HTML 文書のカテゴリ階層への自動割り当て

片山 研一

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

1998 年 2 月 13 日

キーワード： HTML, 階層構造, タグ, ハイパーテキスト, 自動分類.

Abstract

文書のカテゴライズは, 文書の特徴を捉えることによりあらかじめ与えられた分類に沿って文書を整理することである. 大量の文書を整理することによりある目的の文書を後で捜し出す時の手間が軽減できると考えられる. 文書のカテゴリへの割り当てを人手で行うことは非常に手間がかかる作業である. このカテゴリへの割り当てを自動化することで分類作業における手間を軽減でき, さらに客観性を持たせた分類が可能になることが期待される. しかし, 自動化を行うためには, 文書の内容を理解しなければならず, 精度の高い分類は非常に困難である. さらに web 上の文書の場合文字のみでなく画像，音声，プログラム等によって構成されており，問題をさらに難しくしている．特に画像によって文書の主題となるキーワードを埋め込むことも多く文字だけでは情報の欠落が起きていると言える.

本研究では, Web のディレクトリ検索によく見られる文書の階層的な分類を自動的に行なう. この分類では葉ノードのみの分類ではなく, 内部ノードへの分類, 文書の複数ノードへの割り当て, ある階層までしか分類できないという状態も許す. この分類は, 代表的なディレクトリ検索システムである yahoo, yahho などで用いられている分類であり, 現在人手で行なわれている膨大な量の HTML 文書分類の自動化に即利用可能であると考えられる.

文書を割り当てる研究は, ラベル付けされた事例が必要としない教師無しのクラスタリングが, 主流であった. しかし，ディレクトリ検索システムの自動分類化を図るためには，カテゴリにラベルがふられている必要がある．本研究では, 人手であらかじめきめた階層であり，数多くの文書の分類を行っているディレクトリ検索システム yahho のデータを

使うことにより大量の事例を手に入れ，あらかじめ人手で分類した事例を用いた教師有り学習法を用い, 階層への分類を自動的に行なう.

既存の多くの文書自動分類の研究では, 本, 新聞記事, 電子メール, 電子ニュースを対象として扱っていた. 多くの自動分類システムは，情報検索の分野で利用されてきた特徴ベクトルと種々の類似度計算の手法による統計的処理を行って分類している．特徴ベクトルを用いる場合，文書からいかに特徴あるキーワードを取り出すかが問題となる．

HTML 文書を取り扱う分類の研究として, 落谷の WWW ページの分類におけるテキストの特徴分析手法がある．彼は，コサイン距離を用いて類似度計算を行い，特徴素として，形態素，連語，bigram を用いている．しかし，WWW ページの特徴の一つであるハイパーテキスト構造であるリンク情報しか利用していない．本研究では，WWW ページの特徴をより深く利用するために特徴素の取り出しとして HTML 文書特有のタグと呼ばれる文書整形コマンドによる特徴抽出を行う．タグには, タイトルを生成するもの, 語を強調するものなどがあり, 文書の特徴を表すキーワードを見つけ出す手助けと考える．web 階層はノード毎のデータ量に偏りが大きく，このことが，コサイン距離の性能に少なからず影響を与えると考え，コサイン距離以外の距離尺度も実装し比較を行う．また，落谷の手法では，類似度の高い順に k 個のカテゴリに文書を割り当てている．この手法だと類似度が低くてもランクが高ければ必ずカテゴリ割り当ててしまう．本研究で用いる階層の特徴として，上位ノードから下位ノードへ分類していった場合下位ノードとの類似性が高くなければそれ以上下位へはいかない (内部ノードに割り当てる) という状態がある．必ず k 個割り当ててしまう戦略ではこの配置法は行えない．このために類似度のランキングになんらかの閾値を設定してその閾値を越えるカテゴリに割り当て，越えるものがなければ割り当てないという戦略を用いる．さらに，統計的な手法に加え，web 上の文書特有の分類ルールを知識として蓄えて統計的手法と合わせて利用し, その有効性を考える．

評価指標には，情報検索の分野で用いられる，再現率 (recall) と適合率 (precision) を使用する．

本論文では, タグを文書の特徴抽出の道具として, 統計的な方法による語の重み付けと共に用いる特徴抽出の方法を提案し，また，統計的な手法に加え，HTML 文書特有の情報を使いその有効性を考える．従来からの単語の出現頻度のみの情報抽出よりも出現頻度にタグの情報を加えた情報抽出の方法の方が，recall, precision 共に精度の向上がみられた．さらに，各カテゴリ毎に異なる閾値を用いることによって全カテゴリで同じ閾値を使うよりも良い結果が得られた．さらに，URL のドメイン情報を利用した分類ルールを用いることで，統計処理による分類の失敗をカバーすることができた．HTML 文書の分類は，他の分野の文書よりも分類のための特徴抽出に難しさがあるが，HTML 特有の情報を使うことにより結果の向上につながることが示せた．

また，最上位層より完全に分類されるまでの精度を計り，実際の運用と同じモデルでどの程度利用可能であるか調べる．