情報処理学会論文誌 Vol. 50 No. 6 1234–1244 (June 2009)

トピックモデルに基づく文書群の可視化

岩田具治[†1] 山田武士[†1] 上田修功[†1]

トピックモデルに基づく，文書データを潜在的なトピック構造とともに可視化するための手法を提案する．提案法では，文書およびトピックが 2 次元ユークリッド可視化空間に座標を持つと仮定し，それらの座標から文書が生成される過程をモデル化することにより文書群を可視化する．EM アルゴリズムを用いて与えられたデータに最も適合するモデルを推定することにより，文書座標およびトピック座標が得られる．実験により，提案法を用いて可視化したとき，従来法よりも関連文書が近くに配置されることを示す．

Visualizing Documents based on Topic Models

Tomoharu Iwata,[†1] Takeshi Yamada[†1]
and Naonori Ueda [†1]

We propose a method based on a topic model for visualizing documents with the latent topic structure. Our method assumes that both documents and topics have latent coordinates in a two-dimensional Euclidean space, or visualization space, and visualizes documents by considering a generative process of documents as a mapping from the visualization space into the space of documents. A visualization, i.e. latent coordinates of documents, can be obtained by fitting the model to given documents using the EM algorithm. In the experiments, we demonstrate that the proposed model can locate related documents closer together than conventional visualization methods.

1. はじめに

近年，文書や購買ログなどの離散データを解析する手法として，bag-of-words 表現された文書の生成過程を確率的にモデル化したトピックモデルが注目されている．トピックモデルでは，ある文書に含まれる各単語は，文書固有のトピック比率に従ってあるトピックを選択した後，そのトピックに固有の単語出現確率分布に従って生成される，と仮定する．代表例として，Probabilistic Latent Semantic Analysis（PLSA）[10] や Latent Dirichlet Allocation (LDA)[4] があり，情報検索，文書クラスタリング，協調フィルタリングなど，様々な分野に応用されている．

本論文では，離散データの非線形可視化のためのトピックモデル，Probabilistic Latent Semantic Visualization (PLSV)，を提案する．データを可視化することにより，内在する構造的特徴が浮き彫りになり，また，膨大な情報を直感的にブラウジングすることが可能となる．ウェブページ，ブログ，電子メール，特許，論文など，電子的に大量の文書データが日々蓄積されており，文書群可視化法[8),21)] の重要性は益々高まっている．

PLSV では，文書およびトピックが，2 次元もしくは 3 次元ユークリッド可視化空間に座標を持つと仮定し，それらの座標をもとに文書が生成されると考える．具体的には，まず，各文書のトピック比率を，近くに座標を持つトピックが高い確率で選ばれるように決定する．そして，単語は，従来のトピックモデルと同様，文書毎に定めたトピック比率に従ってトピックを選択した後，そのトピックの単語出現確率分にに従って生成する．可視化空間における文書座標を含む PLSV のパラメータは，EM アルゴリズムを用いて与えられた文書データにモデルを適合させることにより推定できる．

これまでに多次元尺度法（MDS）[19]，Isomap[18] など，データ間の距離をできるだけ保存するように低次元空間に埋め込む可視化手法が数多く提案されている．しかし，これらの手法はトピックなど潜在意味を考慮していない．一方，PLSV はトピックモデルに基づき潜在意味を考慮するため，同じ単語を含まない文書であっても，意味的に類似したものであれば，可視化空間において近くに配置することが可能である．

また，PLSA や LDA を用いて，各文書の低次元表現であるトピック比率を推定することができる．しかし，これらの手法では，2 次元空間において 3 トピックまでしか表現できないという問題がある．また，トピック比率はユークリッド空間ではなく単体上に埋め込まれるため，可視化結果が直感的に理解しにくい．それに対し PLSV は，2 次元空間でも任意のトピック数を表現でき，また，人間が直感的に理解しやすいユークリッド空間に埋め込むことができるという特徴を持つ．

PLSA や LDA を用いて推定したトピック比率を，パラメトリック埋め込み法（PE）[11),23)] によりユークリッド空間に埋め込むという方法も考えられる．この方法であれば，任意のト

†1 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

ピック数を表現でき，かつ，ユークリッド空間に埋め込むことができる．しかし，トピック推定過程と可視化過程が分離しているため，トピック推定で蓄積された誤差を可視化において修正することができない，また，可視化空間に文書を埋め込んだきに最適なトピックが推定できるとは限らない，という問題がある．一方 PLSV は，1 つの確率的枠組でトピック推定と可視化を同時に行うため，文書群が可視化空間に埋め込まれたときに最適なトピックを推定することができる．

以下の本文では，まず 2 節で PLSV を定式化し，そのパラメータ推定法を述べる．3 節では，関連研究について述べる．4 節では，文書データ，映画評点データを可視化し，従来法と定量的に比較する．最後に 5 節で，結論と今後の課題を述べる．

2. Probabilistic Latent Semantic Visualization

2.1 モ デ ル

以下では，簡単のため，入力データは文書集合であると考えるが，提案法は購買ログなど他の離散データにも適用できる．

可視化したい対象である N 文書の集合を $\boldsymbol{C} = \{\boldsymbol{w}_n\}_{n=1}^N$ とする．各文書は長さ M_n の単語の系列 $\boldsymbol{w}_n = (w_{n1}, \cdots, w_{nM_n})$ によって表現する．ここで $w_{nm} \in \{1, \cdots, W\}$ は n 番目の文書の m 番目の単語のインデックス，M_n は n 番目の文書の単語数，W は語彙数を表す．

PLSV は，文書集合の可視化空間における座標 $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ を推定するためのトピックモデルである．ここで $\boldsymbol{x}_n = (x_{n1}, \cdots, x_{nD})$ は n 番目の文書の座標，D は可視化空間の次元数を表し，通常 $D = 2$ もしくは $D = 3$ である．また，Z 個のトピックがあり，各トピックは可視化空間において座標 $\boldsymbol{\phi}_z = (\phi_{z1}, \cdots, \phi_{zD})$ を持つとする．文書のトピック比率を，式 (1) のように，可視化空間におけるトピックからのユークリッド距離によって決定する．

$$P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) = \frac{\exp\left(-\frac{1}{2} \parallel \boldsymbol{x}_n - \boldsymbol{\phi}_z \parallel^2\right)}{\sum_{z'=1}^{Z} \exp\left(-\frac{1}{2} \parallel \boldsymbol{x}_n - \boldsymbol{\phi}_{z'} \parallel^2\right)}, \tag{1}$$

ここで $P(z|\boldsymbol{x}_n, \boldsymbol{\Phi})$ は n 番目の文書の z 番目のトピックの比率，$\sum_{z=1}^{Z} P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) = 1$，$\boldsymbol{\Phi} = \{\boldsymbol{\phi}_z\}_{z=1}^Z$ はトピック座標集合，$\|\cdot\|$ は可視化空間におけるユークリッドノルムを表す．トピック比率をこのように計算することで，文書座標 $\boldsymbol{x}_n$ とトピック座標 $\boldsymbol{\phi}_z$ のユークリッド距離が近ければ，そのトピック比率 $P(z|\boldsymbol{x}_n, \boldsymbol{\Phi})$ は高くなる．また，近くに配置されている文書は，トピック座標から同じような距離にあるため，意味的に近く，同じようなトピック比率を持つ．

PLSV では以下の過程により文書集合 $\boldsymbol{C}$ が生成されるとする．

(1) For each topic $z = 1, \cdots, Z$:
  (a) Draw word probability distribution
    $\boldsymbol{\theta}_z \sim \text{Dirichlet}(\alpha)$.
  (b) Draw topic coordinate
    $\boldsymbol{\phi}_z \sim \text{Normal}(\boldsymbol{0}, \beta^{-1}\boldsymbol{I})$.
(2) For each document $n = 1, \cdots, N$:
  (a) Draw document coordinate
    $\boldsymbol{x}_n \sim \text{Normal}(\boldsymbol{0}, \gamma^{-1}\boldsymbol{I})$.
  (b) For each word $m = 1, \cdots, M_n$:
    (i) Draw topic
      $z_{nm}|\boldsymbol{x}_n, \boldsymbol{\Phi} \sim \text{Mult}\left(\{P(z|\boldsymbol{x}_n, \boldsymbol{\Phi})\}_{z=1}^Z\right)$.
    (ii) Draw word
      $w_{nm}|z_{nm}, \boldsymbol{\Theta} \sim \text{Mult}\left(\{P(w|z_{nm}, \boldsymbol{\Theta})\}_{w=1}^W\right)$.

ここで $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_z\}_{z=1}^Z$ は単語出現確率集合，$\boldsymbol{\theta}_z = \{\theta_{zw}\}_{w=1}^W$，$\sum_w \theta_{zw} = 1$，$\theta_{zw} = P(w|z, \boldsymbol{\Theta})$ は z 番目のトピックにおいて w 番目の単語が出現する確率，$\boldsymbol{0}$ は D 次元ゼロベクトル，$\boldsymbol{I}$ は D 次元単位行列を表す．多項分布 (Mult) のパラメータであるトピック毎の単語出現確率 $\boldsymbol{\theta}_z$ は，多項分布の共役事前分布であるディリクレ分布 (Dirichlet) から生成されると仮定した．また，文書座標 $\boldsymbol{x}_n$ およびトピック座標 $\boldsymbol{\phi}_z$ は，可視化結果を安定化させるため，原点を平均とする等分散の正規分布 (Normal) から生成されると仮定した．α，β，γ はハイパーパラメータであり，実験では $\alpha = 0.01$，$\beta = 0.1N$，$\gamma = 0.1Z$ を用いた．

文書座標 $\boldsymbol{x}_n$，トピック座標集合 $\boldsymbol{\Phi}$，単語出現確率集合 $\boldsymbol{\Theta}$ が与えられたときの単語系列 $\boldsymbol{w}_n$ の確率は

$$P(\boldsymbol{w}_n|\boldsymbol{x}_n, \boldsymbol{\Phi}, \boldsymbol{\Theta}) = \prod_{m=1}^{M_n} \sum_{z=1}^{Z} P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) P(w_{nm}|z, \boldsymbol{\Theta}), \tag{2}$$

となる．**図 1** に PLSV のグラフィカルモデルを示す．ここで，塗潰し円は観測変数，中抜き円は潜在変数，矢印は依存関係，矩形は繰り返しを表す．

2.2 パラメータ推定

PLSV の未知パラメータは，最大事後確率（MAP）推定により求めることができる．未知パラメータは文書座標集合 $\boldsymbol{X}$，トピック座標集合 $\boldsymbol{\Phi}$，単語出現確率集合 $\boldsymbol{\Theta}$ であり，全未知パラメータを $\boldsymbol{\Psi} = \{\boldsymbol{X}, \boldsymbol{\Phi}, \boldsymbol{\Theta}\}$ で表現する．なお，トピック数 Z は既知とする．

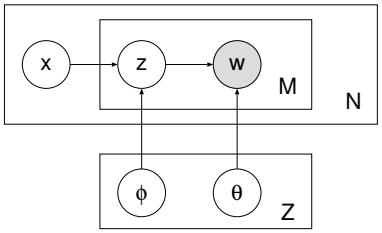


図 1　PLSV のグラフィカルモデル.
Fig. 1　Graphical model representation of PLSV.

文書集合 $\boldsymbol{C}$ が与えられたときのパラメータ集合 $\boldsymbol{\Psi}$ の対数尤度は

$$L(\boldsymbol{\Psi}|\boldsymbol{C}) = \sum_{n=1}^{N}\sum_{m=1}^{M_n}\log\sum_{z=1}^{Z} P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})P(w_{nm}|z,\boldsymbol{\Theta}), \tag{3}$$

となる．事後確率を EM アルゴリズム[5] を用いて最大化することにより，未知パラメータ $\boldsymbol{\Psi}$ を推定する．事前確率を含む完全対数尤度は

$$\begin{aligned} Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}) = &\sum_{n=1}^{N}\sum_{m=1}^{M_n}\sum_{z=1}^{Z} P(z|n,m;\hat{\boldsymbol{\Psi}})\log P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})P(w_{nm}|z,\boldsymbol{\Theta}) \\ &+\sum_{n=1}^{N}\log p(\boldsymbol{x}_n)+\sum_{z=1}^{Z}\log p(\boldsymbol{\phi}_z)+\sum_{z=1}^{Z}\log p(\boldsymbol{\theta}_z), \end{aligned} \tag{4}$$

となる．ここで $\hat{\boldsymbol{\Psi}}$ は現在の推定値，$P(z|n,m;\hat{\boldsymbol{\Psi}})$ は現在の推定値が与えられたときの，n 番目の文書の m 番目の単語のトピック事後確率を表す．E ステップでは，ベイズ則に従い，トピック事後確率を計算する．

$$P(z|n,m;\hat{\boldsymbol{\Psi}}) = \frac{P(z|\hat{\boldsymbol{x}}_n,\hat{\boldsymbol{\Phi}})P(w_{nm}|z,\hat{\boldsymbol{\Theta}})}{\sum_{z'=1}^{Z}P(z'|\hat{\boldsymbol{x}}_n,\hat{\boldsymbol{\Phi}})P(w_{nm}|z',\hat{\boldsymbol{\Theta}})}, \tag{5}$$

ここで $P(z|\hat{\boldsymbol{x}}_n,\hat{\boldsymbol{\Phi}})$ は式 (1) により計算される．M ステップでは，$\sum_{w=1}^{W}\theta_{zw}=1$ という制約のもと，$Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ を θ_{zw} に関して最大化することにより，単語出現確率の推定値 $\hat{\theta}_{zw}$ を求める．

$$\hat{\theta}_{zw} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M_n} I(w_{nm}=w)P(z|n,m;\hat{\boldsymbol{\Psi}})+\alpha}{\sum_{w'=1}^{W}\sum_{n=1}^{N}\sum_{m=1}^{M_n} I(w_{nm}=w')P(z|n,m;\hat{\boldsymbol{\Psi}})+\alpha W}, \tag{6}$$

ここで $I(\cdot)$ は指示関数，つまり A が真ならば $I(A)=1$，偽ならば $I(A)=0$，を表す．文書座標 $\boldsymbol{x}_n$ およびトピック座標 $\boldsymbol{\phi}_z$ の推定値は閉形式で書くことができない．そのため，準ニュートン法[15] などの最適化法を用いて $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ を最大化することにより，座標を推定する．準ニュートン法で必要となる，$Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ の $\boldsymbol{x}_n$ および $\boldsymbol{\phi}_z$ に関する勾配ベクトルはそれぞれ

$$\frac{\partial Q}{\partial \boldsymbol{x}_n} = \sum_{m=1}^{M_n}\sum_{z=1}^{Z}\Big(P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})-P(z|n,m;\hat{\boldsymbol{\Psi}})\Big)(\boldsymbol{x}_n-\boldsymbol{\phi}_z)-\gamma\boldsymbol{x}_n, \tag{7}$$

$$\frac{\partial Q}{\partial \boldsymbol{\phi}_z} = \sum_{n=1}^{N}\sum_{m=1}^{M_n}\Big(P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})-P(z|n,m;\hat{\boldsymbol{\Psi}})\Big)(\boldsymbol{\phi}_z-\boldsymbol{x}_n)-\beta\boldsymbol{\phi}_z, \tag{8}$$

となる．収束するまで E ステップと M ステップを交互に繰り返すことにより，$\boldsymbol{\Psi}$ の局所最適解を推定することができる．

3. 関 連 研 究

3.1 PLSA

PLSV は PLSA[10] をベースとして可視化法に発展させた手法である．相違点を挙げると，PLSV ではトピック比率集合が文書座標集合 $\boldsymbol{X}$ とトピック座標集合 $\boldsymbol{\Phi}$ を用いて式 (1) によって決定されるのに対し，PLSA ではトピック比率集合が未知パラメータ $\boldsymbol{\Lambda}=\{\boldsymbol{\lambda}_n\}_{n=1}^{N}$ として直接推定される．ここで $\boldsymbol{\lambda}_n=\{\lambda_{nz}\}_{z=1}^{Z}$ であり，$\lambda_{nz}=P(z|d_n,\boldsymbol{\Lambda})$ は n 番目の文書の z 番目のトピック比率を表す．従って，D 次元空間において，PLSV では任意のトピック数のトピック比率を表現できるのに対し，PLSA は $D+1$ トピックまでのトピック比率しか表現できない．トピック比率に関するパラメータは，PLSV では $(N+Z)D$，PLSA では $N(Z-1)$ である．一般に $D \ll Z \ll N$ であるため，PLSV は PLSA に比べパラメータが少なく，過学習に陥りにくい．

PLSA の未知パラメータ $\boldsymbol{\Upsilon}$ はトピック比率の集合 $\boldsymbol{\Lambda}$ と単語出現確率の集合 $\boldsymbol{\Theta}$ である．未知パラメータは EM アルゴリズムを用いて以下の尤度を最大化することにより推定できる．

$$L(\boldsymbol{\Upsilon}|\boldsymbol{C}) = \sum_{n=1}^{N}\sum_{m=1}^{M_n}\log\sum_{z=1}^{Z} P(z|d_n,\boldsymbol{\Lambda})P(w_{nm}|z,\boldsymbol{\Theta}). \tag{9}$$

E ステップでは，現在の推定値が与えられたもとでのトピック事後確率を計算する．

$$P(z|d_n,w_{nm};\hat{\boldsymbol{\Upsilon}}) = \frac{P(z|d_n,\hat{\boldsymbol{\Lambda}})P(w_{nm}|z,\hat{\boldsymbol{\Theta}})}{\sum_{z'=1}^{Z}P(z'|d_n,\hat{\boldsymbol{\Lambda}})P(w_{nm}|z',\hat{\boldsymbol{\Theta}})}. \tag{10}$$

M ステップでは，トピック比率を下式により推定し，

$$\hat{\lambda}_{nz} = \frac{\sum_{m=1}^{M_n}P(z|d_n,w_{nm};\hat{\boldsymbol{\Upsilon}})}{\sum_{z'=1}^{Z}\sum_{m=1}^{M_n}P(z'|d_n,w_{nm};\hat{\boldsymbol{\Upsilon}})}, \tag{11}$$

また，単語出現確率を下式により推定する．

$$\hat{\theta}_{zw}=\frac{\sum_{n=1}^{N}\sum_{m=1}^{M_n} I(w_{nm}=w)P(z|d_n,w_{nw};\hat{\boldsymbol{\Upsilon}})}{\sum_{w'=1}^{W}\sum_{n=1}^{N}\sum_{m=1}^{M_n} I(w_{nm}=w')P(z|d_n,w_{nm};\hat{\boldsymbol{\Upsilon}})}. \tag{12}$$

PLSV と同じく，トピック比率および単語出現確率の事前分布として，ディリクレ分布を用いることができる．

3.2　パラメトリック埋め込み法

パラメトリック埋め込み法（PE）[11),23)] は，離散確率分布集合を入力とする非線形可視化法であり，クラス構造や分類器の可視化に用いられている[22)]．PE により，PLSA で推定した任意のトピック数のトピック比率 $\hat{\boldsymbol{\Lambda}}$ を D 次元ユークリッド空間に埋め込むことができる．PE では，下式の KL ダイバージェンスが最小することで，入力の確率分布をできるだけ保存するように低次元空間にデータを埋め込む．

$$E(\boldsymbol{X},\boldsymbol{\Phi})=\sum_{n=1}^{N}\sum_{z=1}^{Z} P(z|d_n,\hat{\boldsymbol{\Lambda}})\log\frac{P(z|d_n,\hat{\boldsymbol{\Lambda}})}{P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})}, \tag{13}$$

ここで $P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})$ は n 番目の文書の座標 $\boldsymbol{x}_n$ が与えられたときの z 番目のトピック比率を表し，PLSV と同様に式 (1) により定義される．未知パラメータである文書座標集合 $\boldsymbol{X}$ とトピック座標集合 $\boldsymbol{\Phi}$ は準ニュートン法などの最適化手法により推定できる．$E(\boldsymbol{X},\boldsymbol{\Phi})$ の $\boldsymbol{x}_n$ および $\boldsymbol{\phi}_z$ に関する勾配ベクトルはそれぞれ

$$\frac{\partial E}{\partial \boldsymbol{x}_n}=\sum_{z=1}^{Z}\Big(P(z|d_n,\hat{\boldsymbol{\Lambda}})-P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})\Big)(\boldsymbol{x}_n-\boldsymbol{\phi}_z), \tag{14}$$

$$\frac{\partial E}{\partial \boldsymbol{\phi}_z}=\sum_{n=1}^{N}\Big(P(z|d_n,\hat{\boldsymbol{\Lambda}})-P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})\Big)(\boldsymbol{\phi}_z-\boldsymbol{x}_n), \tag{15}$$

となる．PLSV と同様に，座標の事前分布として，原点が平均の等分散正規分布を用いることにより，可視化結果を安定化させることができる．

2.2 節で述べた PLSV のパラメータ推定は，PE における座標計算を，PLSA の M ステップに組み込んだものと見なすことができる．推定された座標は，トピック比率を推定するために E ステップで使われる．

3.3　他の関連手法

生成モデルに基づく可視化手法として Generative Topographic Mapping (GTM)[1)] や 12) で提案された手法がある．しかしながら，これらのモデルはトピックモデルではなく，文書データなどトピックのような潜在構造を持つ離散データには適していない．GTM に基づくトピックモデルも提案されている[13)] が，主に可視化ではなく，系列予測のために使われている．Correlated topic model[3)] はトピック間の相関をモデル化することができるトピックモデルである．PLSV も，可視化空間において近くに配置されたトピックは同じ文書から選択されやすくなるため，トピック相関をモデル化できる．PLSV は，トピックを潜在変数として扱う教師なし可視化手法であり，フィッシャー線形判別法[7)] などの教師あり可視化手法とは異なる．

4.　実　　験

4.1　比 較 手 法

PLSV の有効性を評価するため，MDS，Isomap，PLSA，PLSA+PE の 4 つの従来法と比較した．可視化空間は 2 次元 ($D=2$) とした．

MDS は線形次元圧縮法であり，原空間でのデータ間距離が可視化空間でできるだけ保存されるようにデータを埋め込む．1 つの文書を単語頻度ベクトル $\boldsymbol{v}_n=(v_{n1},\cdots,v_{nW})$ として表現し，入力データとした．ここで $v_{nw}=\sum_{m=1}^{M_n} I(w_{nm}=w)$ は n 番目の文書に含まれる単語 w の頻度を表す．なお，文書長の影響をなくすため，L2 ノルムが 1 になるように正規化した．

Isomap は非線形次元圧縮法である．まず，h 近傍グラフを作成し，MDS によりそのグラフにおける最短経路長をできるだけ保存するようにデータを埋め込む．近傍を見つけるための類似度として，文書間の類似度としてよく用いられる，単語頻度ベクトル間のコサイン類似度を用いた．近傍数は $h=5$ とした．

PLSA では，トピック数 $Z=3$ の PLSA で推定したトピック比率を，2 次元単体上に変換し，可視化した．また，PLSA+PE では，PLSA でトピック比率を推定した後，PE を用いて 2 次元ユークリッド空間に埋め込むことで，可視化した．

4.2　デ　ー　タ

NIPS，20News，EachMovie の 3 データを用いた．

NIPS データは 2001 年から 2003 年までの国際会議 Neural Information Processing Systems Conference (NIPS) で発表された論文のデータである．文書数は 593，語彙数は 14,036 であった．各文書は，Neuroscience，Applications など 13 の研究分野に分類されている．

20News データは，20 のグループからなるニュースグループに投稿された約 20,000 記事からなる[14)]．ストップワード，出現回数が 50 未満の単語，単語数が 50 未満の文書を省いた．なお，ストップワードとして文書集合に依存しないものを用いた．語彙数は 6,754 であった．各グループから 50 文書，全 1,000 文書をサンプリングし，可視化した．

EachMovie データは映画評点データであり，しばしば協調フィルタリング問題に用いられる．映画とユーザをそれぞれ文書と単語と見なし，各映画をその映画に評点を付けたユーザの系列として表現する．映画は，アクション，コメディ，ロマンスなど，10 ジャンルに分類されている．50 評点以下の映画は省いた．また，2 つ以上のジャンルに分類される映画がある場合，4.3 節で述べる評価指標が他のデータと一致しなくなるため，2 つ以上のジャンルに分類される映画は省いた．映画数は 764，ユーザ数は 7,180 であった．

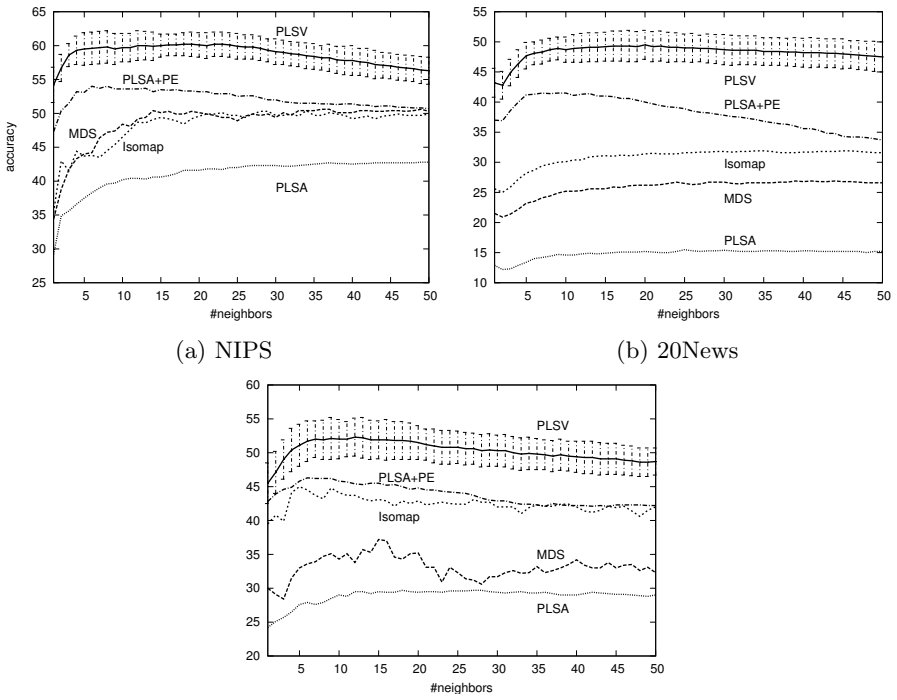
4.3 評価指標

意味的に類似した文書が近くに配置された可視化結果は，文書間の関連を直感的に理解しやすくなり，関連の深い文書をたどりながら目的の文書にたどりつくことが可能となるため，好ましい可視化結果である．同じラベルが付けられた文書は意味的に類似していると考えたとき，同じラベルのサンプルが近くに配置され，異なるラベルのサンプルが遠くに配置されていた場合に高くなる指標が，評価指標として適切である．

この性質を満たす指標として，本論文では可視化空間における k 近傍法によるラベルの予測精度を用いた．k 近傍法では，近傍 k 個のサンプルで最多数のラベルを予測ラベルとするため，近傍に同じラベルのサンプルが多数配置されているとき予測精度が高くなる．ここで近傍はユークリッド距離を用いて選択する．予測精度は $acc(k) = \frac{1}{N}\sum_{n=1}^{N} I\big(y_n = \hat{y}_k(\boldsymbol{x}_n)\big) \times 100$, により計算できる．ここで y_n は n 番目の文書のラベル，$\hat{y}_k(\boldsymbol{x}_n)$ は k 近傍法により予測された座標 $\boldsymbol{x}_n$ の文書のラベルを表す．なお，ラベル情報は可視化の際には用いていない．

4.4 結　　果

各手法による予測精度を**図 2** に示す．PLSV および PLSA+PE のトピック数は $Z = 50$ とした．PLSA のトピック数は $D = 2$ であるため自動的に $Z = 3$ と決まる．20News データでは，ランダムサンプリングにより，30 の評価用データセットを作成し，平均予測精度で評価した．NIPS と EachMovie における PLSV，PLSA+PE，PLSA の予測精度は初期値を変えて 30 回実験したときの平均を表す．ここで，n 番目の文書の m 番目の単語がトピック z である確率 $P(z|n,m)$ の初期値として，0 から 1 の一様乱数を総和が 1 になるように正規化したものを用いた．また座標の初期値として，-0.5 から 0.5 の一様乱数を用いた．PLSV のみ標準偏差を表示する．全てのデータにおいて，PLSV が高い予測精度を達成している．この結果は，PLSV は文書データの特徴を保存したまま，2 次元ユークリッド空間にデータを埋め込むことができることを示唆する．PLSA+PE の予測精度は PLSV よりも低い．これは，トピック推定と可視化がそれぞれ異なる目的関数を持つためであると考えられる．PLSA の予測精度が低い理由として，PLSA は 3 トピックしか 2 次元空間上で



(a) NIPS　(b) 20News

(c) EachMovie

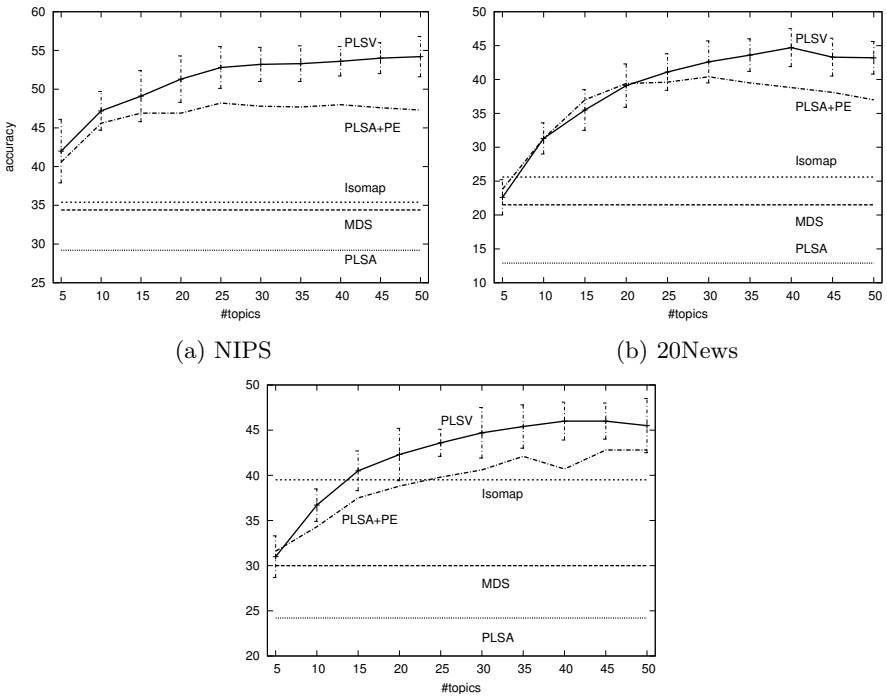
図 2　近傍 k を変化させたときの，2 次元可視化空間における k 近傍法による予測精度．

Fig. 2　Accuracy with the k-nearest neighbor method in the two-dimensional visualization space with different numbers of neighbors k.

表現できないこと，ユークリッド距離はトピック比率の距離として適切でないことが，考えられる．

PLSV と PLSA+PE のトピック数を変化させたときの結果を評価した．最近傍法による予測精度を**図 3** に示す．トピックが少ない場合，PLSV の精度はそれほど高くなく，PLSA+PE と同等である．しかし，トピック数が増加するに従い，PLSV は PLSA+PE に比べ優位になっている．これは，トピック数の増加にともなうパラメータ数の増加が，PLSV は PLSA+PE に比べ小さく，過学習に陥りにくいためであると考えられる．

表 1 に，NIPS データの PLSV 可視化結果に基づいて，最近傍法によりラベルを予測した



(a) NIPS　(b) 20News

(c) EachMovie

図 3　PLSV と PLSA+PE のトピック数を変化させたときの 2 次元可視化空間における最近傍法による予測精度.
Fig. 3　Accuracy with the nearest neighbor method in the two-dimensional visualization space with different numbers of topics Z for PLSV and PLSA+PE.

場合の混同行列（百分率）を示す．‘AP:Applications’ の予測精度が低くなっている．このカテゴリは，特定の技術分野ではなく，画像，文書，センサなど幅広い応用分野を対象としており，特徴的な技術用語が少ない．トピックモデルでは，このようなカテゴリを抽出するには適しておらず，低い精度となっていると考えられる．一方，‘CN: Control & Reinforce Learning’ のように特定の技術分野であり，用いられる単語が特徴的であるカテゴリでは高い予測精度となっている．また，‘BI: Brain Imaging’，‘ET: Emerging Technologies’，‘VM: Vision(Biological)’ のようにサンプル数が少ないラベルの予測精度は低くなっており，トピック推定には，ある程度のサンプル数が必要であると言える．ラベル間の関係を見ると，

表 1　NIPS データにおける，可視化空間における最近傍法により予測した場合の混同行列（百分率）．行は正解ラベル，列は予測ラベルを表す.
Table 1　Confusion matrix with the nearest neighbor classification on NIPS. The row and column represent the true and estimated label, respectively.

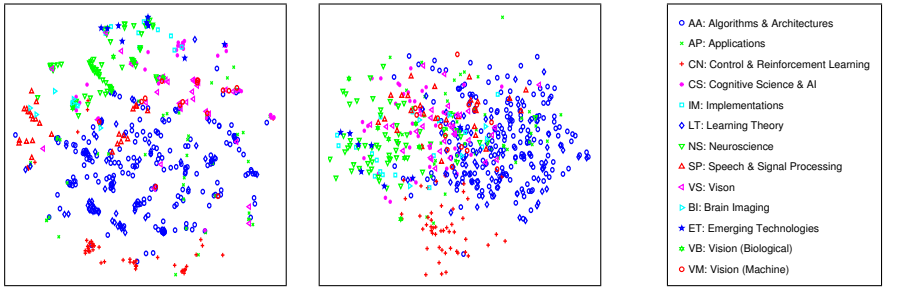
	AA	AP	CN	CS	IM	LT	NS	SP	CS	BI	ET	VM	VB	#samples
AA	67	8	2	2	0	13	1	2	3	0	0	0	0	209
AP	38	17	2	15	0	6	2	4	4	2	0	2	6	47
CN	8	2	80	0	0	4	2	0	0	0	0	4	0	50
CS	10	12	0	40	0	2	25	0	5	2	0	2	0	40
IM	0	0	0	0	50	0	12	0	0	0	38	0	0	16
LT	43	1	3	3	0	47	1	0	1	0	0	0	0	68
NS	0	2	2	21	2	2	55	8	2	5	0	5	0	66
SP	23	8	0	0	0	0	4	62	4	0	0	0	0	26
CS	12	12	0	3	0	6	0	3	33	3	3	6	18	33
BI	14	0	0	0	14	0	14	14	0	29	0	0	14	7
ET	0	0	0	0	56	0	0	0	11	0	33	0	0	9
VM	0	14	14	0	0	0	0	0	43	0	0	29	0	7
VB	7	27	0	0	0	0	0	0	40	0	0	0	27	15

‘CS: Cognitive Science & AI’ と ‘NS: Neuroscience’，‘AA: Algorithms & Archtectures’ と ‘LT: Learning Theory’ など，関連する分野に分類される確率が高くなっており，このことから関連する分野の文書は近くに配置されていると言える．

図 4，**図 5**，**図 6** は PLSV ($Z = 50$)，MDS，Isomap，PLSA ($Z = 3$)，PLSA+PE ($Z = 50$) の NIPS，20News，EachMovie データにおける可視化結果を示す．ここで 1 つの点 1 つの文書 (映画) を表し，その色形はラベル情報を表す．PLSV による可視化結果では，同じラベルの文書が近くに集まっている傾向が強いことが視覚的にも分かる．PLSA では多くの文書が角に集まっており，2 次元空間ではトピック構造を適切に表現できていない．

図 7 に PLSV でトピック数を変化させたときの NIPS データの可視化結果を示す．トピック数が少ない場合は異なるラベルのデータが混在しているが，トピック数 $Z = 40$ では $Z = 50$ とほぼ同等の可視化結果が得られている．

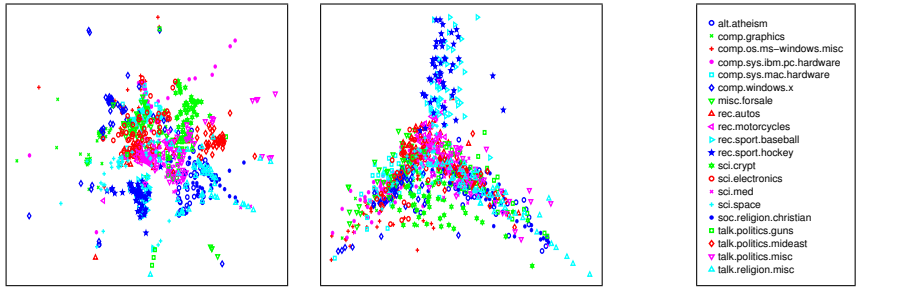
潜在意味を考慮する利点として，同じ単語を含まない文書であっても意味的に類似している場合，近くに配置されるという点がある．この性質を確認するため，20News データにおいて，同じ単語を含まず，かつ，同じラベルである (意味的に類似している) 文書が，最も近傍に配置されている文書の数を各手法で比較した．その結果を**図 8** に示す．PLSV の値が最も高く，潜在意味を適切に抽出できていると言える．MDS，Isomap では文書間類似度が高いものが近くに配置されるため，同じ単語を含まない文書は近くに配置されにくく，低い値



(a) PLSV　(b) MDS

(c) Isomap　(d) PLSA　(e) PLSA+PE

図 4　NIPS データの可視化結果.
Fig. 4　Visualization of documents in NIPS.



(a) PLSV　(b) MDS

(c) Isomap　(d) PLSA　(e) PLSA+PE

図 5　20News データの可視化結果.
Fig. 5　Visualization of documents in 20News.

となっている．また，PLSA の値が低いのは，潜在トピック数が少ないためと考えられる．

PLSV による可視化を詳細に解析する．トピック数 $Z = 50$ のときの PLSV による 20News データの可視化結果を**図 9** に示す．黒円はトピック座標 $\boldsymbol{\phi}_z$，黒バツはラベル毎の平均文書座標を表す．平均文書座標は，N_y をラベル y の文書数としたとき，$\boldsymbol{\mu}_y = \frac{1}{N_y}\sum_{n=1}^{N} I(y_n = y)\boldsymbol{x}_n$ により計算される．黒円付近の数字はトピックのインデックスであり，図 9 下部の表の 1 行目と対応する．下部の表は，各トピックにおいて最も出現しやすい 10 単語を表す．可視化の結果，関連するラベルのデータは近くに配置されている (例えば，rec.sport.baseball と rec.sport.hockey，rec.autos と rec.motorcycles，comp.sys.mac.hardware と comp.sys.ibm.pc.hardware，soc.religion.christian と talk.religion.misc など)．また，ラベル平均の付近に配置されたトピックでは，そのラベルに典型的な単語が高い確率で出現している．例えば，rec.sport の近くのトピック（$z = 1$，$z = 2$）は ‘team’，‘players’，‘game’ の出現頻度が高く，また，comp.graphics の近くのトピック（$z = 5$）では ‘graphics’，‘points’，‘lines’ の出現頻度が高い．

図 10 に PLSV（$Z = 50$）による EachMovie データの可視化結果，および，いくつかの映画タイトルを示す．可視化の結果，同じジャンルの映画は近くに配置されている．例えば，クラシック映画は右下部に配置され，外国映画は上部に配置されている．クラシック映画のみ見るユーザが多くいるため，クラシック映画が近くにかたまっていると考えられる．

これまでの実験では，単語数が少ない文書，および，出現頻度の少ない単語を除いたデータを用いていた．ここでは，PLSV を用いて可視化する際に必要な単語数を見積もるため，20News の全文書を可視化し，単語数と可視化座標の適切さ (最近傍法による予測精度) の関係を調べた．その結果を**図 11** に示す．ここで横軸は単語数，縦軸は単語数 10 毎に文書

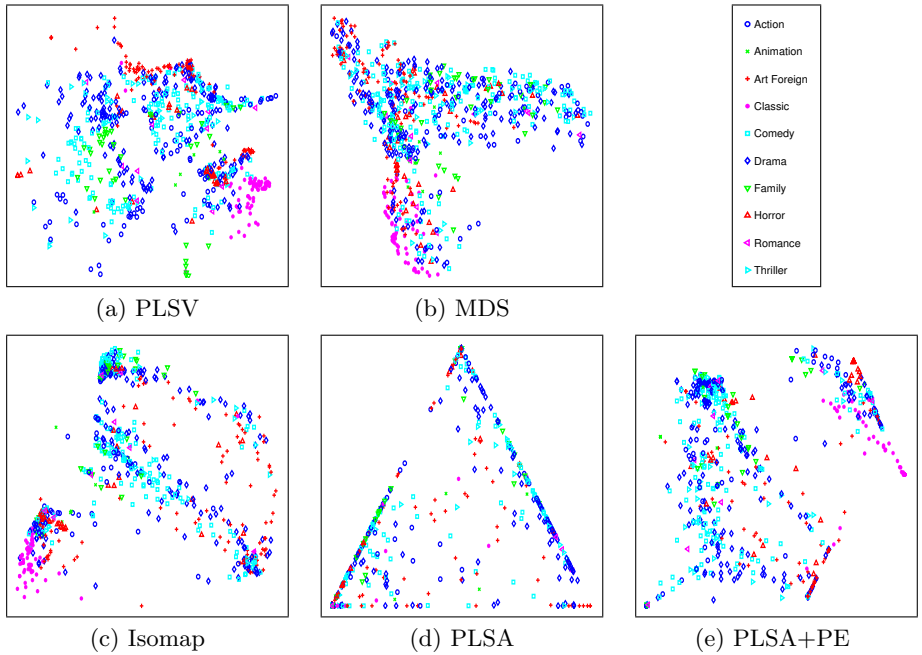


図 6　EachMovie データの可視化結果.
Fig. 6　Visualization of movies in EachMovie.

集合を区切ったときの平均予測精度を表す．単語数が少ない文書は精度が低く，単語数が多い文書は精度が高い．また，100 を越えたあたりで精度の伸びが少なくなっている．この結果より，PLSV により高い精度でトピックを推定するためには，100 以上の単語数が必要であると言える．なお，図 2，図 3 の 20News の実験結果は，単語数 100 以下の文書も含まれるデータであり，その場合でも PLSV は他の比較手法より精度が高く，有効な可視化手法と言える．

今回の実験では，提案法を用いることより，従来法に比べ関連文書をより近くに配置できることを確認した．しかし，ユーザにとって，直感的なデータ理解やブラウジングが，提案法によりどれほど容易になるのか，などの認知的な観点からの有効性の検証も重要であり，今後の課題である．

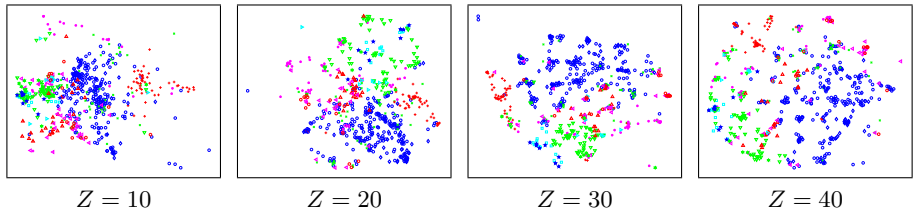


図 7　NIPS データのトピック数を変化させたときの PLSV による可視化結果.
Fig. 7　Visualization by PLSV with different numbers of topics on NIPS.

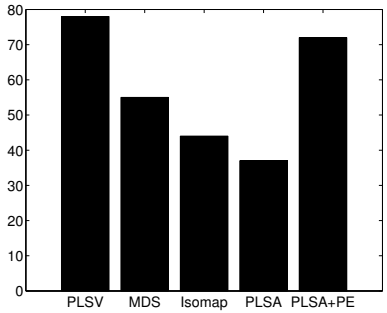
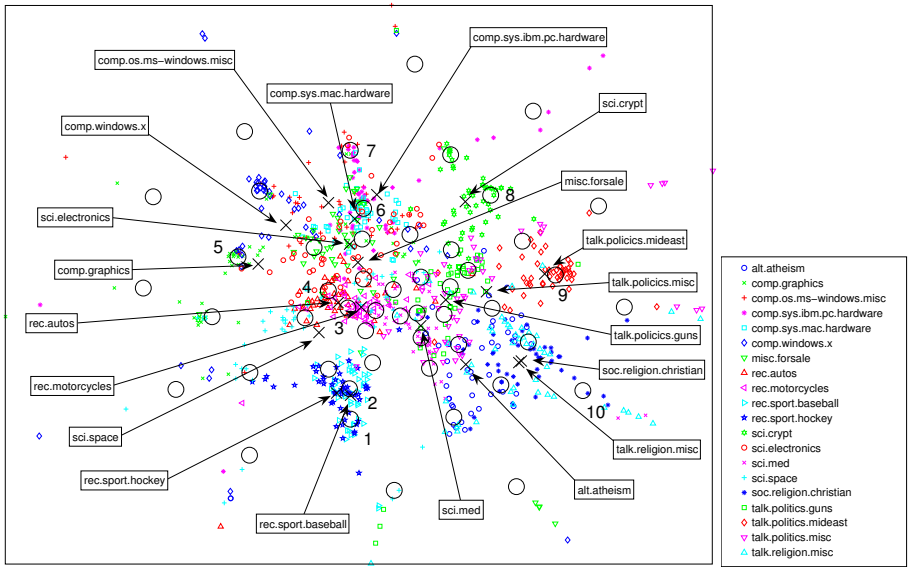


図 8　20News データにおける，同じ単語を含まずかつ同じラベルである文書が，最も近傍に配置されている文書の数.
Fig. 8　The number of documents whose nearest neighbor has the same label and does not have any same words on 20News.

5. おわりに

本論文では，文書などの離散データを可視化するためのトピックモデル Probabilistic Latent Semantic Visualization (PLSV) を提案した．PLSV はトピックモデルであるため，トピックモデルに関連する研究をもとにした拡張が可能である．例えば，文書とともに，画像[2] や時間[20] や著者[16] や引用[6] など，他の情報を扱うためのトピックモデルも提案されており，本論文の枠組みを用いて，これらの情報も同時に可視化することができる．本論文ではトピック数を既知としたが，ディリクレ過程[17] などのノンパラメトリックベイズモデルに拡張することで，トピック数の自動決定も可能となる．また，MAP 推定ではなく，ベ



1	2	3	4	5	6	7	8	9	10
team	game	car	high	graphics	windows	scsi	clipper	israel	god
players	good	bike	engine	points	pc	printer	encryption	israeli	jesus
season	games	big	battery	lines	keyboard	bus	public	militia	ra
league	play	cars	car	point	drive	windows	system	arab	bible
nhl	baseball	water	stuff	line	mouse	drivers	government	jewish	christ
teams	guys	drive	low	reference	mac	hp	keys	jews	heaven
average	win	buy	bought	image	disk	speed	chip	turkish	people
hall	fans	thing	dealer	access	card	local	security	armenian	john
make	division	front	kind	program	memory	network	escrow	armenians	scripture
dave	guy	miles	speed	comp	dos	fonts	nsa	palestine	spirit

図 9　PLSV（$Z = 50$）による 20News データの可視化結果．黒円はトピック座標，黒バツはラベル毎の平均文書座標を表す．下部の表は各トピックにおいて最も出現しやすい 10 単語を表す．

Fig. 9　Visualization of documents in 20News by PLSV with $Z = 50$. Each black circle indicates a topic coordinate, and each black × indicates a label mean. The table at the bottom shows the ten most probable words for ten topics estimated with the visualization.

イズ推定[9] により，より頑健な推定ができると期待できる．被験者実験等による認知的な観点からの有効性の検証は今後の課題である．

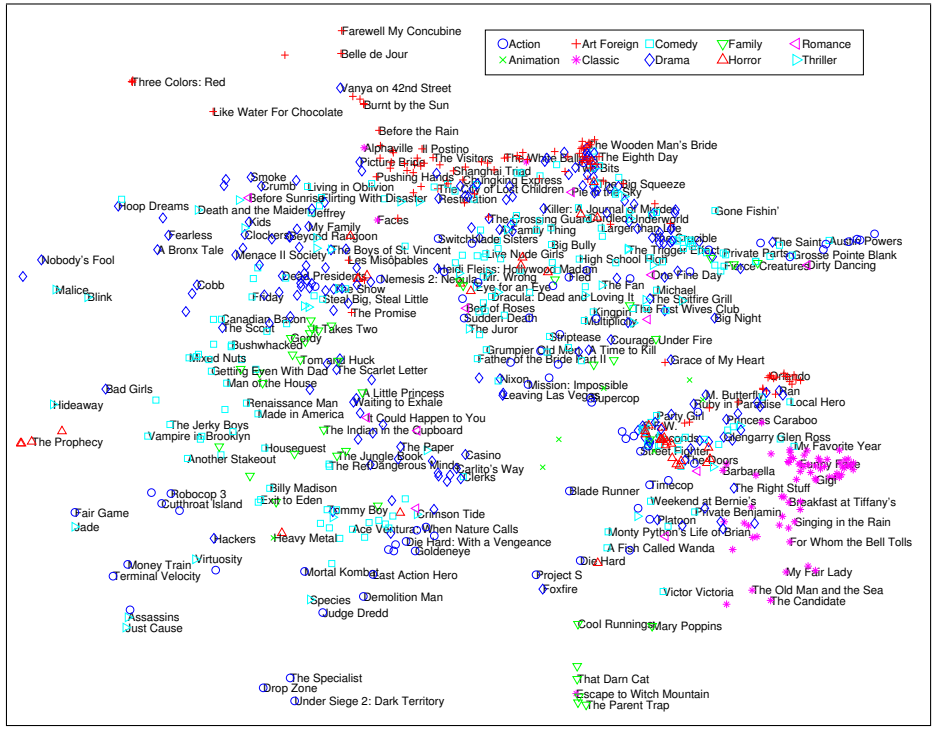


図 10　PLSV（$Z = 50$）による EachMovie データの可視化．いくかの映画タイトルも表示している．

Fig. 10　Visualization of movies in EachMovie by PLSV with $Z = 50$. Some examples of movie titles are also shown.

参　考　文　献

1) Bishop, C.M., Svensen, M. and Williams, C. K.I.: GTM: The Generative Topographic Mapping, *Neural Computation*, Vol.10, No.1, pp.215–234 (1998).
2) Blei, D.M. and Jordan, M.I.: Modeling annotated data, *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR conference on Research and development in information retrieval*, pp.127–134 (2003).
3) Blei, D.M. and Lafferty, J.D.: A Correlated Topic Model of Science, *The Annals of Applied Statistics*, Vol.1, No.1, pp.17–35 (2007).

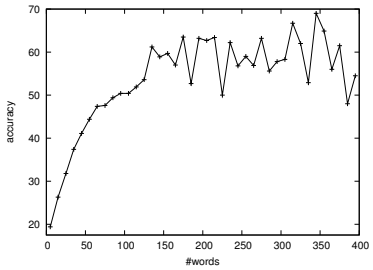
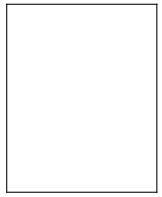


図 11　PLSV（$Z = 50$）により 20News 全文書を可視化したときの単語数と予測精度.
Fig. 11　The number of words and accuracies when all documents in 20News are visualized by PLSV with $Z = 50$.

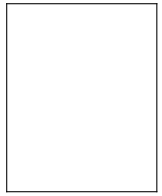
4) Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
5) Dempster, A., Laird, N. and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, Vol.39, No.1, pp.1–38 (1977).
6) Dietz, L., Bickel, S. and Scheffer, T.: Unsupervised prediction of citation influences, *ICML '07: Proceedings of the 24th International Conference on Machine Learning*, pp.233–240 (2007).
7) Fisher, R.A.: The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, Vol.7, pp.179–188 (1936).
8) Fortuna, B., Grobelnik, M. and Mladenic, D.: Visualization of text document corpus, *Informatica*, Vol.29, No.4, pp.497–502 (2005).
9) Griffiths, T.L. and Steyvers, M.: Finding scientific topics, *Proceedings of the National Academy of Sciences*, Vol.101 Suppl 1, pp.5228–5235 (2004).
10) Hofmann, T.: Probabilistic Latent Semantic Analysis, *UAI '99: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, pp.289–296 (1999).
11) Iwata, T., Saito, K., Ueda, N., Stromsten, S., Griffiths, T.L. and Tenenbaum, J.B.: Parametric Embedding for Class Visualization, *Neural Computation*, Vol.19, No.9, pp.2536–2556 (2007).
12) Kabán, A., Sun, J., Raychaudhury, S. and Nolan, L.: On class visualisation for high dimensional data: Exploring scientific data sets, *DS '06: Proceedings of the 9th International Conference on Discovery Science* (2006).
13) Kabán, A.: Predictive Modelling of Heterogeneous Sequence Collections by Topographic Ordering of Histories, *Machine Learning*, Vol.68, No.1, pp.63–95 (2007).
14) Lang, K.: NewsWeeder: learning to filter netnews, *ICML '95: Proceedings of the 12th International Conference on Machine Learning*, pp.331–339 (1995).
15) Liu, D.C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, *Math. Programming*, Vol.45, No.3, pp.503–528 (1989).
16) Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P.: The author-topic model for authors and documents, *UAI '04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp.487–494 (2004).
17) Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: Hierarchical Dirichlet Processes, *Journal of the American Statistical Association*, Vol.101, No.476, pp.1566–1581 (2006).
18) Tenenbaum, J.B., de Silva, V. and Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction., *Science*, Vol.290, No.5500, pp.2319–2323 (2000).
19) Torgerson, W.: *Theory and methods of scaling*, Wiley, New York (1958).
20) Wang, X. and McCallum, A.: Topics over time: a non-Markov continuous-time model of topical trends, *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp.424–433 (2006).
21) Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. and Crow, V.: Visualizing the non-visual: spatial analysis and interaction with information from text documents, *INFOVIS '95: Proceedings of the 1995 IEEE Symposium on Information Visualization*, pp.51–58 (1995).
22) 岩田具治，斉藤和巳：パラメトリック埋め込み法を用いた分類器の視覚的解析，情報処理学会論文誌， Vol.48, No.12, pp.4012–4022 (2007).
23) 岩田具治，斉藤和巳，上田修功：パラメトリック埋め込み法によるクラス構造の可視化，情報処理学会論文誌， Vol.46, No.9, pp.2337–2346 (2005).

(平成 19 年 11 月 28 日受付)
(平成 20 年 2 月 4 日採録)



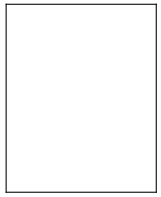
岩田　具治（正会員）

平 13 慶大・環境情報卒．平 15 東大大学院・総合文化・広域科学修士課程了．同年 NTT 入社．平 20 京大大学院・情報学・システム科学博士課程了．博士（情報学）．現在, NTT コミュニケーション科学基礎研究所研究員．機械学習，データマイニング，情報可視化の研究に従事．平 16 船井ベストペーパー賞，平 19 FIT ヤングリサーチャー賞等受賞．電子情報通信学会会員．



山田　武士（正会員）

昭 63 東大・理・数学科卒．同年 NTT 入社．平 8 より 1 年間英国コベントリー大学客員研究員．現在， NTT コミュニケーション科学基礎研究所　創発環境研究グループリーダ．主として統計的機械学習，データマイニング，メタヒューリスティクスによる組合せ最適化等の研究に従事．博士（情報学）．電子情報通信学会，ACM，IEEE 各会員．



上田　修功（正会員）

昭 57 阪大・工・通信工学卒．昭 59 同大学大学院修士課程了．工学博士．同年 NTT 入社．平 5 より 1 年間 Purdue 大学客員研究員．画像処理，パターン認識・学習，ニューラルネットワーク，統計的学習，Web 統計解析の研究に従事．現在，NTT コミュニケーション科学基礎研究所副所長 企画担当主席研究員，奈良先端大客員教授．電気通信普及財団賞受賞、電子情報通信学会論文賞，FIT 船井ベストペーパー賞等受賞．電子情報通信学会，IEEE 各会員．