語の重み付け学習を用いた文書の自動分類

福本文代† 鈴木良弥†

本稿では，文書中に現れる語に対し，重み付けの学習を行った結果を用いて文書の自動分類を行う手法を提案する．我々の手法における学習とは，トレーニングデータにおける各文書の分類の誤り結果から正しい重要語を学習することである．すなわち，トレーニングデータの文書集合に対しクラスタリングアルゴリズムを適用した結果，文書が正しく分類されない場合，頻度による重み付けが適切でないと見なす．これらの文書どうしに対し重要語が判定され重み付けが行われる．本手法では，重要語を判定するために文脈依存の度合いという考え方を用いる．これは，文書中の任意の語が，設定された文脈にどのくらい深く関わっているかという度合いの強さを用いることで主題と関係の深い語を抽出し，これを重要語と見なす方法である．その結果，重要語には高い重み付けを行い，重要でない語（一般語と呼ぶ）に対しては低い重み付けを行う．語の重み付けの学習は，トレーニングデータ中のすべての文書が正しく分類されるまで繰り返される．本手法の有効性を検証するために *Wall Street Journal* を用いて行った実験では 78 個の分野に属する 4,453 の文書に対し，Lewis らによって提案された *Proportional assignment strategy* による Breakeven Points で，0.75 の正解率が得られた．

Term Weight Learning for an Automatic Text Categorisation

FUMIYO FUKUMOTO† and YOSHIMI SUZUKI†

In this paper, we propose a method for term weight learning which is used to characterise texts. In our approach, learning is to learn a true keyword from the error of clustering texts. Parameters of term weighting are then estimated so as to maximise the true keyword and minimise the other words in the text. The characteristic of our approach is that the degree of context dependency is introduced to judge whether a word in a text is a true keyword or not. The experiments using *Wall Street Journal* corpus demonstrate the effectiveness of the method.

1. はじめに

近年，電子化された大量の文書が流通するようになったことを背景に，文書の自動分類に関する研究が情報検索や自然言語処理の研究分野でさかんに行われている[8)~10),16),21),25)]．

文書の自動分類における手法の 1 つに，あらかじめ分類すべき分野を設定し，このうちのいずれかに文書を分類する手法がある．この手法における分類精度は，あらかじめ分類すべき各分野の特徴をどのように表すかに依存して決まり，1. 表記の統計情報を用いた手法，2. 語の意味的な情報を用いた手法に分類できる．

1. は，文書中に出現する語の統計情報を用いる手法であり，ベクトルモデル（Vector model）[20)]や確率モデル（Probabilistic model）[3),11),17)]などが提案されている．ベクトルモデルは，分類を行おうとする文書（テストデータと呼ぶ），およびあらかじめ分類すべき各分野（トレーニングデータと呼ぶ）をそれぞれベクトルで表現し，ベクトルどうしの類似度を用いて分類する手法である．ベクトルの各軸は，文書に現れる語とし，各軸における長さは，語に対し重み付け（term weighting）を行った値が用いられている．

語の重み付けは，情報検索の分野で幅広く研究されており，様々な重み付け手法が提案されている[18)~20)]．Guthrie ら[5)]や湯浅ら[22)]は，文書の自動分類において頻度を用いている．徳永ら[21)]は，重み付け IDF と呼ばれる手法を提案し，これを用いて分類を行っている．

確率モデルは情報検索の研究分野でさかんに用いられている手法であり，語の統計情報を用いて文書 d が分野 c に分類される確率 $P(c \mid d)$ を求める手法である．Iwayama ら[6)]は Single random Variable with Multiple Values（SVMV）と呼ばれる確率モデルを提案し，これを用いて文書の自動分類を行っている．*Wall Street Journal* を用いて行った実験では，Salton らの提案した TF*IDF が Breakeven Points 0.48 に

† 山梨大学工学部コンピュータ・メディア工学科
Department of Computer Science and Media Engineering, Faculty of Engineering, Yamanashi University

対し，SVMV では，0.63 の正解率が得られることが報告されている．しかし，一般に単語表記の統計情報を用いた手法は，処理が簡単で汎用性が高い反面，語の表層的な情報しか扱っていないため，分類精度が低くなることが指摘されている[8]．

2. は，文書中の語に関する意味的な情報を用いて文書の分類を行う手法である[9]．Walker ら[23]は，多義語に注目し，文書に現れる名詞語義の曖昧さを解消した結果を用いて文書の自動分類を行っている．彼らの基本的なアイデアは，多義語の意味を判定することで各分野の特徴が明らかになり，結果的に分野に依存した文書の分類が行われるのではないかということである．しかし，*Wall Street Journal* のように限定された分野では，名詞は同じ意味で用いられていることが多いため，多義語の判定を行った場合と行わずに分類した場合とで顕著な差が生じないことが報告されている[4]．

Blosseville ら[1]は，あらかじめ分類すべき分野の特徴を規則として学習し，その結果を用いて文書の分類を行う手法を提案している．実験では，7 分野に対し約 92%，28 分野に対し約 73%の高い正解率が得られることが報告されている．しかし，規則を生成するときに用いる文書中の重要語は，あらかじめ分類すべき分野に属する文書に対し深い構文・意味解析を行うことで抽出しているため，単語表記の統計情報を用いた手法と比べると高い正解率が得られる反面，対象分野の変更に対しどの程度適用できるかは不明である．

本稿では，1. と 2. の欠点を解消すること，すなわち，分類精度を保ちながら汎用性を高めることを目的とし，文書中に現れる語に対し，重み付けの学習を行った結果を用いて文書の自動分類を行う手法を提案する．本手法では，Guthrie や湯浅らと同様，単語表記の統計的な情報を用いてテストデータの文書をトレーニングデータの特定の分野に分類する．しかし，Guthrie や湯浅らが頻度のみを用いて各分野に属する文書の特徴を表しているのに対し，我々は重み付けの学習を行うことで文書の特徴を表す．我々の手法における学習とは，トレーニングデータにおける各文書の分類の誤り結果から正しい重要語を学習することである．すなわち，トレーニングデータの文書集合に対しクラスタリングアルゴリズムを適用した結果，文書が正しく分類されない場合，頻度による重み付けが適切でないと見なす．これらの文書どうしに対し，重要語が判定され重み付けが行われる．本手法では，重要語を判定するために文脈依存の度合いという考え方を用いる．これは，文書中の任意の語が，設定された文脈にどのくらい深く関わっているかという度合いの強さを用いることで，主題と関係の深い語を抽出し，これを重要語と見なす方法である．その結果，重要語には高い重み付けを行い，重要でない語（一般語と呼ぶ）に対しては低い重み付けを行う．語の重み付けの学習は，トレーニングデータ中のすべての文書が正しく分類されるまで繰り返される．*Wall Street Journal* を用いて行った実験では 78 個の分野に属する 4,453 の文書に対し，Lewis らによって提案された *Proportional assignment strategy* による Breakeven Points で，0.75 の正解率が得られた．

以下，2 章では，語の重み付けの学習を用いることでトレーニングデータ中の各文書を分類する手法について述べる．3 章では学習結果を用いてテストデータを各分野へ分類する手法を示す．4 章では実験について報告し，結果に関する考察を行う．

2. トレーニングデータの分類

我々の手法はトレーニングデータの各文書に対しクラスタリングアルゴリズムを適用した結果，文書が正しくクラスタリングされない場合，頻度による重み付けが適切でないと見なす．これらの文書どうしに対し，重要語が判定され重み付けが行われる．語の重み付けの学習は，トレーニングデータ中のすべての文書が正しく分類されるまで繰り返される．以下では，まず，文脈依存の度合を用いた重要語の抽出とそれを用いて文書に出現する語に対して重み付けを行う手法について述べる．次に語の重み付け学習に基づくクラスタリングアルゴリズムについて述べる．

2.1 重要語の抽出

一般に，主題はテキストの中で論点を示す語である．本稿では名詞を対象とし，主題，あるいは主題と意味的に関係が深い語を**重要語**と呼ぶ．我々は重要語を判定するために文脈依存の度合いという考え方を導入する．**図 1** は，新聞記事（*Wall Street Journal*）の構造を示す．

図 1 において 1 日の新聞は，複数の記事から成り，'経済'，'文化' などいくつかの分野に分類することができる．ここでは分類された各々を**分野**と呼ぶ．ある特定の分野における記事は，いくつかのパラグラフから成る．我々は各々を**パラグラフ**と呼ぶ．Luhn らのキーワード密度方式と同様，「1 つの文献において，主題と関係の深い語は概して文献中に繰り返し出現する」という前提に基づくと，重要語は各パラグラフをまたがり一貫して出現しているととらえることができる[14]．

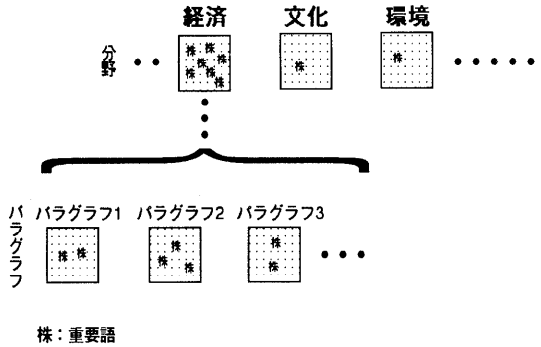

図1 新聞記事の構造
Fig. 1 The structure of newspaper.

我々は重要語に対して重み付けを行うために，図1で示される新聞記事の構造に対し，文脈依存の度合いという考え方を導入する．文脈依存の度合いとは，任意の語が図1で示した特定の**分野**，あるいは特定の**パラグラフ**とどのくらい深く関わっているかという度合いの強さを示す．たとえば，図1において，'経済'に関する分野における重要語を'株'とすると，'株'は，記事の各パラグラフにまたがり出現する．よって，'株'は各パラグラフでの分布の偏りが一般語と同様，小さく，特定のパラグラフに依存する度合いは低い．次に'経済'の分野について考える．一般語は文書中どこにでも現れるため，各分野における分布の偏りはパラグラフでのそれと差はない．一方，'株'の'経済'での依存の度合いは，'株'が'経済'という特定の分野に集中して出現するため，結果的に記事中の特定のパラグラフに依存する度合いよりも強くなると考えられる．我々は χ^2 法を用いて文脈依存の度合いを計算し，この度合いの強弱を利用し，重要語と一般語との区別を行った．

(1) χ^2 法

長尾ら[15]は，任意の語がそれぞれの分野においてその分野を特徴付ける語であるか否かを判定する尺度として χ^2 検定の χ^2 値を用い，この手法がキーワードの抽出に有効であることを検証している．しかし，一般に χ^2 値からは分野全体に対して出現頻度に偏りのある語が抽出できる反面，それぞれの分野においてその分野を特徴付ける語が何であるかは分からない．また文書の量が多い分野の χ^2 の値は大きくなり，少ない分野のそれは小さくなる．したがって，文書ごとに記事の量に偏りがある場合，χ^2 値の大きさだけで語を選べば，比較的記事の量が少ない分野の χ^2 の値は小さくなるため，結果的に重要語を抽出することができない可能性がある．渡辺ら[24]は重要漢字の自動抽出においてこの問題に対処するため，それぞれの分野における出現頻度の理論度数からのずれに注目した．本稿で述べる文脈依存の度合いを示す尺度も，渡辺らと同様，それぞれの文脈における出現頻度の理論度数からのずれを用いる．語（名詞とする）w が i（i は分野，またはパラグラフ）において特定の分野（またはパラグラフ）j に依存する度合いを式(1)に示す．

$$\chi i_{wj}^2 = \begin{cases} \frac{(x_{wj}-m_{wj})^2}{m_{wj}} & \text{if } x_{wj} > m_{wj} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

ここで，

$$m_{wj} = \frac{\sum_{j=1}^{n} x_{wj}}{\sum_{w=1}^{m}\sum_{j=1}^{n} x_{wj}} \times \sum_{w=1}^{m} x_{wj}$$

ただし，

i:	**D**（分野），または **P**（パラグラフ）
m:	名詞の数
n:	i の個数
x_{wj}:	特定の分野またはパラグラフ j における語 w の出現頻度
m_{wj}:	特定の分野またはパラグラフ j における語 w の理想頻度

とする．ここで理想頻度とは，全分野，あるいは全パラグラフに等確率でその名詞が出現した場合の出現頻度である．式(1)において x_{wj} がその理想頻度よりも小さい場合にはゼロとした．

(2) 文脈依存の度合いを用いた重要語の抽出

語 w が特定の分野（またはパラグラフ）j に依存する度合いは χi_{wj}^2 の分散値 χi_w^2 とした．これは，語 w に関する分野，およびパラグラフでの依存の度合いを比較するためである．χi_w^2 はその値が大きいほど語 w が特定の分野，またはパラグラフに強く依存することを示す．語 w の**分野**（**D**）と**パラグラフ**（**P**）における文脈依存の度合いの関係を式(2)に示す．

$$\chi \mathbf{P}_w^2 < \chi \mathbf{D}_w^2 \tag{2}$$

式(2)においてパラグラフにおける語 w の分散値 $\chi \mathbf{P}_w^2$ よりも分野における語 w の分散値 $\chi \mathbf{D}_w^2$ が大きいことから，語 w は特定のパラグラフよりも特定の分野に強く依存することを示す．よって我々は式(2)を満たす語 w を重要語

```
begin
(a)   if D_y と同じ分野に属するような D_y' が存在する
        for all w such that w ∈ D_x ∩ D_y
(a-1)     if χ P_w^2 < χ D_w^2 であり，かつ w ∈ D_x ∩ D_x'，あるいは w ∈ D_y ∩ D_y' である
            then w は重要語であり，α (1 < α) 倍重み付けされる
(a-2)     else if χ P_w^2 ≥ χ D_w^2 であり，かつ w ∈ D_x ∩ D_x'，あるいは w ∈ D_y ∩ D_y' である
            then w は一般語であり，β (0 < β < 1) 倍重み付けされる
          end_if
        end_for
(b)   else
        for all w such that w ∈ D_x ∩ D_y
(b-1)     if χ P_w^2 < χ D_w^2 であり，かつ w ∈ D_x ∩ D_x' である
            then w は重要語であり，α (1 < α) 倍重み付けされる
(b-2)     else if χ P_w^2 ≥ χ D_w^2 であり，かつ w ∈ D_x ∩ D_x' である
            then w は一般語であり，β (0 < β < 1) 倍重み付けされる
          end_if
        end_for
    end_if
end
```

図2 重み付けの学習
Fig. 2 Term weight learning.

と見なした．

2.2 語の重み付け学習

我々の手法は，トレーニングデータ中の各文書に対し，群平均化のクラスタリングアルゴリズムを適用する[7]．もし，異なる分野に属する文書どうしが同一分野に分類された場合，それらの文書に対し重要語が判定され重み付けの学習が適用される．

文書 D_x と $D_{x'}$ が同じ分野に属するとし，D_y は D_x と異なる分野に属するとする．また，D_x と D_y はクラスタリングを行った結果，誤って同じ分野に属すると判定されたとする．クラスタリングを行おうとする文書の集合に D_y と同じ分野に属する D'_y が存在する場合には，D_x と D'_x，および D_y と D'_y がそれぞれ同一分野に分類されるよう，重要語の重み付けを行えばよい．一方，D'_y が存在しない場合には，D_x と D'_x が同一分野に分類されるよう，重み付けを行えばよい．重要語に関する重み付けのアルゴリズムを**図2**に示す．

図2において，(a) はクラスタリングを行おうとする文書の集合に D_y と同じ分野に属する D'_y が存在する場合である．一方，(b) は，D'_y が存在しない場合である．

図2において，(a-1) と (b-1) は正しく重み付けされていない重要語を抽出する処理であり，(a-2) と (b-2) は正しく重み付けされていない一般語を抽出する処理である．図中，$D_x \cap D'_x$ は文書 D_x と D'_x に共通して現れる名詞の集合を示す．語 w が式 (2) を満たし，かつ $D_x \cap D'_x$ の要素である場合に，w は D_x と D'_x が属する分野を特徴付ける重要語であるとした．同様に w が式 (2) を満たし，かつ $D_y \cap D'_y$ の要素である場合に，w は D_y と D'_y が属する分野を特徴付ける重要語であるとした．ただし，**分野**は D_x と D_y から成るとする．

α は重要語に関する重み付け変数を示し，β は一般語に関する重み付け変数を示す．

```
begin
  do Make-Initial-Cluster-Set
  for i := 1 to m(m-1)/2 do
      do Apply-Clustering
       if D_x と D_y が同じクラスタ (分野) に属する
         then do Term-Weight-Learning
                 do Make-Initial-Cluster-Set
                 i := 1
       end_if
  end_for
end
```

図3 学習に基づくクラスタリングアルゴリズム
Fig. 3 Flow of the clustering algorithm based on learning.

2.3 学習に基づくクラスタリングアルゴリズム

トレーニングデータ中の各文書は，語の重み付け学習に基づくクラスタリングアルゴリズムを用いて分類される．クラスタリングアルゴリズムを**図3**に示す．図3においてアルゴリズムは3つの処理，すなわち **Make-Initial-Cluster-Set**，**Apply-Clustering**，**Term-Weight-Learning** から成る．

(1) **Make-Initial-Cluster-Set**

Make-Initial-Cluster-Set では，文書の集合を入力とし，すべての文書対の組合せに対し類似度の値を計算し，すべての文書対と類似度

の値をその値が降順になるように出力する．各文書は式(3)で表される．

$$D_i = (X_{i1}, X_{i2}, \cdots, X_{ix}) \quad (3)$$

x はすべての文書に出現する異なり名詞の個数とし，X_{ij} は文書 D_i に出現する名詞 X_j の頻度とする．文書 $D_1, \cdots, D_m$（ただし，m は文書の個数とする）の任意の組合せに対し，式(4)を用いて類似度を計算する．

$$Sim(D_i, D_j) = \frac{D_i \cdot D_j}{|D_i||D_j|} \quad (4)$$

式(4)はベクトル D_i と D_j の長さの積で正規化した D_i と D_j の内積を示す．式(4)の値が大きいほど，D_i と D_j は類似していることを示す．

(2) **Apply-Clustering**

Apply-Clustering では，**Make-Initial-Cluster-Set** の出力，すなわち，類似度の値が降順に出力された文書対に対し，順次，群平均化のクラスタリングアルゴリズムが適用される．文書 D_x と $D_{x'}$ が同じ分野に属するとし，D_y は D_x と異なる分野に属するとする．クラスタリングの結果，D_x と D_y は誤って同じ分野に属すると判定されたとする．D_x，$D_{x'}$ および D_y に対し，次処理である **Term-Weight-Learning** が適用される．

(3) **Term-Weight-Learning**

Term-Weight-Learning では D_x，$D_{x'}$ および D_y に対し，図2で示される重み付けの学習アルゴリズムが適用される．その結果，D_x，$D_{x'}$，D_y および $D_{y'}$ の各文書を式(5)を用いてベクトル表現として表す．

$$D_i = (X'_{i1}, X'_{i2}, \cdots, X'_{ix}) \quad (5)$$

式(5)において x はすべての文書に出現する異なり名詞の個数を示し，X'_{ij} は以下のとおりとする．

$$X'_{ij} = \begin{cases} 0 & X'_j \text{ が } D_i \text{ に出現しない場合} \\ \alpha \times f(X'_j) & X'_j \text{ が重要語であり } D_i \text{ に出現する場合} \\ \beta \times f(X'_j) & X'_j \text{ が一般語であり } D_i \text{ に出現する場合} \end{cases}$$

ここで $f(X'_j)$ は文書 D_i に名詞 X'_j が出現する頻度とする．

D_x，$D_{x'}$，D_y および $D_{y'}$ の任意の組合せに対して $Sim(D_x, D_{x'})$ と $Sim(D_y, D_{y'})$ の値がそれ以外の組の Sim の値よりも大きくなるように α と β が推定される☆．

D_x，$D_{x'}$，D_y および $D_{y'}$ 以外の文書が式(3)を用いて表され，D_x，$D_{x'}$，D_y および $D_{y'}$ が式(5)を用いて表される．その結果に対し，**Make-Initial-Cluster-Set** が適用され，処理が繰り返される．この処理は，得られたクラスタがすべての文書を含むまで繰り返される．

3. テストデータの分類

トレーニングデータ中の文書 $D_1, \cdots, D_m$（ただし m は文書の個数とする）に対し図3で示すクラスタリングアルゴリズムを適用し，m 個の文書を分野へ分類する．文書 $D_1, \cdots, D_m$，およびテストデータである文書 D をそれぞれ式(3)を用いて表す．D と D_1 から D_m までの各文書との類似度を式(4)を用いて計算し，類似度の値が大きい順に D_1 から D_m までを出力する．D が割り当てられる分野は類似度の値が大きい順に D_1 から D_m が割り当てられた分野とする．

文書 D に割り当てられた分野のうち，上位何位までを抽出し，評価の対象とするかを決定する方法の1つに Lewis[12] によって提案された *Proportional assignment strategy* がある．この手法は，トレーニングデータにおいて，ある特定の分野に割り当てられた文書の数は，テスト文書 D に割り当てられた分野のうち，上位からの分野抽出数に比例するという手法である．このことは，ある特定の分野に割り当てられるトレーニングデータの文書の数が多いほどテスト文書 D に割り当てられた上位からの分野数も多く抽出し，その分野が含まれるか否かを判定するということを示している．たとえば，比例定数が 5.0 の場合には，トレーニングデータの文書数の 2%に相当する文書に分類された分野はテスト文書に割り当てられた分野のうち，上位 10%を抽出し評価の対象とすることを示す．我々は，この手法を用いて文書を各分野に分類した．実験では比例定数を 1.0 とし，各テストデータに対し分野ごとに抽出数を定めた．

4. 実　　験

本章では，1989 年の *Wall Street Journal* を用いて行った2つの実験結果について述べる．まず，我々の手法を用いて実験を行い，文書の分類精度を検証す

☆ 実験では，α の増分および β の減分を 0.001 とし，式(1)および(2)で示される処理，すなわち式(1) α の増分を 0.001 とし，β を固定する，式(2) α の増分を固定し，β の減分を 0.001 とするをそれぞれ交互に繰り返すことで α と β の推定を行った．

る．次にベクトルモデルの一手法として，χ^2 法を用いて語に対し重み付けを行った手法，および確率モデルとして，Iwayama ら[6]の提案した SVMV（Single random Variable with Multiple Values）と本手法とを比較し，本手法の有効性を検証する．

4.1 デ ー タ

実験で用いたデータは 1989 年の *Wall Street Journal* であり，12,380 文書から成る[13]．*Wall Street Journal* は 78 種類の分野が設定されている．実験では，12,380 文書のうち，これらの分野のいずれにも分類されていない文書を除く 8,907 文書を用いた．

8,907 文書に対し品詞のタグ付けを行い[2]，さらに名詞に対して派生処理を行った．ここで，派生処理とは複数形の名詞，名詞の所有格，動名詞など名詞の働きをするものを名詞の単数形で置き換えることをいう．実験では，8,907 文書のうち，約半数に相当する 4,453 文書をテストデータとして，残り 4,454 を重み付け学習のためのトレーニングデータとして用いた．

4.2 文書の分類実験

実験では 4,454 文書から成るトレーニングデータの学習結果に対し，4,453 文書のテストデータを分類した．

文書の分類結果に対する評価の 1 つに再現率（Recall factor）と適合率（Precision factor）を求める方法がある．再現率と適合率を以下に示す．

$$\text{再現率} = \frac{\text{テストデータが分野に正しく分類されている個数}}{\text{テストデータが分類されるべき分野の個数}} \tag{6}$$

$$\text{適合率} = \frac{\text{テストデータが分野に正しく分類されている個数}}{\text{テストデータが分類されている分野の異なり個数}} \tag{7}$$

式 (6) において，再現率の値を高い値にするためには，実験の結果，テストデータに割り当てられた分野の抽出数を増やす，すなわち，1 つの文書に対し評価の対象とする分野の数を増やせばよい．しかし，このことは結果的に適合率を下げてしまう．Breakeven Points は適合率と再現率のバランスをとるためのものであり，両者の値が高く，かつ等しいほど分類精度が高いとされる．我々は，式 (7) で示される '適合率' の分母において，Lewis らによって提案された *Proportional assignment strategy* の比例定数を 1.0 とし，Breakeven Points を求めた．文書の分類実験の結果を**表 1** に示す．表 1 において '分野' は 78 個の分野の中からランダムに '分野' で示された数だけ抽出した個数を示し，'トレーニングデータ' は各分野数に含まれるトレーニングデータの総数を示す．*Wall Street Journal* の文書はその多くが複数の分野に分類されている．テストデータは，分野に属するテストデータの総数を示す．'Breakeven Pts.' は Breakeven Points の値を示す．

表 1 実験結果
Table 1 The result of the experiment.

分野	トレーニングデータ	テストデータ	Breakeven Pts.
10	2,399	1,457	0.80
20	3,893	2,452	0.77
30	5,178	3,508	0.77
40	5,828	3,994	0.76
50	7,344	4,998	0.77
60	8,475	5,976	0.76
70	11,489	6,148	0.75
78	11,649	7,305	0.75

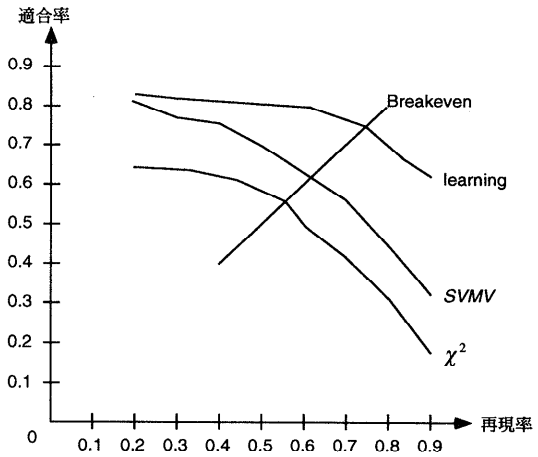

図 4 比較実験結果
Fig. 4 The result of comparative experiment.

表 2 Breakeven Pts.
Table 2 Breakeven Pts.

手法	Breakeven Pts.
Learning	0.75
SVMV	0.64
χ^2	0.56

4.3 他手法との比較

本手法の有効性を検証するため，χ^2 法を用いて語に対し重み付けを行った手法，および，Iwayama らの提案した SVMV と本手法とを比較する．まず，Iwayama らの手法の概略を示し，次に比較実験の結果を示す．

Iwayama らの提案した SVMV において，ドキュメント d が分野 c に分類される確率は式 (8) で示される．

$$P(c \mid d) = P(c) \sum_{t_i} \frac{P(T = t_i \mid c) P(T = t_i \mid d)}{P(T = t_i)} \tag{8}$$

ただし，

表3 本手法により重み付けされた語（上位5語）
Table 3 The first top 5 of the highest weighted words in our method.

	AIR		ARO		BBK		BNK	
No	Word	Wt	Word	Wt	Word	Wt	Word	Wt
1	airline	522.1	aerospace	148.2	share	149.0	bank	84.0
2	mile	136.5	aircraft	143.0	stock	71.9	branch	32.0
3	passenger	120.5	air	730.	company	57.2	credit	30.0
4	revenue	85.0	army	51.0	bank	51.0	tax	24.0
5	air	67.2	jetliner	43.3	security	43.5	letter	16.0
	FOD		STK		ENV		MED	
No	Word	Wt	Word	Wt	Word	Wt	Word	Wt
1	food	140.0	company	50.0	environment	87.0	news	281.0
2	fda	27.0	share	37.7	maquilas	19.0	d&b	108.0
3	general	24.0	stock	31.7	water	12.0	network	69.1
4	cereal	19.0	trading	10.1	plant	10.1	report	69.0
5	health	16.0	investment	9.4	health	9.4	broadcaster	44.8
	ECO		PIP		DIV		CPR	
No	Word	Wt	Word	Wt	Word	Wt	Word	Wt
1	gain	120.5	gas	58.0	cent	85.0	analystic	106.5
2	tax	111.0	pipeline	37.0	share	70.0	IBM	89.8
3	capital	83.4	industry	29.0	company	60.9	machine	69.0
4	rate	79.5	foothill	24.0	dividend	54.6	computer	62.0
5	economy	30.5	oil	7.0	split	46.7	system	48.6

$P(T = t_i \mid c) = \frac{NC_i}{NC}$ NC_i は term t_i が分野 c に現れる頻度数，NC は分野 c に現れる語の総頻度数

$P(T = t_i \mid d) = \frac{ND_i}{ND}$ ND_i は term t_i が文書 d に現れる頻度数，ND は文書 d に現れる語の総頻度数

$P(T = t_i) = \frac{N_i}{N}$ N_i は与えられた文書における term t_i の頻度数，N はトレーニングデータにおける語の総頻度数

$P(c) = \frac{D_c}{D}$ D_c は与えられたトレーニングデータのうち分野 c に分類された文書の数，D はトレーニングデータの文書数

とする．実験結果を**図4**に示す．

図4において，横軸は再現率を示し，縦軸は適合率を示す．'Learning' は本手法の結果を示す．**表2**に各手法の Breakeven Points の値を示す．

図4，および表2において，式(7)で示される '適合率' の分母における比例定数は 1.0 とした．

5. 考　　察

5.1 文書の分類実験

表1によると，78個の分野に対し，7,305 文書のテストデータが存在し，Breakeven Points は 0.75 であった．分野の個数にともなう正解率を比較すると，分野の個数と正解率が必ずしも反比例していないことから本手法は分野の個数に依存せず文書の特徴を表現することができると考えられる．

78個の分野からランダムに12個選択し，それらの分野に含まれる文書に現れる語のうち，重み付けの値が最も大きい語（上位5語）を**表3**に示す．

表4 分野名
Table 4 The domain name.

AIR:	Airlines	ARO:	Aerospace
BBK:	Buybacks	BNK:	Banks
FOD:	Food products	STK:	Stock market
ENV:	Environment	MED:	Media
ECO:	Economic news	PIP:	Pipeline
DIV:	Dividends	CPR:	Computers

表3において，'No' は抽出された語の順位を示し，'Word' は抽出された語を示す．また，'Wt' は語の重み付けの値を示す．ランダムに抽出した12の分野名を**表4**に示す．

表3によると，'FOD' の3位である 'general' 以外の語はそれぞれ各分野の特徴を示す語として妥当であることから語の重み付けの学習は有効であるといえる．

一方，テストデータのうち，最も不正解の多い文書は 'STK' に属する文書であった．'STK' に属する文書は 499 文書存在し，そのうちの約 32%に相当する 159 文書が誤って 'BBK' に属すると判定された．表3によると，'BBK' の上位3語と 'STK' の上位3語は一致し，それらの語に関する重み付けは 'BBK' の方が 'STK' のそれよりも高い値が付与されている．'STK' と 'BBK' のように両者が非常に類似している分野の場合には重み付けの学習を用いても両者の区別をつけることは難しく，この場合，'STK' に分類される文書が誤って 'BBK' に分類されてしまったことから，本

表5 χ^2 法により重み付けされた語（上位5語）
Table 5 The first top 5 of the highest weighted words in χ^2 method.

	AIR		ARO		BBK		BNK	
No	Word	Wt	Word	Wt	Word	Wt	Word	Wt
1	airline	12,109.1	boeing	4,880.0	share	2,348.7	bank	6,196.4
2	ual	5,268.5	force	4,022.3	redemption	1,902.4	bnl	1,517.3
3	passenger	5,142.3	aircraft	3,886.7	devon	1,779.4	bond	1,211.4
4	pilot	4,672.1	defense	2,328.6	hadson	1,641.1	loan	1,023.3
5	flight	4,050.8	missile	2,060.7	buy-back	1,616.4	rate	890.1
	FOD		STK		ENV		MED	
No	Word	Wt	Word	Wt	Word	Wt	Word	Wt
1	spam	3,148.4	stock	7,265.4	ozone	2,650.7	magazine	4,222.3
2	food	2,848.5	share	3,563.2	epa	2,414.0	d&b	3,313.7
3	cereal	2,627.7	buy-back	2,302.0	asbestosis	2,259.0	cable	2,890.1
4	cholesterol	2,518.2	redemption	1,448.5	anthrax	1,483.5	network	2,496.9
5	cooke	2,355.1	big	1,018.6	pollution	1,165.3	broadcaster	1,999.9
	ECO		PIP		DIV		CPR	
No	Word	Wt	Word	Wt	Word	Wt	Word	Wt
1	gain	2,160.5	pipeline	8,521.7	dividend	10,067.7	computer	13,948.8
2	democrat	1,492.0	foothill	5,933.7	share	4,999.4	IBM	8,470.1
3	tax	1,410.6	gas	5,744.4	company	3,666.8	software	4,709.2
4	budget	1,294.5	transcanada	4,984.0	buy-back	2,499.4	cray	3,538.7
5	spending	1,157.3	westcoast	4,494.9	henley	2,166.6	digital	3,291.6

手法の限界であることが分かる．

5.2 他手法との比較

(1) χ^2 法と本手法

χ^2 法に対し，表3で示される分野に含まれる文書に現れる語のうち，重み付けの値が最も大きい語（上位5語）を**表5**に示す．

表5によると，'BBK' の3位と4位である 'devon' と 'hadson'，および 'PIP' の4位と5位である 'transcanada' と 'westcoast' 以外の語はそれぞれ各分野の特徴を示す語として妥当であるといえる．一方，χ^2 法において最も不正解数の多い文書は 'STK' に属する文書であり，これは本手法を用いた場合と同様であった．表5によると，'BBK' と 'STK' の上位5語のうち3語（'share'，'redemption'，'buy-back'）は一致し，それらの語に関する重み付けは，'STK' の方が 'BBK' のそれよりも高い値が付与されている．したがって，χ^2 法の場合にも 'STK' と 'BBK' の区別をつけることは難しいことが分かる．

χ^2 法における正解率が本手法よりも低かった原因として各語に付与された重み付けの値の差が，本手法と比べて小さいことが考えられる．本手法と χ^2 法において，分野ごとに出現する各語の重み付けの値の分散値を**表6**に示す．

表6によると，'BNK' と 'FOD'，および 'MED' 以外はいずれも χ^2 法における分散値が本手法のそれよりも小さいことから，χ^2 法は本手法と比べて分野の特徴が顕著に現れなかったといえる．

表6 本手法および χ^2 による語の重み付けの値の分散値
Table 6 Deviation value of our method and χ^2 method.

分野	本手法	χ^2 法
AIR	4.63	3.64
ARO	4.20	4.12
BBK	3.80	2.57
BNK	2.23	2.25
FOD	2.25	2.72
STK	4.45	2.57
ENV	2.99	2.30
MED	3.89	6.10
ECO	4.44	2.55
PIP	3.94	3.11
DIV	4.93	3.41
CPR	4.50	3.86

(2) SVMV と本手法

表4によると，本手法における Breakeven Points が 0.75 であるのに対し SVMV は 0.64 であった．仮に2つの分野 A，B から成るトレーニングデータにおいて，それぞれの分野に含まれる語の総異なり数が等しいとする．語 w が A と B，およびテストデータにそれぞれ同じ頻度で含まれ，かつ，テストデータは w のみから成るとする．SVMV では，テストデータの文書が A あるいは B に分類される確率が等しくなり，この2つの分野のうち，どちらに

入るかを判定することができない．しかし，本手法では，仮にAの重要語がwであると判定された場合には，wに対して重み付けを行うため，結果的にテスト文書を分野Aに分類することができる．本手法では，統計情報に加え，誤り結果から正しい重要語を学習している．このことがSVMVに比べ各分野の特徴を顕著に表現することができ，その結果，SVMVよりも高い正解率が得られたと考えられる．

6. おわりに

本稿では，文書中に現れる語に対し，重み付けの学習を行った結果を用いて文書の自動分類を行う手法を提案した．*Wall Street Journal*を用いて行った実験では78個の分野に属する4,453の文書に対し，Breakeven Points 0.75の正解率が得られた．

本手法では語の表層的な情報だけを用いて分類を行っているため，テストデータにおける文書の特徴を示す語が学習されたトレーニングデータ中に存在しない場合，正しく分類することができない．今後，この問題に対処するため意味的な類似性を持つ語どうしをクラスタにまとめ，代表語で置き換えるなどの処理をテストデータ，およびトレーニングデータの各文書に対して行う必要がある．

謝辞 本研究は，平成10年度文部省科学研究費補助金，(財) 東電記念科学技術研究所，国際交流助成，および日本学術振興会，平成11年度特定国派遣研究者助成の援助を受けています．ここにそれらを記し謝意に代えさせていただきます．

参考文献

1) Blosseville, M.J., et al.: Automatic document classification: Natural language processing, statistical analysis, and expert system techniques used together, *Proc. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.51–58 (1992).
2) Brill, E.: A simple rule-based part of speech tagger *Proc. 3rd Conference on Applied Natural Language Processing*, ACL, Trento, Italy, pp.152–155 (1992).
3) Fuhr, N.: Models for retrieval with probabilistic indexing, *Information Processing & Retrieval*, Vol.25, No.1, pp.55–72, 1989
4) Fukumoto, F. and Suzuki, Y.: Automatic Clustering of Articles Using Dictionary Definitions, *16th International Conference on Computational Linguistics* (*COLING'96*), pp.406–411 (1996).
5) Guthrie, L. and Walker, E.: DOCUMENT CLASSIFICATION BY MACHINE: Theory and Practice, *15th International Conference on Computational Linguistics*, Kyoto, Japan, pp.1059–1063 (1994).
6) Iwayama, M. and Tokunaga, T.: A Probabilistic Model for Text Categorization: Based on a Single Random variable with Multiple Values, *ANLP'92*, pp.162–167 (1992).
7) Jardine, N. and Sibson, R.: The construction of hierarchic and non-hierarchic classifications, *Computer Journal*, pp.177–184 (1968).
8) 河合敦夫：意味属性の学習結果に基づく文書自動分類方式，情報処理学会論文誌，Vol.33, No.9, pp.1114–1122 (1992).
9) 亀田弘之，藤崎博也：テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム，情報処理学会論文誌，Vol.28, No.11, pp.1103–1111 (1987).
10) Kupiec, J., Pedersen, J. and Chen, F.: A trainable document summarizer, *Proc. SIGIR'95*, pp.68–73, (1995).
11) Kwok, K.L.: Experiments with a component theory of probabilistic information retrieval based on single terms as document components, *ACM Trans. Information Systems*, Vol.8, No.4, pp.363–386, 1989
12) Lewis, D.: An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task, *SIGIR'92*, pp.37–50 (1992).
13) Liberman, M. (Ed.): *CD-ROM I*, Association for Computational Linguistics Data Collection Initiative, University of Pennsylvania (1991).
14) Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information, *IBM journal*, Vol.1, No.4, pp.307–319 (1957).
15) 長尾 真，水谷幹男，池田浩之：日本語文献における重要語の自動抽出，情報処理，Vol.17, No.2, pp.110–117 (1976).
16) Paice, C.D.: Constructing literature abstracts by computer:Techniques and prospects, *Information Processing and Management*, Vol.26, pp.171–186 (1990).
17) Robertson, S.E. and S. Jones, K.: Relevance weighting of search terms, *Journal of the American society for Information Science*, No.27, pp.129–146 (1976).
18) Salton, G. and Yang, C.S.: On the specification of term values in automatic indexing, *Journal of Documentation*, Vol.29, No.4, pp.351–372 (1973).

19) Salton, G.: Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley (1988).
20) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
21) 徳永健伸，岩山 真：重み付き IDF を用いた文書の自動分類について，情報処理学会自然言語処理研究会，Vol.100, No.5, pp.33–40 (1994).
22) 湯浅夏樹，上田 徹，外川文雄：大量文書データ中の単語間共起を利用した文書分類，情報処理学会論文誌，Vol.36, No.8, pp.1819–1827 (1995).
23) Walker, D. and Amsler, R.: The Use of Machine-Readable Dictionaries in Sublanguage analysis, *Analyzing Language in Restricted domains*, Grishman and Kittredge (Ed.), pp.69–84, Lawrence Erlbaum, Hillsdale, NJ (1987).
24) 渡辺靖彦，竹内雅人，村田真樹：χ^2 法を用いた重要漢字の自動抽出，電子情報通信学会技術研究報告，Vol.94, No.24, pp.15–22 (1994).
25) Zechner, K.: Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences, *Proc. 16th COLING*, pp.986–989 (1996).

(平成 9 年 9 月 2 日受付)
(平成 11 年 1 月 8 日採録)

福本 文代（正会員）

1986 年学習院大学理学部数学科卒業．同年沖電気工業（株）入社．総合システム研究所勤務．1988 年より 1992 年まで（財）新世代コンピュータ技術開発機構へ出向．1993 年マンチェスター工科大学計算言語学部修士課程修了．同大学客員研究員を経て 1994 年より山梨大学工学部助手，1999 年同学部助教授，現在に至る．自然言語処理の研究に従事．理学博士．ACL，言語処理学会各会員．

鈴木 良弥

1986 年山梨大学工学部計算機科学科卒業．1988 年同大学大学院工学研究科計算機科学専攻修了．同年木更津工業高等専門学校助手．1993 年東京工業大学大学院総合理工学研究科博士後期課程修了．1994 年より山梨大学工学部助手，現在に至る．音声言語処理の研究に従事．工学博士．電子情報通信学会，日本音響学会，言語処理学会各会員．