ユーザによる既存分類に基づくナイーブベイズを用いた文書分類ツール

情報科学科　中洲 利基　　指導教員：粕谷 英人

1　はじめに

インターネットを介したメールなどによるファイルのやり取りは現代では一般的である．それが度重なったり，取得したファイルを整理せずにしておくと一部のディレクトリに蓄積され，整理することに無気力になってしまう．また，ファイル名から内容が推測できないものが含まれているとさらに整理をしなくなり，悪循環に陥る．それらの整理されていないファイルの分類について，自動分類ができると各ファイルを確認する手間を省くことができる．

本論文では，ユーザが持つ文書の分類構造に基づいた文書の分類手法を提案する．また，既存の分類構造に該当しない文書に対して新たな分類を提案し，ユーザへの有益なフィードバックを用意する．

2　文書分類

文書分類で考えられる手法は，分類の手本を学習しそれに基づいて分類する教師あり文書分類と，手本なしで全ての文書を客観的に観察しいくつかのクラスタに分類する教師なし文書分類の 2 通りの方法がある．教師あり文書分類ではどのような文書が各クラスタに含まれるかを学習して別の文書を分類し，教師なし文書分類では文書同士の類似性を元にいくつかのクラスタに分類する．

3　提案手法

本手法では予めディレクトリ毎に分類した文書群と，未分類の文書群を入力データとする教師あり文書分類を用いる．読み取ることができる全ての文書を要素に分割して Bag-of-words にしたものから，各要素の出現回数の TF-IDF 値で文書のベクトル化を行う．全てのベクトルデータと文書が含まれているディレクトリから，ナイーブベイズ分類器を作成し，未分類文書を通して分類確率を計算する．

また，分類結果だけでなく，ディレクトリを特徴付ける要素と文書を示すことで，‘なぜそのような分類になったのか’ が分かりやすく，文書を探す手助けや，既存分類の中の適さない文書の確認にもなる．

3.1　要素に分割

文書を要素に分割する際，本手法では名詞バイグラムを用いる．意味を持つ最小の単位である形態素に分割してしまうと文脈をたどることができなくなってしまう．そこで N-gram の手法を用いて，形態素の N 要素ずつに分割すると要素数も少なく，文脈をたどった要素群を作ることができる．$N = 2$ のときのバイグラムで，更にその中でも形態素の名詞のみを取り出した名詞バイグラムによる学習精度が高くなることは明らかにされているため，この手法を本論文でも使用する [1].

3.2　フィルタリング

本手法は未分類の文書を分類器に通したときに得られるナイーブベイズ確率が閾値未満であれば，どの既存分類にも適さない文書として新たな分類とする．与えられた分類済文書から分類器を作成して，それに再び分類済文書を通して既存分類との正誤判定を行い，得られた各文書のナイーブベイズ確率で正誤判定をソートした結果は，確率が高いほど正解判定数が多くなり安定する．しかし下位になると次第に正解判定数が少なくなるため，確率で降順にソートされた正解判定を上位から確認し，初めて失敗判定したときのナイーブベイズ確率を閾値とする．

3.3　ディレクトリ構造の学習

本論文ではディレクトリの入れ子構造は無視している．入れ子のディレクトリ中文書は全て上位ディレクトリに存在するものとする．なぜなら，ディレクトリの階層が深くなるほどファイルが少なくなりやすく，学習がうまくいかないことが考えられるためである．

4　評価

本手法の分類性能と，実使用についての 2 種類の評価を行う．分類性能については，株式会社ロンウイットの livedoor ニュースコーパス (総文書数 7367, カテゴリ数 9, 平均ファイルサイズ 3514byte)[2] を用いて，10 回交差検定を行った．評価の指標として適合率, 再現率, F 値の値を取った．いくつかの Bag-of-words について実験した結果を表 1 に示す．3.1 節で述べた通り，名詞バイグラムを使った結果が最も高精度となった．

実使用については，実際の環境で文書の幅が多い場合や少ない場合でどれ程の精度で分類できるかや，感想を集計する．

表 1　交差検定結果

%	適合率	再現率	F 値
バイグラム	77.18	73.43	71.09
形態素 (全品詞)	82.23	79.84	77.88
形態素 (名詞のみ)	84.38	81.84	81.31
形態素バイグラム	81.61	78.05	76.74
名詞バイグラム	85.15	82.53	81.51

5　おわりに

本論文では比較的パラメータも少なく，ユーザでも扱いやすいナイーブベイズ分類器を使用した．また，既存の分類構造に該当しない文書に対して新たな分類を提案し，ユーザへの有益なフィードバックを用意した．

今後の課題として本手法は分類にナイーブベイズを用いたが，分類器の変更は実装上容易に可能であるので調査する．ディレクトリの入れ子構造の学習についても，現状ではフラットなディレクトリとして処理しているので，深層まで学習できるように調査を進める．

参考文献

[1] 西野文人, et al. 日本語テキスト分類における特徴素抽出. 情報処理学会研究報告自然言語処理 (NL), 1996, 1996.27 (1995-NL-112): 95-102.

[2] RONDHUIT(2006 年)「ダウンロード - 株式会社ロンウイット, livedoor ニュースコーパス」<https://www.rondhuit.com/download.html#ldcc>, (2018 年 1 月 12 日参照)