2P-4

語の関係性を抽出した特徴ベクトルによる文書分類の提案

Proposal of Document Classification with Feature Vector of Term Correlation

今井 智宏 †
Tomohiro Imai

望月 久稔 †
Hisatoshi Mochizuki

1. はじめに

ウェブ上では多種多様な人によって非常に高い頻度で様々な文書が更新される．これを自動で解析することで市場調査や動向調査などの利用が期待できる．解析法の多くは頻度情報を用い, 人による解析よりも高速であるが，語の関係性を解析することは難しい．

そこで語の共起関係からグラフを構築し，PageRank [1] を用いて解析することで特徴ベクトルを抽出した．そして頻度情報を用いて特徴ベクトルを抽出する TF-iDF と，最近傍法による文書分類の精度を比較した [5].

本論では提案手法において，解析する文書の複雑さと分類精度の依存関係について分析する．また k 近傍法を用いて分類することにより，提案手法の性質について評価する．

2. 特徴ベクトルによる文書分類

はじめに，文書から特徴ベクトルを抽出する方法を説明し，次に文書の分類方法について述べる．

2.1. 特徴ベクトルの抽出

図 1 に特徴ベクトルを抽出するまでの流れを示す．はじめに解析対象とする文書を形態素解析し，文書の特徴を表す上で大きな役割を果たすと考えられる名詞を抽出する．また，本論では同じ 1 文中に存在する語同士の関係を共起と定義する．共起関係にある語同士は意味的なつながりがあると考え，双方向にリンクを持つとする．抽出した名詞を節点，語同士のリンクを辺とした無向グラフを構築し，語の関係性を文書グラフとして捉える．

次に，文書グラフ G を式 (1) に示す [1]．無向グラフを表す行列 H に対して，定常ベクトルを算出できることを保証するために，原始性を付加するテレポーテーション行列 e^T を加え，さらに α によって H への依存率を決定する．

$$G = \alpha H + (1-\alpha)e^T/n \tag{1}$$

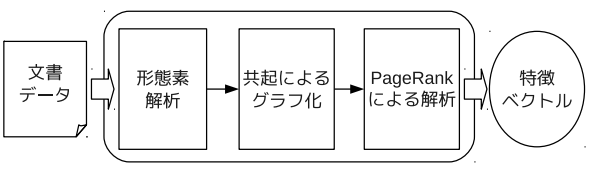


図 1: 特徴ベクトル抽出の流れ

続いて構築したグラフを PageRank [1] で解析する．文書グラフ G からべき乗法によって定常ベクトルを抽出する．計算式を式 (2)，式 (3) に示す．i は収束までの繰り返し回数，π は収束すると G に対する定常ベクトル，e は単位ベクトルを表す．

$$\pi^{(i+1)T} = \pi^{(i)T}G \tag{2}$$

$$\pi^{(i+1)T}e = 1 \tag{3}$$

抽出した定常ベクトルは文書の特徴を表す特徴ベクトルであり，ベクトルの値は文書中に出現した語のその文書における重要度を表す．そして値が大きい語ほど文書の特徴を表し，本論ではその語を重要語と呼ぶ．また，パラメータ α とべき乗法の収束条件の大きさによって，精度と計算量が変化する．

提案手法の抽出精度は解析対象とする文書グラフのリンク構造に大きく依存している．そこで文書構造の複雑さは文書の大きさに相関があると考え，次章では文書の大きさと抽出精度の依存関係について検証する．

2.2.k 近傍法を用いた文書分類

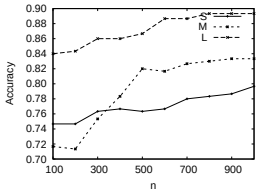
分類手法には k 近傍法を使用する．k 近傍法は分類対象データの近傍を k 個求め，それぞれが投票した結果，最も票数が多いクラスを対象データのクラスとする．SVM のような近年の学習器と比べると分類精度は低いが，データセットのクラス分布からの影響は SVM に比べて小さい [2]．近傍を求める際にベクトル間の距離はユークリッド距離を用いる．

k 近傍法は近傍によって分類するため，文書の特徴を正確に抽出できるほど，各文書の特徴ベクトルがより個別化されて，より分類精度が高くなると考えられる．したがって，分類精度は特徴ベクトルの抽出精度と相関があると考えられる．次章ではこの点について提案手法を評価する．

3. 評価

特徴ベクトルを用いて文書を分類し，その精度を評価することで，特徴ベクトルの抽出精度について評価する．実験は Intel Celeron G550 2.60GHz，Memory 4GB，CentOS6.4 上で行う．形態素解析器は JUMAN [3] を使用する．実験データとして日経 BP 社の ITpro [4] の management, network, security の記事計 21955 件を使用する．パラメータ α を 0.90，べき乗法の収束条件を 0.00000001 とする．また，対象手法として特徴ベクトルの抽出に TF-iDF を使用する．評価データを各クラス 100 件ずつ無作為に選出し，それ以外のデータを訓練データとする．また使用する訓練データの数を変動させて，それぞれの分類精度について考察する．

†大阪教育大学, Osaka Kyoiku University



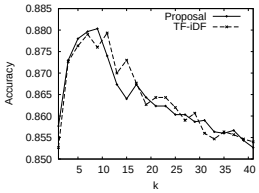
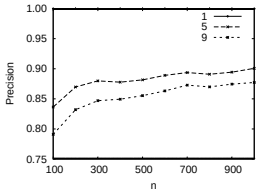


図 2: $k = 1$ の分類精度　図 3: 両手法の分類精度



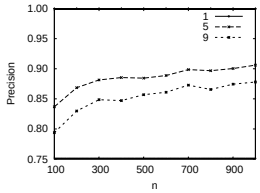


図 4: 提案手法の適合率　図 5: 対象手法の適合率

文書の大きさと分類精度の関係性を見るために，実験データ全体を文書の大きさが 0.3kB 以上 2.0kB 未満，2.0kB 以上 3.0kB 未満，3.0kB 以上の三つに分け，それぞれをデータセット S，M，L とする．ここで分類精度を式 (4) と定義する．

$$\text{分類精度} = \frac{\text{分類正解数}}{\text{分類対象文書数}} \tag{4}$$

各データセットに対し，提案手法を用いて特徴ベクトルを抽出し，k 近傍法を用いて分類した結果から分類精度を検証する．はじめに，図 2 に $k = 1$ で提案手法によって解析した結果を示す．縦軸は分類精度，横軸は各クラスの訓練データ数を表し，グラフは各データセットの分類精度である．

データ数が 400 以降は文書の大きさと分類精度の間に正の相関が見られる．また k が 50 までに関して，すべてにおいて L が最も精度が高いことを確認した．よって文書の大きさが大きい文書ほど，提案手法は文書分類の精度が高いため，文書の複雑さと分類精度には正の相関があると考えられる．また，訓練データの数が大きいほど，比較するサンプルが多くなり，より似通った文書が訓練データに含まれる可能性が高まるため，より高い精度で分類することができる．しかし，k 近傍法を用いるため，訓練データが大きくなると比較する対象が増えることにより，処理時間は大きくなる．

一方，対象手法においても文書の大きさと分類精度の間に正の相関が見られた．1 文書あたりから得られるサンプル数が大きくなったためであると考えられる．

両手法間の差が生じなかったため，文書の大きさだけでの評価は不十分であることがわかった．

続いて，データセット L を用いて近傍数を変動させたときの分類精度について比較する．図 3 に両手法の分類精度を示す．縦軸は分類精度，横軸は近傍数 k を表す．また，分類精度は近傍数ごとの訓練データ数 100 から 1000 における分類精度のマクロ平均である．

提案手法は，解析対象である文書において，頻度情報だけでなく名詞間のリンク構造を解析することによって，頻度は低いが重要語であると考えられる語を抽出できる．そのため，文書の特徴をより正確に表した特徴ベクトルを抽出し，近傍となるデータをより正確に選出できると考えられる．図 3 において提案手法は，対象手法に比べて近傍数が小さいときの分類精度 $k = 10$ までが常に上回っている．

一方，対象手法は $k = 17$ 付近まで近傍数が大きくなっても比較的精度が高い．対象手法は，文書単体における頻度情報を訓練データ全体による頻度情報で汎化することで，提案手法に比べてより平均的な分布をしているためと考えられる．

次に分類に成功した文書に関して，その投票状態から適合率を式 (5) と定義する．

$$\text{適合率} = \frac{\text{正解クラスからの投票数}}{\text{総投票数}} \tag{5}$$

図 4，図 5 に各手法の適合率を示す．縦軸は適合率，横軸はクラスごとのデータ数を表し，グラフは近傍数を 1，5，9 と変動させた適合率である．両手法ともに分類精度のピークから精度が徐々に減少しているのは，近傍数が増えることによって，両手法ともに適合率が低下したことが一因である．図 4，図 5 では近傍数が増加するにつれて分類に対するノイズが多くなるため，適合率は減少している．両手法ともに近傍数 50 までの検証を行ったところ，適合率の減少は収束していくことを確認し，同様の傾向にあることが分かった．

4. おわりに

本論は，文書の大きさが大きいほど提案手法の分類精度が向上することを確認した．さらに k 近傍法を分類に使用することによって，対象手法に比べてより正確に近傍を抽出できることを示した．これらの結果から提案手法が対象手法に比べ，解析対象とする文書の特徴に特化した特徴ベクトルが抽出できると考えられる．

今後は文書の複雑さを共起の密度も考慮し，提案手法との依存関係についてより詳細に分析する．

参考文献

[1] Sergey Brin,Lawrence Page,The anatomy of a large-scale hypertextual Web search engine,Computer Networks and ISDN Systems,30,1-7, pp.107-117, 1998.

[2] 新納浩幸，佐々木稔，k 近傍法とトピックもデルを利用した語義曖昧性解消の領域適応，自然言語処理，Vol.20，No.5，pp707-726，2013.

[3] JUMAN,http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN, 2013.

[4] ITpro,http://itpro.nikkeibp.co.jp/,2013.

[5] 今井智宏，望月久稔，語の共起による文書グラフの構築と PageRank を導入した重要語抽出法，FIT2013，第 2 分冊，pp109-110，2013.