The 27th Symposium on Information Theory
and Its Applications (SITA2004)
Gero, Gifu, Japan, Dec. 14-17, 2004

単語の潜在的意味を考慮した文書分類手法
Text Categorization Method based on Latent Meaning of Words

松井 治樹* 石田 崇* 平澤 茂一*
Haruki MATSUI Takashi ISHIDA Shigeichi HIRASAWA

Abstract— Text categorization is the technology which automatically assigns the given document to the decided category. Recently, various methods have been proposed, but most methods consider an individual word as a symbol, and the meaning of word is not taken into consideration. The Naive Bayes classifier, which is the most typical technique by using probability, also say it. On the other hand, the technique of extracting the meaning of words is studied actively. Recetly, as the technique of extracting the latent meaning of a word by using probability, PLSA(Probabilistic Latent Semantic Analysis) and SAM(Semantic Aggregate Model) have been proposed. This paper propose by using Naive Bayes Classifer and SAM, a new text categorization method based on latent meaning of words.

Keywords— Text Categorization, Naive Bayes, SAM

1 はじめに

文書分類は与えられた文書をあらかじめ決められた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ン，確率を用いた手法など多くの手法が提案されている [5].

確率を用いる最も代表的な文書分類手法に Naive Bayes 分類 [1] がある．この手法ではカテゴリが与えられた下で文書を互いに独立な単語の系列とみなす．ここで単語は単に記号として扱われ，その意味や概念は考慮されない．

一方，単語の意味や概念を抽出する技術が盛んに研究されている．その手法としてシソーラスなどの大規模辞書を用いた手法，従来の自然言語処理技術を応用した LSA(Latent Semantic Analysis)，確率モデルに基づいた PLSA(Probabilistic Latent Semantic Analysis)，SAM(Semantic Aggregate Model)[2][3] などがある．

本稿では，SAM を用いて単語の潜在的意味を考慮した文書分類手法を提案する．

2 確率を用いた文書分類手法

本節では確率を用いた代表的な文書分類手法である Naive Bayes 分類について説明する．

2.1 Naive Bayes 分類 [1]

Naive Bayes 分類では，カテゴリ $k \in K$ が与えられた下で各単語 $w \in V$ は互いに独立という仮定を置く．このときカテゴリ k の下で文書 d が生起する確率を

$$P(d|k) = P(|d|)|d|! \prod_w \left(\frac{P(w|k)}{n(w)!} \right)^{n(w)} \tag{1}$$

としている．$P(|d|)$ は長さ $|d|$ の文書が生起する確率であり，$n(w)$ は文書 d における単語 w の頻度である (文書 d の長さとは文書 d における全単語の頻度である)．ここである文書 d がカテゴリ k に帰属する確率 $P(k|d)$ は

$$\begin{aligned} P(k|d) &= \frac{P(k)P(d|k)}{P(d)} = \frac{P(k)P(d|k)}{\sum_{k'} P(k')P(d|k')} \\ &= \frac{P(k)P(|d|)|d|! \prod_w \left(\frac{P(w|k)}{n(w)!} \right)^{n(w)}}{\sum_{k'} P(k')P(|d|)|d|! \prod_w \left(\frac{P(w|k')}{n(w)!} \right)^{n(w)}} \\ &= \frac{P(k) \prod_w P(w|k)^{n(w)}}{\sum_{k'} P(k') \prod_w P(w|k')^{n(w)}} \end{aligned} \tag{2}$$

と簡略化される．Naive Bayes 分類では，文書 d が生起したとき確率 $P(k|d)$ をカテゴリごとに算出し，最も高いカテゴリに分類する．

2.2 Naive Bayes 分類と特徴選択

学習で得られる全単語の集合を V_{total} としたとき，Naive Bayes では分類に用いる単語を選択して V_{total} を縮小し，$V(\subset V_{total})$ とすることで分類性能が向上するとされている [1]．以下この単語の選択を特徴選択とよぶ．一般に特徴選択の選択基準として相互情報量 (式 (3)) が用いられる [1].

$$I(K;w) = \sum_{k \in K} P(w,k) \log \left(\frac{P(w,k)}{P(w)P(k)} \right) \tag{3}$$

3 単語の意味抽出技術

単語の潜在的な意味を抽出する技術として SAM が持橋らによって提案されている [2][3]．SAM は単語の意味の概略を数学的に見通しよく扱うことを目的として提案された手法である．ここでは SAM の概要を説明する．

3.1 SAM(Semantic Aggregate Model)[2][3]

SAM のグラフィカルモデルは図 1 のようになる．これは，ある単語 w とある単語 w' が共起する際，共通の潜在的な意味クラス $c \in C$ の存在を仮定したモデルである．このとき，単語 w と単語 w' の生起確率を

$$P(w,w') = \sum_c P(w|c)P(w'|c)P(c) \tag{4}$$

と考える．$N(w,w')$ を実際のデータにおける共起回数の観測値とすれば，データの尤度

$$L = \sum_{w,w'} N(w,w') \log P(w,w')$$

を最大にする $P(c)$，$P(w|c)$ は EM アルゴリズムにより次のように最尤推定できる．

E-Step

$$P(c|w,w') = \frac{P(w|c)P(w'|c)P(c)}{\sum_{c'} P(w|c')P(w'|c')P(c')}$$

*〒 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 Okubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555 Japan. E-mail: matsui@hirasa.mgmt.waseda.ac.jp

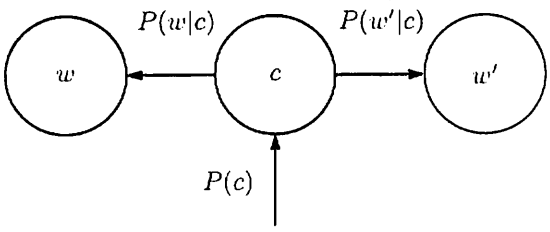

図 1: SAM のグラフィカルモデル

M-Step

$$P(c) = \sum_{w,w'} P(c|w,w')N(w,w')$$

$$P(w|c) = \frac{\sum_{w,w'} P(c|w,w')N(w,w')}{P(c)}$$

3.2 SAM の性質

SAM は単語のもつ潜在的な意味を確率として表現することができる．しかし単語 w と単語 w' の共起を (w,w') としたとき，共起の系列 $\{(w,w')\}$ をどのように観測するかを一意に決めることができないという問題がある．すなわち共起の系列の長さをどう決めるか，それに対し頻度をどう取るか，これらを確率的に考えるための最適な方法に議論の余地が残されている [3]．

4 単語の潜在的意味を考慮した文書分類手法の提案

Naive Bayes 分類をはじめ，通常の文書分類手法は単語の意味を考慮せず単語を記号的に扱う．本研究では SAM を用いて単語の潜在的な意味を考慮し，Naive Bayes モデルを拡張した文書分類手法を提案する．

4.1 SAM の文書分類への適用

SAM はその目的から，一般的な日本語の文章から共起 (w,w') の頻度を観測し，単語の潜在的な意味を抽出する．本研究では文書分類への適用を見据え，カテゴリごとに共起 (w,w') の頻度を観測して SAM により必要な確率を推定する．

SAM を用いてカテゴリごとに確率の推定を行う利点は，以下の 3 つである．

共起単位で見た利点として，あるカテゴリ k における共起 (w,w') に対して，意味を考慮した確率値 $P(w,w'|k)$ を得ることができる．次に SAM では各共起 (w,w') は互いに独立としているが，2 単語を考えているため，ある系列 $\boldsymbol{x}$ は $\boldsymbol{x} = \{\cdots,(w_i,w_j),\cdots,(w_j,w_k),\cdots\}$ というように共起の要素は重なりをもつ．よって一系列の確率を考えたとき，系列全体の意味を考慮した値が得られると考えられる．

最後に SAM では各共起 (w,w') は互いに独立と仮定しているため，Naive Bayes 分類で用いられる分類規則を自然に適用できる．詳しくは 4.3 節，4.4 節で述べる．

4.2 SAM を用いたカテゴリごとの確率の推定

カテゴリごとに共起データを観測した場合，図 1 の SAM モデルは図 2 のようなグラフィカルモデルとして考えることができる．今 $P(x|k) = P_k(x)$ というように表記すると，カテゴリ k の下で潜在的意味を考慮した共起確率 $P_k(w,w')(= P(w,w'|k))$ は

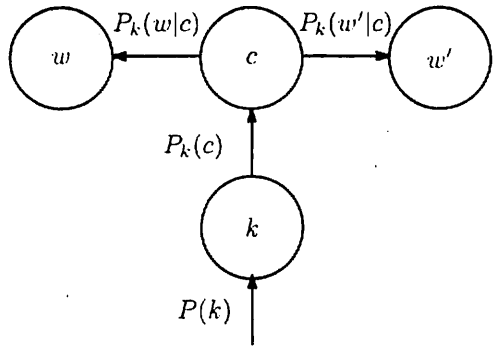

図 2: 提案のグラフィカルモデル

$$P_k(w,w') = \sum_c P_k(w|c)P_k(w'|c)P_k(c) \tag{5}$$

であり，これは SAM 同様，EM アルゴリズムを用いて以下のように推定できる．

E-Step

$$P_k(c|w,w') = \frac{P_k(w|c)P_k(w'|c)P_k(c)}{\sum_{c'} P_k(w|c')P_k(w'|c')P_k(c')}$$

M-Step

$$P_k(c) = \sum_{(w,w')} P_k(c|w,w')N_k(w,w')$$

$$P_k(w|c) = \frac{\sum_{(w,w')} P_k(c|w,w')N_k(w,w')}{P_k(c)}$$

$N_k(w,w')$ はカテゴリ k において観測された共起 (w,w') の頻度である．

4.3 分類規則

Naive Bayes 分類同様，カテゴリごとに SAM を用いる提案モデルは，カテゴリ k を与えられた下で共起 (w,w') を互いに独立としている．このときカテゴリ k の下で文書 d が生起する確率は，文書 d を共起の系列 $\boldsymbol{d} = \{(w,w')\}$ として

$$P(d|k) = P(|d|)|d|! \prod_{(w,w')} \left(\frac{P(w,w'|k)}{n(w,w')!}\right)^{n(w,w')} \tag{6}$$

である．なお $n(w,w')$ は文書 d における共起 (w,w') の頻度であり，$P(w,w'|k)$ は潜在的な意味を考慮した確率 (式 (5)) を用いている．ここで文書 d が生起したとき，カテゴリ k に帰属する確率 $P(k|d)$ は，

$$P(k|d) = \frac{P(k)\prod_{(w,w')} P(w,w'|k)^{n(w,w')}}{\sum_{k'} P(k')\prod_{(w,w')} P(w,w'|k')^{n(w,w')}} \tag{7}$$

である．この帰属確率最大基準に基づき，文書を分類する．

4.4 Naive Bayes 分類との比較

Naive Bayes モデルの概念図は一般に図 3 で表される．これはカテゴリ k の下で各単語 w は互いに独立である

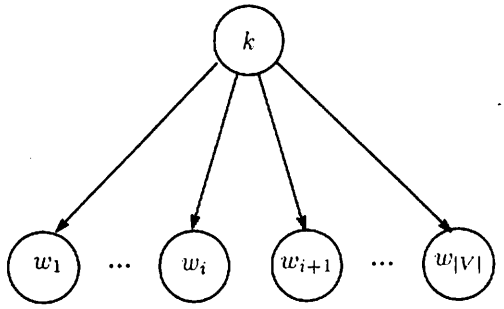

図 3: Naive Bayes モデル

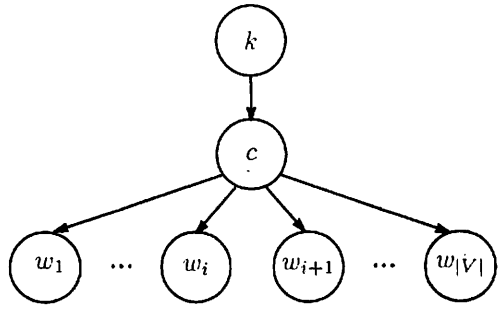

図 4: 提案モデル

ことを表している．一方，提案手法のモデルの概念図は，カテゴリ k のある意味クラス c に注目した場合，図 4 のように表現できる．これはカテゴリ k の下で単語 w は各意味クラス c を介して互いに条件付独立であることを意味している (なお，カテゴリ k の下で意味クラス c の独立性は仮定していない)．この仮定により，カテゴリ k の下での共起確率 $P(w, w'|k)$ を，単語 w が各意味クラス c に所属する確率 $P(w|c)$ の分布を考慮して算出できる．提案手法は Naive Bayes モデルの表現力を高め，その分類規則を自然に適用した手法といえる．

5 実験方法

5.1 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，学習に約 9500 文書，テスト用の文書として約 4500 文書を用いた．

単語として用いる品詞は名詞のみとし，学習で得られた全単語数は 44460 であった．

5.2 評価指標

分類性能の評価指標として，F 値のマイクロ平均とマクロ平均 [4] を用いる．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平等とする場合，以下のように定義される．

$$F\text{ 値} = \frac{\text{適合率} \times \text{再現率} \times 2}{\text{適合率} + \text{再現率}}$$

F 値のマイクロ平均は，今回ひとつの文書を必ず 1 カテゴリに分類するため一般に正解率と呼ばれる指標と同等となる．

$$\text{正解率} = \text{正分類数}/\text{全テスト文書数}$$

一方，マクロ平均は各カテゴリの F 値を平均したもので，カテゴリごとのテスト文書数の違いを問わない指標である．

以降，F 値のマイクロ平均を正解率，F 値のマクロ平均を平均 F 値とよぶ．

5.3 スムージング法

スムージング法として，Naive Bayes 分類で用いる $P(w|k)$ はラプラス法 [6] を用いた．提案手法の $P(w, w'|k)$ は推定により得るため，単純にラプラス法を適用できない．そこで 1/(単語数 × 単語数) を加えて合計が 1 となるようにスムージングを行った．

5.4 提案手法の意味クラス数，収束条件

各カテゴリで EM アルゴリズムを実行する提案手法は，学習に相当の時間を必要とする．そのため，意味クラス数を各カテゴリ 10 と固定し，EM アルゴリズムの収束条件はステップ数を 150 と決めて 3 回行い，最も尤度が高かった結果を用いた．

6 予備実験

本実験を行う前に提案した分類手法を，共起の観測の仕方，特徴選択の 2 つに関して以下の予備実験を行った．予備実験では全 44460 語のうち，特徴選択で 30 % の単語を用いる．

6.1 共起の観測法

提案手法において，SAM と同様に確率の推定で問題となる共起の観測法について実験する．

共起 (w, w') の系列の長さは文書単位とし，頻度の観測法として以下の 2 つの方法を試した．例として，ある文書において単語 w_1, w_2, w_3, w_4 がそれぞれ 1, 3, 5, 2 回出現したとして説明する．

[観測法 1] 出現した共起の頻度を 1 とする方法

w_1, w_2, w_3, w_4 のすべての組み合わせに対して，その頻度を 1 とする．これは一文書で出現する各共起をすべて同等に扱う観測法である．

[観測法 2] 単語の頻度に基づく方法

単語の組を考えたとき，その頻度の小さい方の値を共起頻度とする．すなわち w_1 と w_2, w_3, w_4 のそれぞれの組み合わせは頻度 1，(w_2, w_3) の頻度は 3，(w_2, w_4) の頻度は 2 となる．これは 2 つの単語が組として何回出現したかという考えに基づいた観測法である．

6.2 特徴選択

Naive Bayes 分類では単語は単なる記号として扱われるため，$P(w, K)$ の分布が偏った特徴的な単語を選択することが重要である．その基準として相互情報量 (式 (3)) が用いられる．

本手法では共起を基にその潜在的な意味を考えるため，用いる単語が少ない場合，相互情報量のような特徴選択ではうまく共起を観測することが難しいと考えられる．そのため相互情報量とは別に，単純に各カテゴリで出現頻度の高い単語から累積頻度率を閾値として選択して合成するという特徴語選択法を試す．累積頻度率とは，N_k をカテゴリ k における全単語の頻度の合計，N'_k を選択した単語の頻度の合計として，N'_k/N_k である．

6.3 予備実験結果

予備実験の結果を表 1 に示す．特徴選択法の欄の「頻度」は累積頻度率を閾値にして頻度に基づき単語を選択した方法である．

表 1: 予備実験結果

共起観測法	特徴選択法	正解率	平均 F 値
観測法 1	頻度	0.816	0.762
観測法 1	相互情報量	0.813	0.757
観測法 2	頻度	0.813	0.759
観測法 2	相互情報量	0.810	0.756

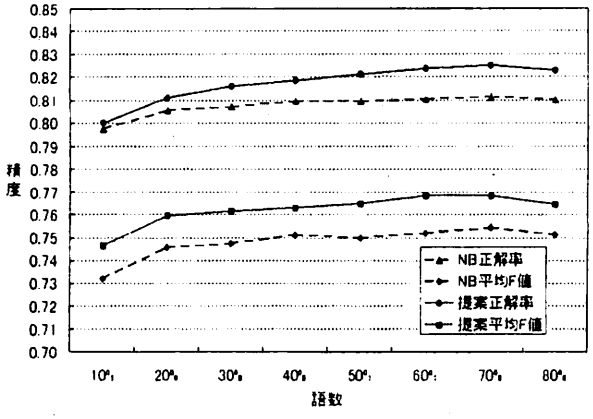

図 5: Naive Bayes と提案手法の精度比較

表 2: 単語数 70 % 時の精度比較

手法	正解率	平均 F 値
Naive Bayes	0.812	0.754
提案手法	0.825	0.769

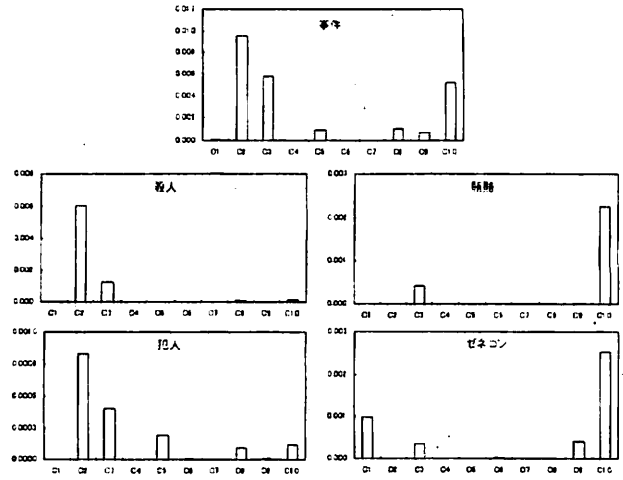

図 6: 社会における各単語の $P(w|c)$

表 1 より最も精度がよかったのは，観測法 1，頻度に基づく特徴選択を用いた場合であった．共起の観測法は観測法 2 よりも観測法 1，特徴選択法は相互情報量よりも頻度に基づく選択法の方がよい精度を示した．

7 本実験

予備実験で最もよい性能を見せた，共起の観測法 1，頻度に基づく特徴選択法を用いた手法を提案手法として Naive Bayes 分類と比較する．

7.1 実験結果

実験結果を図 5 に示す．横軸は全単語数 44460 語のうち分類に用いた語数の割合，縦軸は精度を示す．提案手法が Naive Bayes 分類よりも正解率，平均 F 値の両方において精度がよいのがわかる．

両者とも最良の精度を示した単語数 70 % 時の Naive Bayes 分類と提案手法の分類精度を表 2 に示す．正解率で 1.3 %，平均 F 値で 1.5 % の精度の向上が見られた．

また予備実験の結果 (表 1) より，どの共起の観測法，特徴選択法を用いても，提案手法は単語数 30 % で単語数 70 % を用いた Naive Bayes 分類の平均 F 値を上回っている．少ない語数で Naive Bayes 分類の最良の精度と同等となったのはモデルの表現力を高めたためといえる．また，共起の観測法，特徴選択法により提案手法の精度がひどく不安定になることはないといえる．

7.2 考察

提案手法における「潜在的意味の考慮」に関して考察を行う．今回実験に用いた実験データで，カテゴリ「社会」において SAM を用いて推定した確率のうち，単語「事件」,「賄賂」,「ゼネコン」,「殺人」,「犯人」の $P_k(w|c)$ の分布を図 6 に示す．

図 6 より,「事件」は意味クラス c_2，c_3，c_{10} で高い値を示している．これは分野「社会」において,「事件」は大きく 3 つの意味で用いられると考えられる．ここで，単語「殺人」,「犯人」が意味クラス c_2，c_3 で高い値をもつのに対し,「賄賂」,「ゼネコン」は意味クラス c_2 ではほとんど値をもたず，意味クラス c_{10} で高い値を示す．このように提案手法では他の単語との共起を見ることで，どのような「事件」なのか，おおまかな意味を考慮できたため精度が向上したと考えられる．ただし今回は意味クラス数が少なかったためか，意味クラス c_3 のように一見して解釈が難しいものも多々見られた．

8 まとめと今後の課題

SAM と Naive Bayes モデルを用いて，単語の潜在的意味を考慮した文書分類手法を提案した．実験により分類精度の向上が見られた．

しかし，依然 SAM における共起頻度に関する問題を解決しておらず，意味クラス数の最適な与え方，また共起を考えた特徴選択法など，まだまだ改善すべき点も多い．計算時間も合わせてこれらの点に関しては今後の課題としたい．

9 謝辞

著者の一人である松井は，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．

参考文献

[1] A.McCallum, 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).
[2] 持橋大地，松本祐治: "意味の確率的表現", 情報処理学会研究報告, 自然言語処理研究会, 2002-NL-147, pp.77-84, (2002).
[3] 持橋大地，松本祐治: "PLSA による確率的概念空間の評価", 情報処理学会研究報告, 自然言語処理研究会, 2002-NL-153, pp.41-47, (2003).
[4] Y.Yang,"An Evaluation of Statistical Approaches to Text Categorization", *Journal of Information Retrieval*, Vol.1, No.1/2, pp.67-88, (1999).
[5] 金明哲ほか: 統計科学のフロンティア 10 言語と心理の統計, 岩波書店, (2003).
[6] 北 研二: 確率的言語モデル, 東京大学出版会, (1999).
[7] CD：毎日新聞'94, 毎日新聞社, 日外アソシエーツ.