Document Classification Using Non-content Features

馬場, 隆寛

（別紙様式 2）

氏　　名　：馬場隆寛

論 文 名　：Document Classification Using Non-content Features
（非内容的特徴による文書分類）

区　　分　：甲

論　文　内　容　の　要　旨

コンピュータおよびインターネットの普及により大量の文書がデジタルデータとしてアクセス可能になった．それぞれの文書に書かれている内容に加え，大量の文書についての傾向から新しい知識が得られる場合がある．例えば，COVID-19 に関する 20 万編を超える研究論文を機械的に解析する試みでは，特定の病態への影響を持つ遺伝子変異の推定などを，複数の論文での統計的情報から見つけることが期待されている．統計的解析によって新たな知識を得るための文書中の特徴は，個々の文書においては主題として記述される内容には直接関係がない場合がある．複数の文書についての統計的な解析によって，著者の感情・意図や文書の価値などの「非内容的情報」の取得が期待できる．

本研究では，文書の非内容的情報による分類に取り組む．非内容的情報は，例えば，著者の精神状態である．文書分類では，出現語句に対する機械学習による効率的な手法が確立されている．一般に，機械学習による文書分類では，分類されるべきカテゴリの情報（ラベル）を持つ文書データ（訓練データ）が大量に必要である．しかし，非内容的情報に関するラベルの付いた文書は，内容に関するラベルの付いた文書に比べて大量に収集することが難しい．例えば，人間の精神状態を説明する文書を集めることは文書の用途や精神状態に関する語句の検索などにより可能だが，精神状態の情報が付与された文書を集めるには付加的な作業が必要である．そこで，本研究では，内容によって収集した文書から内容に関する特徴を取り除くことによって，目的の非内容的情報に関するラベルを持つ訓練データを作成する．例えば，精神疾患を持つ患者による発言を収集するために，グループカウンセリングサイトでの投稿を用いる．この文書の内容はカウンセリングに関するものであるが，内容に関する特徴を取り除くことにより，精神疾患を持たない著者の一般的な内容の文書との比較が可能になる．

本研究で行う文書分類では，内容を表す語を一般化することによって内容に関する特徴を取り除く．文書の内容は，主に内容を表す語の出現を調べることで，ある程度の粒度まで特定できるので，単純な語句の出現に基づく特徴を用いて分類した場合，内容の影響を大きく受けた分類になってしまう．本研究では，内容を表す語を品詞名などに一般化し，語や語の連なりを特徴量とする．内容を表す語は，具体的には，当該分野の専門用語や，一般に「内容語」と呼ばれる名詞，形容詞，動詞，副詞を用いる．これにより，文書の内容に依存しない特徴による分類が可能となる．

本稿では具体的な問題について実験を行い，提案手法の有効性を検証した．ラベルの付与された訓練データを大量に取得することが困難である 3 つ問題について検証を行った．

第一に，研究論文の被引用数を推定し，文書の属性を分類できることを検証した．研究論文の非内容的情報として，被引用数で表される学術的インパクトによる分類を試みた．被引用数が付与さ

れた研究論文の抄録は大量に取得することができるが，単純に内容的特徴を用いて分類を行うと，論文そのものよりも研究分野による傾向を見つけてしまうことを示した．大量の抄録に対し提案手法を適用した結果，研究分野と関係の無い語句を用いて高い精度で被引用数を推定することができた．

第二に，特定の言語で記された文書から著者の母国語の推定を行い，文書そのものではなく著者の属性を分類できることを検証した．著者の母国語が明示される文書は少ないが，多言語で執筆される研究論文の多くに英語の抄録が添えられる点を利用して提案手法を適用した．研究論文抄録の非内容的情報として著者の母国語を想定した．提案手法を適用することによって，研究内容に依存しない特徴による高精度な母国語検知を実現した．

第三に，ソーシャルネットワークシステム（SNS)のコメントの投稿者の精神状態の推定を行い，文字数が少ない文書でも分類できることを検証した．精神疾患を持つ患者をひとりひとり確認した場合，取得できるコメントの量が限られる．これに対し，グループカウンセリングサイトのコメントを用いることで大量のコメントを得た.このコメントの非内容的情報は投稿者の精神状態である．提案手法により，カウンセリングサイトのコメントを正例，一般的な SNS のコメントを負例とする分類から，内容に依存しない特徴による高精度な推定を実現した．

以上から，提案手法が幅広い応用問題に対して適用可能であることがわかった．