自然言語処理 146−15
(2001. 11. 21)

教師なし文書集合分類における単語選択

川前徳章　青木輝勝　安田　浩

東京大学 先端科学技術研究センター
〒153-8904 東京都目黒区駒場 4-6-1
TEL:(03)5452-5277 FAX:(03)5452-5278
email:kawamae, aoki, yasuda@mpeg. rcast. u-tokyo. ac. jp

本稿は，文書を内容の類似性に基づき高い精度で分類するために必要な単語の選択基準を提案する．文書の分類は文書内に出現した単語を用いて行われるが，単語にはノイズが含まることと，単語の数が文書数に比較して圧倒的に多いことが，分類結果の精度と計算量に影響を与えている．この問題を解決するために，本稿は文書の類似性を測定する単語の選択方法を提案する．提案手法は単語のノイズを考慮した重み付けと文書の分類には因子分析を用いる．提案した重み付けの値によって単語を選択することができる．提案手法を適用した結果，全ての単語を利用する場合に比較しても遜色のない文書集合の分類が実現できることが明らかになった．

情報検索 単語選択 概念検索 特異値分解 因子分析 潜在的意味

Words Selection in Unlabeled Document Classification

NORIAKI KAWAMAE, TERUMASA AOKI and HIROSHI YASUDA

Research Center for Advanced Research and Technology University of Tokyo
4-6-1, Komaba, Meguroku, Tokyo, 153-8904, JAPAN
TEL:+81-3-5452-5277 FAX:+81-3-5452-5278
email:kawamae, aoki, yasuda@mpeg.rcast.u-tokyo.ac.jp

This paper proposes the effect of prior words selection in unlabeled document classification based on its contents similarity with high precision. The document classification is performed by using words in document. It has a number of deficits, mainly due to word's noise and number. Because they affect on the precision of document classification and computation cost. To solve these problems, we suggest prior words selection that measures the document similarity based on its contents. Our method weights words to remove their noise and introduces factor analysis to classify document based on its content.We can select words by using the value of our suggested weighting. The experimental results reveal that selected words classify documents no less than all words.

Information Retrieval, Words selection, Conceptual Search, Singular value decomposition, Factor Analysis, Latent Semantic

1. はじめに

多くのWebページや電子化された各種資料・文書のようなテキスト文書がオンラインで入手可能となってきている．そのために，ユーザにとって必要な情報を探す効率的な文書検索技術がますます必要になってきている．有望な文書検索技術の一つに文書分類がある．文書分類により内容の類似性に基づいて文書が分類されることで，概念検索などが実現することが期待される[6]．しかし，現在の文書検索技術がユーザにとって効率的なものになるために次の問題を解決しなければならない．

(1) 単語の数
検索システムが文書に出現した単語を全てインデックスとして利用しているために，単語の数は膨大なものとなっている．文書の高い分類精度を達成するために大量の単語を用いる場合，過学習あるいは計算時間の問題が生じることになる．

(2) 単語の意味と表記
単語の利用には個人差があり，同じ内容の文書でも利用される単語は異なってくる．例えば，ゆらぎや表記の違いなどである．

† 東京大学先端科学技術研究センター
Research Center for Advanced Research and Technology The University of Tokyo

(3) キーワードマッチングの問題
現在の検索システムはユーザの入力した単語の有無で検索を行っている．従ってユーザの検索へのニーズに合った内容の文書が存在しても，その文書に含まれるキーワードをユーザが入力しなければ検索できない．その結果，適合率は高いが，再現率は低くなる傾向がある．

本稿は，文書を内容の類似性に基づいて高い精度で分類することを実現するのために必要な単語の選択基準を提案した．本研究の新規性は文書をノイズを含んだ単語の集合と定義し，このノイズを考慮した単語の重み付け，文書の分類に因子分析を導入したことで現在の文書検索技術の問題を解決した．本稿では単語のノイズを単語の潜在的意味と表現の差異と定義する．提案手法を Web で閲覧できる新聞記事に適用した結果，次の成果が明らかになり，手法の有効性を確認した．

(1) 単語の選択
提案した単語の重み付けは，文書分類に有効なだけでなく，その分類に必要な単語を文書集合に出現した全単語から抽出できる．その結果，計算量の逓減を実現した．

(2) 潜在的意味空間
文書の類似度を単語の背後にある潜在的意味によって測定した．その結果，内容の類似性に基づき，単語のゆらぎや表記の違いに頑健な文書の分類が実現できた．

(3) 類似文書の検索
潜在的意味によって定義された文書の類似性が内容の類似性になっていることで，高い適合率と再現率を持つことが実現できた．

論文の構成は次のようになっている．2 章は文書分類と単語選択に関する既存研究の問題点を振り返り，3 章は単語の重み付けと，それを利用して文書分類を行うために因子分析の導入を提案する．4 章では実験によって提案手法の有効性を確認し，最終章でまとめとする．

2. 既存の文書分類と単語選択に関する研究

文書分類は大きく分けて二通りある．一つはあらかじめ設定されたカテゴリに分類するものであり，もう一つは文書を内容の類似性によって相対的に分類するものである．表題の教師は文書のカテゴリの意味である．本稿は次の理由から後者の分類がユーザにとって効率的な文書検索技術になると考える．前者の分類の場合はカテゴリ付の文書からの学習を必要とするが，存在する文書の大半はカテゴリが付与されていないことと，多くの文書が毎日のように更新あるいは作成されるために，対象とする分類が流動的である．従って学習によって獲得した分類手法の有効な期間も短くなると考えられるためである．

文書の分類を行うのにベクトル空間モデル (Vector-Space Model:VSM)[3] がよく用いられる．これは単語を軸とする空間において，各文書を単語の重みによって配置することで文書間の類似度を測定するものである．空間の軸の数は分類する文書集合に含まれる単語の総数に等しくなる．これに対して LSA(Latent Semantic Analysis)[1] は特異値分解 (singular value decompostion;SVD) に基づいて，元の単語の軸よりも少数の軸で構成される空間で文書間の類似性を測定する手法である．LSA での軸はベクトル空間の軸を合成した軸であり，単語の潜在的意味の軸であると考えられている．

LSA の問題点としてモデルの存在を前提としていないことがある[6]．つまり仮説検証的なアプローチになっていないので，その最適性が保証されないことにある．文書分類において単語の選択を比較したものには Taira[4] や Yang[5] があるが，本稿で対象とする文書分類とは教師がある点で異なっている．

3. 潜在的意味空間の抽出と単語選択

本稿は文書を内容の類似性によって分類するために次の提案を行う．文書の分類には因子分析を導入し，その因子分析で潜在的意味を抽出するのに最適な単語の重み付けを提案する．この重み付けの値によって分類に利用する単語を選択することができる．

3.1 文書単語行列

文書単語行列 A は文書を行，単語を列とする行列である．文書を d_i，単語を w_j とすると，行列の要素 (i, j) は各文書 d_i の単語の重み a_{ij} で表現できる[8]．単語の重みは大局的重み付けと局所的重み付けの二通りがあるが，これらを組み合わせて利用する．他の従来の重み付けには[5],[7]，がある．提案する文書の類似度は単語を軸とするベクトル空間でなく潜在的意味を軸とする空間で測定する．従って単語の重みは座標値でなく因子分析によって潜在的意味を抽出するのに有効な重みである．

3.1.1 局所的重み付け (Local Weighting)

局所的重み付けは文書 d_i 内の単語の出現頻度を利用して重み付けを行う．単語の出現確率とそれを対数

化したものの積によって高頻度語の影響を緩和することができるように次の重み付けを提案する．log の底は 2 とした．

L_{ij}:単語の出現頻度を利用した重み付け

$$L_{ij} = -P_{ij} log(1 + P_{ij}) \quad (1)$$

P_{ij}:文書 d_i における単語 w_j の出現確率

3.1.2 大局的重み付け (Global Weighting)

大局的重み付けは文書集合全体に渡って単語の重み付けを行う．文書に出現する確率に偏りがある単語は他の文書と区別する有効な情報が多いと考えられるので次の重み付けを提案する．

G_j:単語の相対頻度のエントロピー

$$H_j = -\frac{1}{\log N} \sum_{i=1}^{N} P_{ij} \log P_{ij} \quad (2)$$

N:文書集合に含まれる文書の総数

P_{ij}:文書 d_i における単語 w_j の相対出現確率

エントロピーであるから単語が等確率で出現する場合，最大値となる．ここでは $\log N$ によって最大が 1 となるように正規化を行った．重み付けとして利用する場合，1 から引く．その結果，出現確率が偏りがある単語は 1 に近づき，等確率であれば 0 に近づくように重みを付けられる．

3.2 因子分析の導入

因子分析は観測されたデータから，それらの原因となる少数の因子を発見する手法である．ここでは単語の重み付けが観測されたデータに，潜在的意味が因子に相当する．因子分析を導入した理由は，潜在的意味を単語集合からから推定し，定式化した統計モデルの検証が可能なことによる．因子分析を用いて単語の重み付けと潜在的意味を定式化すると次のようになる．

$$a_{ij} = c_{i1}w_{1j} + c_{i2}w_{2j} + \ldots + c_{im}w_{mj} + u_i v_j \quad (3)$$

a_{ij}:文書 d_i における単語 w_j の観測値

先に挙げた文書単語行列 A の重みに対応

c_{im}:因子得点．文書 d_i における潜在的意味 c_m の得点

w_{mj}:因子負荷量．単語 w_j と因子得点 c_m の相関

u_i:独自因子得点．文書 d_i に固有な得点

v_j:独自因子負荷量．単語 w_j と独自因子得点 u_i の相関

因子分析における概念と単語の関係を説明する．例えば，情報量，推定，検定，仮説という単語から構成された文書を単語を要素としたベクトルで表現すると次のようになる．

文書 d = (情報量，推定，検定，仮説，近江商人，江戸，明治)

それぞれの単語を有無によって重みをつけると次のようになる．

文書 d = (1，1，1，1，0，0，0)

因子分析によって統計，歴史という潜在的意味が抽出されること文書ベクトルは次のように表現できる．

文書 d = (統計，歴史)

それぞれの潜在的意味に対しての重みをつけると次のようになる．

文書 d = (1，0)

従って，因子分析により文書のベクトルを小さくするだけでなく，本質的な文書の類似関係を測定することができるようになる．

文書単語行列 $A(N \times M)$ 型行列は次の形式で表現できる．

$$A = CW + VU \quad (4)$$

C:因子得点行列，$(N \times m)$ 型行列

W:因子負荷量行列，$(m \times M)$ 型行列

V:独自因子得点行列，$(N \times M)$ 型行列

U:独自因子負荷量行列，$(M \times M)$ 型行列．対角成分の j 番目が単語 w_j の独自因子負荷量，他の成分は 0

これから文書間の類似度行列は次のように求められる．

$$AA^T \fallingdotseq \{CW\}\{CW\}^T \quad (5)$$

一方，行列 A の特異値分解のモデルは次のように定義される．

$$A = U\Sigma V^T \quad (6)$$

LSA において行列 A を Σ において特異値の大きい順に m 個を選んで再構成した場合，その行列を A_m とすると次のような形に書き直せる．

$$A = U\Sigma V^T \fallingdotseq U_m \Sigma {V_m}^T = A_m \quad (7)$$

この A_m を利用することで文書間の類似度行列は次のように求められる．

$$AA^T \fallingdotseq \{U_m\Sigma_m\}\{U_m\Sigma_m\}^T \quad (8)$$

式 5 と比較すると特異値分解に基礎を置いた従来の研究は，観測された文書単語行列をそのまま用いていることが分かる．従って文書分類の精度が文書単語行列に依存し，単語のノイズに影響を受けることになる．

3.3 単語の選択

4. 潜在的意味空間における文書分類

本章では提案手法による文書分類と単語選択を，単語の重み付け，潜在的意味空間それぞれの場合についての実験結果について述べる．

4.1 実験の準備

分類に利用した文書は web で閲覧できる新聞記事で全部で 210 ある．各記事は経済，芸能，情報技術，政治，社会，スポーツ，世界情勢についてのカテゴリから構成されていて，それぞれ 30 ある．各記事のカテゴリは検索結果の評価においてのみ利用する．評価としては再現率と適合率を用いた．用意した記事に対して形態素解析を行い，単語毎に分割する．その中で分類で用いた単語の品詞は名詞と未知語である．未知語とは形態素解析に用いた辞書に登録されていない単語である．ニュース記事のカテゴリを表す単語は分類には利用しなかった．形態素解析には茶筅[9] を用いた結果，利用する単語集合は 7863 の異なり語から構成されていた．

4.2 単語の重み付け

表 1 は各重み付け毎と空間の組み合わせにおける文書の類似度を比較したものである．比較は文書間の類似度行列を求め，各カテゴリと全体において各文書と類似度が 0．7 以上の文書を検索結果として利用する場合，検索される文書数，再現率と適合率の平均値で評価した．

L1 は文書の出現頻度，L2 はそれを log を用いて平滑化を行い，$L3$ は文書毎の単語の出現エントロピー[7]，L4 は提案した局所的重み付けである．G1 は G3 を相対化しなかったもの，G2 は単語の出現エントロピーで次のように求める．

$$G2_j = -P_j \log P_j - (1 - P_j) \log(1 - P_j) \quad (9)$$

N:文書集合に含まれる文書の総数

P_j:単語 w_j が含まれる文書の頻度

G4 は DF，G3 は提案した重み付けである．

通常のベクトル空間は用いた 7863 の単語から構成される軸であり，LSA，因子分析は共に最も高い再現率と適合率を得たときの軸の数は 7 である．この結果から，局所的重み付けとしては L3 と L4，大局的重み付けでは G1，G3 と G4 がほぼ同じ精度を示し，空間についてはベクトル空間に比較して SVD，因子分析が圧倒的に高い精度を示している．SVD あるいは因子分析によって単語そのものより潜在的意味で測定した文書の類似度が内容の類似性を反映することが分かる．

4.3 単語選択基準の比較評価

図 1，2，は空間と重み付け毎に文書から利用する単語を増加させていった場合の break even point の変化を示したものである．利用する単語は大局的重みの G1 あるいは G3 の値が高い順に選択し，局所的重みの L3 と L4 を組み合わせた場合の変化を調べた．G4 は単語数が 2500 以下では計算によって求めることができなかったので結果から除外した．全ての単語を分類に用いた場合では SVD と因子分析ほぼ同じ結果を示したが，利用する単語が少ない場合，因子分析は SVD に比較して若干，高い値を示していることが分かる．この結果から実際に出現した単語のうち同じ分類の精度を出すならば半分で済むことが言える．局所的重みについては L4 を組み合わせることで精度の高い分類が実現することが確認できる．単語を軸とする VSM では単語の数、重み付けによって大きな差異は見られなかった．単語とそれに関しての重みによっては文書の類似度が困難なことが分かる．以上の結果から文書の分類に有効な重み付けが分類に必要な単語を選択していることが明らかになった．

4.4 考　察

文書の分類を単語そのものでなく，その潜在的意味で分類するために，因子分析を導入し，単語の重み付けを提案し，その重みを単語選択に適用した実験を行った．その結果，次のようなことが明らかになった．因子分析，SVD は共に最終的にはもとの単語よりも少数の潜在的意味で測定した文書間の類似性が内容の類似性を反映できるが，出発となる文書・単語行列の時点で提案した単語選択の基準を用いて行列の圧縮を行ってから潜在因子を抽出しても結果に大きな変化が見られなかった．従って文書間の類似性を測定するのに有効な単語を選択できることが分かる．この原因として，文書・単語行列はほとんどの要素が 0 の疎な行列であるが，提案した単語選択に用いた大局的重み付けはこの行列から分類に有効となる単語の列を選択したことが考えられる．その結果，情報損失をできるだけ少なく文書・単語行列の単語方向の圧縮することができた．

5. ま と め

本稿は文書を内容の類似性によって高い精度で分類するために必要な単語の選択基準を提案した．実験の

表 1　文書・単語行列と構成空間の違いによる文書類似度の平均
Table 1　Documents Precision and Recall with Each Space on Document Word Matrix

カテゴリ	経済	芸能	情報技術	政治	スポーツ	社会	世界情勢	平均
A	N P R	N P R	N P R	N P R	N P R	N P R	N P R	N P R
L1G1	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/99.7/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	28.7/96.6/92.7	30.0/98.9/98.7	30.0/98.7/98.7	30.0/99.7/99.6	30.1/99.8/100.0	30.3/98.3/99.1	29.1/98.7/95.8	29.7/98.7/97.8
	28.7/96.3/92.4	30.0/98.9/98.9	30.2/98.6/99.1	30.0/100.0/99.8	29.9/99.8/99.6	30.3/98.1/99.1	29.0/98.7/95.6	29.8/98.6/97.8
L1G2	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/99.9/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	28.3/97.0/91.8	29.7/100.0/99.1	29.5/98.7/97.1	30.0/100.0/99.8	29.9/100.0/99.8	29.6/99.4/98.2	20.1/98.5/66.2	28.2/99.1/93.1
	27.8/96.7/90.0	29.7/100.0/99.1	29.8/98.8/98.0	29.9/100.0/99.8	29.7/100.0/98.9	29.1/99.4/96.7	18.9/98.6/62.2	27.8/99.1/92.1
L1G3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/99.8/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	28.9/96.9/93.6	30.0/99.3/99.3	30.1/98.9/99.3	29.9/99.8/99.6	30.1/99.8/100.0	30.1/98.8/99.1	29.1/98.7/95.8	29.8/98.9/98.1
	28.8/96.5/92.9	30.1/99.4/99.8	30.2/98.7/99.3	30.0/100.0/99.8	30.1/99.8/100.0	30.1/98.8/99.1	29.0/98.7/95.6	29.8/98.8/98.1
L1G4	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/99.8/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	29.6/97.8/96.4	30.1/99.7/100.0	30.2/99.0/99.6	30.1/99.8/100.0	30.1/99.7/100.0	30.0/99.5/99.6	30.0/99.2/99.3	30.0/99.2/99.3
	29.7/97.5/96.4	30.2/99.5/100.0	30.2/98.8/99.6	30.1/100.0/100.0	30.0/99.5/99.6	30.0/99.5/99.3	30.1/99.1/99.3	30.0/99.1/99.2
L2G1	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/99.7/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	28.7/96.6/92.7	29.9/99.0/98.7	30.0/98.7/98.7	30.0/99.7/99.6	30.1/99.8/100.0	30.2/98.4/99.1	29.1/98.7/95.8	29.7/98.7/97.8
	28.7/96.3/92.4	30.0/99.0/98.9	30.2/98.6/99.1	30.0/100.0/99.8	29.9/99.8/99.6	30.3/98.2/99.1	29.0/98.7/95.6	29.7/98.6/97.8
L2G2	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/99.9/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	28.3/97.0/91.8	29.7/100.0/99.1	29.6/98.8/97.6	30.0/100.0/99.8	29.9/100.0/99.8	29.6/99.4/98.2	20.3/98.6/66.9	28.2/99.1/93.3
	27.8/96.8/90.2	29.7/100.0/99.1	29.8/98.8/98.2	29.9/100.0/99.8	29.7/100.0/98.9	29.1/99.4/96.7	18.9/98.6/62.2	27.9/99.1/92.2
L2G3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/99.8/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	29.0/97.0/94.0	30.0/99.4/99.3	30.1/98.9/99.3	29.9/99.8/99.6	30.1/99.8/100.0	30.1/98.9/99.1	29.1/98.7/95.8	29.8/98.9/98.2
	28.9/96.5/93.1	30.1/99.4/99.8	30.2/98.7/99.3	30.0/100.0/99.8	30.1/99.8/100.0	30.1/98.8/99.1	29.0/98.7/95.6	29.8/98.8/98.1
L2G4	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/99.8/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	29.6/97.8/96.4	30.1/99.7/100.0	30.2/99.0/99.6	30.1/99.8/100.0	30.1/99.7/100.0	30.0/99.5/99.6	30.0/99.2/99.3	30.0/99.2/99.3
	29.6/97.6/96.4	30.2/99.5/100.0	30.2/98.8/99.6	30.1/100.0/100.0	30.2/99.5/100.0	30.0/99.5/99.3	30.0/99.2/99.3	30.0/99.1/99.2
L3G1	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
L3G2	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	30.0/99.9/99.8	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0	29.1/100.0/97.1	29.9/100.0/99.6
	29.9/99.9/99.6	30.2/99.2/100.0	30.0/99.9/100.0	30.0/100.0/100.0	28.8/99.0/95.3	30.0/100.0/100.0	17.7/100.0/58.9	28.1/99.7/93.4
L3G3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
L3G4	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
L4G1	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
L4G2	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	30.0/99.9/99.8	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0	29.1/100.0/97.1	29.9/100.0/99.6
	29.9/99.9/99.6	30.2/99.2/100.0	30.0/99.9/100.0	30.0/100.0/100.0	28.8/99.0/95.3	30.0/100.0/100.0	17.7/100.0/58.9	28.1/99.7/93.4
L4G3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
L4G4	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.0/100.0/3.3	1.1/100.0/3.6	1.0/100.0/3.4
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/100.0	30.0/100.0/100.0
	30.0/100.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/0.0/100.0	30.0/99.9/100.0	30.0/100.0/100.0	30.0/100.0/10s0.0	30.0/100.0/100.0

N:検索された文書数
P:適合率
R:再現率
各セルの数値は上から通常のベクトル空間，SVD，因子分析を用いたもので，左から検索された文書数，適合率，再現率に対応している．

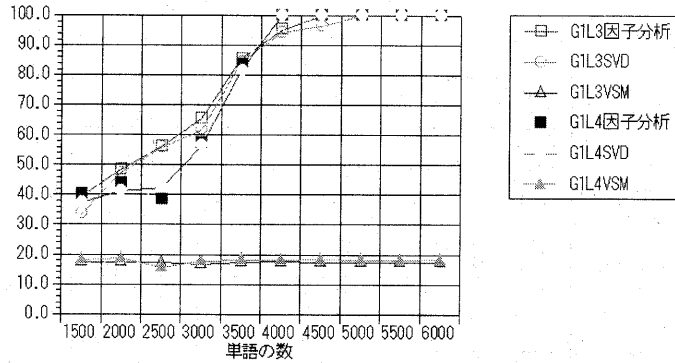

図 1 単語選択として G1 を用いた場合の Break Even Point の変化
Fig. 1 Precision and Recall with Word Selection on Factor Analysis

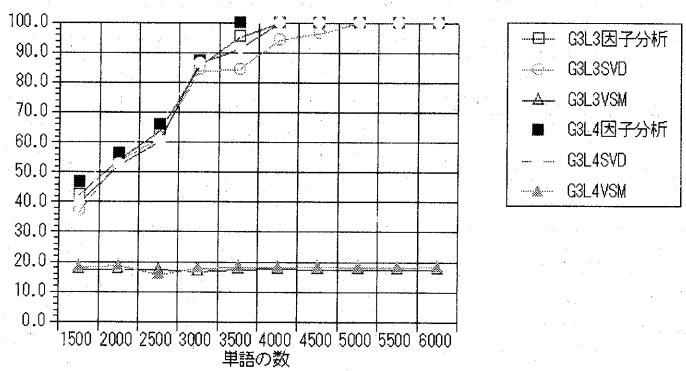

図 2 単語選択として G3 を用いた場合の Break Even Point の変化
Fig. 2 Precision and Recall with Word Selection on LSA

結果，一部の単語を用いても文書の類似性が変化しないことから，提案した大局的重み付けは文書の分類に有効なだけでなく，その値によって文書分類に必要な単語を選択することが明らかになった．

参 考 文 献

1) Deerwester, S., Dumais, S. T., Furnas, G.W., Landaure, T.K., and Harshman, R.: Indexing by latent semantics analysis, Journal of the American Society for Information Science, 1990.
2) Schutze, H. & Pedersen, J.: A vector model for syntagmatic and paradigmatic relatedness, In Processings of the 9th Annual Conference of the University of Waterloo Centre for the New OED and Text Research 1993.
3) Salton, G., McGill, M. J.: Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
4) Taira, H., Haruno, .M.: Feature Selection in SVM Text Categorization, Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-97), 1997.
5) Yang, Y. & Pedersen, J.O.: A Comparative Study on Feature in Text Categorization, Processings of the Fourteenth International Conference on Machine Learning (ICML-97), 1997.
6) 川前徳章, 青木輝勝, 安田浩: 情報理論的モデルを用いた情報検索, 信学技報 DE2001-57,2001.
7) 川前徳章, 青木輝勝, 安田浩: 単語のノイズを除去した教師なし文書の分類と検索, 信学技報 NLC2001-48,2001.
8) 北研二: 確率的言語モデル, 東京大学出版会,1999.
9) 茶筅:http://chasen.aistnara.ac.jp/index.html.ja
10) 徳永健伸: 情報検索と言語処理, 東京大学出版会,1999.