E-025

単語出現頻度の偏りを考慮した文書分類
Text Classification Considering Deviation of Word Frequency of Appearance

山岸 英貴 † 松井 治樹 † 平澤 茂一 †
Hidetaka Yamagishi Haruki Matsui Shigeichi Hirasawa

1. はじめに

近年，電子化された文書量の増大に伴い文書の分類に関する研究が注目を集めている．計算機の普及と性能の向上により文書を計算機上で扱うことが可能となり，文書を自動的に分類することへの需要が高まっている．

文書の自動分類手法は，既存の分野に新たな文書を割り当てる方式と，分野を持たない文書集合から分類構造を生成する方式の 2 つに大別される [1].

本研究では文書構成の統計的特徴を考慮した，前者の方式の分類手法を対象とする．従来,「出現頻度からみた語の重要性」に着目した分類手法が構築されている [2].

そこで本研究では,「出現頻度からみた語の重要性」と χ 2 乗値を用いた「出現頻度の偏りからみた語の重要性」の 2 つの視点から分野を特徴付ける単語を抽出し，その単語に基づいて文書を分類する手法を提案する．

また，本手法を新聞記事データ [4] に適用し，その有効性を示す.

2. 従来の文書自動分類方法

既存の分野に新たな文書を割り当てる分類手法として,各分野における出現頻度の高い単語に着目した手法が提案されている [2]．ここではその手法の概要を説明する．

2.1 高出現頻度の単語に着目した文書分類手法 [2]

学習文書から分野ごとの特徴ベクトルを作成し，それに基づいて新規文書を分類する．

2.1.1 分野の特徴ベクトルの作成

学習文書内の異なり総単語数を N とし，ある分野 k $(k = 1, 2, \ldots, K)$ に出現した単語 i $(i = 1, 2, \ldots, N)$ の出現率 Y_{ik} を以下のように定義する．

定義 1（出現率）

$$Y_{ik} = \frac{w_{ik}}{\sum_i w_{ik}} \quad (\text{ただし} \sum_i Y_{ik} = 1) \tag{1}$$

w_{ik} は分野 k における単語 i の出現頻度を表す．□

各分野において Y_{ik} を降順に並べ，N の 40%にあたる単語数 n $(n = 0.4N)$ 個までを重要語 (次元数 : n) として抽出する．これは，分野の特徴ベクトルの要素となる単語数の削減にもつながる．

ここで，各分野 k における全ての重要語 i に対して，算出した Y_{ik} をもとに X_{ik} を以下の式で計算する．

$$X_{ik} = \frac{Y_{ik}}{\sum_k Y_{ik}} \tag{2}$$

X_{ik} は Y_{ik} を分野間で基準化したものとなっている．

X_{ik} が大きな値を取る単語 i は，出現頻度が高くかつ特定の分野 k に集中して出現する単語を表しており，このような単語は分野の識別力が高いと考えられる．

以上により求めた X_{ik} を用いて，分野 k の特徴ベクトル $\boldsymbol{x_k}$ を以下のように定義する．

定義 2（分野の特徴ベクトル）

$$\boldsymbol{x_k} = (X_{1k}, X_{2k}, X_{3k}, \ldots, X_{nk}) \in R^n \tag{3}$$

□

2.1.2 新規文書の分類

以上で得られた特徴ベクトル $\boldsymbol{x_k}$ をもとに，以下の手順で新規文書を分類する．

1. 新規文書について形態素解析を行い，各単語の出現率 Z_i $(i = 1, 2, \ldots, n)$ を求める．

$$Z_i = \frac{z_i}{\sum_i z_i} \quad (\text{ただし} \sum_i Z_i = 1) \tag{4}$$

z_i は新規文書中の単語 i の出現頻度を表す．

2. Z_i から新規文書の特徴ベクトル $\boldsymbol{u}$ を構成する．

$$\boldsymbol{u} = (Z_1, Z_2, Z_3, \ldots, Z_n) \in R^n \tag{5}$$

3. 全ての分野 k に対して，$\boldsymbol{u}$ と各特徴ベクトル $\boldsymbol{x_k}$ との内積値 $C_k = (\boldsymbol{u}, \boldsymbol{x}_k)$ を求める．
4. C_k の最も大きい分野に新規文書を分類する．

□

2.2 従来手法の問題点

従来手法において重要単語として抽出され，分野の特徴ベクトルを構成している単語は，その分野の中で単純に出現頻度が高い単語である．しかしこの手法では，その出現頻度の高さがどの程度偏っているかということが考慮されていない.

3. 提案手法

本研究では，単語 i に対して χ 2 乗値を用いて理論的な出現頻度からの偏りを考慮する．

出現頻度が高くかつ偏って出現することで分野を特徴付ける単語を抽出し，それを用いて文書を分類する手法を提案する．

3.1 χ 2 乗値について

本手法では，単語の出現頻度から，その偏りの度合いを示す指標 ${\chi^2}_{ik}$ 値を以下で与える．

ある単語 i $(i = 1, 2, \ldots, N)$ が，特定の分野 k に依存する度合いを表す.

定義 3（${\chi^2}_{ik}$ 値）

$${\chi^2}_{ik} = \frac{(w_{ik} - m_{ik})|w_{ik} - m_{ik}|}{m_{ik}} \tag{6}$$

$$\text{ここで，} m_{ik} = \frac{\sum_{k=1}^{K} w_{ik}}{\sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}} \sum_{i=1}^{N} w_{ik} \tag{7}$$

N : 学習文書内の異なり総単語数
K : 分野数
w_{ik} : 分野 k における単語 i の出現頻度
m_{ik} : 分野 k における単語 i の理論頻度

□

†早稲田大学大学院理工学研究科経営システム工学専攻

これは各単語の出現頻度と，その単語が全分野に等確率で出現した場合の出現頻度の差分を表している．すなわち $Y_{ik} - m_{ik} > 0$ となる場合，この単語はその分野に対して特徴的であると言える．これにより，偏って出現している単語ほど ${\chi^2}_{ik}$ 値が高くなる [3].

このように，${\chi^2}_{ik}$ 値を用いることにより関連しない分野に対しては，関連しないということを負の値を与えることにより表現できる．

3.2 重要語の抽出

「国際」分野の単語について，出現率と ${\chi^2}_{ik}$ 値を降順に並べたものの例を表 1，2 に示す．

表 1 では「査察」や「問題」といった単純に出現率の高い単語が，表 2 では「査察」や「北朝鮮」のように出現頻度の偏りが大きい単語が上位に位置するのが分かる．

表 1: 出現率の順位

順位	単語	出現率
1	査察	0.0217
2	問題	0.0148
3	北朝鮮	0.0118
4	米	0.0118
5	大統領	0.0098
6	施設	0.0098
⋮	⋮	⋮

表 2: ${\chi^2}_{ik}$ 値の順位

順位	単語	${\chi^2}_{ik}$ 値
1	査察	48.99
2	北朝鮮	26.72
3	大統領	22.36
4	朝	15.58
5	IAEA	15.58
6	通常	11.50
⋮	⋮	⋮

以上より，${\chi^2}_{ik}$ 値を用いることで出現頻度の偏りから見て重要な単語を抽出できると考えられる．

提案手法では表 1，2 でそれぞれの上位順位の単語を重要と考えたとき，そのどちらにも出現する単語，例えば「査察」,「北朝鮮」,「大統領」のような単語を重要語とする。

つまり，出現頻度が高くかつ偏って出現している単語を重要語として抽出する．

3.3 提案手法の文書分類手順

以下で提案手法の手順を示す.

1. 重要語の抽出.
   各分野で，出現率と ${\chi^2}_{ik}$ 値の上位 n 単語内に存在する共通の単語を重要語として抽出する.
2. 分野 k の特徴ベクトル $\boldsymbol{x_k}' \in R^n$ を構成.
   抽出した単語に対し (2)，(3) 式を用いて $\boldsymbol{x_k}'$ を求める.
3. 新規文書の特徴ベクトル $\boldsymbol{u} \in R^n$ を構成.
4. 新規文書の分類.
   $\boldsymbol{x_k}'$ と $\boldsymbol{u}$ との内積値 C_k が最も大きい分野に新規文書を分類する.

4. シミュレーションによる評価

実験データは毎日新聞の 1 年分の記事データ [4] を利用した．今回の実験では 9 分野に限定し，学習文書を 9000 件，テスト文書を 4500 件用いて，以下の基準によって評価を行った.

$$\text{分類誤り率} = \frac{\text{誤って分類された文書数}}{\text{テスト文書の総文書数}}$$

$$\text{分類適合率} = \frac{\text{ある分野に正しく分類された文書数}}{\text{ある分野に分類された総文書数}}$$

$$\text{分類再現率} = \frac{\text{ある分野に正しく分類された文書数}}{\text{ある分野のテスト文書数}}$$

実験結果を表 3，4 に示す．提案手法と比較している従来手法は 2 節で述べた手法である.

表 3: 平均分類誤り率の比較

	誤り率
従来手法	0.292
提案手法	0.273

表 4: 分類適合率と分類再現率の比較

	適合率		再現率	
	従来	提案	従来	提案
国際	0.812	0.837	0.782	0.782
経済	0.865	0.858	0.788	0.786
家庭	0.741	0.732	0.746	0.768
文化	0.685	0.686	0.274	0.280
読書	0.382	0.390	0.742	0.801
科学	0.378	0.356	0.827	0.950
芸能	0.726	0.773	0.798	0.800
スポーツ	0.860	0.862	0.852	0.854
社会	0.756	0.833	0.558	0.512
平均	0.689	0.703	0.707	0.726

4.1 考察

4.1.1 分野間での再現率・適合率の違い

表 4 より,「読書」,「科学」の適合率が,「文化」の再現率が他分野に比べて低い.「科学」は少数の単語の出現頻度が高い．こういった分野は少数の専門的な単語が多く出現すると考えられる．また「読書」,「文化」は他分野にも出現する単語が多い．このような分野は一般的な単語が多く含まれると考えられる．こうした分野に対しては，分野の特徴ベクトルを構成する重要語集合の一部の単語のみが特徴的であったり，全単語とも特徴的でなかったと考えられる.

4.1.2 提案手法の効果

表 4 より，適合率・再現率ともに向上した「芸能」,「スポーツ」などは，重要語集合において出現頻度が低い単語でも偏りが大きかった．提案手法により，そういった特徴的な単語を重要語として多く抽出できた結果と言える.

また「科学」については，単純な出現頻度だけでなく，χ 2 乗値も少数の専門的な単語の値のみ高くなった．このためこうした単語に重みが偏ってしまい，再現率は向上したが，他の分野へ分類されるべき文書もこうした単語の重みにより,「科学」に分類され，適合率は低下したと考えられる.

5. むすび

本研究では,「出現頻度からみた語の重要性」と「出現頻度の偏りからみた語の重要性」の両視点から分野を特徴付ける単語を利用し，分類誤り率を低減させる手法を提案し有効性を示した.

今後は単語の重み付け方法の検討と，今回重要語とした単語との共起を考え，出現頻度は低いが分野を特徴付けている単語の抽出などを考えたい.

参考文献

[1] 徳永健伸，情報検索と言語処理, 財団法人東京大学出版会, 1999.

[2] 呉勇，山田祥，岸本陽次郎, “ 名詞頻度を使った分類用辞書の構築と評価 ”，電子情報通信学会論文誌，No.2，Vol.J84-D-I，pp.213-221，2001.

[3] 河合敦夫, “ 意味属性の学習結果にもとづく文書自動分類方式 ”，情報処理学会論文誌，No.9，Vol.33，pp.1114-1122，1992.

[4] 毎日新聞社，CD-毎日新聞'94, 日外アソシエーツ，1995.