トピック特有因子分析法と文書分類への応用

川 谷 隆 彦†

高精度な文書分類を図るには，各クラスに特有な特徴を抽出して使用することが重要である．本論文では，与えられた 2 つの文書集合の一方が他方に対して有する特有な話題の分析方法を提案する．各文書を文ベクトルの集合で表現したとき，提案手法では，全文ベクトルの射影値の 2 乗和に関する着目文書集合と他方の文書集合との比を最大にするような射影軸を一般固有値問題の固有ベクトルとして求め，着目文書集合に特有な話題を表す因子とする．このような因子をある着目クラスの文書集合と，既存の分類系で着目クラスに誤分類された文書集合との間で求めることにより，着目クラスでは出現するが他のクラスでは出現しにくい特徴，反対に他のクラスでは出現するが着目クラスには出現しにくい特徴の抽出に用いることができる．既存の分類系を kNN とし，このような特徴を用いる分類系を併用した結果，Reuters-21578 のテストデータに対する F 値は kNN 単独の 83.69%から 87.27%に向上した．

Topic Distinctiveness Factor Analysis and its Application to Text Categorization

Takahiko Kawatani†

To improve performance in text categorization, it is important to extract distinctive features for each class. This paper proposes a method to extract topic distinctiveness factors that a given document set possesses compared to another document set. Suppose all sentence vectors that compose each document are projected onto projection axes. The method obtains the projection axes that maximize the ratio between the document sets as to the sum of squared projections by solving a generalized eigenvalue problem. By applying the method to the document set that belongs to a given class and a set of documents that are misclassified as belonging to that class by an existent classifier, we can obtain features that take large values in the given class but small ones in other classes, as well as features that take large values in other classes but small ones in the given class. A classifier was constructed applying the above features to complement the kNN classifier. As the results, the micro-averaged F_1 measure for Reuters-21578 improved from 83.69 to 87.27%.

1. まえがき

近年文書分類の研究がさかんに行われている．これまでに数々の手法が提案されてきたが，Yang らの比較実験によれば[1]，kNN[1)~3)]，サポートベクターマシン（SVM）[1),4)]，LLSF[5)] が優れた性能を示している．そのほか，AdaBoost[6)] も高い性能を有することが報告されている．しかしながら，これらの技術はすでに深く検討されてきており，今後個々の技術の改善で性能を飛躍的に向上させていくのは困難と思われる．さらなる性能の向上には新たなアプローチが必要と考えられる．

ところで，多くの分類法では，文書クラスに関する情報を何らかの形で記述し，入力文書と照合している．これをクラスモデルと呼べば，クラスモデルは，たとえば，ベクトル空間モデルでは各クラスに属する文書の平均文書ベクトルにより，kNN では各クラスに属する文書の文書ベクトルの集合により，AdaBoost では単純な仮説の集合により表現されている．正確な分類を図るにはクラスモデルは各クラスを正確に記述したものでなければならない．現在まで提案されている分類法も高度なものほどクラスモデルは各クラスを正確に記述しているといってよいであろう．しかしながら，多くの分類法ではクラスモデルの記述の正確さは指向しているが，クラスモデルにクラス間の重なりがあることには配慮していないようである．kNN に

† メディアドライブ株式会社
Mediadrive Co.

本研究は，日本ヒューレット・パッカード（株）ヒューレット・パッカード研究所にて行われたものである．

せよ，AdaBoost にせよ，あるクラスのクラスモデルには他のクラスとマッチする情報も含まれてしまっている．クラスモデル間に重なりが存在すれば，ある入力文書とその入力文書が属さないクラスとの間に不必要な類似性が生ずることになり，誤分類の原因となりうる．誤分類の原因を取り除くためには，クラスモデルがクラス間で重ならないよう，各クラスに特有な情報を用いてクラスモデルを記述することが望まれる．

本論文ではこの問題に焦点を当てる．まず 2 つの文書集合の一方が他方に対して有する特有な特徴の抽出を試みる．本論文では，このような特徴を，文書を文ベクトルの集合で表したうえで，全文ベクトルの射影値の 2 乗和に関する両文書集合の比を最大にするように求められた射影軸に文ベクトルを射影することにより求める．これにより，一方の文書集合に着目すれば，その文書集合には現れるが他方の文書集合には現れにくい特徴，その文書集合には現れにくいが他方の文書集合には現れる特徴を求めることができる．上記射影軸は着目する文書集合に特有なトピックを反映するものなので，トピック特有因子（Topic Distinctiveness Factor: TDF），また，その手法をトピック特有因子分析（Topic Distinctiveness Factor Analysis: TDFA）と呼ぶこととする．文献 7)〜9) においてはこれらはトピック差分因子（Topic Difference Factor: TDF），トピック差分因子分析（Topic Difference Factor Analysis: TDFA）と呼ばれていたが，上記因子は文書集合間の差分というより着目文書集合の特有性を表すといった方がより適切と考えられるので上記のように呼ぶこととした．次に，この TDFA を文書分類に応用し，着目クラスには現れるが他クラスでは現れにくい特徴，他クラスでは現れるが着目クラスでは現れにくい特徴を求める．本論文では，既存の分類法に基づくメインの分類系に対する相補的分類系でこのような特徴を用いることを試みる．相補的分類系では，メインの分類系で得られた入力文書の各クラスに対する類似度に対し補正を行う．

以下，2 章では，TDF の求め方，その解釈などについて述べた後，簡単な例について TDF がどのように求められるのかを示す．3 章では，相補的分類系のための TDF の求め方，類似度の補正方法などについて述べる．4 章では，メインの分類系として kNN を採用したときの実験方法と結果について述べ，コーパスとして Reuters-21578 を用いたとき F 値が大幅に向上することを示す．

2. トピック特有因子（TDF）

2.1 アプローチ

2 つの文書集合を考え，一方の文書集合が他方に対して有する特有な特徴の抽出を試みる．本論文では，両方の文書集合の話題は非常に近いことを想定しなければならない．話題の近い 2 つの文書集合の特有な特徴の抽出のためには，両方の文書集合をできるだけ正確に表現する必要がある．文書集合の表現が曖昧であれば，得られる結果も曖昧にならざるをえない．従来は，多くの場合，文書は各単語の文書内の頻度を成分とする文書ベクトルにより表現されていた．この表現では，文書にどのような単語が現れるかは正しく表現される．しかしながら，長い文書の場合，文書ベクトル内でゼロでない成分どうしが文書内で互いに関連があって意味のある単語共起を反映しているとは限らず，文書の表現として曖昧さが残されるように思われる．そこで，本論文では各文書は文の集合からなるとし，文ベクトルの集合で文書を表すこととする．文ベクトルは各単語の文内の頻度，もしくは有無を成分とするベクトルである．各文ベクトル内で値がゼロでない成分どうしは意味のある単語共起となるケースが多くなり，文書内の単語共起に関する情報がより明確になることが期待できる．

図 1 は単語空間における文書集合 D，T の分布，および求めるべき射影軸 $\boldsymbol{\alpha}$ ($||\boldsymbol{\alpha}|| = 1$) を示している．ここで図のように，文書集合 D，T の全文ベクトルを $\boldsymbol{\alpha}$ へ射影したとする．もし，射影軸 $\boldsymbol{\alpha}$ が文書集合 D の特有な特徴を反映していれば，文書集合 D に含まれる文の射影値は大きく，文書集合 T のそれは小さいであろう．そこで，文書集合 D，T の全文ベクトルを $\boldsymbol{\alpha}$ へ射影したときの射影値の 2 乗和をそれぞ

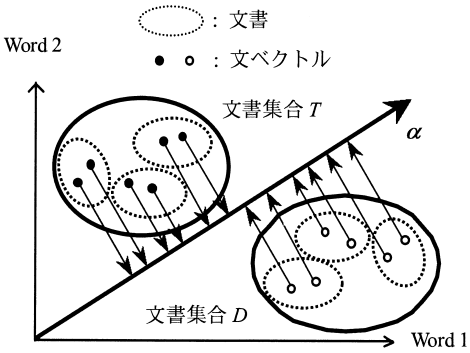

図 1 単語空間における文書，文ベクトルの分布と射影軸 $\boldsymbol{\alpha}$ の例
Fig. 1 Example of distributions of documents and sentence vectors and projections of sentences onto $\boldsymbol{\alpha}$.

れ P_D，P_T として，$\boldsymbol{\alpha}$ に反映される両文書間の違いの程度を表す評価基準 $J(\boldsymbol{\alpha})$ を

$$J(\boldsymbol{\alpha}) = \frac{P_D}{P_T} \tag{1}$$

で定義する．$J(\boldsymbol{\alpha})$ を最大化する $\boldsymbol{\alpha}$ を求めると，文書集合 D の文ベクトルの射影値の 2 乗和は大きく，文書集合 T のそれは小さくなるはずなので，$\boldsymbol{\alpha}$ は文書集合 D には多く存在するが文書集合 T には存在しにくい特徴を反映する射影軸となる．文書集合 D から見ると，この $\boldsymbol{\alpha}$ は存在すべき特徴を反映することになるので，これを文書集合 D の正のトピック特有因子（Positive Topic Distinctiveness Factor: P-TDF）と呼ぶこととする．

また，もう 1 つの求めるべきベクトルを $\boldsymbol{\beta}$ として評価基準 $J(\boldsymbol{\beta})$ を

$$J(\boldsymbol{\beta}) = \frac{P_T}{P_D} \tag{2}$$

により定義し，$J(\boldsymbol{\beta})$ を最大にする $\boldsymbol{\beta}$ を求める．$\boldsymbol{\beta}$ は文書集合 T には多く存在するが文書集合 D には存在しにくい特徴を反映する射影軸となる．$\boldsymbol{\beta}$ は文書集合 T の正のトピック特有因子となるが，文書集合 D から見ると存在すべきでない特徴を反映するので，文書集合 D の負のトピック特有因子（Negative Topic Distinctiveness Factor: N-TDF）と呼ぶこととする．

2.2 トピック特有因子の算出と用い方

文書集合 D を $\{D_1,..,D_M\}$，T を $\{T_1,..,T_N\}$ により表す．さらに，文書 D_m，T_n の k 番目の文ベクトルをそれぞれ $\boldsymbol{d}_{mk}$ $(k = 1,..,C(D_m))$，$\boldsymbol{t}_{nk}$ $(k = 1,..,C(T_n))$ とする．$C(D_m)$，$C(T_n)$ はそれぞれ文書 D_m，T_n の文数である．そうすると，前節における P_D，P_T は以下のように求めることができる．

$$P_D = \sum_{m=1}^{M} \sum_{k=1}^{C(D_m)} (\boldsymbol{d}_{mk}{}^T \boldsymbol{\alpha})^2 = \boldsymbol{\alpha}^T \boldsymbol{S}_D \boldsymbol{\alpha} \tag{3}$$

ただし，$\boldsymbol{S}_D = \sum_{m=1}^{M} \sum_{k=1}^{C(D_m)} \boldsymbol{d}_{mk} \boldsymbol{d}_{mk}{}^T$

$$P_T = \sum_{n=1}^{N} \sum_{k=1}^{C(T_n)} (\boldsymbol{t}_{nk}{}^T \boldsymbol{\alpha})^2 = \boldsymbol{\alpha}^T \boldsymbol{S}_T \boldsymbol{\alpha} \tag{4}$$

ただし，$\boldsymbol{S}_T = \sum_{n=1}^{N} \sum_{k=1}^{C(T_n)} \boldsymbol{t}_{nk} \boldsymbol{t}_{nk}{}^T$

上付きの添字 T は転置を表す．また，行列 $\boldsymbol{S}_D$，$\boldsymbol{S}_T$ は文書集合 D，T の平方和行列と呼ぶこととする．式 (3)，(4) を用いると，評価基準 $J(\boldsymbol{\alpha})$，$J(\boldsymbol{\beta})$ は

$$J(\boldsymbol{\alpha}) = \frac{P_D}{P_T} = \frac{\boldsymbol{\alpha}^T \boldsymbol{S}_D \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{S}_T \boldsymbol{\alpha}} \tag{5}$$

$$J(\boldsymbol{\beta}) = \frac{P_T}{P_D} = \frac{\boldsymbol{\beta}^T \boldsymbol{S}_T \boldsymbol{\beta}}{\boldsymbol{\beta}^T \boldsymbol{S}_D \boldsymbol{\beta}} \tag{6}$$

で表されることになる．式 (5)，(6) は実は線形判別分析における評価基準と形式的に同じであり，$\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ の解も線形判別分析と同じように，それぞれ

$$\boldsymbol{S}_D \boldsymbol{\alpha} = \lambda \boldsymbol{S}_T \boldsymbol{\alpha} \tag{7}$$

$$\boldsymbol{S}_T \boldsymbol{\beta} = \lambda \boldsymbol{S}_D \boldsymbol{\beta} \tag{8}$$

なる一般固有値問題の固有ベクトルで与えられる[10)~12)]．あるいは，$\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ はそれぞれ，行列 $\boldsymbol{S}_T{}^{-1}\boldsymbol{S}_D$，$\boldsymbol{S}_D{}^{-1}\boldsymbol{S}_T$ の固有ベクトルで与えられるといってもよい．

式 (7) の場合を考える．式 (7) の解として得られる固有値を $\lambda_1, \lambda_2, .., \lambda_i$ $(\leq \lambda_{i+1}), ..,$ 固有値 λ_i に対応する固有ベクトルを $\boldsymbol{\alpha}_i$ とする．$\lambda_i = \boldsymbol{\alpha}_i{}^T \boldsymbol{S}_D \boldsymbol{\alpha}_i / \boldsymbol{\alpha}_i{}^T \boldsymbol{S}_T \boldsymbol{\alpha}_i$ なので，λ_i は $\boldsymbol{\alpha}_i$ を用いたときの評価基準 $J(\boldsymbol{\alpha}_i)$ の値そのものである．したがって，λ_i は因子 $\boldsymbol{\alpha}_i$ に関する文書集合 D の特有さの程度と解釈できる．式 (8) の場合も同様である．

式 (3)，(4) においては，文の長さの違いの影響を排除するため，$\boldsymbol{d}_{mk}$，$\boldsymbol{t}_{nk}$ を $\hat{\boldsymbol{d}}_{mk} = \boldsymbol{d}_{mk}/||\boldsymbol{d}_{mk}||$，$\hat{\boldsymbol{t}}_{nk} = \boldsymbol{t}_{nk}/\|\boldsymbol{t}_{nk}\|$ のように正規化して用いてもよい．この場合には，評価基準は各文ベクトルと $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ との類似度の 2 乗和に関する両文書集合の比と等価となる．

上記のように求められる TDF は特徴変換のために用いることになるが，これには以下のような特性がある．文ベクトル $\boldsymbol{x}$ に次式を適用し，求められた固有ベクトル $\{\boldsymbol{\alpha}_i\}$ もしくは $\{\boldsymbol{\beta}_i\}$ が張る空間 Z に写像したとする．

$$z_i = \boldsymbol{\alpha}_i{}^T \boldsymbol{x} \tag{9}$$

$$z_i = \boldsymbol{\beta}_i{}^T \boldsymbol{x} \tag{10}$$

ここで，文書集合 D，T の文ベクトルは原単語空間で**図 2** (a) のように分布しているとする．式 (9) による写像の場合，空間 Z では図 2 (b) のように，各文ベクトルは文書集合 T では原点の周りに，文書集合 D では原点から離れた領域に分布するようになる．一方，式 (10) による写像の場合，空間 Z では図 2 (c) のように各文ベクトルは文書集合 D では原点の周りに，文書集合 T では原点から離れた領域に分布するようになる．このように TDFA は 2 つの文書集合の分離を容易にするような特徴変換である．

2.3 正 則 化

文書集合 D の P-TDF を求める場合を考える．P-

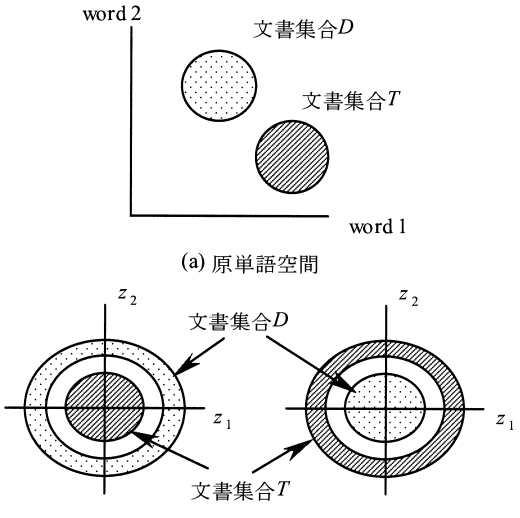

図 2 写像前後の文書 D, T の文の分布
Fig. 2 Sentence distributions of document set D and T before and after projection.

TDF は $\boldsymbol{S}_T{}^{-1}\boldsymbol{S}_D$ の固有ベクトルで与えられるので，P-TDF が求められるためには行列 $\boldsymbol{S}_T$ に対して逆行列が存在しなければならない．それには行列 $\boldsymbol{S}_T$ は正則行列である必要がある．しかし，実際には文書数が単語数よりも少ない，あるいは特定の単語対がつねに共起するような場合には $\boldsymbol{S}_T$ は正則行列として求められない．また，文書集合 D では出現し，文書集合 T では出現しない単語（文書集合 D の特異単語と呼ぶ）が存在する場合も問題となる．このような場合，$\boldsymbol{S}_T$ を正則化する必要があるが，その方法として次のような方法が知られている[14),15)]．すなわち，σ^2 をパラメータ，I を単位行列として

$$\hat{\boldsymbol{S}}_T = \boldsymbol{S}_T + \sigma^2 I \tag{11}$$

を $\boldsymbol{S}_T$ として用いる方法である．

式 (11) は $\boldsymbol{S}_T$ の対角成分に σ^2 を加えることを意味する．各単語に対し，その単語に対応する成分のみ σ，他は 0 となる単語ベクトルを用意したとする．$\boldsymbol{S}_T$ の対角成分に σ^2 を加えることは，全単語ベクトルを文書集合 T に加えることを意味し，評価基準は，全単語ベクトルが文書集合に加わるので，式 (1) ではなく，

$$J(\boldsymbol{\alpha}) = P_D/(P_T + \sigma^2) \tag{12}$$

としたことに相当する．また，これにより文書集合 T は文書集合 D の有する特異単語を σ 個含むことになる．この場合，σ の値を大きく設定すると，文書集合 D の特異単語も見かけ上文書集合 T で多く存在することになるので，特異単語の TDF への寄与は小さくなる．その結果，文書集合 D で頻度が高く，真に特有性の高い特異単語が過小評価されることになる．反対に σ の値を小さく設定すると，特異単語の TDF への寄与は大きくなる．その結果，文書集合 D で頻度が低く雑音と見なされうる特異単語が過大に評価されるようになる．したがって，σ の値を適切に設定することは重要な問題となる．

2.4 例

ここでは簡単な例により TDF がどのように求められるのかを見る．いま**表 1** に示すように，文書集合 D は D_1 と D_2，T は T_1 と T_2 の各々2 個の文書で構成され，各文書は 2 つの 5 次元文ベクトルによって表されるものとする．文書集合 D，T の相違点，共通点は以下のとおりである．

- 文書集合 D では文書 D_1 の文 1 で単語 5 が現れるが，文書集合 T では現れない．
- 文書集合 D では文書 D_2 の文 1 で単語 2 と 3，文 2 で 1 と 4 が共起する．
- 文書集合 T では文書 T_2 の文 1 で単語 1 と 3，文 2 で 2 と 4 が共起する．
- 単語 1 と 2，および 3 と 4 の共起は共通である．

このような文書集合に対して $\sigma^2 = 0.1$ として TDF を求めてみた．**表 2** に，文書集合 D の P-TDF として式 (7) の固有値 λ_n $(n = 1, .., 5)$ と各固有値に対応する固有ベクトル $\boldsymbol{\alpha}_n = (\alpha_{n1}, .., \alpha_{n5})^T$ の各成分を示す．同様に，**表 3** に N-TDF として式 (8) の解を与える固有値 μ_n $(n = 1, .., 5)$ と各固有値に対応する固有ベクトル $\boldsymbol{\beta}_n = (\beta_{n1}, .., \beta_{n5})^T$ を示す．また，**表 4** には各文ベクトルを式 (9)，(10) により射影したときの射影値（$n = 1, 2$）を示す．これらから以下がいえる．

(1) 表 2 において固有ベクトル $\boldsymbol{\alpha}_1$ では，α_{11} と α_{14} がともに負，α_{12} と α_{13} がともに正の値をとっており，文書集合 D における単語 2 と 3，および 1 と 4 の共起が反映されていることが分かる．そのため，表 4 からも分かるように，文書 D_2 の文ベクトルの $\boldsymbol{\alpha}_1$ への射影値の絶対値は大きくなっている．また，文書 D_2 以外の文書の文ベクトルの射影値は 0 となっている．

(2) 表 2 における $\boldsymbol{\alpha}_2$ は α_{25} のみが大きな値を持ち，文書 D における単語 5 の存在を反映している．事実，表 4 で文書 D_1 の文 1 からの射影値のみが大きな値をとっている．

(3) 以下同様に，表 3 における $\boldsymbol{\beta}_1$ には文書集合 T における単語 1 と 3，および 2 と 4 の共起が反映されている．そのため，表 4 で文書 T_2 の文

表 1 想定する文書集合 D, T
Table 1 Sentence vectors in document set D and T.

文書	文	文ベクトル	文書	文	文ベクトル
D_1	1	11001	T_1	1	11000
	2	00110		2	00110
D_2	1	01100	T_2	1	10100
	2	10010		2	01010

表 2 文書集合 D の P-TDF
Table 2 P-TDF's of document set D.

n	λ_n	α_{n1}	α_{n2}	α_{n3}	α_{n4}	α_{n5}
1	20.00	-0.50	0.50	0.50	-0.50	0.00
2	10.74	0.04	0.04	-0.01	-0.01	1.00
3	0.97	-0.01	-0.01	-0.71	-0.71	0.02
4	0.22	-0.39	-0.39	0.13	0.13	0.81
5	0.00	0.50	-0.50	0.50	-0.50	0.00

表 3 文書集合 D の N-TDF
Table 3 N-TDF's of document set D.

n	μ_n	β_{n1}	β_{n2}	β_{n3}	β_{n4}	β_{n5}
1	20.00	0.50	-0.50	0.50	-0.50	0.00
2	2.78	-0.43	-0.43	0.14	0.14	0.77
3	0.97	0.01	0.01	-0.71	-0.71	-0.02
4	0.00	-0.31	0.31	0.31	-0.31	0.79
5	0.00	0.28	-0.28	-0.28	0.28	0.83

表 4 各文ベクトルの射影値
Table 4 Projections of each document.

文書	文	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
D_1	1	0.00	1.07	0.00	-0.08
	2	0.00	-0.02	0.00	0.27
D_2	1	1.00	0.02	0.00	-0.29
	2	-1.00	0.02	0.00	-0.29
T_1	1	0.00	0.07	0.00	-0.85
	2	0.00	-0.02	0.00	0.27
T_2	1	0.00	0.02	1.00	-0.29
	2	0.00	0.02	-1.00	-0.29

からの射影値のみが値を有している．

(4) 表 3 における $\boldsymbol{\beta}_2$ には，文書 D_1 の文 1 の文ベクトルと文書 T_1 の文 1 の違いが反映され，文書 T_1 の文 1 からの射影値が大きな値をとっている．

(5) 表 2，表 3 における固有ベクトル $\boldsymbol{\alpha}_n$ $(n > 2)$，および $\boldsymbol{\beta}_n$ $(n > 2)$ は対応する固有値が小さく，TDF として有効ではない．

これらの観察により，

- TDF により 2 つの文書集合間の単語の出現傾向の違いだけでなく，単語間の共起傾向の違いも反映されること，
- 図 1 に示されるような写像が実際に得られること，

が分かった．前者には文書を文ベクトルの集合として表現したことの効果が現れている．この例では，もし従来のように文書を単一の文書ベクトルで表した場合には，文書 D_1 で $(1\ 1\ 1\ 1\ 1)^T$ となる以外は他のすべての文書で $(1\ 1\ 1\ 1\ 0)^T$ となり，単語 5 に関する出現傾向の違いは分かるものの，単語 1〜4 における単語共起の違いは求められない．このように文書の文ベクトルの集合による表現を活かせたのは，平方和行列のような 2 次統計量を導入したことにある．2 次統計量では文または文書ベクトルの各成分のクロスタームが生じ，自然な形で単語の共起が反映される．

3. 文書分類への応用

3.1 アプローチ

1 章で述べたように，本論文では TDF のみを用いた分類系の構築はねらわずに，既存の分類法を用いたメインの分類系に対する相補的認識系の中で TDF を用いるようにしている．その理由は以下のとおりである．

(1) クラス l に属する文書集合を文書集合 D，クラス l 以外に属する文書集合を文書集合 T として TDF を求める場合を考える．TDF のみを用いる分類系を構築しようとすると，文書集合 T にはクラス l 以外に属する全文書を含めなければならない．そうすると，数としては文書集合 T に含まれる文書の方が圧倒的に多くなり，かつクラス l とは類似性を持たない文書も多く含まれるようになるため，クラス l に非常に紛らわしい他のクラスに属する文書とクラス l との違いが TDF に正しく反映できるかどうか疑問である．メインの認識系でクラス l に誤った，もしくは誤りそうになった文書集合を文書集合 T とした方が有効な TDF が求められると考えられる．

(2) TDF のみを用いて分類系を構成するよりも，性能の高い既存の分類法と組み合わせた方がより容易に高い性能を実現できると考えられる．

このような考えから，本論文では，メインの分類系では何らかの類似度尺度を用いて分類されていることを前提に，相補的分類系では，各クラスの類似度に対し，そのクラスに現れるべき特徴が入力文書に現れた場合にゲインを，現れるべきでない特徴が現れた場合

にペナルティを与えるようにした．

3.2 相補的分類系

クラス l における TDF の求め方についてまず述べる．まず，メインの認識系においてすべての訓練用文書の分類を行い，各訓練データについてクラス l に対する類似度を求める．ついで，閾値 γ をしかるべき方法で決めておき，上記類似度が γ 以上の文書から，クラス l に属する文書の集合 D，他のクラスに属する文書の集合 T を求める．文書集合 T はクラス l に近い文書の集合であり，このような文書を対抗文書と呼ぶこととする．クラス l の P-TDF は式 (7) の解を与える固有ベクトル $\{\boldsymbol{\alpha}_n\}$，N-TDF は式 (8) の解を与えるベクトル $\{\boldsymbol{\beta}_n\}$ によって与えられる．

また，文ベクトルの集合 $\{\boldsymbol{x}_1,..,\boldsymbol{x}_K\}$ で与えられる入力文書 X に対するクラス l のゲインを $g(X)$，ペナルティを $p(X)$ とすると，これらは，

$$g(X)=\sum_{i=1}^{L_G}\sum_{k=1}^{K}(\boldsymbol{x}_k{}^T\boldsymbol{\alpha}_i)^2 \tag{13}$$

$$p(X)=\sum_{i=1}^{L_P}\sum_{k=1}^{K}(\boldsymbol{x}_k{}^T\boldsymbol{\beta}_i)^2 \tag{14}$$

により与えられる．L_P，および L_G は $p(X)$，$g(X)$ において何個の固有ベクトルを用いるかを示すパラメータであり，最適な値は実験的に決める必要がある．メインの分類系における入力文書 X のクラス l に対する類似度を $sim(X)$ とすると，補正後の類似度 $sim_C(X)$ は

$$sim_C(X)=sim(X)+ag(X)-bp(X) \tag{15}$$

により与えられる．ここで，a，b は値が正のパラメータであるが，これらの値も実験的に決める必要がある．$sim_C(X)$ の算出は $sim(X)$ が閾値 γ より大きいときにのみ行われ，γ 以下であれば無条件に入力文書 X はクラス l に属さないと判断される．$sim_C(X)$ が閾値 δ よりも大きければ，入力文書はクラス l に帰属すると判定される．

なお，式 (13)，(14) の定義では，長い文ほど $g(X)$，$p(X)$ の値が大きくなりがちなので単語数で正規化することも考えられる．また，式 (3)，(4) で $\boldsymbol{d}_{mk}$，$\boldsymbol{t}_{nk}$ を正規化して TDF を算出するときには，式 (13)，(14) においても $\boldsymbol{d}_{mk}$，$\boldsymbol{t}_{nk}$ の正規化ベクトルを用いる必要があるほか，文書の長さの影響を避けるため，文の数で正規化することも考えられる．

4. 分類実験

4.1 実験条件

文献 1) は種々の文書分類法の比較を行った論文としてよく知られている．そこで，本論文においても実験条件を文献 1) にほぼ合わせるようにし，実験結果を比較できるようにした．用いられたコーパスは Reuters-21578 である．訓練データ，テストデータの振り分けは ApteMod に従った．結局，実験に用いた文書は，90 クラスで訓練データ 7,770 文書，テストデータ 3,019 文書であった．Reuters-21578 には複数のラベル（帰属するクラス）が付与された文書が少なからず存在し，ここでも 1 つの文書の複数のクラスへの帰属を許容するマルチラベルの分類実験を行った．

また行った主な前処理は，文切り出し，各単語の活用の基本形への変換（lemmatization），ストップワード除去，単語選択である．文切り出しは，本論文では文書を文ベクトルの集合として扱うことから必要になるもので，通常はピリオドを文の境界として切り出しを行った．しかし，コーパスの中には単語が一定間隔で並んだ表のような文書があり，そのような文書に対しては 1 行を 1 つの文と見なして強制的に切り出した．単語選択については，χ^2 統計量を用いた手法[16]により行い，文献 1) に合わせ 2,500 単語を選択した．また，単語 i に対する重み w_i は，*tf-idf* に基づき以下のように決定した．

$$w_i=(1+\log f_i)\log(N_D/n_i) \tag{16}$$

ここで，f_i は単語 i の着目文における頻度，n_i は単語 i の現れる文書数，N_D は文書の総数である．

4.2 実験方法

メインの分類系としては，処理が単純で高い性能が得られることで知られている kNN を選択した．kNN では，まず，入力文書はすべての訓練文書との間で類似度（余弦類似度）が求められ，類似度が大きい k 個のデータが選択される．入力文書とクラス l との類似度 $sim(X)$ は，選択された k 個のデータのうちクラス l に属する訓練文書の入力文書間との余弦類似度の総和で与えられる[1]．k の値は文献 1) に従い，45 とした．さらに，$sim(X)$ があらかじめ決められた閾値よりも大きければ入力文書はクラス l に帰属すると決定する．この閾値はクラスごとに全訓練データを用いて F 値が最大になるように決定した．再現率（着目クラス l に帰属する文書のうち，クラス l に帰属すると決定された文書の割合）を r，精度（着目クラス l に帰属すると決定された文書のうち，クラス l

表 5 Reuters-21578 に対する分類結果の比較
Table 5 Performance comparison for Reuters-21578.

分類法	再現率	精度	F値
SVM	81.20	91.37	85.99
kNN	83.39	88.07	85.67
kNN(本論文)	81.57	85.93	83.69

に帰属する文書の割合）を p として，F 値の定義は $F = 2rp/(r+p)$ を用いた．**表 5** に，文献 1) に述べられている kNN，SVM のテストデータに対する再現率，精度，F 値（%），および本論文でのそれらを示す．表から分かるように，本論文での kNN の結果は文献 1) の結果よりも F 値が約 2%劣っている．

3.2 節における文書集合 D，T を求めるための閾値 γ としては各クラスで上記 F 値を最大にする閾値よりも低く設定した．文書集合 D は，クラス l の場合，クラス l に属しかつクラス l との類似度が γ 以上の文書となるので，クラス l の中で類似度の低い例外的な文書を除外することができる．これは，例外的な文書の TDF への影響の排除をねらいに行ったものである．対抗文書の集合 T はクラス l に誤った，もしくは誤りそうになった文書から構成されることになる．Reuters-21578 はマルチラベルなので，複数のラベルを持つ文書はそれぞれのクラスの文書集合に属することになる．また，複数のクラスに紛らわしい文書は複数のクラスの対抗文書に属することになる．

また，式 (13)，(14) における L_P，L_G，式 (15) における a，b の決定には次のような問題があった．これらのパラメータの値は，本来は TDF を求めた訓練データやテストデータとは異なる第 3 のデータを用いて決定すべきである．しかし，Reuters-21578 にはそのようなデータは用意されていない．訓練データを用いて決定することも考えられるが，TDF は訓練データを用いて求められるので訓練データにチューンされている．そのような TDF を用いてさらに訓練データを用いてパラメータの値を決定すると，これらの値は訓練データに 2 重にチューンされてしまうことになる．そのため，上記のようにパラメータの値を決定してテストデータの評価に用いても真の評価にはならないと考えられる．そこで，本論文ではテストデータを用いて交差検定を行うこととした．具体的には，テストデータを N 分割し，$N-1$ 組のデータをパラメータ決定用データに用い，残り 1 組を真のテスト用データに用いた．そして，データを回転させながら N 回の実験を行い，テスト用データに対する結果の総計をテストデータ全体に対する結果とした．このような実験では，$N-1$ 組のデータと残り 1 組とは独立なので，結果はテストデータにチューンされないはずである．また，テストデータ中の各文書は必ず 1 回真のテスト用データになるので，得られた結果はテストデータ全体に対する実力を表すものとなる．

上記のパラメータの決定は具体的には以下のように行った．各クラスごとに，まず，L_P，L_G に 15 以内で適当な値を与えた後，式 (15) における a，b を線形判別分析[10)~12)] を用いて決定し，閾値 δ を F 値が最大になるように決定する．線形判別分析は，各文書を $sim(X)$，$g(X)$，$p(X)$ の各値を要素とする 3 次元のベクトルで表し，クラス l の文書集合とその対抗文書集合の間で実行した．線形判別分析を適用することにより，クラス l の文書集合とその対抗文書集合とを最適に分離することができる．これをあらゆる L_P，L_G の値の組合せに対して実行し，結果が最も良い組合せを選択した．

また，文ベクトルの正規化の要否を決める実験を行った結果，文ベクトルを正規化し，さらに式 (13)，(14) を各文書の文の数で正規化する方法が良い結果を与えることを確認した．次節で述べる実験結果は，σ^2 や γ などのパラメータの値を変えながら行った実験の中で最も良い結果である．

4.3 実験結果

図 3 に，着目クラスを “earn” として，テストデータにおいて kNN によって着目クラスに正しく分類された文書集合，対抗文書集合のゲイン $g(X)$ に対する分布を示す．図 3 において，横軸は $z = g(X)$ であり，縦軸は次式で与えられる確率密度分布 $Prob(z)$ を示す．すなわち，

$$Prob(z_k) = n(z_k) / \sum_k n(z_k) \tag{17}$$

である．ここで，$n(z_k)$ は $g(X)$ の値が z_k となる文書数である．また**図 4** はペナルティ $p(X)$ に対する同様の分布を示す．なお，着目クラスに属する文書数は 1,077，対抗文書数は 51 であった．また，式 (13)，(14) における L_P，L_G はともに 5 としている．対抗文書は kNN において “earn” に紛らわしい文書であり，“earn” に属する文書とは分けにくいはずであるが，図 3，4 では多少の重なりはあるものの着目クラスの文書集合とよく分離していることが分かる．

図 5 は式 (12) における正則化の効果を示す図である．横軸 τ は式 (12) における σ^2 の値を決めるパラメータであり，行列 $\boldsymbol{S}_T$ の対角成分の平均値に τ を乗じた値を σ^2 としている．また，縦軸は，$N = 20$

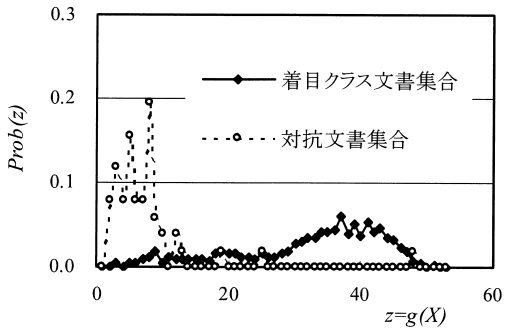

図 3 クラス “earn” におけるテストデータのゲインの分布
Fig. 3 Probability density function of $g(x)$ for correctly classified and misclassified documents by kNN as belonging to class “earn”.

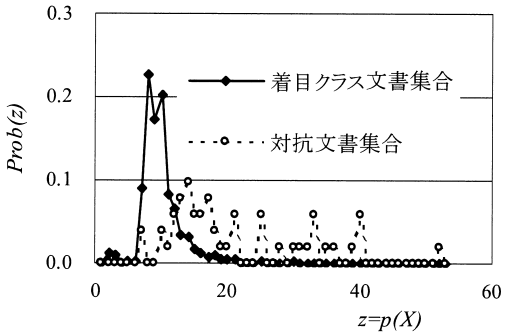

図 4 クラス “earn” におけるテストデータのペナルティの分布
Fig. 4 Probability density function of $p(x)$ for correctly classified and misclassified documents by kNN as belonging to class “earn”.

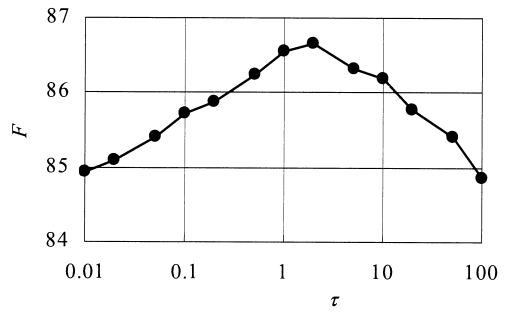

図 5 F 値と正則化パラメータとの関係
Fig. 5 Relationship between F measure and regularization parameter.

として式 (15) において $g(X)$ のみで類似度の補正を行ったときの F 値を示す．図から分かるように F 値は $\tau = 2.0$ のときに最大値をとっている．$\tau < 2.0$ のときに F 値が低いのは雑音の影響を受けやすい不安定な TDF が求められたためと考えられる．また，$\tau > 2.0$ のときに F 値が下がるのは，式 (12) の分母において σ^2 の値が大きくなったために射影軸が P_T の値の変化に対して鈍感になり有効性が減少した結果と考えられる．

表 6 補正前の分類精度
Table 6 Performance before similarity correction.

データ	再現率	精度	F値
訓練データ	90.69	89.02	89.85
テストデータ	81.57	85.93	83.69

表 7 交差検定を行わない場合の分類精度
Table 7 Performance without cross validation.

	評価データ	パラメータの決定データ	再現率	精度	F値
a	訓練データ	訓練データ	94.15	95.06	94.60
b	テストデータ	テストデータ	85.44	93.59	89.33
c	テストデータ	訓練データ	84.74	87.02	85.87

また，**表 6** には類似度の補正前の訓練データ，テストデータの分類結果を示す．**表 7** は交差検定を行わなかった場合の類似度補正後の訓練データ，テストデータの分類結果を示す．表 7 では，(a) 式 (13)，(14)，(15) のパラメータを訓練データで決定し，訓練データを評価した結果，(b) パラメータをテストデータで決定し，テストデータを評価した結果，(c) パラメータを訓練データで決定し，テストデータを評価した結果の 3 通りを示している．パラメータを訓練データで決定した場合，訓練データの F 値は著しく向上しているが，それでも 95%以下にとどまっている．訓練データすら 95%以下の F 値にとどまっていることは，多くのクラスの間で境界がはっきりせず，クラス領域が重なっていることを暗示している．テストデータでパラメータを決定したときのテストデータの F 値は 89.33%，訓練データでパラメータを決定したときは 85.87%となっており，いずれも kNN 単独の F 値 83.69%を上回っている．しかし両者の値は著しく異なっている．前者はテストデータに対する過度のチューニング，後者は訓練データに対する過度のチューニングを起こした結果と考えられ，これらはテストデータに対する適正な評価結果とはなりえない．

表 8 に，テストデータの分割数 N の値を 2，5，10，20 としたときに，式 (15) において $sim(X)$ を $g(X)$ のみを用いて補正して $sim_C(X)$ を決定した場合の F 値，$p(X)$ のみを用いた場合の F 値，$g(X)$，$p(X)$ の両者を用いて補正した場合の F 値，精度，再現率を示す．この表から以下がいえる．

- ペナルティおよびゲインの両方が性能を向上させるうえで有効である．

表 8 交差検定を行ったときの補正後の分類結果
Table 8 Performance after similarity correction with cross validation.

$sim_C(X)$	尺度	N			
		2	5	10	20
$sim(X)+ag(X)$	F値	86.44	86.70	86.67	86.66
$sim(X)-bp(X)$	F値	84.55	85.04	85.09	85.18
$sim(X)+ag(X)$ $-bp(X)$	F値	86.64	86.95	87.10	87.27
	精度	90.03	90.59	91.03	91.28
	再現率	83.49	83.60	83.49	83.60

- ペナルティを与えるよりもゲインを与える方が有効である．これは各クラスに存在すべき特徴を抽出する方が，存在すべきでない特徴を抽出するよりも容易であることを示している．原因は，各クラスの対抗文書集合には様々なクラスの文書が含まれるため，対抗文書集合の P-TDF（各クラスの N-TDF）には顕著な傾向が現れにくかったためと考えられる．
- $g(X)$, $p(X)$ の両者を用いて類似度の補正を行った場合，F 値は 87.27%に達し，kNN 単独の 83.69%に比べて著しく改善されている．ただし，この F 値はテストデータ全体に対する値ではあるが，テストデータを用いた交差検定を行っているため実質的な訓練データは Reuters-21578 の正規の訓練データよりも多くなっており，表 5 における既存手法の F 値とは対等な比較はできない．対等な比較を行うには，メインの分類系用の訓練データ，パラメータ決定のためのデータの両方を Reuters-21578 の正規の訓練データから選択する必要がある．そのためには何らかの工夫が必要となるが，85.87%と 87.27%の間の F 値が得られるものと思われる．

なお，上記で $N = 20$ としたとき，L_P，L_G の平均はそれぞれ 1.7，3.0 であった．また，P-TDF，N-TDF の算出において，式 (7)，(8) で 15 の固有ベクトルを求めるための所要時間は，120 Mhz の WS を用いたとき，約 60 min であった．計算コストは実用の範囲内にあるといえる．

5. ま と め

以上，本論文をまとめると以下のようになる．

(1) 2 つの文書集合間のトピックの違いを反映したベクトル（TDF）を，各文書の文ベクトルの射影値の 2 乗和に関する両文書間の比を最大にするベクトルとして求める方法（TDFA）を提案した．簡単な評価実験を通じ，TDF には単語の出現傾向の違いだけでなく単語共起の違いも反映されていることを確認した．

(2) 各クラスの文書集合とその対抗文書集合との間で TDFA を適用することにより，各クラスにつき，着目クラスには出現するが他のクラスには出現しにくい特徴，他のクラスには出現するが着目クラスには出現しにくい特徴を求め，kNN を用いた分類系の相補的分類系で用いる方法を提案した．

(3) Reuters-21578 を用いた分類実験により，F 値は kNN 単独の場合に比べ，著しく向上した．

本論文で提案した TDF が文書分類に有効であったという事実は，とりもなおさず TDF には各クラスに特有な話題が忠実に反映されていたことを示す．したがって TDFA は 2 つの内容の似通った文書もしくは文書集合の間で，一方に含まれ，他方には含まれない話題の抽出などに効果を発揮するものと考えられる．このような能力をふまえ，TDFA の文書分類以外の応用を見い出していくことは今後の重要な課題である．

参 考 文 献

1) Yang, Y. and Liu, X.: Re-examination of Text Categorization, *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'99*), pp.43–49 (1999).
2) Masand, B., Linoff, G. and Walts, D.: Classifying News Stories Using Memory Based Reasoning, *Proc. 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'92*), pp.59–64 (1992).
3) Yang, Y.: Expert network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval, *Proc. 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'94*), pp.13–22 (1994).
4) Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *European Conference on Machine Learning* (*ECML*) (1998).
5) Yang, Y. and Chute, C.G.: An Example-based Mapping method for Text Categorization, *ACM Trans. Information Systems* (*TOIS*), Vol.12, no.3, pp.252–277 (1994).
6) Schapire, R.E. and Singer, Y.: Boost Boost-Texter: A Boosting-based System for Text Categorization, *Machine Learning*, 39, pp.135–168 (2000).

7) 川谷隆彦：文書集合間の差異検出法と文書分類への応用，情報処理学会自然言語処理研究報告，2002-NL-148, pp.1–8 (2002).
8) 川谷隆彦：トピック差分因子分析法による文書間の相違性の評価，情報処理学会自然言語処理研究報告，2002-NL-150, pp.43–48 (2002).
9) Kawatani, T.: Difference Factor Extraction between Two Document Sets and its Application to Text Categorization, *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'2002*), pp.137–144 (2002).
10) Duda, R.O. and Hart, P.E.: *Pattern Classification and Scene Analysis*, John Wiley & Sons Inc. (1973).
11) 奥野忠一，久米　均，芳賀敏郎，吉澤　正：多変量解析法（改訂版），日科技連 (1981).
12) 石井健一郎，上田修功，前田英作，村瀬　洋：わかりやすいパターン認識，オーム社 (1998).
13) Fukunaga, K.: *Introduction to Statistical Pattern Recognition* (*2nd Edition*), Academic Press Inc. (1990).
14) McLachlan, G.J.: *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, Inc. (1992).
15) Friedman, J.H.: Regularized Discriminant Analysis, *J. Amer. Statist. Assoc.*, 84, pp.165–175 (1989).
16) Yang, Y. and Pederson, J.P.: Feature Selection in Statistical Learning for Text Categorization, *14th International Conference on Machine Learning*, pp.412–420 (1997).

(平成 16 年 6 月 29 日受付)
(平成 17 年 3 月 1 日採録)

川谷　隆彦（正会員）

1944 年生．1967 年東京大学工学部物理工学科卒業．同年日本電信電話公社（現 NTT）電気通信研究所入所．ホログラム，文字認識の研究実用化に従事．1993 年ヒューレット・パッカード日本研究所入所．引き続き文字認識，テキスト処理の研究実用化に従事．2004 年メディアドライブ（株）入社．工学博士．電子情報通信学会，IEEE Computer Society，ACM 各会員．