

５．まとめ

本稿では，文書の自動分類に関する研究として字面解析を用いた自動分類法を提案した．また，不要語の削除を行うことにより，精度の低いキーワード抽出手法でもかなり高い精度で分類を確認した．

今後の課題として，分類をあらかじめ設定しない，完全自動化された文書分類手法を考察する．

参考文献

[１]西野文人：日本語テキスト分類における特徴素抽出，自然言語処理,112-14,pp.95-102, 1996.

[２]福本文代，鈴木良弥，福本淳一：辞書の語義文を用いた文書の自動分類，情報処理学会論文誌,Vol.37,No.10,pp.1789-1799,1996.

[３]林淑隆：複合語文法規則を利用した自動キーワード抽出に関する研究，平成６年度徳島大学修士論文.

[４]長尾　真編：自然言語処理　第一版，岩波書店 1996.

[５] 湯浅夏樹，上田徹，外川文雄：大量文書データ中の単語間共起を利用した文書分類，情報処理学会論文誌，Vol.36,No.8,pp.1819-1827,1995.