自 然 言 語 処 理 148－1
（２００２．３．４）

文書集合間の差異検出法と文書分類への応用

川谷隆彦

日本ヒューレット･パッカード㈱　ヒューレット･パッカード研究所

takahiko_kawatani@hp.com

本報告では、先ず 2 つの文書集合間の差異検出法を提案する。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乗和に関する両文書集合間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、文書集合間のトピックの差を表す因子とする。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、着目クラスで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で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 に対し F 値は kNN 単独の 83.69%から 87.14%に向上した。

A Method to Extract Differences between Text Sets and its Application to Text Categorization

Takahiko KAWATANI

Hewlett-Packard Labs Japan, Hewlett-Packard Japan

takahiko_kawatani@hp.com

This paper proposes a method to extract differences between two text sets. Each text is represented as a set of sentence vectors. Suppose all sentence vectors are projected on a vector. The proposed method obtains the vectors which maximize the ratio between the text sets as to the sum of squared projections by solving a generalized eigenvalue problem. By applying the method to the text set belonging to a given class and a set of texts missclassified as belonging to the class by an existent classifier, we can obtain the features which frequently appear in the given class but rarely in other classes, and the features which frequently appear in other classes but rarely in the class. By using the features in the complementary classifier to kNN, the micro averaged F value for Reuters-21578 was improved from 83.69% to 87.14%.

1. まえがき

近年文書分類の研究が盛んに行われている。これまでに数々の手法が提案されてきたが、Yang の比較実験によれば[1]、kNN[1, 2, 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。その他、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。更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、どのような分類法も、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、例えば、ベクトル空間モデルでは各クラスに属する文書の平均ベクトルにより、kNNでは各クラスに属する文書の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云ってよいであろう。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ないようである。KNNにせよ、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類似性が存在することになり、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、各クラス固有の情報を用いてクラスモデルを記述する必要がある。

本報告ではこの問題に焦点を当てる。先ず2つの文書集合間の差異を最もよく表す特徴の抽出を試みる。本報告では、このような特徴を、文書を文ベクトルの集合で表した上で、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。上記射影軸は両文書集合のトピックの違いを反映するものなので、トピック差分因子(Topic Difference Factor : TDF)と呼ぶこととする。また、その手法をトピック差分因子分析(Topic Difference Factor Analysis: TDFA)と呼ぶこととする。次に、このTDFAを文書分類に応用することにより、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。本報告では、このような特徴を既存のメインの分類系に対する相補的分類系で用いる分類法を提案する。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.**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.**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.**では、メインの分類系としてkNNを採用した時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. トピック差分因子 (TDF)

2.1 アプローチ

文書集合$D=\{D_1,..,D_M\}$、$T=\{T_1,..,T_N\}$を考える。各文書は文ベクトルの集合により成るものとし、文書D_m、T_nのk番目の文ベクトルをそれぞれ$\boldsymbol{d}_{mk}(k=1,..,K_D(m))$、$\boldsymbol{t}_{nk}(k=1,..,K_T(n))$とする。ここで求めるべきベクトルを$\boldsymbol{\alpha}$とする。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$P_D$、$P_T$とすると、これらは以下のように求められる。

$$P_D = \sum_{m=1}^{M}\sum_{k=1}^{K_D(m)}(\boldsymbol{d}_{mk}{}^T\boldsymbol{\alpha})^2 = \boldsymbol{\alpha}^T\boldsymbol{S}_D\boldsymbol{\alpha} \quad (1)$$

$$P_T = \sum_{n=1}^{N}\sum_{k=1}^{K_T(n)}(\boldsymbol{t}_{nk}{}^T\boldsymbol{\alpha})^2 = \boldsymbol{\alpha}^T\boldsymbol{S}_T\boldsymbol{\alpha} \quad (2)$$

ここで、添え字Tは転置を表す。また、$\boldsymbol{S}_D$、$\boldsymbol{S}_T$は$\sum_{m=1}^{M}\sum_{k=1}^{K_D(m)}\boldsymbol{d}_{mk}\boldsymbol{d}_{mk}{}^T$、$\sum_{n=1}^{N}\sum_{k=1}^{K_T(n)}\boldsymbol{t}_{nk}\boldsymbol{t}_{nk}{}^T$によりそれぞれ定義される行列である。これらを平方和行列と呼ぶこととする。ここで$\boldsymbol{\alpha}$に反映される両文書間の差異の程度を表す評価基準$J(\boldsymbol{\alpha})$を、

$$J(\boldsymbol{\alpha}) = \frac{P_D}{P_T} = \frac{\boldsymbol{\alpha}^T\boldsymbol{S}_D\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{S}_T\boldsymbol{\alpha}} \quad (3)$$

により定義する。$J(\boldsymbol{\alpha})$を最大にする$\boldsymbol{\alpha}$は、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、文書集合Dにはよ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の正のトピック差分因子(Positive Topic Difference Factor : P-TDF)と呼ぶこととする。式(3)で与えられる評価基準は実は線形判別分析におけるそれと形式的に同じであり、$\boldsymbol{\alpha}$の解も線形判別分析と同じように、

$$\boldsymbol{S}_D\boldsymbol{\alpha} = \lambda\boldsymbol{S}_T\boldsymbol{\alpha} \qquad (4)$$

なる一般固有値問題の固有ベクトルで与えられる[7, 8, 9]。$\boldsymbol{\alpha}$はS_T^{-1} S_Dの固有ベクトルで与えられると云ってもよい。

また、求めるべきベクトルを$\boldsymbol{\beta}$とし、評価基準$J(\boldsymbol{\beta})$を、

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

とすると、$J(\boldsymbol{\beta})$を最大にする$\boldsymbol{\beta}$は文書集合Tにはよ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。これを文書集合Dの負のトピック差分因子(Negative Topic Difference Factor : N-TDF)と呼ぶこととする。この場合には$\boldsymbol{\beta}$は、

$$\boldsymbol{S}_T\boldsymbol{\beta} = \lambda\boldsymbol{S}_D\boldsymbol{\beta} \qquad (6)$$

の固有ベクトルで与えられる。

また、式(1)、(2)において、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なる。

2.2 解釈

式(4)の場合を考える。式(4)の解として得られるi次の固有値、固有ベクトルをλ_i、$\boldsymbol{\alpha}_i$とする。$\lambda_i = \boldsymbol{\alpha}_i^{\mathrm{T}} S_D \boldsymbol{\alpha}_i / \boldsymbol{\alpha}_i^{\mathrm{T}} S_T \boldsymbol{\alpha}_i$なので、$\lambda_i$は$\boldsymbol{\alpha}_i$を用いた時の評価基準の値そのものである。また、式(4)のような形の固有値問題の固有ベクトルは2段階の写像によって求められ、また、この写像によって行列$\boldsymbol{S}_D$、$\boldsymbol{S}_T$が同時に対角化されることが知られている[9, 10]。即ち、式(4)の場合には、

(1) S_Tのj次の固有値、固有ベクトルを$\rho_j(\geq \rho_{j+1})$、$\boldsymbol{\phi}_j$とする。固有値ρ_jは文書集合Tの全文ベクトルの$\boldsymbol{\phi}_j$への射影値の2乗和である。

(2) 文書集合Dの文ベクトル$\boldsymbol{d}_{mk}$が空間Yのベクトル$\boldsymbol{y}_{mk}$に写像されたとする。但し、$\boldsymbol{y}_{mk}$のj番目の成分は次式で与えられる。

$$y_{mkj} = \boldsymbol{\phi}_j^T \boldsymbol{d}_{mk} / \sqrt{\rho_j} \qquad (7)$$

(3) 空間Yで文書集合Dの主成分（平方和行列の固有値の大きい固有ベクトルの方向）を求めると、これが原空間の$\boldsymbol{\alpha}$と対応する。

式(7)で示されるように空間Yへの写像に際しては$\boldsymbol{d}_{mk}$と$\boldsymbol{\phi}_j$との内積をρ_jの平方根で割っている。そのため、空間Yでは文書集合Dは、文書集合Tの主成分方向に沿う成分は圧縮され、反対に固有値の小さい固有ベクトルに沿う成分は伸張される。$\boldsymbol{\alpha}$はこのようにした上で求められる文書集合Dの主成分と対応するので、文書集合Tの文ベクトルの射影値の2乗和は小さく、文書集合Dからのそれは大きくなる。このため

$$a_{mk} = \sum_{i=1}^{L} (\boldsymbol{d}_{mk}^T \boldsymbol{\alpha}_i)^2 \qquad (8)$$

を求めると、$\boldsymbol{d}_{mk}$が文書集合Dに含まれ文書集合Tには含まれない内容を記述していれば、a_{mk}の値は大きくなる。なお、Lはa_{mn}の算出に際し用いる固有ベクトルの数であるが、これは実験的に適切な値を決める必要がある。

2.3 正則化

再び式(4)の場合を考える。式(4)において固有ベクトルが求められるためには行列$\boldsymbol{S}_T$は正則行列でなければならない。しかし、実際にはサンプル数が単語数よりも少ない、特定の単語対が常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。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[11]。即ち、σ^2をパラメータ、Iを単位行列として

$$\hat{S}_T = S_T + \sigma^2 I \qquad (9)$$

を$\boldsymbol{S}_T$として用いる方法である。

式(9)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する。$\boldsymbol{S}_T$の固有ベクトルは文書集合Tに含まれる全文ベク

トルの射影値の 2 乗和を最大にする射影軸である。この場合、全単語ベクトルの任意の単位ベクトルへの射影値の 2 乗和は常にσ^2であるから、$\boldsymbol{S}_T$の対角成分にσ^2を加えたことによって、固有ベクトルの方向は変わらず、固有値がσ^2のバイアスを加えられることになる。また、評価基準は、全単語ベクトルが文書集合に加わるので、式(3)ではなく、

$$J(\boldsymbol{\alpha}) = P_D /(P_T + \sigma^2) \qquad (10)$$

としたことに相当する。

式(3)の評価基準では、求められた$\boldsymbol{\alpha}$がP_Tの値を非常に小さくする場合には、そのP_Tの値は文書集合Tの雑音の影響を受けている可能性がある。逆に云えば式(3)の評価基準では$\boldsymbol{\alpha}$は文書集合Tの雑音の影響を受けやすい。式(10)を用いれば分母の値にバイアスが加えられるので、文書集合Tの雑音の影響を受けにくくなる。

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、各文書が 1 つの文ベクトルで構成され、文書集合Dは D-1~ D-4、Tは T-1~ T-4 の各々4 個の 5 次元文ベクトルで構成されているものとする。文書集合D、Tの違い、共通点は以下のとおりである。

- 文書集合Dでは D-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 D-3 で単語 2 と 3、及び D-4 で 1 と 4 が共起する。
- 文書集合Tでは T-3 で単語 1 と 3、及び T-4 で 2 と 4 が共起する。
- 単語 1 と 2、及び 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σ^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(4)の固有値と固有ベクトルを示す。表 2 では n 次固有値λ_nと n 次固有ベクトル$\boldsymbol{\alpha}_n$=(α_{n1},.., α_{n5})の各成分を示す。同様に、表 3 に N-TDF として式(6)の解を与える n 次固有値μ_nと n 次固有ベクトル$\boldsymbol{\beta}_n$=(β_{n1},.., β_{n5})を示す。また、表 4 には各文ベクトルの 1、2 次の固有ベクトルへの射影値を示す。これらから以下が云える。

(1) 表 2 における 1 次固有ベクトル$\boldsymbol{\alpha}_1$では、α_{11}とα_{14}が共に負、α_{12}とα_{13}が共に正の値をとっており、文書集合Dにおける単語 2 と 3、及び 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、文書集合Dの文ベクトル D-3、D-4 の$\boldsymbol{\alpha}_1$への射影値の絶対値は大きくなっている。また、D-3、D-4 以外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、文書集合Dにおける単語 5 の存在を反映して

表 1　想定する文書集合

#	D	T
1	11001	11000
2	00110	00110
3	01100	10100
4	10010	01010

表 2　文書集合 D の P‐TDF

n	λ_n	α_{n1}	α_{n2}	α_{n3}	α_{n4}	α_{n5}
1	20.00	-1.58	1.58	1.58	-1.58	0.00
2	10.74	0.11	0.11	-0.04	-0.04	3.05
3	0.97	-0.01	-0.01	-0.40	-0.40	0.01
4	0.22	-0.41	-0.41	0.14	0.14	0.84
5	0.00	0.35	-0.35	0.35	-0.35	0.00

表 3　文書集合 D の N‐TDF

n	μ_n	β_{n1}	β_{n2}	β_{n3}	β_{n4}	β_{n5}
1	20.00	1.58	-1.58	1.58	-1.58	0.00
2	2.78	-0.72	-0.72	0.23	0.23	1.31
3	0.97	0.01	0.01	-0.40	-0.40	-0.01
4	0.00	-0.25	0.25	0.25	-0.25	0.65
5	0.00	0.24	-0.24	-0.24	0.24	0.70

表 4　各文ベクトルの射影値

text	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
D-1	0.00	3.27	0.00	-0.13
D-2	0.00	-0.07	0.00	0.46
D-3	3.16	0.08	0.00	-0.49
D-4	-3.16	0.08	0.00	-0.49
T-1	0.00	0.22	0.00	-1.44
T-2	0.00	-0.07	0.00	0.46
T-3	0.00	0.08	3.16	-0.49
T-4	0.00	0.08	-3.16	-0.49

いる。事実、表４で D-1 からの射影値のみが大きな値を取っている。

(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、及び 2 と 4 の共起が反映されている。そのため、表 4 で T-3、T-4 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、文ベクトル D-1 と T-1 の違いが反映され、T-1 からの射影値が大きな値を取っている。

(5) 表 2、3 における 3 次以上の固有ベクトルは固有値が小さく、TDF として有効ではない。

これらの観察により、TDF には 2 つの文書集合間の出現単語の違いのみならず、単語間の共起の違いも反映されることが分かった。本報告では文書を文ベクトルの集合として表現しているが、これは単語間の共起の違いが TDF に正確に反映されるようにするためである。

3. 文書分類への応用

3.1 アプローチ

1.で述べたように、本報告では TDF のみを用いた分類系の構築は狙わずに、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。その理由は以下の通りである。

(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、本報告では、メインの分類系では類似度を用いて分類を行うことを前提に、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

各クラスにおける TDF の求め方について先ず述べる。先ず、メインの認識系において全ての訓練用文書の分類を行い、各訓練データについて各クラスの類似度を求める。次いで、クラス毎に閾値γを然るべき方法で決めておき、類似度がγ以上でクラス 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 に誤った、もしくは誤りそうになった文書の集合であり、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(4)の解を与える固有ベクトル$\{\boldsymbol{\alpha}_n\}$、N-TDF は式(6)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、

$$g(X)=\sum_{i=1}^{L_G}\sum_{k=1}^{K}(\boldsymbol{x}_k{}^T\boldsymbol{\alpha}_i)^2 \qquad (11)$$

$$p(X)=\sum_{i=1}^{L_P}\sum_{k=1}^{K}(\boldsymbol{x}_k{}^T\boldsymbol{\beta}_i)^2 \qquad (12)$$

により与えられる。L_P、及び L_G は何次の固有ベクトルまで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

$$sim_C(X)= sim(X) + a\,g(X) - b\,p(X) \qquad (13)$$

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より大きい時にのみ行われ、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(11)、(12)の定義では、長い文書ほど $g(X)$、$p(X)$の値が大きくなってしまうので単語数で正規

化することも考えられる。また、$\boldsymbol{d}_{mk}$ 、$\boldsymbol{t}_{nk}$ を正規化して TDF を算出する時には、式(11)、(12)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

実験条件は[1]にほぼ沿っている。用いたコーパスは 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、87 カテゴリで訓練データ 7770 文書、テストデータ 3019 文書であった。Reuters-21578 には複数のラベル（帰属するカテゴリ）が付与された文書が少なからず存在し、ここでは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、lemmatizing、ストップワード除去、単語選択である。文切り出しは、本報告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、通常はピリオドで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 行がひとつの文とみ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[12]により行い、2500 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。

$$w_i = (1+\log f_i)\log(N_D / n_i) \qquad (14)$$

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、先ず、入力文書は全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 に属する訓練文書と入力文書間の余弦類似度の総和で与えられる。k の値は[1]に従い、45 とした。さらに、$sim(X)$ が予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。この閾値はクラス毎に F 値が最大になるように決定した。再現率を r、精度を p として、F 値の定義は $F=2rp/(r+p)$ を用いた。表 5 に、[1]に述べられている kNN、SVM のテストデータに対するマイクロ平均による再現率、精度、F 値、及び本報告でのそれらを示す。表から分かるように、本報告での kNN の結果は[1]の結果よりも F 値が約 2%劣っている。原因は不明である。

また、式(11)、(12)における L_P、L_G、式(13)における a、b の決定には次のような問題があった。これらの値は、本来は訓練データを用いて決定すべきものである。しかし、訓練データを用いて求められた TDF は訓練データにチューンされている。そのような TDF を用いてさらにこれらの値を決定すると、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。そのため訓練データによってこれら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。そこで、テストデータを用いて交差検定を行った。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。

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。まず、各クラス毎に、L_P、L_G に 15 以内で適当な値を与えて式(13)における a、b を線形判別分析[7, 8, 9]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 の文書集合と他のクラスに属する文書集合の間で実行した。これをあらゆる L_P、L_G の値の組み合わせに対して実行し、結果が最もよいものを採用した。

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(11)、(12)を各文書の文の数で正規化する方法がよ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よい結果である。

4.3 実験結果

図 1 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 1 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 $Pro(z)$を示す。即ち、

$$Pro(z) = n(z) / \int n(z) \qquad (15)$$

である。ここで、$n(z)$は $g(X)$ の値が z となる文書数である。また図 2 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 1077、対抗文書数は 51 であった。また、式(11)、(12)における L_P、L_G は共に 5 としている。対抗文書は kNN において"earn"に誤分類された文書であり、"earn"に属する文書とは分けにくい筈であるが、図 1、2 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。

また、表 6 に、テストデータの分割数 N の値を 2、5、10、20 とした時に、式(13)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。これらは何れもマイクロ平均である。この表から以下が云える。

- ペナルティ及びゲインの両方が性能を向上させるうえで有効である。
- 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ほうが、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。これは、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 0.8714 に達し、kNN 単独の 0.8369 に比べて著しく改善されている。

なお、上記で N=20 とした時、L_P、L_G の平均は各々 1.5 であった。また、P-TDF、N-TDF の算出において、式(4)、(6)で 15 の固有ベクトルを求めるた

表 5　Reuters-21578 に対する分類結果

方法	再現率	精度	F値
SVM	0.8120	0.9137	0.8599
kNN	0.8339	0.8807	0.8567
本報告	0.8157	0.8593	0.8369

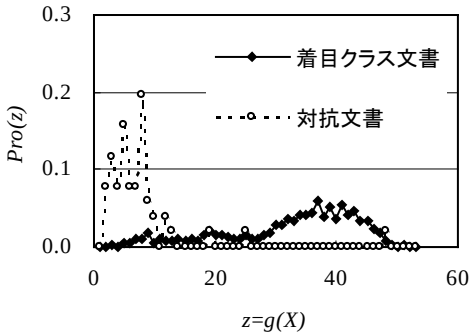

図 1　クラス"earn"におけるテストデータのゲインの分布

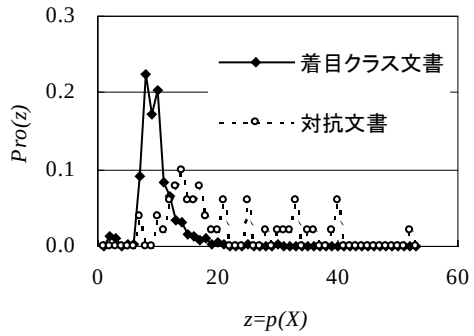

図 2　クラス"earn"におけるテストデータのペナルティの分布

表 6　類似度補正後の分類結果

		N			
		2	5	10	20
$sim(X)+ag(X)$	F値	0.8644	0.8660	0.8668	0.8676
$sim(X)-bp(X)$	F値	0.8455	0.8516	0.8521	0.8530
$sim(X)+ag(X)$ $-bp(X)$	F値	0.8652	0.8682	0.8709	0.8714
	精度	0.8908	0.8953	0.8995	0.9003
	再現率	0.8410	0.8426	0.8440	0.8440

めの所要時間は、120Mhz の WS を用いたとき、約 60min であった。計算コストは高くないと云える。

5. まとめ

以上、本報告を纏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。

(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 単独のそれに比べ著しく向上した。

本報告で提案した TDF が文書分類に有効であったという事実はとりも直さず TDF にはクラス間のトピックの差が忠実に反映されていたことを示す。従って、TDFA を任意の 2 つの文書集合間のトピックの違いの検出に用いても良好な結果を与えるものと思われ、TDFA の文書分類以外の応用を探すことも今後の重要なテーマである。ひとつの文書集合が表すコンセプトそのものより他の文書集合との差異が重要となるような場合には TDFA は効果を発揮するものと思われる。また、その際に、本報告では試みなかったが、ベクトルとしての TDF が何を表すか解釈できるようにすることが望ましい。これも重要なテーマである。

参考文献

[1] Y. Yang and X. Liu. Re-examination of Text Categorization. In *Proceedings of the 22nd Annual International ACM SIGIR Conference.on Research and Development in Information Retrieval (SIGIR'99),* pp.43-49, 1999.

[2] B. Masand, G. Linoff and D. Walts. Classifying News Stories Using Memory Based Reasoning. In *Proceedings of the 15nd Annual International ACM SIGIR Conference.on Research and Development in Information Retrieval (SIGIR'92),* pp.59-64, 1992.

[3] Y. Yang. Expert network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. In *Proceedings of the 17nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94),* pp.13-22, 1994.

[4] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning (ECML),* 1998.

[5] Y. Yang and C.G. Chute. An Example-based Mapping method for Text Categorization. *ACM Transaction on Information Systems (TOIS),* 12(3), pp.252-277, 1994.

[6] R. E. Schapire and Y. Singer. Boost BoostTexter: A Boosting-based System for Text Categorization. *Machine Learning,* 39, pp.135-168, 2000.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis,* John Wiley & Sons Inc., 1973.

[8] 奥野, 久米, 芳賀, 吉澤. 多変量解析法（改訂版）. 日科技連（1981）.

[9] 石井, 上田, 前田, 村瀬. パターン認識. オーム社（1998）.

[10] K. Fukunaga. *Introduction to Statistical Pattern Recognition (Second Edition),* Academic Press Inc., 1990.

[11] G.J.McLachlan. *Discriminant Analysis and Statistical Pattern Recognition,* John Wiley & Sons, Inc., 1992.

[12] Y. Yang and J.P.Pederson. Feature Selection in Statistical Learning for Text Categorization. In *The Fourteenth International Conference on Machine Learning,* pp.412-420, 1997.