The 29th Symposium on Information Theory
and its Applications (SITA2006)
Hakodate, Hokkaido, Japan, Nov. 28 – Dec. 1, 2006

単語の共起を考慮に入れたナイーブベイズモデルによる文書分類
Text Categorization Using Naive Bayes Model Based on Co-Occurrence words

津田 裕一* 八木秀樹† 平澤 茂一*
Yuichi TSUDA Hideki YAGI Shigeichi HIRASAWA

Abstract— Text categorization is an important technique for automatically assigning a given document to its category. The Naive Bayes classification is one of the most typical document classification method which uses the probability. In this method, the document is considered as a set of independent words. However the occurrence of the words is not necessarily independent in general. In this study, we proposes a new document classification method by taking account of the co-occurrence of two words. We show the effectiveness of the proposed text categorization method by simulation.

Keywords— text categorization, naive bayes, mutual information, co-occurrence words

1 はじめに

文書分類とは与えられた文書を既存のカテゴリに自動的に割り当てる技術である．その手法としてベクトル空間モデルに基づく手法，決定木，サポートベクターマシーン，確率を用いた手法など多くの手法が提案されている [1]．確率を用いる手法の中で最も代表的なものにナイーブベイズ分類法 [2] がある．この手法ではカテゴリが与えられたもとで文書は互いに独立な単語の系列とみなされるが，一般には単語の生起は独立であるとはいえない．

一方，文書分類において，分類に用いる効果的な単語を選択する処理は重要なステップである．これを特徴選択といい，この選択指標として一般に相互情報量が用いられる．相互情報量による特徴選択は，カテゴリに偏って出現する単語を選択するが，選択された単語の中には分類に用いる単語として適切でない場合がある [3]．そのような単語は特定の複数のカテゴリを特徴付けていると考えられ，分類多岐語と呼ばれている [3]．分類多岐語は，共起を考慮する事で単一のカテゴリを特徴付けることができると考えられる．

本研究では，分類多岐語とそれ以外の単語の 2 単語間の共起を考慮した文書分類手法を提案する．また，提案手法を新聞記事データ [7] に適用し，その有効性を示す．

* 〒 169–8555 東京都新宿区大久保 3–4–1 早稲田大学理工学部経営システム工学科 School of Science and Engineering,Waseda University 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan. E-mail: tuda@hirasa.mgmt.waseda.ac.jp
† 〒 169–8050 東京都新宿区西早稲田 1–6–1 早稲田大学メディアネットワークセンター，Media Network Center,Waseda University, Nishi Waseda 1–6–1, Shinjuku-ku, Tokyo, 169–8050 Japan.

2 文書分類手法

単語 w_i の集合を W_{total} で表し，$N = |W_{total}|$ とする．文書分類とは，与えられた文書 d_j をその内容にしたがってあらかじめ決められた同じ概念をもつカテゴリ $c_k \in C$ に自動的に分類する手法を指す．以下に文書分類の基本的な手続きを示す．

1. 文書を単語などを要素とした多次元ベクトルで表現する．
2. 学習データ (人手によりあらかじめカテゴリが付与された文書集合) を用いて各カテゴリの特徴を表現する．
3. 与えられた文書を最も類似したカテゴリに分類する．

本研究では，文書 $d_j = (w_1, w_2, \ldots, w_N)$ に対し，カテゴリ $c_k \in C$ の事後確率 $P(c_k|d_j)$ を最も大きくするカテゴリ c_k を選択する手法である．すなわち

$$c_{\hat{k}} = \arg\max_{c_k} P(c_k|d_j) = \arg\max_{c_k} \frac{P(d_j|c_k)P(c_k)}{P(d_j)} = \arg\max_{c_k} P(d_j|c_k)P(c_k) \quad (1)$$

となる $c_{\hat{k}}$ を推定する．ここで，$P(d_j|c_k)$ はカテゴリ c_k のもとでの文書 d_j の生起確率である．

3 従来手法

3.1 ナイーブベイズ分類 [2][4]

ナイーブベイズ分類では，$P(d_j|c_k)$ を推定する際，図 1 に示すように，カテゴリが与えられたもとで文書 d_j における各単語 w_i は互いに独立に生起するという仮定をおいている．

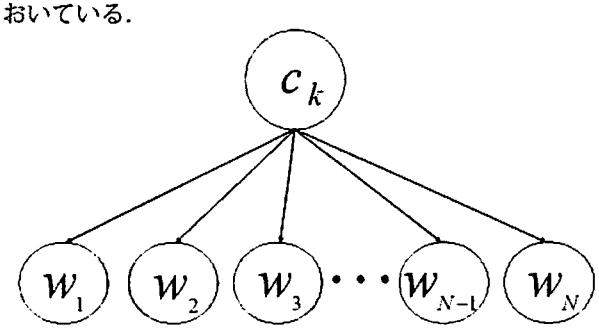

図 1: 文書における単語の生起

すなわち，$P(d_j|c_k)$ を次式のように仮定する．

$$P(d_j|c_k) = P(w_1, \ldots, w_N|c_k) = \prod_{i=1}^{N} P(w_i|c_k). \quad (2)$$

以上から，(1) 式は次式のように書き換えられる．

$$c_{\hat{k}} = \arg\max_{c_k} P(c_k) \prod_{i=1}^{N} P(w_i|c_k) \quad (3)$$

文書 d_j は (3) 式を最大化するカテゴリ $c_{\hat{k}}$ に分類される．

3.2 特徴選択

ナイーブベイズ分類では学習で得られる全単語の集合を W_{total} としたとき，W_{total} を縮小した $W(\subset W_{total})$ を分類に用いる単語とすることで計算量の増加や過学習などが避けられる [2]．以下，この単語集合 W を選択することを特徴選択と呼ぶ．一般に特徴選択の基準として相互情報量が用いられ [2]，その値が大きい単語を $N'(N' \leq N)$ 個選択する．相互情報量は

$$I(w_i;C) = \sum_{c_k \in C} P(w_i, c_k) \log \frac{P(w_i, c_k)}{P(w_i)P(c_k)} \quad (4)$$

で定義される．ここで，$P(w_i, c_k)$ は単語 w_i とカテゴリ c_k の同時確率，$P(w_i)$ は単語 w_i の生起確率，$P(c_k)$ はカテゴリ c_k の生起確率を表す．

3.3 従来手法の問題点

ナイーブベイズ分類ではカテゴリのもとで各単語の生起は独立であると仮定しているが，自然言語において各単語の生起の間には完全な独立性は成り立たないと考えられる．そこで，本研究ではナイーブベイズ分類に単語の共起を考慮に入れた手法を提案する．しかし，単語の組合せは膨大に存在するため，すべての組合せを考慮する事は困難である．そこで次節で述べる分類多岐語とその他の単語の 2 単語の組合せについてのみ共起を考慮することとする．

4 分類多岐語

相互情報量に基づく特徴選択では，出現頻度の偏りの大きい単語が抽出される傾向がある．ただし，必ずしも上位の単語がカテゴリを特徴づけるとは限らない．出現頻度の大きい単語では複数のカテゴリに同程度の頻度で出現する場合でも，出現頻度の偏りが大きくなる傾向がある．

表 1: 単語出現数の例

相互情報量	単語	経済	家庭	芸能	スポーツ
大	観客	3	2	71	77
小	野球	2	1	3	24

表 1 は相互情報量により選択された単語を示した例である．相互情報量による特徴選択ではあるカテゴリに特徴的に出現し，かつ出現文書数の大きい単語であるほど，選択されやすいという性質がある．表 1 を見ると「観客」の順位は高いが，芸能とスポーツ両カテゴリにおいて出現する単語数がほぼ同等であり，特定のカテゴリを特徴付ける単語といえない．一方「野球」は特定のカテゴリを特徴付ける単語といえるが，全カテゴリでの出現頻度が小さいため，(4) 式より相互情報量の値は小さくなる．この例のように「観客」は複数のカテゴリを特徴付けるが，「野球」のような単語との共起を見ることで唯一のカテゴリを特徴付けることができ，そのような組を分類規則に含めればより分類精度の向上が期待できる．

そこで，「観客」のように複数のカテゴリにおいて頻出する単語を分類多岐語と定義する．

定義 1 (分類多岐語).
単語 w_i に対し，$V(w_i)$ を

$$V(w_i) = \left(\frac{1}{\sum_k n_{w_i c_k}} \right)^2 \left(n^{(1)}_{w_i c_{k_1}} - n^{(2)}_{w_i c_{k_2}} \right)^2 \quad (5)$$

と定義する．ただし，

$n^{(1)}_{w_i c_{k_1}}$: ある単語 w_i について最も出現単語数が大きいカテゴリ c_{k_1} での出現単語数

$n^{(2)}_{w_i c_{k_2}}$: ある単語 w_i について 2 番目に出現単語数が大きいカテゴリ c_{k_2} での出現単語数

$\sum_k n_{w_i c_k}$: ある単語 w_i の総出現単語数

$V(w_i)$ がある閾値 ε $(\varepsilon \geq 0)$ 以下の単語 w_i を分類多岐語と呼ぶ． □

特に分類多岐語である事を強調したい場合は y_l と表記する．また，分類多岐語の集合を Y とする．

5 提案手法

5.1 共起単語の相互情報量

分類多岐語とそれ以外の単語との 2 単語の共起に関して相互情報量を計算し，最も大きい 1 単語を共起を考慮する単語とする手法を提案する．単語 w_i と w_j の 2 単語の共起に関する相互情報量は

$$I(w_i, w_j; C) = \sum_{c_k \in C} P(w_i, w_j, c_k) \log \frac{P(w_i, w_j, c_k)}{P(w_i, w_j)P(c_k)} \quad (6)$$

で計算できる．ここで，$P(w_i, w_j, c_k)$ は単語対 w_i, w_j とカテゴリ c_k の同時確率，$P(w_i, w_j)$ は単語 w_i と単語 w_j の同時確率，$P(c_k)$ はカテゴリ c_k の生起確率を表す．

5.2 提案アルゴリズム

step1 [相互情報量の計算]
単語を相互情報量の降順に並べ，その上位 N' 個の単語を選択する．その時の相互情報量の閾値を $z(z \geq 0)$ とする．

step2 [分類多岐語の抽出]

選択された単語 N' 個から $V(w_i) < \varepsilon$ となる分類多岐語 y_l を抽出する.

step3 [共起単語の選択]

抽出された分類多岐語 y_l に対し，それ以外のすべての単語 w_i との組について相互情報量 $I(y_l, w_i; C)$ を計算し，閾値 z 以上，かつ，最も相互情報量の大きい単語 w_{i_l} を共起を考慮する単語とする．すなわち

$$i_l = \arg\max_i \{I(y_l, w_i; C) | \, w_i \in W_{total} \setminus Y\} \quad (7)$$

とする.

5.3 分類規則

共起 (y_l, w_{i_l}) を考慮に入れるため，$P(d_j|c_k)$ を次式のように仮定する.

$$P(d_j|c_k) = \prod_{w_i \in W_{total} \setminus Y} P(w_i|c_k) \times \prod_{y_l \in Y} P(y_l|w_{i_l}, c_k). \quad (8)$$

$$c_{\hat{k}} = \arg\max_{c_k} P(c_k) \prod_{w_i \in W_{total} \setminus Y} P(w_i|c_k) \times \prod_{y_l \in Y} P(y_l|w_{i_l}, c_k). \quad (9)$$

(8) 式の確率を最大とするカテゴリ c_k に文書を分類する.

6 実験方法

6.1 実験データ

実験データとして 94 年毎日新聞記事データ [7] を用いる．扱うカテゴリ数は 9 カテゴリ，単語として用いる品詞は名詞のみとし，学習に約 9,000 文書，テスト用の文書として約 4,500 文書を用いた．学習で得られた全単語数 $N = 39,129$ であった.

6.2 評価指標

[正解率]

$$\text{正解率} = \frac{\text{正しく分類された文書数}}{\text{全テスト文書数}}. \quad (10)$$

[F 値]

F 値はカテゴリごとに得られる値で，適合率と再現率の評価重みを等しくおく場合，以下のように定義される.

$$F \text{ 値} = \frac{\text{適合率} \times \text{再現率} \times 2}{\text{適合率} + \text{再現率}}. \quad (11)$$

分類性能の評価指標として，正解率と 9 カテゴリの平均 F 値 [6] を用いる．以降 9 カテゴリの平均 F 値を単に F 値と呼ぶ.

6.3 スムージング法

スムージング法として，$P(w_i|c_k)$ と $P(w_i, w_j|c_k)$ の推定にラプラス法 [5] を用いた．B_{ki} を学習文書でのカテゴリ c_k における単語 w_i を含む文書数，$B_{ki,j}$ を学習文書でのカテゴリ c_k における単語 w_i と単語 w_j を同時に含む文書数，$|c_k|$ を学習文書でのカテゴリ c_k に含まれる文書数とすると，

$$P(w_i|c_k) = \frac{1 + B_{ki}}{2 + |c_k|}. \quad (12)$$

$$P(w_i, w_j|c_k) = \frac{1 + B_{ki,j}}{2 + |c_k|} \quad (13)$$

である.

6.4 予備実験

本実験を行う前に従来手法と提案手法で用いる選択単語数を決定するため，選択単語数による従来手法の精度の変化を調べた．図 2 のように単語数 $N' = 4,000$ 語で

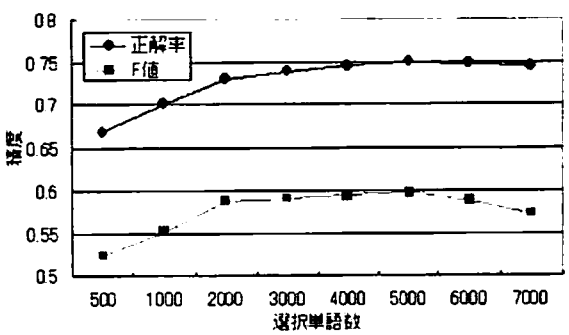

図 2: 選択単語数による従来手法の精度の変化

ほぼ横ばいになり，それ以上では精度の低下が見られた．本研究では選択単語数を $N' = 4,000$ とした.

7 実験結果と考察

7.1 実験結果

図 3 に共起を考慮する分類多岐語数を閾値 ε を 0.001 刻みで 0.001 から 0.015 まで変化させ抽出した結果を示す（この時の分類多岐語数は 203 ～ 738 に対応している）．ただし，共起を考慮する分類多岐語数 0 は従来手法を表す．また，表 2 に F 値が最大の時の従来手法と提案手法の精度を示す.

表 2: F 値が最大の時の従来手法と提案手法の比較

	従来	分類多岐語数 403(閾値 0.004)
正解率	0.746	0.762
F 値	0.593	0.617

7.2 考察

1. 表 2 より正解率，F 値がそれぞれ 1.6，2.4 ポイント向上し，提案手法の有効性が示された．表 3 は実験中に得られた共起単語の例である．表 3 において，

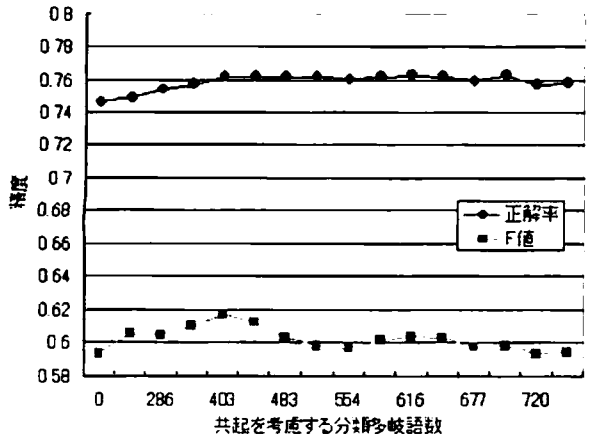

図 3: 共起を考慮する分類多岐語数による精度の変化

表 3: 共起単語の例

分類多岐語	共起単語	分類多岐語	共起単語
著作権	音楽	ミュージカル	演出
消費者	価格	先生	小学校
金	メダル	希望	小売価格
治療	患者	文献	研究
江戸	時代	酒	おおさじ
関西	大阪	阪神	打撃
EU	貿易	ヒット	商品
芝	馬	ハウス	ワシントン

例えば「ヒット」は一見“野球のヒット”を想像するが,「商品」との共起を考慮すると“ヒット商品”の意味になり，経済カテゴリに出現するような意味となる．また，その他の分類多岐語でも共起を見ることでどのような意味で用いられるかが想像できるようになったと考えられる．

2. 図 3 を見ると共起を考慮する分類多岐語数を増やす（閾値 ε を上げる）とある一定までは精度が向上するが，それ以上では向上が見られなくなることが分かる．これは閾値を上げると唯一のカテゴリを特徴付ける単語までも分類多岐語として抽出されるため，共起を考慮しても精度向上につながらないためと考えられる．
3. 表 4 に従来手法，提案手法における正解文書数，不正解文書数を示す．従来手法では誤分類されたが，提案手法で正しく分類できた 156 文書について 1 例を示す．d =(記念，G 3，芝，連，馬) という文書は従来手法ではカテゴリ社会に誤分類されるが，提案手法ではスポーツカテゴリに正しく分類できる．これは表 3 に示したように，単語「芝」,「馬」が共起することにより，競馬のトピックであることが明らかになり，スポーツカテゴリの意味であることが明確になったためといえる．
4. 文書分類の前処理に当る学習ステップにおいて共起頻度を調べる計算量は増加したが，分類処理の部分においては従来も提案も計算量は大きくは変わらなかった．

表 4: 従来・提案手法における正解・不正解文書数

		従来不正解	従来正解
提案	不正解	986	86
(共起多岐語数 403)	正解	156	3272

8 まとめと今後の課題

特定の複数のカテゴリを特徴付ける分類多岐語を定義し，ナイーブベイズ分類に単語の共起を考慮に入れた分類手法を提案した．これにより，分類多岐語の意味を限定し，分類精度を向上させることができる．また，実験により分類精度の向上を示した．

今後は 2 単語間の共起だけではなく，複数単語間の共起なども考える予定である．

9 謝辞

著者の一人である津田は，本研究を行うにあたり，数多くのご助言，ご支援を賜りました早稲田大学平澤研究室の各氏に感謝いたします．本研究の成果の一部は早稲田大学特定課題研究助成 $No.2005B-189$ による．

参考文献

[1] 北研二: 確率的言語モデル，東京大学出版会，1999.
[2] A. McCallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification”, *In Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pp. 41-48, 1998.
[3] 津田裕一，山岸英貴，石田崇，平澤茂一，“相互情報量に基づく特徴選択を用いた文書自動分類，” FIT2005(第 4 回情報科学技術フォーラム) 論文集，D-029，2005 年.
[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，pp. 101-106，2005 年.
[5] Karl-Michael Schneider, “On Word Frequency Information and Negative Evidence in Naive Bayes Text Classification，” *EsTAL 2004, LNAI 3230*, pp. 474-485, 2004 年.
[6] Y. Yang, “An Evaluation of Statistical Approaches to Text Categorization，” *Journal of Information Retrieval*, Vol. 1, No. 1/2, pp. 67-88, 1999.
[7] CD：毎日新聞'94，毎日新聞社，日外アソシエーツ.