DEIM Forum 2009 B6-6

Naïve Bayes分類における 高頻度語と低頻度語の扱いについて

–ニュース記事に於ける–

岡本　遼† 内田　晋† 塩谷　勇† 三浦　孝夫††

† 産業能率大学情報マネジメント学部　〒 259–1197 神奈川県伊勢原市上粕屋 1573
†† 法政大学工学部　〒 184–8584 東京都小金井市梶野町 3–7–2

あらまし　この論文では Naïve Bayes による日本語ニュース記事の分類に於ける高頻度語と低頻度語の扱いについて述べる。従来、Zipf の法則から分類に有効でない高頻度語の上限と低頻度語の下限を決定している。この論文では、Naïve Bayes 分類に於いて Bayesian Information Criterion (BIC) の基準から、高頻度語と低頻度語の扱いを統一的に論じる。その結果、BIC による高頻度語と低頻度語の選択が、将来の記事の分類精度を高めることが実験によって示され、本手法の有効性が示された。

キーワード　Naïve Bayes 分類, Bayesian Information Criterion, モデル選択

Ryo OKAMOTO†, Susumu UCHIDA†, Isamu SHIOYA†, and Takao MIURA††

† Sanno University　Kamikasuya 1573, Isehara, 259–1197 Japan
†† Hosei University　Kajinocho 3–7–2, Koganei, 184–8584 Japan

1. ま え が き

本研究の目的は、分類が既知の過去と現在のニュース記事を学習して、分類が未知の将来のニュース記事を精度高く分類することである。従来、Naïve Bayes [1] などの分類手法が用いられているが、時間の推移を考慮なしに分類精度を高める試みがされている。この論文では、過去と現在の記事の時間差の対応関係を考慮して分類学習することで、将来のニュース記事の分類法について論じる。高頻度語は網羅性が高く、時間的な変動が小さいと推測できる。一方、低頻度語は特定性が高く、時間的な変動が高いと推測できる。高頻度語と低頻度語の上限と下限を変化させて Naïve Bayes [1] で学習して Bayesian Information Criterion(BIC) [9] で最適化することで将来の記事の分類精度が実験によって高くなる事から、本手法の有効性を示す。すなわち、ニュース記事などの時間推移の考慮と高頻度語と低頻度語の扱いを組み入れた Naïve Bayes 分類手法を BIC という統一基準で最適化する将来のニュース記事分類モデルを提案し、その有効性を確認する。

情報の検索や分類はわれわれの抱えている問題を解決してくれる有用なツールになると期待されている [11]。自然言語の重要な特性の一つが Zipf の法則で、言語の差異に大きく依存することなく、広い意味で成立することが示されている。検索や分類で単語の出現頻度を基準に扱うベクトル型の確率モデルでは文書に出現する頻度の高い語と頻度の低い語を文書の索引から除外する。この理由は文書の網羅性や特定性を応用問題の適度な水準に維持することが目的で、文書の意味を考慮しない検索や分類でも比較的に高い性能が得られている。これらのツールの設計は、ヒューリステックな要素が含まれ、特に不要語や高頻度語・低頻度語の扱いで、検索や分類の性能が大きく左右される [2]。高頻度語・低頻度語の特性に経験則である Zipf の第 1 法則と第 2 法則が知られている [11]。Zipf の第 1 法則と第 2 法則から中頻度語を求める手法が知られている [11] が、高頻度語の上限と低頻度語の下限は経験的に決められている。この論文では過去のニュース記事の学習から、 現在と将来のニュース記事を確率モデル Naïve Bayes によって既知のカテゴリに分類する場合に、高頻度語と低頻度語の扱いについて BIC [9] の立場から統一的に論じる。

SPAM メイルの判定やニュース記事の分類は情報を分類する意味で重要な応用技術である。通常の教師あり分類は既知の文書を学習して分類が未知の記事を分類するが、ニュース記事の分類は、過去の記事を学習記事とし、現在や未来の記事を分類する。すなわち、正しく分類された過去の記事を学習し、新しい記事を分類しする時間推移を考慮した教師あり学習である。例えば、「首相、昨日永田町 · · · 」が政治、「日本経済は本日 · · · 」が経済の記事に分類されていると仮定する。では、「本日、日本の · · · 」の記事は、何れに分類すべきか。Naïve Bayes 法を使

用した分類ではニュース記事を単語とその品詞に分解後、不要語と共に単語の品詞や出現頻度から分類に有用でない単語が除外される。Naïve Bayes 法は教師あり学習法の一つであり、正しく分類されているニュース記事の教師データから、分類が未知のニュース記事を教師データに沿って分類する。この手法は記事の中の各単語の出現が独立であると仮定することで、カテゴリと記事の同時出現確率の計算が容易になり、同時確率を最大にするカテゴリに記事を割り当てる。記事中の各単語の出現が独立(注1)という強い仮定にもかかわらず、比較的高い精度が得られる。また、その処理もデータベースと整合性がよく、比較的高いパフォーマンスが得られる。

通常、Naïve Bayes 法では、教師データを学習して、テストデータで評価を行う。この論文は Naïve Bayes 法の用い方に特徴があり、過去の教師データで学習した結果が時間の新しいテストデータで分類を最適化することで、過去のデータと新しいデータの対応関係を学習して将来のデータの分類をするモデルである。このときに最適化に用いる基準が BIC である。われわれは Naïve Bayes 法を用いるが、分類の基準は BIC である。なお、時間軸上でカテゴリ変動を考慮した Näive Bayes 分類 Dynamic Naive Bayesian Classifier [?] が知られているが、本研究ではカテゴリ変動がないと仮定している。

われわれの問題は、古いニュース記事を学習データとして、時間的に新しい記事を分類する問題を扱う理由から、記事の分類性能が高いというよりも、過去の分類既知の記事の学習から将来の新しい分類未知の記事を高い性能で分類するモデル選択の問題と考えることができ、AIC (Akaike Information Criterion) [7] の考えと整合する。AIC は高頻度語と低頻度語の扱いの変更による単語数の変化がモデル選択に反映されないために、われわれの問題に適さない。MDL [10] や Bayesian Information Criterion (BIC) [9] のようなデータ量がモデル選択に反映される基準が適している。

モデル選択の立場から、AIC [7] と共に、MDL [10]、AIC の修正、Bayesian Information Criterion (BIC) [9] など、多くの研究がなされている。BIC [8], [9] は Bayes の考えに基づいており、将来に発生するデータに適したモデルという意味で、単語頻度を考えたニュース分類の応用に適した手法と考えられる。ここでは、Bayes の基づく Bayesian Information Criterion (BIC) [9] によって、高頻度語と低頻度語の問題をモデル選択の問題と考えて論じる。

この論文では次節以降、Naïve Bayes、BIC について述べ、つぎに高頻度語と低頻度語の扱いについて述べる。そして、評価実験について述べ、最後に結論を述べる。

2. Naïve Bayes

Naïve Bayes によるニュース記事のカテゴリの分類法について述べる。Naïve Bayes [1] は教師あり学習の一つであり、あらかじめカテゴリが既知の記事 (教師データ) と未知の記事であるテストデータが排他的に与えられ、教師データによって学習して、テストデータで評価を行う。

教師データを D、学習全記事数 N、カテゴリ $C_1, .. , C_k$、各カテゴリの単語の頻度 $N_1 = |C_1|, .. , N_k = |C_k|$ とする。各カテゴリ C_i の中の単語 $w_{i,j}$ の頻度 $|w_{i,j}|$ とすると、カテゴリ C_i が与えられたときの単語 $w_{i,j}$ の条件付確率は $p(w_{i,j}|C_i) = \frac{p(w_{i,j} \cap C_i)}{p(C_i)} = \frac{|w_{i,j}|}{N_i}$ で計算される。記事に出現しない単語の確率 0 を排除するため (ゼロ頻度問題 [5]) に、ラプラス法によるスムージングを用いると $p(w_{i,j}|C_i) \fallingdotseq \frac{|w_{i,j}|+1}{N_i+N_i'}$ となる。ここで、N_i' は C_i の単語の異なり数を表す。

分類する記事を $Q = q_1, .. , q_n$ として、q_i は記事の単語とする。各カテゴリの学習記事と、真のカテゴリに属する記事の特性に大きな変動がないとすると、近似的に $p(q_k|C_i) \fallingdotseq p(w_{i,j}|C_i)$ で計算できる。ここで、q_k=$w_{i,j}$ とする。分類する記事 Q の各単語 q_i が独立と仮定をすると、Q がカテゴリ C_i である同時確率は $p(q_1, .., q_n, C_i) = p(C_i) \times \Pi_k p(q_k|C_i)$ で与えられる。$p(C_i)$ はカテゴリ C_i の割合を表し、$p(C_i) = \frac{N_i}{N}$ で求める。従って、記事の分類問題は $\arg\max_{C_i} p(q_1, .., q_n, C_i)$ で表される。すなわち、$p(q_1, .., q_n, C_i)$ を最大にするカテゴリ C_i を見つける問題である。

3. Bayesian Information Criterion

Bayesian Information Criterion (BIC) [9] は当初、最適なモデル選択が可能であるとされたが、動機が異なるが MDL と等価なモデル選択として知られている。BIC は次の式を最小にするパラメタ θ を求める問題となる。

$$\mathrm{BIC} = -2 \cdot \log\ p(x/\theta) + m \cdot \log\ n$$

ここで、$p(x/\theta)$ は最大尤度、m はパラメータ数、n はデータ数を表す。AIC と異なる点はモデル選択の中に、データ数に応じたモデル選択の項がある。本研究の目的である、高頻度語と低頻度語の問題は、データ数に応じたモデル BIC が適している。

Naïve Bayes に BIC を適用する場合、同時確率 $p(q_1, .., q_n, C_i)$ は、分類に用いる高頻度語の上限 z_{high}、低頻度語の下限 Z_{low} とすると、$p(q_1, .., q_n, C_i/z_{low}, z_{high})$ と表される。従って、この場合のモデル選択のパラメータ数が 2 であり、データ数が単語数 w となる。従って、Naïve Bayes における BIC は以下の式を最小にする z_{low}, z_{high} を見つける問題となる。

$$\mathrm{BIC} = -2 \cdot \log\ p(q_1, .., q_n, C_i/z_{low}, z_{high}) + 2 \cdot \log\ w$$

$q_1, .., q_n$ は一つの評価記事に対応し、BIC による高頻度語と低頻度語の最適なモデルは、各評価記事に依存して選択しなければならない。より正確には、正解カテゴリに割り当てられれば、BIC の意味で最適なモデル選択が可能である。$p(q_1, .., q_n, C_i/z_{low}, z_{high})$ の計算は Naïve Bayes と同じ方法で行うことができる。

4. 高頻度語と低頻度語の扱い

文書の特定性と網羅性の観点から、低頻度語と高頻度語の削除が行われてきた。これらの値の選択は経験的に選ばれてきた。Zipf の法則は多くの文書で成立することが知られており、Zipf

(注1)：bag of words.

の第 1 法則、第 2 法則から [11], [16] では中頻度語を求めている。Zipf の第 1 法則は高頻度語で成立する法則で、単語の頻度 f と頻度の順位 r との積が定数 C なる法則である

$$f \times r = C$$

Zipf の第 2 法則は出現頻度 f の単語数 F_f と頻度 1 の単語数 F_1 の間に成立する関係である。

$$\frac{F_1}{F_f} = \frac{f(f+1)}{2}$$

中頻度の語は両方が成立する中頻度語を $f = f_k$ で求める。そして、頻度 1 の語を手がかりにすると次の式を得る。

$$f_k = \frac{\sqrt{8F_1 + 1} - 1}{2}$$

f_k が中頻度の語であるとして、低頻度語と高頻度語の削除すべきそれぞれの境界を決定する方法である。本研究では、中頻度語を直接使用しないが、実験によって将来の記事の分類精度が高い低頻度語と高頻度語の中心付近が Zipf の法則で求めた中頻度語に近いことが示される。

5. ニュース記事分類モデル

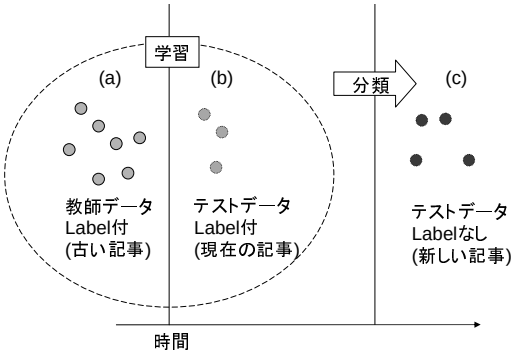

図 1　ニュース記事の分類モデル
Fig. 1　News document classification model.

本論文のニュース分類のデモルは、時間の推移、高頻度語と低頻度語の扱いを考慮に入れた、将来の記事を分類するモデルで、BIC という統一基準で最適化を行う (図 1)。カテゴリ分類が既知の古い記事 (a) と現在の記事 (b) を用意し、(a) を学習して Naïve Bayes から (b) を分類する。(b) の BIC の平均値が高くなる z_{low} と z_{high} を求める。すなわち、BIC の基準で分類の最適化をパラメータ z_{low} と z_{high} について行う。学習結果から (b) のカテゴリを推定して、その結果を再び学習する Naïve Bayes に EM を組み合わせて再び学習することも可能である。最後に、この学習結果を将来の記事 (c) で評価を行う。

このモデルの特徴的な事は、時間推移の変化を学習して、将来の記事の分類に用い、その基準を将来の予測に適している BIC モデルで統一的に議論する点にある。

6. 実　　験

6.1 準　　備

カテゴリに分類されている古いニュース記事から、新しいニュース記事を分類する問題に於いて、学習データの中に同一の記事があれば、学習データと同じ分類にする。少しでも異なれば正しく分類できない。一方、計算機による意味の処理はコストが大きく、大量の記事に対して二つの記事の意味が同じあるかの判定は非常に難しい。分類しなければならない記事と似ている記事が学習データの中にあれば、その学習データと同じ分類にする。意味が扱い得ないから、似ているという判断はさらに難しい。通常、出現する単語が似ていると解釈し、日本語を意味のある単語に「分かち書き」によって区切り、それらの品詞を得るために、形態素解析のソフトウエア ChaSen [3](または mecab [4]) でかなり高精度に単語に区切り、単語に対応する品詞を表示できる。本論文でも、ChaSen を使用した。

ニュースの記事が類似していると、形態素解析から得られる単語と品詞、それらの頻度が似ているならば、似ている記事と判断するベクトル空間モデルを用いる。記事の中に出現する単語で、「の」、「が」などの助詞を含め、記事の特定に有用でない単語を形態素解析の結果から削除する。さらに、出現頻度の高い単語は、記事の特徴として利用でないために除外する。同様に、出現頻度の低い単語は特定性が高くなり、除外する。

ニュースの記事は、あらかじめ単語を属性によって分類してされているとする。その際、時間や人名は、個々の記事の依存性が高く、カテゴリへの依存性が低いためにニュース記事のカテゴリ分類には適さない [6]。[6] では、名詞を組織 (名詞-固有名詞-組織)、人名 (名詞-固有名詞)、地域 (名詞-固有名詞-地域)、時間 (日付時刻、名詞-副詞可能)、一般用語 (名詞-一般) の 5 つのグループに分けて、ニュース記事のカテゴリ分類をしている。この論文では一般用語を選び、記事の分類に用いる。

Naïve Bayes において、分類する記事に出現する単語が、学習記事の中にない場合は削除して、学習記事の中に出現する単語のみにする。また、ゼロ頻度問題を確率場と考えて、各カテゴリに出現しない単語は最小の確率を割り当て、カテゴリ間の単語数の調整も最小確率で行う。Naïve Bayes 法の処理はデータをテーブルで表現して、テーブルの演算で表現できるためにデータベースの利用ができる。

BIC に関して、クエリ $Q = q_1, .., q_n$ に応じて頻度語の境界を変更するとクエリに応じて学習データを作り直す必要がある。ここでは、学習データのカテゴリ割合がテストデータでも変わらないと仮定して、カテゴリ割合に比例配分して z_{low} と z_{high} を決め、評価実験を行う。

6.2 実験データ

news.ceek.jp サイト [14] より、2008 年 11 月 3 日から 2008 年 11 月 12 日まで、スポーツニュース記事の RSS を約 9,000 件取得した。各記事は、タイトル (title) と本文 (RSS の description) などからなり、タイトルと本文の両方が記述されている記事 2,788 件を選び出した。このうち、(a)3-10 日の 2,188 件を学習データ、(b)11 日の 290 件を z_{low} と z_{high} の推定に、(c)12 日

表 1　カテゴリ別名詞-一般の頻度

Table 1　Number of Noun-general word frequencies in sport categories.

カテゴリ	(a)	(b)	(c)	合計
ボウリング	1	0	0	1
競泳	1	0	0	1
競艇	2	0	0	2
シンクロ	3	0	0	3
レスリング	2	0	1	3
競輪	1	2	1	4
プロレス	9	0	0	9
スケート	13	0	0	13
アメフト	12	3	0	14
バレー	15	0	0	15
F1	16	1	0	17
ラグビー	1	18	0	19
カーリング	20	0	0	20
バドミントン	1	18	4	23
格闘技	16	6	5	27
競馬	27	0	1	28
アイスホッケー	45	0	3	48
陸上	41	3	6	50
バスケ	47	8	7	62
ボクシング	50	8	6	64
テニス	41	15	11	67
フィギュア	71	1	1	73
ゴルフ	117	12	5	134
相撲	127	35	32	194
その他	233	29	35	297
サッカー	509	49	97	655
野球	749	101	95	945
合計	2,188	290	310	2,788

の 310 件を評価用に用いた (表 1)。このように時間で区切った理由は、AIC や BIC は将来の情報推定に用いるために、時間の新しい情報を推定するモデル選択を z_{low} と z_{high} に反映させるためである。

実験にはタイトルを使用せずに本文のみを用いた。[14] サイトの RSS は記事本文の先頭 283 バイトのみを配信している理由から、これをそのまま用いた。ChaSen によって本文から「名詞-一般」のみを選択した。ChaSen の辞書は IPADIC のみを用い、未知語は [6] の観点からも処理することなく除外した。

スポーツ記事のカテゴリは手作業で、非専門家が行った。カテゴリは排他的に、野球、サッカーなどを含む 27 のカテゴリに分けた。

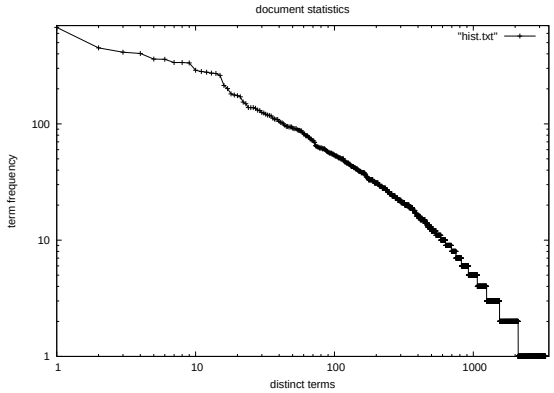

図 2　すべてのスポーツ記事の出現頻度と異なり語

Fig. 2　Distinct terms v.s. term frequencies of sport news documents.

すべての記事から 256,818 語 (ChaSen の結果をそのまま利用しているため、英数字は 1 文字が 1 語) を取り出し、名詞-一般のみの 4,789 語 (異なり数 2,913) を取り出した (図 2)。異なりとその頻度の関係を調べると図 2 となり、Zipf の法則に従うと仮定した。未知語の 1,042 語 (異なり数) も除外した。

表 2　分 類 精 度

Table 2　The precisions of ceek.jp news classifications.

	頻度の下限と上限	最良推定	BIC
	z_{low}	1	6
	z_{high}	279	636
ケース	組み合わせ	精度	精度
(1)	(a)L+(b)T	0.793	0.709
(2)	(a)L+(c)T	0.713	0.741
(3)	(a)(b)L+(c)T	0.716	0.7411
(4)	(a)(b')L+(c)T	0.687	0.730
(5)	(a)(b")L+(c)T	0.719	0.725

L:学習, T:評価, (b'):(b) の正解記事のみ,
(b"):(b) に推定カテゴリの割り当て

表 3　Zipf の法則による中頻度からの左右のバランス

Table 3　Word frequency balance of Lower and higher regions based on Zipf's law.

低頻度区間	単語数	高頻度区間	単語数
1-46	17,095	47-276	9,769
6-46	12,422	47-636	13,612

6.3　実 験 結 果

分類の精度は、microaveraging などを用いることなく、単純に評価記事数の中の正解記事数の割合を示した (表 2)。(a) のデータを学習させて (b) のカテゴリを推定する最も正解率の高い z_{low} と z_{high} を求めた結果、表 2 に示した $z_{low} = 1$, $z_{high} = 279$ となった (最良推定)。同じデータに対して、BIC の値が最小になる z_{low} と z_{high} を求めると表 2 のようにそれぞれ、z_{low}=6 と z_{high}=636 となった。この学習のときの (b) の正解率が最良推定では 0.793、BIC では 0.709 となった (表 2 ケース (1))。この学習結果を (c) のデータで評価すると、最良推定では 0.713、BIC では 0.741 となる (表 2 ケース (2))。

また、(a) と (b) の両方を学習データとした場合の (c) の正解率が最良推定では 0.716、BIC では 0.7411 となった (表 2 ケース (3))。(b) のデータを、(a) から推定した正解データ (b') のみを用いると最良推定では 0.687、BIC では 0.730 となった (表 2 ケース (4)。この (b) を (a) からの推定データをみを用いたデータ (b") では、最良推定が 0.719、BIC では 0.725 となった (表 2 ケース (5))。これは Nigam [15] のラベルなしの 1 ステップのみの EM の場合に対応する。この結果から、将来の記事に対する有効性が示された。

Zipf の法則による高頻度語と低頻度語のそれぞれの上限と下限を計算すると次のようになる。学習データの頻度 1 の単語数は 1,114、4. より $f_k = 46.71$ (約 47) を求める。表 3 に Zipf の法則から求めた中頻度語の低頻度側と高頻度側の単語数を数えると、最良推定よりも、BIC による領域の方が、中頻度語からの左右のバランスが取れている。

6.4　考　　察

表 2 の Naïve Bayes 法を用いた学習のケース (1) では (b) の

推定精度が最大になるように z_{low} と z_{high} を決めているために、精度が高くなっている。しかし、この学習結果を (c) に適用すると、精度が低い。これは Naïve Bayes 法の過学習の結果と思われる。[2] では低頻度語の注目して分類精度の改善を行っており、表 2 の最良推定の結果では低頻度語の利用によって分類精度が高くなっているが、将来の記事に対する分類精度が低くなっている。この学習結果から (b) をカテゴリを推定して正解の記事のみを取り出し、(a)(b) を合わせて再学習しても、精度は高くならない。同様に、(b) の記事の推定カテゴリのみを用いて学習して (c) を評価すると、精度が少し改善される。

一方、BIC の場合はケース (1) の場合を除いて、最良推定よりも精度が高くなっている。

低頻度領域と高頻度領域で分類の用いた単語数を比較した ceek.jp のスポーツニュースを用いた評価実験表 3 より、最良推定よりも、BIC の場合が左右のバランスが良い。

Naïve Bayes 法の多くの研究がある [17]。本研究の Naïve Bayes 法の使用法は、分類の既知の記事から分類が未知のニュース記事を分類することにあるが、分類が未知の記事は時間的に新しい。そして、このときに、Naïve Bayes 法によって BIC の基準が高くなる高頻度語と低頻度語の扱いを選ぶことで、将来の記事の分類を高くすることができることが実験的に示された。その意味で、本研究のアプローチの新しい点は、過去と現在の両方の記事のラベルが既知のときに、過去の記事から現在の記事のラベル推定精度を BIC の基準で測定することで、将来の記事の分類精度が高くなる。この理由は、BIC という基準が将来の記事の分類を考えたメカニズムが埋め込まれていると考えられる。

別の観点から言えば、Naïve Bayes 法の過学習を BIC で抑制していると考えられる。しかし、BIC の基準の中にはデータ数に応じて BIC の基準が変わるという第 2 項の効果から、高頻度語と低頻度語の選択基準が BIC に反映され点が単なる過学習の抑制と異なる。

本手法に於いて、学習過程で最小の BIC の計算に多くの計算が必要となる。特に、パラメータの数が 2 次元であるために、効率の良い最適化法が必要である [13]。しかし、Zipf の法則から中頻度語を中心として頻度度語と低頻度語がバランスしていると仮定するならば、計算量は低く抑えられる。今後、多くのデータについて、検証を行いたい。

7. あとがき

Naïve Bayes による時間的に新しいニュース記事の分類学習に低頻度語と高頻度語の両方を考慮に入れた BIC という統一基準による分類モデルを提案した。また、実験によってその有効性を確認した。Naïve Bayes 法は、SQL による表の検索で分類がほぼ記述でき、処理性能が高い。各ニュース記事が属するカテゴリが、排他的な場合と重複がある場合が考えられるが、本報告では排他的な場合のみを扱っており、重複の場合は今後検討をしたい。

低頻度語、高頻度語の領域では、それぞれ語の特定性、網羅性が高くなり、分類性能が低下する対策として TF-IDF などによる語の重みによって、分類性能を高める事ができる。しかし、重み付けと分類手法は依存関係がなく、応用領域に応じて性能の良い組み合わせを選択する必要がある。本研究では頻度語の許容する上限と下限を設けるアプローチで Näive Bayes 分類を行う場合は、BIC の基準の意味で最適な頻度語の上限と下限を選ぶことができ、Näive Bayes 分類の過学習の回避ができ、加えて EM または EM 的な手法によって、学習性能をたけめることができる。今後、語の重み付けを導入して、Bayes 的なアプローチに本手法を拡張する。本研究ではラプラス法によるスムージングを行っており、性能の高い手法を今後、検討する [19]。

文　献

[1] P. Domingos and M. Pazzani, On the optimality of the simple bayesian classifier under zero-one loss, Machine Learning, 29, 103-130, 1997.

[2] 相沢、低頻度語の利用によるテキスト分類性能の改善と評価、情処論誌、44, 7, 1720-1730, 2003.

[3] 形態素解析システム茶筌、http://chasen.naist.jp/hiki/ChaSen/, 2008.

[4] mecab、http://mecab.sourceforge.net/, 2008.

[5] 北研二、確率的言語モデル、東大出版、1999.

[6] Juha Makkonen, Investigations on Event Evolution in TDT, Student Research Workshop, 43-48 Proceedings of HLT-NAACL, 2003.

[7] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automatic Control, AC-19, 716-723, 1974.

[8] H. Akaike, A Bayesian Analysis of the Minimum AIC Procedure, Ann. Inst. Statist. Math., 30, Part A, 9-14, 1978.

[9] G. Schwarz, Estimating the dimension of a model, Ann. Statist., 6, 461-464, 1976.

[10] J. Rissanen, Stochastic Complexity, J. Roy. Statist. Soc., Ser. B 49, 223-139, 1987.

[11] 徳永、情報検索と言語処理、東大出版、1999.

[12] I. Shioya and H. Oh'uchi and T. Miura, Document Retrieval using Projection by Frequency Distribution, International Journal on Artificial Intelligence Tools (IJAITS), 16, 647-659, 2007.

[13] I. Shioya and H. Inaba, Maximum Likelihood A-R Parameter Estimation with Missing Data, 13-th JAACE Symposium on Stochastic Systems, Kyoto, Oct. 28-30, 1981.

[14] http://news.ceek.jp/, 2008.

[15] K. Nigam, A. MaCallum, S. Thrun and T. Mitchell, Text Classificartion from labeled and unlabeled document using EM, Machine Learning, 3, 103-134, 2000.

[16] 大内、三浦、塩谷、ランダムプロジェクションによるテキストストリームの検索、DEWS 2004, 3-C-02, 2004.

[17] F. Sebastiani, Machine Learning in Automated Text Categorization, ACM Com. Survey, 34, 1-47, 2002.

[18] Sucar2003 Héctor Hugo Avilés-Arriaga, Luis Enrique Sucar and Carlos Eduardo Mendoza, Visual Recognition of Gestures using Dynamic Naive Bayesian Classifier, IEEE RO-MAN, Preliminary version, 2003.

[19] Chengxiang Zhai and Jhon Lafferty, A Study of Smoothing Methods for Language Models Applied to Information Retrieval, ACM Trans. Information Systems, 22, 2, 179-214, 2004.