# 文書分類を対象としたダイス係数に基づくカテゴリ情報付き重み付け法に関する研究

1X08C119-0　三木 智広
指導教員　後藤 正幸

## 1　研究背景・目的

近年, 情報技術の発展と普及に伴い,WWW や電子図書館においても膨大な文書が蓄積され, 効率的な知識獲得のため, より高精度な文書の自動分類が必要とされている.

一般的に，文書の分類問題を対象とした研究は大きく分けて特徴空間の構成法，分類器の構築法の二つの視点で行われていることが多い．前者では，文書を分類する際に有効な単語を抽出し，単語に重みを付与する方法が研究され，後者では分類器を構成する方法が研究されている．

現状，より高精度な分類器を構築しようとする研究が多い中，単語に対して重みを付与する方法についての研究も注目されている．重みとは単語の重要度であり，重みの付与を適切に行うことで分類精度の向上にも繋がる．そこで，本研究では, この重み付け方法に着目する．

従来，重み付けの方法として，カテゴリに関する情報を用いた頻度差分量 [1] や PWI[2] などが提案されている．しかし，これらの手法は単語共起を考慮していない．単語共起とは，任意の文書や文において，ある単語とある単語が同時に出現することである．単語の共起関係の重要性は，情報検索の分野ですでに明らかになっており，単語間の共起の程度を測る尺度としてダイス係数 [3] が報告されている. しかし，ダイス係数自体は単語間に付与される尺度であり，1 つの単語の重みを与えていない．また，カテゴリに関する情報も考慮されていない．本研究で対象とする文書分類問題では，学習データのカテゴリ情報が予め与えられているため，その有効活用により，より優れた重み付け手法を構築できる可能性がある．そこで本研究では，カテゴリ情報を用いた単語の共起による重み付け法を提案すると共に，提案手法が分類性能の改善につながることを示す.

## 2　従来手法と本研究への展開

### 2.1　ダイス係数

ダイス係数 [3] とは，2 つの単語間の共起関係を，各単語の出現文書数と共起文書数を用いて定量化する方法である．単語 $w_i$ の出現文書数を $df(w_i)$，単語 $w_i$ と単語 $w_j$ の共起頻度を $df(w_i, w_j)$ としたとき，ダイス係数 $Dice(w_i, w_j)$ は

$$Dice(w_i, w_j) = 2 \times \frac{df(w_i, w_j)}{df(w_i) + df(w_j)}, \tag{1}$$

で定義される．ダイス係数は，前述の通り情報検索の分野で考案された方法で，単語間の共起関係を測るものであり，カテゴリに関する情報が付与されていない.

### 2.2　頻度差分量

文書分類問題に適した重み付けの方法として頻度差分量 [1] が報告されている．頻度差分量では，まず学習フェーズにおいて，所属カテゴリが与えられている学習用文書集合に対してカテゴリ別に単語 $w_i$ の出現頻度を求める．いま，$c_k$ をある一つのカテゴリ，$K$ を総カテゴリ数として，カテゴリ集合を $\mathcal{C} = \{c_1, ..., c_K\}$ と定義する．また，$tf(w_i)$ を学習用文書における単語 $w_i$ の出現頻度，$tf_p(w_i)$ を $p$ 番目に出現頻度が高いカテゴリにおける単語 $w_i$ の出現頻度とすると，頻度差分量 $NDF(w_i; \mathcal{C})$ は，式 (2) のように定義される．

$$NDF(w_i; \mathcal{C}) = \frac{tf_1(w_i) - tf_2(w_i)}{tf(w_i)}. \tag{2}$$

また，$NDF$ では，各単語の所属カテゴリを一意に定め，それを基に文書分類を行う．$tf(w_i, c_k)$ を，カテゴリ $c_k$ に含まれる単語 $w_i$ の延べ出現頻度とし，各語の所属カテゴリは

$$\hat{k}_i = \arg\max_k tf(w_i, c_k). \tag{3}$$

となる．次に，分類フェーズにおいて新たなカテゴリが未知の文書 $d$ の分類基準を式 (4) で定める．

$$\begin{aligned} \hat{c} &= \arg\max_{c_k} \sum_{w_i | k=\hat{k}_i} f_d(w_i) NDF(w_i; \mathcal{C}) \\ &= \arg\max_{c_k} \sum_{w_i | k=\hat{k}_i} f_d(w_i) \frac{tf_1(w_i) - tf_2(w_i)}{tf(w_i)}. \end{aligned} \tag{4}$$

ただし，$f_d(w_i)$ を，カテゴリが未知の文書 $d$ に含まれる単語 $w_i$ の延べ出現頻度とする．(4) 式ではカテゴリが未知の文書 $d$ に含まれる各単語 $w_i$ の出現頻度と頻度差分量の積を，単語 $w_i$ の所属カテゴリごとに和の形で蓄積させている．これにより，Naive Bayes 法 [4] などで問題となるゼロ頻度問題の影響を受けないという利点がある.

## 3　提案手法

前述の通り，ダイス係数は情報検索の分野で考案された方法であるためカテゴリ情報は考慮されておらず，重み付け手法としてそのまま用いることはできない．そのため，本研究では，カテゴリ情報を用いてダイス係数の重み付け手法への拡張を与える．単語の共起情報を考慮するダイス係数の和をカテゴリ毎に求めることにより，あるカテゴリにおいて特徴的に出現する単語に重みを付与できると考える.

### 3.1　カテゴリ情報付きダイス係数への拡張

本研究では，カテゴリ情報付きダイス係数 $cDice(w_i|c_k)$ を式 (5) により定義する.

$$cDice(w_i|c_k) = 2 \times \frac{\sum_j Dice(w_i, w_j|c_k)}{F_{c_k}(w_i)}. \tag{5}$$

(5) 式の $\sum_j Dice(w_i, w_j|c_k)$ では，単語ごとの特徴量を算出するため，同一カテゴリ内における，ある単語とその他全ての単語についてのダイス係数を算出し，その和をとっている．他の多くの単語と共起する単語は，その文章が論じているトピックの代表的なキーワードを表すので，重要度として適切である．しかし，共起する単語の種類数が増えるにつれ $\sum_j Dice(w_i, w_j|c_k)$ は大きくなる傾向にあるので，$F_{c_k}(w_i)$ で割ることにより規準化し，これにより，ある単語の重要度を定める．ここで，$F_{c_k}(w_i)$ は，カテゴリ $c_k$ における同一文書内で単語 $w_i$ と共起する異なり単語種類数とする．また，ダイス係数ではカテゴリ情報は考慮されておらず，文書分類問題の特性を活かせていない．そのため，カテゴリが付与された学習用文書を用い，各カテゴリにおける各単語の重みを算出する.

### 3.2　分類方法

カテゴリが未知の文書 $d$ をカテゴリ $c_k$ へ分類するための分類基準を (6) 式で定める．ここでは，ゼロ頻度問題の影響を受けない鈴木の分類方法 [1] を用いる．

$$\begin{aligned}\hat{c} &= \arg\max_{c_k} \sum_i f_d(w_i) cDice(w_i|c_k) \\ &= \arg\max_{c_k} \sum_i f_d(w_i) \frac{2 \times \sum_j Dice(w_i, w_j|c_k)}{F_{c_k}(w_i)}.\end{aligned} \tag{6}$$

(6) 式により，分類対象文書 $d$ について各カテゴリ $c_k$ のカテゴリ評価値を算出し，文書 $d$ をカテゴリ $\hat{c}$ へ分類する．

## 4　評価実験

新聞記事データを用いた文書分類を行い，提案手法の有効性を検証する．

### 4.1　実験 1-実験方法

提案手法の有効性を検証するために, 各手法を用いて重みの高い単語，上位 10 単語を抽出する評価実験を行った．また，抽出された単語の頻度の比較に上位 10 単語の平均値と中央値を用いる．データは，読売新聞 2000 年の記事データを用いた．カテゴリは政治を使用した．合計 500 件の記事をランダムに抽出し，これを実験データとする．

実験データに対し, 以下の 3 つの手法を適用した．

- 頻度差分量 [1]　(NDF)
- PWI[2]　(PWI)
- カテゴリ情報付きダイス係数　(提案手法)

### 4.2　実験 2-実験方法

提案手法の有効性を検証するために, 各手法を用いて分類精度の評価実験を行った．データは，読売新聞 2000 年，毎日新聞 2000 年の記事データを用いている．カテゴリは政治, スポーツ, 犯罪・事件, 生活, 社会, 科学, 経済, 文化の 8 カテゴリを使用した．各カテゴリ 500 件ずつ, 合計 4000 件の記事をランダムに抽出し，これを学習データとする．同様に，各カテゴリ 50 件ずつ，合計 400 件をテストデータとする．

実験データに対し, 以下の 3 つの手法を適用した．

- 頻度差分量を重みとして用いた手法 [1]　(NDF)
- PWI を重みとして用いた手法 [2]　(PWI)
- カテゴリ情報付きダイス係数　(提案手法)

### 4.3　実験結果及び考察

実験 1，実験 2 の結果はそれぞれ表 1，図 1 の通りである．

**表 1．実験 1-単語の重みの比較**

| | PWI | | NDF | | 提案手法 | |
|---|---|---|---|---|---|---|
| 順位 | 単語 | 頻度 | 単語 | 頻度 | 単語 | 頻度 |
| 1 | 議員 | 2 | 選 | 921 | 参院 | 141 |
| 2 | 入党 | 10 | 選挙 | 758 | 公明党 | 127 |
| 3 | 見込める | 4 | 自民党 | 393 | 県連 | 118 |
| 4 | 省庁 | 124 | 投票 | 344 | 出馬 | 105 |
| 5 | 脱出 | 6 | 候補 | 460 | 補選 | 85 |
| 6 | 下院 | 2 | 拡大 | 140 | 擁立 | 69 |
| 7 | 先行き | 11 | 参院 | 141 | 議案 | 59 |
| 8 | 国会 | 125 | 公明党 | 127 | 投開票 | 53 |
| 9 | 外交 | 9 | 導入 | 120 | 社民 | 52 |
| 10 | 法令 | 5 | 法人 | 118 | 保守党 | 37 |
| 平均値 | **29.8** | | **352.5** | | **84.6** | |
| 中央値 | **7.5** | | **242.5** | | **77.0** | |

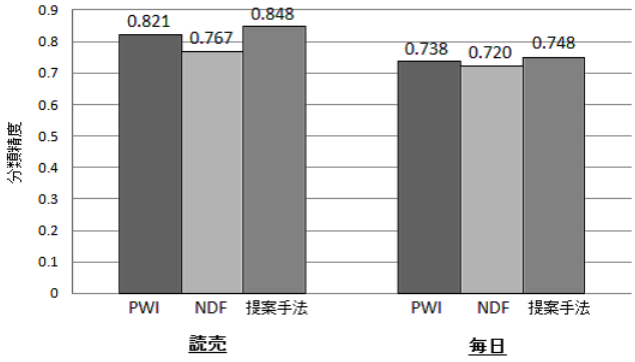


**図 1．実験 2-分類精度**

実験 1 の結果より，PWI の中央値は 7.5，NDF の中央値は 242.5, 提案手法の中央値は 77.0 となった．また，PWI の平均値は 29.8，NDF の平均値は 352.5, 提案手法の平均値は 84.6 となった．PWI は高い頻度の単語も抽出されているが，相互情報量を用いているため，頻度の低い単語の重みが大きくなっていることが分かる．NDF では他と比べ比較的高い頻度の単語の特徴量が高くなっていることから，頻度情報の影響によって重みが大きくなっていることが分かる．他の手法と比べ提案手法では，そのカテゴリに連想されるような単語の重みを大きくすることができていることが分かる．これは，提案手法では単語の共起情報を用いることで，頻度情報の影響を受けすぎず，政治カテゴリを特徴付ける単語を取り出せているからだと考えられる．

また，実験 2 の結果より，PWI や頻度差分量と提案手法を比較すると，提案手法の分類精度が全ての実験で優れていることが分かる．分類精度の向上の要因として，単語の共起情報に対しカテゴリの情報を用いることで，カテゴリを特徴づける単語に対し大きい重みを付与できているからだと考えられる．

## 5　まとめと今後の課題

本研究では, カテゴリ別ダイス係数に基づく単語の重み付け手法を提案した．提案手法を実際の新聞記事データの分類に適用して実験を行った結果，従来手法よりも分類精度が向上することを明らかにし，その有効性を示すことができた．

今後の課題は, 本提案を用いて Web ページや論文などの他の文書データに対する有効性を検証することを考えている．

## 参考文献

[1] 鈴木 誠,“ カテゴリ間の単語頻度の差分を用いたテキストの自動分類, ”日本経営工学会論文誌, vol59, No.4, pp.355-362 (2008).

[2] A. Aizawa. “ The feature quantity: An Information Theoretic Perspective of Tfidf-like Mesures,” *Proc. of the 23th ACM Inc. Conf. on Research and Development in Information Retrieval*, pp.104-111 (2000).

[3] 岸田 和明,“ 文書検索におけるクエリーの拡張方法：大域的分析と局所的分析の実証比較, ” 情報処理学会研究報告, vol67, pp.55-62 (2001).

[4] 花井 拓也, 山村 毅,“ 単語間の依存性を考慮したナイーブベイズ法によるテキスト分類, ” 情報処理学会研究報告,vol.2005,No.22, 2005-NL-166-14 , pp.101-106(2005)