ニューラルネットを用いたテキストの特徴語抽出

斉藤和巳

Extracting Characteristic Words of Text Using Neural Networks

Kazumi SAITO

ABSTRACT

In this paper, we discuss methods for detecting an adequate topic of documents and extracting characteristic words of such topics, by using two types of neural networks formalized as statistical models. The main features of these models are that their learning algorithms utilize an objective function that maximizes posterior probabilities for topic detection, and that characteristic words are extracted based on the magnitude of resulting parameter values. Through the experiments using a set of real Web pages, we evaluate the methods in the aspect of topic detection performances and extraction capabilities of characteristic words.

アブストラクト

本稿では, 統計モデルとして定式化した 2 種類のニューラルネットを用いて, 文書が属すると思われるトピックの推定とトピックを特徴づける単語群の抽出法を論じる. これら推定と抽出法の特徴は, 事後確率の最大化によるトピック抽出に着目した目的関数を採用して学習し, その結果として得られたパラメータの大きさに基づいて特徴語を抽出することである. Web 上のテキストへの適用事例を通して, トピック抽出性能と特徴語抽出能力を評価する.

キーワード

ニューラルネット, ナイーブベイズ, テキストマイニング.

1 はじめに

文書とそのトピックのペアからなるサンプル集合が与えられたとき, 各トピックに特徴的に現れる単語を求めることは基本的な課題の一つである. 単純には, トピック毎に文書群を分類して, 各グループでの単語頻度を求めて降順にソートし, 上位に現れる単語をそのトピックの特徴語とすることができる. これに対して, 複数のトピックに高い頻度で出現する単語は望ましくないとして, 各トピック固有に出現する識別的な単語を抽出したいケースも考えられる.

このような特徴語抽出を学習法の観点から見れば, 単純な単語頻度に基づく方法は, 例えば, トピック間の識別を陽に意識しないモデルを土台とするナイーブベイズ (naive Bayes) 法 (Duda, Hart, & Stork, 2000) が対応すると考えられる. 一方, 識別的な単語を抽出可能とするには分類学習法が必要となる. このようなテキスト分類には, SVM (Support Vector Machine) 法 (e.g., Joachims, 1998) が比較

的頻繁に適用されているものの, その学習結果を明示的に理解するのは一般に困難と思われる. これに対して, ニューラルネットを用いれば, 既に多様なルール抽出法 (e.g., Saito & Nakano, 2002) が提案されいるので, 特徴語を求めるのに利用できると考えられる. しかしながら, 対象とする語彙数は一般に数万以上となるので, 標準的な多層パーセプトロンをランダムな初期値から学習するのでは, 学習サンプルにオーバーフィットしてしまい, 適切な特徴語を求めることは困難になると思われる.

本論文では, ニューラルネットを用いたテキストの特徴語抽出に向けた試みとして, 比較的単純な 2 種のニューラルネットについて考察し, それらを用いた特徴語抽出の初期実験結果について報告する. これらニューラルネットと抽出法の特徴は, 事後確率の最大化によるトピック抽出に着目した目的関数を採用して学習し, その結果として得られたパラメータ値の大きさに基づいて特徴語を抽出することである. 現実の www ページを用いた実験を通して, 提案手法のトピック抽出性能と特徴語抽出能力を評価する.

2　学習アルゴリズム

テキスト表現法と既存のナイーブベイズ法を説明した後, 統計モデルとして定式化した 2 種類のニューラルネットとその学習アルゴリズムについて述べる.

2.1　フレームワーク

まず, テキスト表現法について説明する. $\{(\mathbf{x}_n, \mathbf{y}_n) : n = 1, \cdots, N\}$ を文書とトピックのペアからなるサンプル集合とする. ただし, $\mathbf{x}_n = (x_{n,1}, \cdots, x_{n,V})^T$ であり, V は全テキストでの異なる単語数 (語彙規模) を, $x_{n,v}$ は文書 $\mathbf{x}_n$ にて v 番目の単語が現れた頻度を表す. ここで $\mathbf{x}^T$ はベクトル $\mathbf{x}$ の転置を表す. すなわち, 文書中に現れる単語順序を無視し, その出現頻度のみに着目してテキストを記述する BOW (bag-of-words) 表現 (e.g., Manning & Schütze, 1999) を採用する. 一方, $\mathbf{y}_n = (y_{n,1}, \cdots, y_{n,K})$ であり, K は異なるトピック数を, $y_{n,k}$ は 1 か 0 で文書 $\mathbf{x}_n$ が k 番目のトピックを持つかどうかを表す. 以下では, $\mathbf{y}_n$ はどれか一つだけのトピックを持つとする. すなわち, $y_{n,1} + \cdots + y_{n,K} = 1$ である.

本論文では, 各トピックに特徴的に現れる単語群を求めることを課題とする. 特徴語の定義としては, 目的や用途に応じて多様に設定できると考えられるが, ここでは比較的単純に, トピックに頻繁に出現する単語, あるいは, 各トピック固有に高い頻度で現れる単語を想定する. 以下では, このような特徴語の抽出に有効と考えられる学習モデルとアルゴリズムについて述べる.

2.2　ナイーブベイズモデル

本稿で対象とするような BOW に基づくテキスト分類などで, ナイーブベイズ (naive Bayes) 法 (Duda, Hart, & Stork, 2000) は, シンプルでロバストな手法の一つとして広く採用されている. ナイーブベイズでは, サンプル $\mathbf{x}$ の生成モデルとして, トピック k を持つ文書は, 語彙に関する多項 (multinomial) 分布により生成されると仮定する. すなわち, 第 v 単語の生起確率を $\theta_{k,v}$ とすれば,

$$P(\mathbf{x} \mid k) \propto \prod_{v=1}^{V} \theta_{k,v}^{x_k}, \quad \theta_{k,v} > 0, \quad \sum_{v=1}^{V} \theta_{k,v} = 1, \tag{1}$$

に基づいてトピック k の文書 $\mathbf{x}$ が生成されるとする. したがって, サンプル集合 $\{(\mathbf{x}_n, \mathbf{y}_n) : n = 1, \cdots, N\}$ に対して, 多項分布パラメータ $\theta_{k,v}$ を求めるための対数尤度項は以下となる[*1]

$$L_1(\mathbf{\Theta}) = \log \prod_{n=1}^{N} \prod_{k=1}^{K} P(\mathbf{x}_n \mid k)^{y_{n,k}} = \sum_{n=1}^{N} \sum_{k=1}^{K} y_{n,k} \log P(\mathbf{x}_n \mid k)$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} y_{n,k} \sum_{v=1}^{V} x_{n,v} \log(\theta_{k,v}). \tag{2}$$

ここで, $\theta_k = (\theta_{k,1}, \cdots, \theta_{k,V})^T$ とし, $\mathbf{\Theta}$ は θ_1 から θ_K の各要素を並べて構成したベクトルを表す.

しかしながら, $L_1(\mathbf{\Theta})$ を直接最大化する結果は, 一般に学習サンプル集合にオーバーフィットしてしまう傾向があり, 汎化性能 (未学習データに対する性能) の観点では望ましくない. この問題に対しては, パラメータ θ_k はディリクレ (Dirichlet) 分布に従う事前分布を持つとし, 次式の学習目的関数を最大化する MAP (maximum a posteriori) 推定が広く採用されている.

$$J_1(\mathbf{\Theta}) = L_1(\mathbf{\Theta}) + \lambda_1 \sum_{k=1}^{K} \sum_{v=1}^{V} \log(\theta_{k,v}), \tag{3}$$

ただし, λ_1 (> 0) はハイパーパラメータを表す. 式 (3) で定義した学習目的関数には, 式 (1) で規定されるパラメータに対する制約があることを考慮し, ラグランジェ (Lagrange) 未定乗数法に基づき最適解を求めれば以下を得る.

$$\widehat{\theta}_{k,v} = \frac{\sum_{n=1}^{N} y_{n,k} x_{n,v} + \lambda_1}{\sum_{n=1}^{N} \sum_{v=1}^{V} y_{n,k} x_{n,v} + V\lambda_1}. \tag{4}$$

この解が式 (1) の制約を満たすことは容易に検証できる.

単語頻度ベクトルのみが与えられた文書 $\mathbf{x}$ に対して, トピックを推定するには事後確率 (posterior probability) が標準的に用いられている. すなわち, $\mathbf{x}$ に対して次式が最大となるトピック k を推定結果とする.

$$P(k \mid \mathbf{x}) = \frac{P(k)P(\mathbf{x} \mid k)}{\sum_{j=1}^{K} P(j)P(\mathbf{x}|j)} = \frac{P(k) \exp(\sum_{v=1}^{V} x_v \log(\widehat{\theta}_{k,v}))}{\sum_{j=1}^{K} P(j) \exp(\sum_{v=1}^{V} x_v \log(\widehat{\theta}_{j,v}))}. \tag{5}$$

ただし, $P(k)$ はトピック k の文書の生起確率を表し, $P(k) = (y_{1,k} + \cdots + y_{N,k})/N$ として推定できる. 式 (5) の右辺の分母は, 全トピックの和となっているので, 結局以下のようにして文書 $\mathbf{x}$ のトピックを求めても同値である.

$$k = \arg\max_j \{\log(P(j)) + \sum_{v=1}^{V} x_v \log(\widehat{\theta}_{j,v})\}. \tag{6}$$

すなわち, 入力空間 $\mathbf{x}$ において, 識別境界は区分線形となる.

$y_{n,k} \in \{0, 1\}$ より, 第 k トピックに属す文書群全体での第 v 単語の出現回数は $\sum_{n=1}^{N} y_{n,k} x_{n,v}$ に他ならない. よって, λ_1 が定数より, 式 (4) を考慮すれば, 各トピック毎に第 v 単語の出現頻度の総数とパラメータ $\widehat{\theta}_{k,v}$ の大小関係は完全に一致する. すなわち, ナイーブベイズ法で求めたパラメータの

[*1] 厳密には多項係数に起因する定数の加算が必要であるが, パラメータ $\mathbf{\Theta}$ とは独立なので簡略化のため無視している.

大小に基づく特徴語抽出は単純頻度法になることが分かる. 明らかに, $\{\widehat{\theta}_{k,v}\}$ の値が大きければ, 第 v 単語はトピック k の頻度ベース特徴語と見なすことができる. しかしながら, このようにして求めた結果では, 複数のトピックに渡って頻度の高い単語が抽出される可能性があるので, トピック固有の特徴語とは言い難い.

2.3 ニューラルネットモデル

ナイーブベイズでは, 生成モデルに対する尤度項である式 (2) を最大化することによりパラメータを求めている. これに対して, 式 (5) に基づいてトピック抽出を行なうことに着目すれば, 学習用の単語頻度ベクトルに対する事後確率とトピックに関する尤度項を用いて, 次式の最大化によりパラメータを求めるアプローチを考えることができる.

$$L_0(\Theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k} \log P(k \mid \mathbf{x}_n). \tag{7}$$

すなわち, 負のクロスエントロピー (cross-entoropy) の最大化であり, 多クラス分類問題でのニューラルネットの学習目的関数として頻繁に採用されている (e.g., Bishop, 1995).

いま, 以下のパラメータ変換を考える.

$$w_{j,v} = \log(\theta_{j,v}). \tag{8}$$

このとき, 式 (7) で定義した尤度項は以下のようにパラメタライズされる.

$$L_2(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k} \log\{\frac{p(k)\exp(\sum_{v=1}^{V} w_{k,v}x_{n,v})}{\sum_{j=1}^{K} p(j)\exp(\sum_{v=1}^{V} w_{j,v}x_{n,v})}\}. \tag{9}$$

よって, 式 (1) で規定されるパラメータに対する制約を無視すれば, 式 (9) は, 入力ベクトル $\mathbf{x}$ の線形重み付け和に soft-max 関数が施された, クロスエントロピーに基づく学習目的関数に他ならない. つまり, 式 (9) をニューラルネット学習問題と見ることができる. なお, 導出原理は異なるが, 式 (9) は最大エントロピー (maximum entropy) アプローチ (Nigam, Lafferty, & McCallum, 1999) の目的関数とも等価になる.

このケースでも $L_2(\mathbf{w})$ を直接最大化した結果は, 一般に学習サンプル集合にオーバーフィットしてしまう傾向があり, 汎化性能の観点では望ましくない. よって, 自乗値ペナルティ (weight-decay) 項を用いた次式の学習目的関数を考える.

$$J_2(\mathbf{w}) = L_2(\mathbf{w}) - \lambda_2 \sum_{k=1}^{K}\sum_{v=1}^{V} w_{k,v}^2. \tag{10}$$

ただし, λ_2 (>0) はハイパーパラメータを表す. なお, $\lambda_2 > 0$ ならば, $J_2(\mathbf{w})$ の Hesse 行列は負定値 (negative definite) であり, 常に大域的最適解を学習で得ることが保証される. すなわち, $\mathbf{w}$ と同じ次元の任意のベクトル $\widetilde{\mathbf{w}}$ に対して, 以下のように, Hesse 行列の 2 次形式は負の値となる.

$$\begin{aligned}\widetilde{\mathbf{w}}^T \frac{\partial^2 J_2(\mathbf{w})}{\partial \mathbf{w}\partial \mathbf{w}^T}\widetilde{\mathbf{w}} &= \left.\frac{d^2 J_2(\mathbf{w}+\varepsilon\widetilde{\mathbf{w}})}{d\varepsilon^2}\right|_{\varepsilon=0} \\ &= -\sum_{n=1}^{N}\sum_{k=1}^{K} p(k|\mathbf{x}_n)\{\sum_{v=1}^{V}\widetilde{w}_{k,v}x_{n,v} - \sum_{j=1}^{K} p(j|\mathbf{x}_n)\sum_{v=1}^{V}\widetilde{w}_{j,v}x_{n,v}\}^2 - \lambda_2\sum_{k=1}^{K}\sum_{v=1}^{V}\widetilde{w}_{k,v}^2 < 0. \end{aligned} \tag{11}$$

頻度ベクトルのみが与えられた文章 $\mathbf{x}$ に対するニューラルネットのトピック推定式は以下となる.

$$k = \arg\max_j \{\log(P(j)) + \sum_{v=1}^{V} \widehat{w}_{j,v} x_v\}. \tag{12}$$

一方, $\{\widehat{w}_{k,v}\}$ の値が大きければ, 式 (9) より, 第 v 単語はトピック k の識別的な特徴語と見なすことができる. すなわち, 複数のトピックに渡って頻度の高い単語に対する重みは, 事後確率最大化を行なうので, 大きな値になることが抑制される.

2.4 制約付きニューラルネットモデル

ナイーブベイズでは, 式 (1) に示したように, パラメータ に関して $\sum_v \theta_{j,v} = 1$ の制約がある. しかしながら, 式 (10) を最適化して得られる $\widehat{w}_{j,v}$ を式 (8) で逆変換しても, もはや一般にこの制約は成り立たない. そこで, 式 (8) に替えて, 以下のパラメータ変換を考える.

$$\theta_{j,v} = \frac{\exp(u_{j,v})}{\sum_{v=1}^{V} \exp(u_{j,v})}. \tag{13}$$

明らかに, 任意の $\mathbf{u}$ に対して, 式 (1) の $\theta_{j,v}$ に関する制約は常に満たされる.

いま, 全学習サンプルの頻度ベクトルを $\sum_{v=1}^{V} x_{n,v} = 1$ で正規化するとすれば, 式 (7) で定義した尤度項は以下のようにパラメタライズされる.

$$L_3(\mathbf{u}) = \sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k} \log \frac{q_{n,k}}{\sum_{j=1}^{K} q_{n,j}}, \quad q_{n,k} = P(k)\exp(\sum_{v=1}^{V} u_{k,v}x_{n,v} - \log\sum_{v=1}^{V}\exp(u_{k,v})). \tag{14}$$

よって, 式 (10) と同様にして, 学習目的関数 $J_3(\mathbf{u})$ を以下のように定義できる.

$$J_3(\mathbf{u}) = L_3(\mathbf{u}) - \lambda_3 \sum_{k=1}^{K}\sum_{v=1}^{V} u_{k,v}^2 \tag{15}$$

ここでも, 自乗値ペナルティ項を採用し, λ_3 (> 0) をそのハイパーパラメータとする. ただし, $\lambda_3 > 0$ でも, $J_3(\mathbf{u})$ の Hesse 行列が負定値となる保証はない.

頻度ベクトルのみが与えられた文章 $\mathbf{x}$ に対する制約付きニューラルネットのトピック推定式は以下となる.

$$k = \arg\max_j \{\log(P(j)) + \sum_{v=1}^{V} \widehat{u}_{j,v} x_v - \log(\sum_{v=1}^{V} \exp(\widehat{u}_{j,v}))\}. \tag{16}$$

一方, $\{\widehat{u}_{k,v}\}$ の値が大きければ, 式 (14) より, 第 v 単語はトピック k の識別的な特徴語と見なすことができる. ただし, $\{\widehat{w}_{k,v}\}$ とは異なる性質を持つと予想される.

2.5 ニューラルネット学習アルゴリズム

ナイーブベイズでは, 最適解を解析的に求めることができるが, 上述した 2 種類のニューラルネットの解を求めるには反復学習法が必要となる. このような非線形最適化には多様な手法の適用が可能である. ただし, 一般に語彙規模 V は数万であり, 数十のトピック数を想定すれば, 数十万ものパラメータ群を学習することになるため, 標準的なニュートン法の適用は困難である. 一方, 単純な最急降

下法では収束までの効率が問題となる. そこで, 代表的なニューラルネット学習法と比較して優れた収束性を示した BPQ 法 (Saito & Nakano, 1997) を採用する.

BPQ は, 準ニュートン (quasi-Newton) 法を土台とした非線形最適化手法である. 詳細には, 探索方向を少記憶 BFGS 法で計算し, 探索幅を 2 次近似の最小値として求める. よって, BPQ の記憶容量を定めるパラメータを適切に設定することにより, ニューラルネットの重み数が非常に多い大規模問題への適用も可能となる. なお, 上述したニューラルネット目的関数のように, 自乗値ペナルティ項を付加すれば, さらに収束性能が向上することも報告されている (Saito & Nakano, 2000).

3　評価実験

定式化したモデルを評価するための実験に関して, そのフレームワークと結果について述べる.

3.1　実験設定

あるディレクトリー型ポータル検索サイトからリンクが張られる実際の www ページを用いた評価実験を行った. このサイトは “Arts & Humanities”, “Business & Economy” などの 14 のトップカテゴリを有し, さらに各カテゴリ毎に数十の第 2 レベルカテゴリを持つ. 本実験では, “Arts & Humanities” に着目し, その第 2 レベルカテゴリをトピックとして用いた.

www ページのデータ収集には Gnu Wget と呼ばれるインターネットロボットを用い, この検索サイトから直接リンクされるサイト外の www ページの実際のテキストと, それが “Arts & Humanities” の第 2 レベルのどのトピックを持つかのラベル情報を得た. ただし, www ページは一般に複数のトピックを持つものの, 前節で述べたモデルでの特徴語抽出に関する基本性能を評価するため, 単一のトピックを持つ www ページのみを抽出して実験を行なった. ページの総数は 9,644 であった. 本実験のため収集した www ページの単語には, 動詞の活用などを基本形に統一する語末処理を施し, 冠詞などに代表される 571 個の不要語 (stop words) を削除し, そして頻度 10 未満の単語を無視した. その結果, 単語の総数は 18,818 となった. また, トピックの総数は 26 であった.

3.2　トピック抽出性能評価

ナイーブベイズ, ニューラルネット, および制約付きニューラルネットでのピック抽出性能を評価した. ここでは, 6,000 サンプルをランダムに選択して学習データを作成し, 残りの 3,644 サンプルをテストに用いた. ただし, このような一度のサンプリング (hold-out) に基づく性能評価では, 信頼性が高いとは言えない. そこで, さらに独立なサンプリングを行ない, 学習とテスト用のサンプル集合のペアを 5 組用意して評価を行なった.

図 1 では, ナイーブベイズとニューラルネットでのテストデータに対する性能 (正当率 %) を比較する. ただし, ハイパーパラメータは 10^{-6} から 10 倍ずつして 10 までの範囲で変化させた. 図より, ナイーブベイズでは, $\lambda_1 = 1$ のラプラススムージングに対応するハイパーパラメータで最良の結果を得られていることが分かる. 一方, ニューラルネットでは, ハイパーパラメータが $\lambda_2 = 10^{-3}$ で最良の結果を得られたことが分かる. ナイーブベイズの最良結果と比較すれば, ニューラルネットの性能は 6% 強程度向上している. これは事後確率に基づく学習目的関数を採用することで, 汎化性能の改善ができたと考えられる. また, それぞれの曲線を比較すれば, ピークの位置と高さは異なるものの,

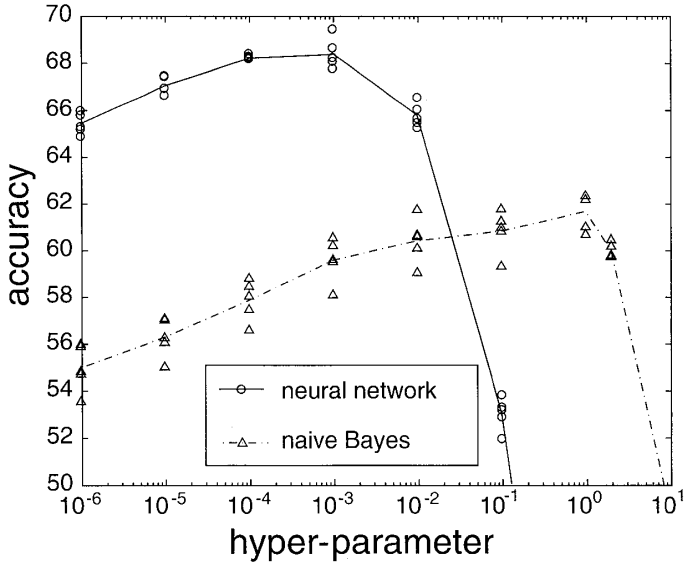

図 1　ナイーブベイズとニューラルネットの性能比較.

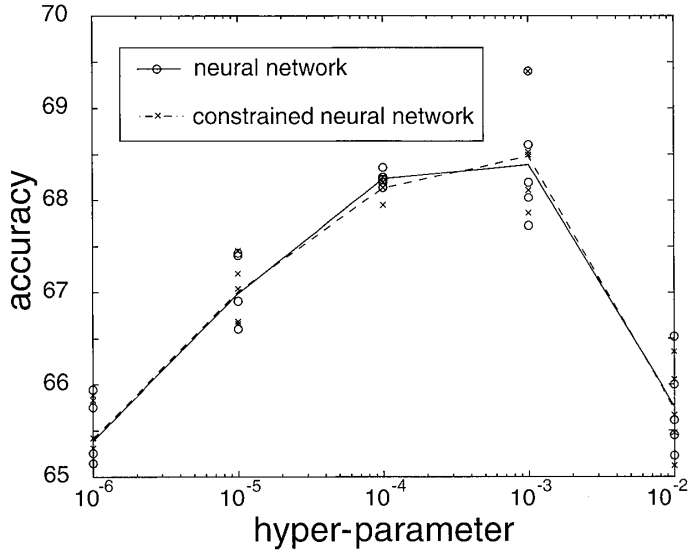

図 2　制約なしと付きのニューラルネットの性能比較.

比較的類似した挙動を示していることが分かる.

図 2 では, ニューラルネットと制約付きニューラルネットでのテストデータに対する性能を比較する. ただし, ハイパーパラメータは 10^{-6} から 10^{-2} の範囲で比較している. 図より, これら 2 つの性能は殆んど同等であることが分かる.

3.3　収束性能評価

既に述べたように, ニューラルネットの学習には BPQ を採用している. 標準的な BP (back propagation) と比較して, BPQ の収束性能を評価した. BP では, 一般に収束性能が向上すると考えら

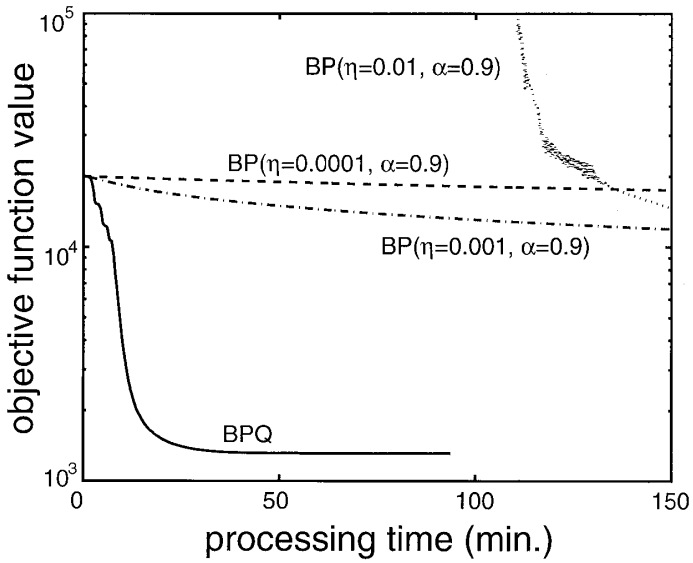

図 3　ニューラルネットの収束性能比較.

れる慣性 (momentum) 項を係数 $\alpha = 0.9$ として付加し, 学習定数 η を試行錯誤で $0.01, 0.001, 0.0001$ の 3 種に設定した. ニューラルネット重みの初期値は, 一つの妥当な候補として, ナイーブベイズで求まる頻度パラメータを用いた. 反復終了条件は, 勾配ベクトルの全要素の絶対値が 10^{-4} 以下になるか, 反復回数が 1000 を越えるときとした. この実験では収集した全 www ページを学習に用いた.

図 3 に, BPQ と学習定数 η を変えた 3 種の BP の収束性能を比較する. ただし, 実験の例題としては, 式 (10) で定義されるニューラルネットを用い, ハイパーパラメータ値は最良と考えられる $\lambda_2 = 10^{-3}$ を採用した. また, ここでは目的関数値に -1 を乗じて最小化問題として扱っている. $\eta = 0.01$ のときは, 学習の初期段階で目的関数値が急激に増えだし, 後半で目的関数値が減りだしても挙動が不安定であることが分かる. すなわち, 学習定数が大き過ぎることが示唆される. 一方, $\eta = 0.0001$ のときは, 目的関数値の減少が極めて遅く, 逆に学習定数が小さ過ぎることが示唆される. しかるに, 両者の中間の $\eta = 0.001$ のときでも, 目的関数値が効率良く減少していないことが分かる. これら 3 種の BP では, いずれも最大反復回数が 1000 を越えて終了した.

これに対して, BPQ を用いれば, 目的関数値が効率良く減少していることが分かる. なお, 反復回数が 366 のとき, BPQ は勾配ベクトルの全要素の絶対値を 10^{-4} 以下にして終了した. すなわち, 最適解をほぼ正確に求めることができている. また, 初期段階で目的関数値が急激に減少していることより, 勾配ベクトルに対する終了条件を緩和して処理時間を短縮することができる. 本実験では, 語彙規模が約 20,000 でニューラルネット重みの総数は約 500,000 となるが, サンプル数が約 10,000 に対して 100 分程度で学習が完了することより, 十分に実用的な計算時間であると考えられる. なお, 本実験には 1GHz Pentium のパソコンを用いた.

3.4　特徴語抽出能力評価

3 種の手法から比較的容易に抽出できる特徴語を比較した. つまり, 3 手法で得られたそれぞれのパラメータ群 $\{\widehat{\theta}_{k,v}\}$, $\{\widehat{w}_{k,v}\}$, および $\{\widehat{u}_{k,v}\}$ のそれぞれに着目し, 各パラメータ (単語とトピックのペアに対応) の値が大きい上位 20 ペアを抽出した.

表1　ナイーブベイズによる特徴語.

	word	topic
1	book	Design Arts
2	design	Arts and Crafts
3	www	Employment
4	design	Artists
5	art	Web Directories
6	art	Thematic
7	award	Awards
8	plan	Arts and Crafts
9	book	Awards
10	art	Art History
11	architectur	Artists
12	docum	Museums Galleries and Centers
13	architect	Artists
14	paint	Thematic
15	home	Arts and Crafts
16	book	Arts and Crafts
17	design	Education
18	new	Museums Galleries and Centers
19	docum	Criticism and Theory
20	docum	Arts Therapy

表1に, ナイーブベイズによる抽出結果を示す. ただし, 既に述べたように単語には語末処理が施されている. 表から分かるように, “book”, “design”, “art”, “docum” の4単語は, それぞれ異なるトピックで3回現れた. これらの単語は頻繁に使われると言う点では頻度ベースの特徴語と見なせるが, 各トピック固有の特徴語とは言い難い.

表2に, ニューラルネットによる抽出結果を示す. ナイーブベイズによる結果と比較すれば, 双方に共通に現れたのは “award”, “architect” の2単語だけであり, 表1で複数回現れた “book” などの単語は全く現れなかった. 表2において, 慎重な検証が必要とはされものの, 直観的には各トピック固有の特徴語が抽出されていると思われる.

表3に, 制約付きニューラルネットによる抽出結果を示す. まず, ナイーブベイズによる結果と比較すれば, “award”, “architect” に加えて, 2回現れた “book” が共通であった. 一方, ニューラルネットによる結果と比較すれば, まさに半数の単語が共通に現れた. よって, 表3に現れた単語全体の概要としては, ナイーブベイズとニューラルネットによる結果の中間的なものと考えられる. なお, このように異なる性質の特徴語ではあるものの, 学習結果のトピック抽出性能が殆んど同等なのは興味深い点と思われる.

表2　ニューラルネットによる特徴語.

	word	topic
1	architect	Artists
2	council	Cultural Policy
3	health	Arts Therapy
4	night	Events
5	fashion	Chats and Forums
6	attack	Museums Galleries and Centers
7	cancer	Arts Therapy
8	award	Awards
9	museum	Museums Galleries and Centers
10	pharmaceut	Arts Therapy
11	bead	Crafts
12	cemeteri	Reference
13	actor	Performing Arts
14	recruit	Employment
15	perform	Performing Arts
16	jazz	Arts Therapy
17	african	Cultures and Groups
18	antiqu	Arts and Crafts
19	batik	Crafts
20	censorship	Censorship

4　議論と今後の課題

本論文では, トピック学習モデルと特徴語抽出法を対応させて議論することにより, 学習結果のパラメータ値の大きさに基づいて, トピックと単語のペアの抽出のみを試みた. 本研究の次のステップとしては,「各トピックを特徴付ける特徴語はそれぞれ何か?」という基本的な問いにも十分に答えられるようにすることである. これにより, 文書クラスター群に対して適切なアノテーションを与えることができる. すなわち, spherical k-means 法 (Dhillon & Mohda, 2001) や混合多項分布モデル (Nigam et al., 2000) などで求まる文書クラスタリング結果を適切に解釈するのに役立つと考えられる.

本論文では, 文書のトピックは高々一つと仮定したが, 現実の多様な文書を対象とするには多重トピックへの拡張が不可欠となる. このような多重性を扱う確率モデルには, 文書トピックが既知ならば PMM (parametric mixture model) と呼ばれるモデル (Ueda & Saito, 2002) があり, 文書トピックが未知ならば LDA (latent Dirichlet allocation) と呼ばれるモデル (Blei, Ng, & Jordan, 2003) などがある. 多重トピック文書のクラスタリング結果に対する適切なアノテーションも重要な研究課題と考えている.

別途著者らは, 多変量データから実数指数の多項式型法則式を求める研究を進めている (e.g., Saito & Nakano, 1997). そこでは, 入力変数に対数変換を施し, 中間ユニットの活性化関数が指数関数の

表3　制約付きニューラルネットによる特徴語.

	word	topic
1	antiqu	Arts and Crafts
2	architect	Artists
3	wildlif	Thematic
4	recruit	Employment
5	council	Cultural Policy
6	book	Design Arts
7	lesson	Education
8	bead	Crafts
9	book	Arts and Crafts
10	african	Cultures and Groups
11	deco	Organizations
12	death	Institutes
13	histori	News and Media
14	perform	Performing Arts
15	survivor	Web Directories
16	anim	Visual Arts
17	cemeteri	Reference
18	fashion	Chats and Forums
19	award	Awards
20	aesthet	Criticism and Theory

ニューラルネットを用いている. すなわち, 対数変換と指数関数を用いてニューラルネットを構成する点で, 本稿で議論した式 (9) と類似した構造であるため, 両者を統合する枠組への研究発展が期待できる.

本研究のさらなる発展としては, 単語出現頻度に基づく文章からのトピックや特徴語抽出だけでなく, 一般の頻度分布に基づくサンプルに対する特徴名義値の抽出と考えている. 一つの例としては, www サイト間のハイパーリンク数の頻度に着目した www コミュニティの発見や, コミュニティを代表する www ページの抽出である. さらには, 異なる複数の頻度データを統合する枠組への発展も考えられる. 例えば, 文書のトピック抽出と www コミュニティの発見を同時に行なう枠組である.

5　おわりに

本論文では, 統計モデルとして定式化した 2 種類のニューラルネットを用いて, 文書が属すると思われるトピックの推定とトピックを特徴づける単語群の抽出法を論じた. これら推定と抽出法の特徴は, 事後確率の最大化によるトピック抽出に着目した目的関数を採用して学習し, その結果として得られたパラメータの大きさに基づいて特徴語を抽出することである. Web 上のテキストへの適用事例を通して, トピック抽出性能と特徴語抽出能力を評価した.

本実験からは, ニューラルネットを用いたトピック抽出と特徴語抽出法の有望性を示唆する結果が得られたものの, この主張をサポートするには, さらに多くの検証実験が不可欠である. また, 多様な

文書データを用いた提案手法のさらなる評価実験, 文書クラスタに対する適切なアノテーション法の検討, さらに, 多重トピック文書を扱うための枠組の拡張など今後の課題も多く残されている. しかしながら, トピック学習モデルと特徴語抽出法を対応させて議論することにより, 本研究にて一つの興味深い初期実験結果が得られたと考えている.

参考文献

[1] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Clarendon Press.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

[3] Dhillon, I. S. & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*, 143–175.

[4] Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification (2nd Ed.)*. John Wiley & Sons.

[5] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the European Conference on Machine Learning (ECML'98)* (pp. 137–142).

[6] Luenberger, D. G. (1984). *Linear and nonlinear programming*. Addison-Wesley.

[7] Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

[8] Nigam, k., Lafferty, J., & McCallum, A. (1999) Using maximum entropy for text classification. Presented at the *IJCAI-1999: Workshop on Machine Learning for Information Filtering*

[9] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103–134.

[10] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*, 130–137.

[11] Saito, K. & Nakano, R. (1997). Partial BFGS update and efficient step-length calculation for three-layer neural networks. *Neural Computation*, *9*, 123–141.

[12] Saito, K., & Nakano, R. (1997). Law discovery using neural networks. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 1078–1083).

[13] Saito, K. & Nakano, R. (2000). Second-order learning algorithms with squared penalty term. *Neural Computation*, *12*, 709–729.

[14] Saito, K. & Nakano, R. (2002). Extracting regression rules from neural networks. *Neural Networks*, *15*, 1279–1288.

[15] Salton, G. (1988). *Automatic text processing*. Addison-Wesley.

[16] Ueda, N. and Saito, K. (2002). Single-shot detection of multiple topics using parametric mixture models. *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD'02)* (pp. 626–631).