日本応用数理学会論文誌
Vol. 18, No. 3, 2008, pp. 363～388

PMM 型主成分分析を用いた文書ストリームの主要潜在トピック抽出

木村 昌弘 * 斉藤 和巳 †
* 龍谷大学理工学部電子情報学科
† 静岡県立大学経営情報学部

概要. 文書ストリームデータにおける主要潜在トピックの抽出を，文書の BOW 表現に基づいて効率よく行う，PMM-PCA 法と呼ぶ新たな教師なし学習法を提案する．PMM-PCA 法は，PCA 法と異なり，単語頻度ベクトル群の時系列として表現された文書ストリームデータに対して，その適切な確率的生成モデルに従うという性質を有している．実際の文書ストリームデータを用いた実験により，提案法の有効性を実証する．

Extracting Major Latent Topics in Document Streams Using PMM-PCA

Masahiro Kimura* Kazumi Saito†
*Department of Electronics and Informatics, Ryukoku University
†School of Administration and Informatics, University of Shizuoka

Abstract. We propose a new unsupervised learning method called *PMM-PCA* for efficiently extracting the major latent topics in a document stream based on the "bag-of-words" (BOW) representation of a document. Unlike PCA, PMM-PCA follows a suitable probabilistic generative model for the document stream represented as time-series of word-frequency vectors. Using real data of document streams on the Web, we experimentally demonstrate the effectiveness of the proposed method.

1. はじめに

World Wide Web を対象とした数理モデリングやデータマイニングの研究は，学習理論に関する様々な問題を提起し，計算知能や機械学習における新たな研究分野として注目されている [1,5–7,10]．Web 空間は，コミュニケーションの新たなメディアとして発展し続けており，今や，巨大な情報貯蔵庫を形成しているとともに，人間社会の縮図とも考えられうるようになりつつある．したがって，Web 空間内で日々流通する大量情報に基づいて，注目すべき社会現象や潜在トピックを発見し社会の動向を理解する手法を構築することは，興味深い研究課題と考えられる．本論文では，新聞ニュース記事や研究企画書や政治ブログなどのような Web 空間における比較的長い文書からなる文書ストリームデータを対象として，主要潜在トピックを抽出しその主要アクティブ期間を同定する問題を考察

する．ここに，潜在トピックの主要アクティブ期間とは，その潜在トピックが顕著に存在する期間のことである．

これまで，Swan と Allan [13] は，ニュース記事ストリームデータから主要な出来事を抽出するとともに，その出来事が大きな話題となっていた期間を同定する手法を提案している．さらに，Kleinberg [8] は，Swan らの研究を拡張し，主要な出来事に関する時間的階層構造の構築を試みている．しかしながら，Swan らの研究や Kleinberg の研究では，文書がある一つのキーワードやキーフレーズを含むか否かに注目して，主要な出来事を主要トピックとして抽出することを行っていた．ここに，自然言語処理技術を駆使してコーパスから固有表現や名詞句をキーワードやキーフレーズとして抽出するという，前処理が必要であったことに注意しておく．一方，我々は，文書を “bag-of-words”(BOW) 表現 [9] し，複数の単語群の出現に注目することにより，主要潜在トピックを抽出する手法を探究する．ここに，文書の BOW 表現とは，文書中に現れる単語の順序を無視し，それら単語の出現頻度のみに着目して，文書を単語頻度ベクトルという超高次元ベクトルで表現するものである．すなわち，我々のアプローチは，自然言語処理技術に大きく依存しないシンプルな統計的アプローチといえる．

さて，BOW 表現された文書群データから主要潜在トピックを抽出するには，主成分分析 (PCA) 法 [4] や潜在的意味解析 (LSA) 法 [3] を適用することが考えられる．しかしながら，これらのアプローチでは，文書の確率的生成モデルを考慮していないので，性能限界があると考えられる．例えば，BOW 表現された文書群データに PCA を適用する場合，そのデータはガウス分布によって生成されていると仮定していることになるが，BOW 表現された文書データは頻度情報であり整数値のベクトルであるので，ガウス分布モデルは文書の確率的生成モデルとしては不適切であると考えられる．さらに，これらのアプローチは，時間情報を扱う手法を提供していないので，主要潜在トピックの主要アクティブ期間同定には単純に適用できない．

本論文では，単語頻度ベクトル群の時系列（BOW 表現された文書群の時系列）として表現された文書ストリームデータにおける，主要潜在トピック抽出とその主要アクティブ期間同定を効率よく行うために，多重トピック文書の数理モデルであるパラメトリック混合モデル (PMM) [14, 15] に基づいた，そのような文書ストリームデータの潜在変数つき確率的生成モデルを導入し，PMM 型主成分分析 (PMM-PCA) 法と呼ぶ，この生成モデルに従った教師なし学習法を提案する．ここに，文書ストリームデータ（BOW 表現された文書群の時系列）と抽出する主要潜在トピック数 L が与えられたとき，PMM-PCA 法を適用することにより，任意の $\ell, (1 \leq \ell \leq L)$ に対して，第 ℓ 主要潜在トピックの主要アクティブ期間とそのトピックを代表する文書のランキング結果が出力される．特に，抽出した各主要潜在トピックのタイトルをその上位にランキングされた文書群から作成し，主要潜在トピックのタイトルと主要アクティブ期間を組にして表にまとめることを通じて，大規模文書ストリームデータの理解促進を可能にする．Web 空間内の実際の大規模な文書ストリームデータを用いた実験により，提案手法の有効性を実証する．

2. 文書ストリームの確率的生成モデル

単語頻度ベクトル群の時系列として表現された文書ストリームデータにおける主要潜在トピックを抽出しその主要アクティブ期間を同定するために，そのような文書ストリームデータの潜在変数つき確率的生成モデルを導入する．

2.1 文書ストリーム

本論文では，文書群の時系列（文書ストリーム）データ $\mathcal{D}$,

$$\mathcal{D} = \bigcup_{t=1}^{T} D(t);\ \ D(t) = \{d(t,n);\ n = 1, \cdots, N(t)\},\ (t = 1, \cdots, T)$$

を考える．ここに，T は文書ストリームの時間ステップの総数，$D(t)$ は時間ステップ t での文書群，$N(t)$ は時間ステップ t での文書の総数，$d(t,n)$ は時間ステップ t での第 n 番目の文書をそれぞれ表す．BOW 表現に従って，文書 $d(t,n)$ を単語頻度ベクトル，

$$\boldsymbol{x}(t,n) = (x_1(t,n), \cdots, x_V(t,n))$$

で表現する．ここに，各 $x_i(t,n)$ は，想定する語彙（単語）集合[*1] $\mathcal{W} = \{w_1, \cdots, w_V\}$ に対して，文書 $d(t,n)$ における語彙 w_i の出現回数を表す．ただし，V は想定する語彙の総数である．

2.2 確率的生成モデル

文書ストリームデータ $\mathcal{D}$ における主要潜在トピックの出現と消滅をモデル化するために，時系列データ $\{D(t); t = 1, \cdots, T\}$ の潜在変数つき確率的生成モデルを導入する．

任意の t に対して文書群 $D(t)$ を，それに属するすべての文書を単純にマージすることにより構成された長い 1 つの文書と同一視し，BOW 表現に従って単語頻度ベクトル

$$\boldsymbol{X}(t) = (X_1(t), \cdots, X_V(t))$$

で表現する．ここに，

$$\boldsymbol{X}(t) = \sum_{n=1}^{N(t)} \boldsymbol{x}(t,n)$$

[*1] コーパスにおける有意な単語全体の集合である．例えば，低頻度語や stop words と呼ばれる文書の内容に関与しない語は削除されている．また，英語の場合では，3 人称単数現在形や過去形などで変化した語は同一視されている．

が成り立つことに注意する．我々は，ナイーブベイズモデル[*2]を仮定することにより，$D(t)$ の生成過程を次のようにモデル化する．すなわち，特徴ベクトル $\boldsymbol{X}(t)$ は，多項分布

$$P(\boldsymbol{X}(t)) \propto \prod_{i=1}^{V} \psi_i(t)^{X_i(t)} \tag{2.1}$$

に従って確率的に生成されるとモデル化する．ここに，$P(\boldsymbol{X}(t))$ は $\boldsymbol{X}(t)$ の生起確率をあらわしている．また，各 $\psi_i(t)$ は，時間ステップ t において $D(t)$ 内に語彙 w_i が出現する確率であり，

$$\psi_i(t) \geq 0,\ (i = 1, \cdots, V);\ \sum_{i=1}^{V} \psi_i(t) = 1$$

を満たしている．本論文では，多項分布のパラメータベクトル

$$\boldsymbol{\psi}(t) = (\psi_1(t), \cdots, \psi_V(t))$$

を，時間ステップ t における“単語生起確率ベクトル”と定義する．

長い 1 つの文書と見なした $D(t)$ は，複数の主要潜在トピックをもつ文書と考えられるので，その生成過程を，ナイーブベイズモデルに準拠した多重トピック文書の確率的生成モデルである，パラメトリック混合モデル (PMM) [14, 15] に基づいてモデル化することを考える．PMM では，多重トピックをもった文書中の単語は，その多重トピックに属す各単一トピックに特徴的な単語の混合からなると仮定し，多重トピックをもつ文書の単語頻度ベクトルの分布を，基底ベクトルの混合により構成される単語生起確率ベクトルをもつ多項分布によってモデル化する．ここに，各基底ベクトルは，各単一トピックに関する多項分布の単語生起確率ベクトルに対応している．ゆえに，PMM に従って，我々は，時間ステップ t での単語生起確率ベクトル $\boldsymbol{\psi}(t)$ を，

$$\boldsymbol{\psi}(t) = \left(1 - \sum_{\ell=1}^{L} h_\ell(t)\right)\bar{\boldsymbol{\psi}} + \sum_{\ell=1}^{L} h_\ell(t)\boldsymbol{\phi}_\ell \tag{2.2}$$

とモデル化する．ここに，文書ストリーム $\mathcal{D}$ には，L 個の主要潜在トピックとともに，1 つの通常トピックが存在すると仮定している．そして，各 $\boldsymbol{\phi}_\ell = (\phi_{\ell,1}, \cdots, \phi_{\ell,V})$ は第 ℓ 主要潜在トピックの単語生起確率ベクトル[*3]であり，$\bar{\boldsymbol{\psi}} = (\bar{\psi}_1, \cdots, \bar{\psi}_V)$ は通常トピックの単語生起確率ベクトルである．また，各 ℓ に対して，第 ℓ 主要潜在トピックの主要アクティブ

[*2] ナイーブベイズモデルに基づく手法は，BOW 表現に基づいた大規模文書群データの分類等において，基本的かつ有効であることが知られている [9].

[*3] 本論文では，各主要潜在トピックの単語生起確率ベクトル $\boldsymbol{\phi}_\ell$ およびその主要アクティブ期間を，はずれ値（特殊な文書データ）の影響をできるだけ受けず安定的に推定するために，時刻 t での文書群を単純にすべてマージした $\boldsymbol{X}(t)$ を考えている．そして，このようにマージした文書は，多重トピックをもつ文書であるので PMM を仮定している．

期間を $[s_\ell, e_\ell]$ とし，

$$h_\ell(t) = \begin{cases} c_\ell, & \forall t \in [s_\ell, e_\ell], \\ 0, & \text{otherwise}, \end{cases} \tag{2.3}$$

と仮定する[*4]．ただし，各 c_ℓ は，$0 < c_\ell \leq 1$，$\sum_{\ell=1}^{L} c_\ell \leq 1$ となる定数であり，第 ℓ 主要潜在トピックの重みを表している．

3. PMM 型主成分分析

文書ストリームデータ $\mathcal{D}$ における主要潜在トピックの抽出とその主要アクティブ期間の同定を，確率的生成モデル (2.1), (2.2), (2.3) に基づいて行うことを考える．すなわち，各 ℓ に対して，第 ℓ 主要潜在トピックの主要アクティブ期間 $[s_\ell, e_\ell]$ を推定し，さらに，その潜在トピックを表す文書群 $\{d(t,n)\}$ をランキングにより抽出することを考える．

一般に，文書ストリームの観測データ $\mathcal{D}$ から，PMM に基づいたその確率的生成モデル (2.1), (2.2), (2.3) を学習することは，推定すべきパラメータ数が極めて多いので困難である．したがって，別のアプローチとして，主成分分析 (PCA) 法のような主軸（成分）抽出アプローチを考える．すなわち，大雑把に言えば，「特徴空間において文書データ点群の主軸（すなわち，分散が大きい方向）を抽出し，文書データの主軸上への射影値に基づいて，主要潜在トピックを表す文書群を抽出する．」というアプローチを考える．以下に，このような主軸抽出アプローチに従って，確率的生成モデル (2.1), (2.2), (2.3) に基づき $\mathcal{D}$ の主要潜在トピックを抽出する PMM-PCA 法を提案する．

まず，いくつかの記号法を準備する．V 次元ユークリッド空間 $\mathbf{R}^V$ 内の $(V-1)$ 次元標準単体を，

$$\Delta^{V-1} = \left\{ (y_1, \cdots, y_V) \in \mathbf{R}^V ; 0 \leq y_1, \cdots, y_V \leq 1, \sum_{i=1}^{V} y_i = 1 \right\}$$

とする．各 t, n に対して，$d(t,n)$ および $D(t)$ 内の総単語数を，それぞれ，$M(t,n)$, $M(t)$ とする．すなわち，

$$M(t,n) = \sum_{i=1}^{V} x_i(t,n), \quad M(t) = \sum_{i=1}^{V} X_i(t)$$

である．ここに，各 ℓ, t, n に対して，

$$\bar{\boldsymbol{\psi}}, \ \boldsymbol{\phi}_\ell, \ \frac{\boldsymbol{x}(t,n)}{M(t,n)}, \ \frac{\boldsymbol{X}(t)}{M(t)} \ \in \ \Delta^{V-1}$$

が成り立つことに注意する．

[*4] 同じトピックが何度か現れるということもあるが，本論文ではまず第一歩目として，主要アクティブ期間が一つという最も単純な場合に限定している．

3.1 提案法の概要

主要潜在トピックは通常トピックに比べて，ある複数の単語群がよく出現すると考えられる．すなわち，主要潜在トピックに特徴的な単語群が存在すると考えられる．我々は，特徴的な単語群がほとんど重ならないような独立なトピック群を主要潜在トピック群と考える．したがって，主要潜在トピックの単語生起確率ベクトル $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$ と，通常トピックの単語生起確率ベクトル $\bar{\boldsymbol{\psi}}$ に対して，$k \neq \ell$ ならば $\boldsymbol{\phi}_k - \bar{\boldsymbol{\psi}}$ と $\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\psi}}$ は直交すると仮定する（Fig. 1 を参照）．すなわち，

$$(\boldsymbol{\phi}_k - \bar{\boldsymbol{\psi}}) \cdot (\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\psi}}) = 0 \quad \text{if } k \neq \ell \tag{3.1}$$

と仮定する．ここに，“$\cdot$” は $\mathbf{R}^V$ における内積である．

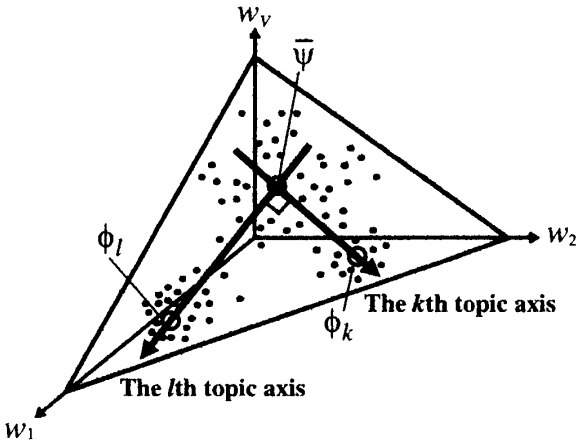

Fig. 1. A conceptual illustration of the relationship between the topic axes and the sample data $\{\boldsymbol{X}(t)/M(t); t = 1, \cdots, T\}$. The big trinagle represents the simplex Δ^{V-1}, and the filled circles indicate samples $\boldsymbol{X}(t)/M(t)$'s. $\boldsymbol{\phi}_k$, $\boldsymbol{\phi}_\ell$ and $\bar{\boldsymbol{\psi}}$ are indicated by the unfilled circles. The thick arrows indicate the topic axes.

今，文書ストリーム $\mathcal{D}$ の確率的生成モデルとして，時間ステップ t における単語生起確率ベクトル $\boldsymbol{\psi}(t)$ が t に依存しないモデルを仮定する．すなわち，$\boldsymbol{\psi}(t) = \bar{\boldsymbol{\theta}}$ というモデルを仮定する．このとき，単語生起確率ベクトル $\bar{\boldsymbol{\theta}} = (\bar{\theta}_1, \cdots, \bar{\theta}_V)$ は，式 (2.1) から最尤推定法により，

$$\bar{\theta}_i = \frac{\sum_{t=1}^{T} X_i(t)}{\sum_{t=1}^{T} M(t)} = \frac{\sum_{t,n} x_i(t,n)}{\sum_{i,t,n} x_i(t,n)}, \quad (i = 1, \cdots, V) \tag{3.2}$$

と推定される．PMM-PCA 法では，文書ストリームにおける通常トピックの単語生起確率ベクトル $\bar{\boldsymbol{\psi}}$ を，それに属する文書の時間情報を無視し，すべての文書が単一トピックのもとで生成されたと仮定したときに，最尤推定法で推定される単語生起確率ベクトル $\bar{\boldsymbol{\theta}}$ で近似する．すなわち，

$$\bar{\boldsymbol{\psi}} = \bar{\boldsymbol{\theta}}$$

と近似する．

単体 Δ^{V-1} 上で $\bar{\boldsymbol{\theta}}$ を通る直線を“軸”と呼ぶ．また，$\boldsymbol{\phi}_\ell$ を通る軸を，“第 ℓ トピック軸”と呼ぶ．Δ^{V-1} 上での $\mathcal{D}$ のサンプルデータ

$$\boldsymbol{D} = \left\{ \frac{\boldsymbol{X}(t)}{M(t)};\ t = 1, \cdots, T \right\} \tag{3.3}$$

を考える．PMM-PCA 法では，“Δ^{V-1} 上でサンプルデータ $\boldsymbol{D}$ の射影値の分散を最大にする軸（第 1 主軸）”として第 1 トピック軸を推定し，“Δ^{V-1} 上で第 1 トピック軸に直交する軸でかつ $\boldsymbol{D}$ の射影値の分散を最大にする軸（第 2 主軸）”として第 2 トピック軸を推定し，さらに，これを続けて第 L トピック軸を推定する（Fig. 1 を参照）．次に，各 ℓ に対して，$\boldsymbol{D}$ の第 ℓ トピック軸上への射影値の分布の時間変化に基づいて，その主要アクティブ期間 $[s_\ell, e_\ell]$ を推定する．さらに，その主要アクティブ期間における文書群 $\{d(t,n);\ t \in [s_\ell, e_\ell], n = 1, \cdots, N(t)\}$ を，“第 ℓ トピック度”に基づいてランキングすることにより，文書ストリームデータ $\mathcal{D}$ における第 ℓ 主要潜在トピックを表す文書群を抽出する．ただし，“単体 Δ^{V-1} 上でサンプルデータ $\boldsymbol{D}$ の射影値の分散を最大にする軸”という概念については 3.3 節で定義し，文書の“第 ℓ トピック度”については 3.5 節で定義する．また，トピック軸の推定法は 3.3 節，主要アクティブ期間の推定法は 3.4 節，主要潜在トピックを表す文書群抽出のための文書ランキング法は 3.5 節で，それぞれ詳説する．

3.2 頻度ベクトル射影値のガウス分布近似

文書ストリームの確率的生成モデルは多項分布を基本にしているが，処理の容易さの観点から，中心極限定理に基づいて，多項分布に従う量をガウス分布の言葉で表現することを考える．ただし，$\boldsymbol{X}(t)$ をガウス分布でモデル化するのではなく，任意の実 V 次元単位ベクトル $\boldsymbol{u}$ に対して，$\boldsymbol{X}(t)/M(t)$ のベクトル $\boldsymbol{u}$ 方向への射影値を，PMM を仮定し中心極限定理に基づいてガウス分布でモデル化することを考える．

まず，長い一つの文書とみなした $D(t)$ を，BOW 表現において用いる V 個の語彙以外の単語を抜きさった，単語の羅列（単語出現の順序集合）として，

$$D(t) = \left\langle w_{\lambda_1(t)}, \cdots, w_{\lambda_{M(t)}(t)} \right\rangle$$

と表現する．ここに，$\lambda_m(t)$ は $D(t)$ における初めから m 番目の単語 ID，すなわち，$w_{\lambda_m(t)}$ は $D(t)$ における初めから m 番目の単語を表している．このとき，任意の時間ステップ t と任意の実 V 次元単位ベクトル $\boldsymbol{u} = (u_1, \cdots, u_V)$ に対して，

$$A(t; \boldsymbol{u}) = \frac{1}{M(t)} \sum_{m=1}^{M(t)} u_{\lambda_m(t)}$$

なる量を考える．ここで，文書ストリーム $\mathcal{D}$ に対する我々の確率的生成モデルでは，各 $\lambda_m(t)$ は $\boldsymbol{\psi}(t)$ をパラメータベクトルとする多項分布に従って独立に生成されるので，各

$u_{\lambda_m(t)}$ も同様に，$\boldsymbol{\psi}(t)$ をパラメータベクトルとする多項分布に従って独立に生成されることになる．したがって，$\left\{u_{\lambda_m(t)}; m = 1, \cdots, M(t)\right\}$ は統計的独立かつ同一分布に従う (i.i.d.) 確率変数族となる．さて，すべての t に対して，$D(t)$ の長さ $M(t)$ は十分大きいとする．このとき，i.i.d. 確率変数族の平均である確率変数 $A(t;\boldsymbol{u})$ は，中心極限定理より，平均 $\mu(t;\boldsymbol{u})$ で分散 $\sigma(t;\boldsymbol{u})^2$ のガウス分布 $\mathcal{N}(\mu(t:\boldsymbol{u}), \sigma(t;\boldsymbol{u})^2)$ に従うと近似できる．ここに，

$$\mu(t;\boldsymbol{u}) = \boldsymbol{\psi}(t)\cdot\boldsymbol{u} = \sum_{i=1}^{V}\psi_i(t)u_i \tag{3.4}$$

$$\sigma(t;\boldsymbol{u})^2 = \frac{1}{M(t)}\left\{\sum_{i=1}^{V}\psi_i(t){u_i}^2 - \mu(t;\boldsymbol{u})^2\right\} \tag{3.5}$$

である．一方，$A(t;\boldsymbol{u})$ の定義より，

$$A(t;\boldsymbol{u}) = \frac{1}{M(t)}\boldsymbol{X}(t)\cdot\boldsymbol{u} \tag{3.6}$$

であることが容易に示される．

3.3 トピック軸の推定法

まず，各 ℓ に対して，第 ℓ トピック軸を推定することを考える．我々は，文書ストリームデータ $\mathcal{D}$ が，時間に依存しない単一トピックをもつ多項分布モデルで生成されたと仮定したときに，異常な軸をトピック軸として抽出するというアプローチをとる．すなわち，$\boldsymbol{\psi}(t) = \bar{\boldsymbol{\theta}}$ という単一トピックの多項分布モデルに基づいて，“単体 Δ^{V-1} 上でサンプルデータ $\boldsymbol{D}$ の射影値の分散が最大となる軸” という概念（式 (3.3) を参照）を定義し，トピック軸を抽出する．

任意の t に対して，$\boldsymbol{\psi}(t) = \bar{\boldsymbol{\theta}}$ と仮定する．このとき，式 (3.4), (3.5) より，任意の実 V 次元単位ベクトル $\boldsymbol{u}$ に対し $A(t;\boldsymbol{u})$ は，平均 $\mu(\boldsymbol{u})$ で分散 $\sigma(t;\boldsymbol{u})^2$ のガウス分布 $\mathcal{N}(\mu(\boldsymbol{u}), \sigma(t;\boldsymbol{u})^2)$ に従うと近似できる．ただし，

$$\mu(\boldsymbol{u}) = \bar{\boldsymbol{\theta}}\cdot\boldsymbol{u} = \sum_{i=1}^{V}\bar{\theta}_i u_i \tag{3.7}$$

$$\sigma(t;\boldsymbol{u})^2 = \frac{1}{M(t)}\left\{\sum_{i=1}^{V}\bar{\theta}_i{u_i}^2 - \mu(\boldsymbol{u})^2\right\} \tag{3.8}$$

である．さて，

$$A_0(t;\boldsymbol{u}) = A(t;\boldsymbol{u}) - \bar{\boldsymbol{\theta}}\cdot\boldsymbol{u} \tag{3.9}$$

とおくと，式 (3.6) より，$A_0(t;\boldsymbol{u})$ は，Δ^{V-1} 上でサンプルデータ $\boldsymbol{X}(t)/M(t)$ を，ベクトル $\boldsymbol{u}$ によって決定される軸へ射影した値を表す．また，式 (3.7), (3.8), (3.9) より，確率変数

$A_0(t;\boldsymbol{u})$ の確率密度関数 $p(A_0(t;\boldsymbol{u}))$ は,

$$p\left(A_0(t;\boldsymbol{u})\right) = \frac{1}{\sqrt{2\pi\bar{\sigma}(\boldsymbol{u})^2/M(t)}} \exp\left(-\frac{A_0(t;\boldsymbol{u})^2}{2\bar{\sigma}(\boldsymbol{u})^2/M(t)}\right)$$

とモデル化できる．ここに，$\bar{\sigma}(\boldsymbol{u})$ (>0) は $\boldsymbol{u}$ に依存したパラメータである．

以上より，Δ^{V-1} 上でサンプルデータ $\boldsymbol{D}$ をベクトル $\boldsymbol{u}$ によって決定される軸へ射影したとき，その射影値の分散を最大化するとは，$\bar{\sigma}(\boldsymbol{u})^2$ の推定値を最大化することと定義する．ところで，最尤推定法により，$\bar{\sigma}(\boldsymbol{u})$ は,

$$\bar{\sigma}(\boldsymbol{u})^2 = \frac{1}{T}\sum_{t=1}^{T} M(t)A_0(t;\boldsymbol{u})^2 \tag{3.10}$$

と推定される．したがって，式 (3.9), (3.10) より，トピック軸を見つける問題は，関数 $E(\boldsymbol{u})$,

$$E(\boldsymbol{u}) = \sum_{t=1}^{T} M(t)\left\{\left(\frac{1}{M(t)}\boldsymbol{X}(t) - \bar{\boldsymbol{\theta}}\right)\cdot\boldsymbol{u}\right\}^2, \tag{3.11}$$

を，$\boldsymbol{u}\cdot\boldsymbol{u}=1$ の下で最大にする $\boldsymbol{u}$ を求める問題となる．すなわち，第 1 トピック軸を決定する V 次元ベクトル $\boldsymbol{u}_1$ は，$E(\boldsymbol{u})$ を $\boldsymbol{u}\cdot\boldsymbol{u}=1$ の下で最大にする $\boldsymbol{u}$ として推定する．次に，第 2 トピック軸を決定する V 次元ベクトル $\boldsymbol{u}_2$ は，$E(\boldsymbol{u})$ を $\boldsymbol{u}\cdot\boldsymbol{u}=1$ かつ $\boldsymbol{u}\cdot\boldsymbol{u}_1=0$ の下で最大にする $\boldsymbol{u}$ として推定する．そしてこれを続けて，第 L トピック軸を決定する V 次元ベクトル $\boldsymbol{u}_L$ を推定する．ところで，Lagrange 乗数法により式 (3.11) から容易に，$\boldsymbol{u}_\ell$, $(1\le\ell\le L)$ は，次の $V\times V$ 実対称行列 $B=(b_{i,j})$ の長さ 1 の第 ℓ 固有ベクトルによって求められることがわかる．ここに,

$$b_{i,j} = \sum_{t=1}^{T} M(t)\left(\frac{X_i(t)}{M(t)} - \bar{\theta}_i\right)\left(\frac{X_j(t)}{M(t)} - \bar{\theta}_j\right),\ (i,j=1,\cdots,V)$$

である．$\{\boldsymbol{u}_\ell;\ \ell=1,\cdots,L\}$ は，例えば，パワー法により符号を除いて効率よく推定できることに注意しておく．

3.4 主要アクティブ期間の推定法

次に，推定したトピック軸に基づいて，文書ストリームデータ $\mathcal{D}$ における，各主要潜在トピックの主要アクティブ期間を推定することを考える．各 ℓ に対して，$\boldsymbol{u}_\ell$ を，3.3 節で推定した第 ℓ トピック軸を規定する V 次元単位ベクトルとする．$\mathcal{D}$ の時間ステップ t における単語生起確率ベクトル $\boldsymbol{\psi}(t)$ は，式 (2.2), (2.3) で与えられているとする．

さて，3.2 節で，各 ℓ に対して，確率変数 $A(t;\boldsymbol{u}_\ell)$ はガウス分布 $\mathcal{N}(\mu(t;\boldsymbol{u}_\ell),\sigma(t;\boldsymbol{u}_\ell)^2)$ に従うと近似できることを見た（式 (3.4), (3.5) を参照）．特に,

$$\mu(t;\boldsymbol{u}_\ell) = \boldsymbol{\psi}(t)\cdot\boldsymbol{u}_\ell \tag{3.12}$$

に注意しておく．今，$\boldsymbol{u}_\ell$ は $\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}}$ に平行であり，$\bar{\boldsymbol{\psi}} = \bar{\boldsymbol{\theta}}$ と近似しているので，式 (2.2), (3.1), (3.12) より，

$$\mu(t;\boldsymbol{u}_\ell) = \boldsymbol{\theta}\cdot\boldsymbol{u}_\ell + h_\ell(t)(\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}})\cdot\boldsymbol{u}_\ell \tag{3.13}$$

と近似できる．ゆえに，式 (3.13), (2.3) より，

$$\mu(t;\boldsymbol{u}_\ell) = \begin{cases} \bar{\boldsymbol{\theta}}\cdot\boldsymbol{u}_\ell + c_\ell(\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}})\cdot\boldsymbol{u}_\ell, & \forall t \in [s_\ell, e_\ell], \\ \bar{\boldsymbol{\theta}}\cdot\boldsymbol{u}_\ell, & \text{otherwise}, \end{cases} \tag{3.14}$$

と近似できる．すなわち，ガウス分布に従う確率変数 $A(t;\boldsymbol{u}_\ell)$ は，第 ℓ 主要潜在トピックの主要アクティブ期間とそれ例外で平均値が異なると考えられる．よって，確率変数 $A(t;\boldsymbol{u}_\ell)$ は，$t \in [s_\ell, e_\ell]$ においてはガウス分布 $\mathcal{N}(\mu_\ell, \sigma_\ell{}^2/M(t))$ に従い，$t \notin [s_\ell, e_\ell]$ においてはガウス分布 $\mathcal{N}(f_\ell, g_\ell{}^2/M(t))$ に従うという近似を行う．ただし，$\mu_\ell, \sigma_\ell, f_\ell, g_\ell, s_\ell, e_\ell$ はパラメータであり，最尤推定法により文書ストリームデータ $\mathcal{D}$ から推定する（ただし，$\sigma_\ell > 0, g_\ell > 0, 1 \le s_\ell < e_\ell \le T$ である）．ここに，ガウス分布の分散については，ロバストに推定することを考慮し，式 (3.5) に基づいて，第 ℓ 主要潜在トピックがアクティブな期間では $\sigma_\ell{}^2/M(t)$ であり，そうでない期間では $g_\ell{}^2/M(t)$ であるという近似を導入した．

まず，s_ℓ と e_ℓ を固定すると，その下での $\mu_\ell, \sigma_\ell, f_\ell, g_\ell$ の最尤推定値は，それぞれ，

$$\mu_\ell = \frac{1}{\sum_{t=s_\ell}^{e_\ell} M(t)} \sum_{t=s_\ell}^{e_\ell} M(t)A(t;\boldsymbol{u}_\ell)$$

$$\sigma_\ell{}^2 = \frac{1}{e_\ell - s_\ell + 1} \sum_{t=s_\ell}^{e_\ell} M(t)(A(t;\boldsymbol{u}_\ell) - \mu_\ell)^2$$

$$f_\ell = \frac{1}{\sum_{t=1}^{T} M(t) - \sum_{t=s_\ell}^{e_\ell} M(t)} \left\{ \sum_{t=1}^{T} M(t)A(t;\boldsymbol{u}_\ell) - \sum_{t=s_\ell}^{e_\ell} M(t)A(t;\boldsymbol{u}_\ell) \right\}$$

$$g_\ell{}^2 = \frac{1}{T - e_\ell + s_\ell - 1} \left\{ \sum_{t=1}^{T} M(t)(A(t;\boldsymbol{u}_\ell) - f_\ell)^2 - \sum_{t=s_\ell}^{e_\ell} M(t)(A(t;\boldsymbol{u}_\ell) - f_\ell)^2 \right\}$$

と計算される．したがって，s_ℓ と e_ℓ が与えられたならば尤度が計算できるので，基本的には s_ℓ と e_ℓ を全探索をすることにより，パラメータ μ_ℓ と $\sigma_\ell, f_\ell, g_\ell, s_\ell, e_\ell$ の最尤推定値が計算できる．特に，第 ℓ 主要潜在トピックの主要アクティブ期間 $[s_\ell, e_\ell]$ が推定できる．ここに，s_ℓ と e_ℓ の推定は，他のパラメータの推定とは独立であることに注意しておく．

3.5 トピック文書のランキング法

次に，各 ℓ に対し，推定したトピック軸と主要アクティブ期間に基づいて，文書ストリームデータ $\mathcal{D}$ における第 ℓ 主要潜在トピックを表す文書群の抽出を考える．我々は，文書 $d(t,n)$ の第 ℓ トピック度 $r_\ell(d(t,n))$ を定義しそれに基づいて，第 ℓ 主要潜在トピック

の主要アクティブ期間における文書群 $\{d(t,n); t \in [s_\ell, e_\ell], n = 1, \cdots, N(t)\}$ をランキングすることにより，そのような文書群を抽出する．

まず，V 次元単位ベクトル $\boldsymbol{v}_\ell$ を，

$$\boldsymbol{v}_\ell = \begin{cases} \boldsymbol{u}_\ell, & \text{if } \mu_\ell \geq f_\ell, \\ -\boldsymbol{u}_\ell, & \text{if } \mu_\ell < f_\ell, \end{cases}$$

と定義する．ただし，$\boldsymbol{u}_\ell$ は，3.3 節で推定した第 ℓ トピック軸を規定する V 次元単位ベクトルであり，μ_ℓ, f_ℓ は，それぞれ，3.4 節で推定したガウス分布 $A(t; \boldsymbol{u}_\ell)$ の主要アクティブ期間とそれ以外での平均値である．このとき，式 (3.14) より，$\boldsymbol{v}_\ell$ は単体 Δ^{V-1} 上で $\bar{\boldsymbol{\theta}}$ から $\boldsymbol{\phi}_\ell$ に向かう単位ベクトルと見なせることに注意する．実際，式 (3.14) より，$\mu_\ell \geq f_\ell$ ならば $(\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}}) \cdot \boldsymbol{u}_\ell \geq 0$ であり，$\mu_\ell < f_\ell$ ならば $(\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}}) \cdot \boldsymbol{u}_\ell < 0$ だからである．

さて，3.3 節と同様な異常検出アプローチにより，文書 $d(t,n)$ の第 ℓ トピック度 $r_\ell(d(t,n))$ を定義する．すなわち，$\mathcal{D}$ 内のすべての文書は時間に依存しない単一トピックをもつ多項分布モデルで生成されたと仮定したとき，第 ℓ トピック軸への射影値が異常な文書ほど第 ℓ 主要潜在トピックを表す文書であるというアプローチを考える．今，任意の t に対し $\boldsymbol{\psi}(t) = \bar{\boldsymbol{\theta}}$ とおき，$\mathcal{D}$ 内のすべての文書は，$\bar{\boldsymbol{\theta}}$ をパラメータベクトルとする多項分布で生成されたと仮定しよう．$\mathcal{D}$ 内の文書 $d(t,n)$ に対して，その正規化データ $\boldsymbol{x}(t,n) / M(t,n) \in \Delta^{V-1}$ を単位ベクトル $\boldsymbol{v}_\ell$ によって向き付けられた第 ℓ トピック軸へ射影した値，

$$A_1(t,n;\boldsymbol{v}_\ell) = \left\{ \frac{1}{M(t,n)} \boldsymbol{x}(t,n) - \bar{\boldsymbol{\theta}} \right\} \cdot \boldsymbol{v}_\ell \tag{3.15}$$

を考える．我々は比較的長い文書からなる文書ストリームを対象としているため，各 $M(t,n)$ はある程度大きい（すなわち，$M(t,n) \geq 200$ 程度）と仮定している．よって，3.3 節と同様な議論を行うことにより，$A_1(t,n;\boldsymbol{v}_\ell)$ はガウス分布 $\mathcal{N}(0, \bar{\sigma}_\ell^2 / M(t,n))$ に従う確率変数とモデル化でき，$\bar{\sigma}_\ell$ (> 0) は最尤推定法により，

$$\bar{\sigma}_\ell^2 = \frac{1}{\sum_{t=1}^{T} N(t)} \sum_{t=1}^{T} \sum_{n=1}^{N(t)} M(t,n) A_1(t,n;\boldsymbol{v}_\ell)^2 \tag{3.16}$$

と推定できることがわかる（式 (3.10) を参照）．そこで我々は，文書 $d(t,n)$ の第 ℓ トピック度 $r_\ell(d(t,n))$ を，

$$r_\ell(d(t,n)) = \frac{A_1(t,n;\boldsymbol{v}_\ell)}{\bar{\sigma}_\ell / \sqrt{M(t,n)}} \tag{3.17}$$

と定義する．ただし，$\bar{\sigma}_\ell$ は式 (3.16) で与えられるものである．

ここで，文書 $d(t,n)$ の第 ℓ トピック度 $r_\ell(d(t,n))$ の定義の意味について考える．まず，通常トピックを表す文書 $d(t,n)$ では，一般に $\boldsymbol{x}(t,n)/M(t,n)$ は $\bar{\boldsymbol{\theta}}$ の近くに分布すると考えられ，上記の議論より，$r_\ell(d(t,n))$ の値はガウス分布 $\mathcal{N}(0, 1)$ に従って分布すると考えられる．一方，第 ℓ 主要潜在トピックを表す文書 $d(t,n)$ では，$\boldsymbol{x}(t,n)/M(t,n)$ は $\boldsymbol{\phi}_\ell$ の近くに

分布すると考えられるので，式 (3.15), (3.17) より，$r_\ell(d(t,n))$ の値は，ガウス分布 $N(0,1)$ に従った分布に比べて正の方向に大きくシフトすると考えられる．したがって，$r_\ell(d(t,n))$ は，$\boldsymbol{x}(t,n)/M(t,n)$ が Δ^{V-1} 上で $\bar{\boldsymbol{\theta}}$ から $\boldsymbol{\phi}_\ell$ 方向へどのくらい遠ざかっているかを測定していると考えられる．また，式 (3.17) より，単体 Δ^{V-1} 上での位置 $\boldsymbol{x}(t,n)/M(t,n)$ が同じ文書 $d(t,n)$ でも，文書の単語数 $M(t,n)$ が大きいものほどトピック度が高くなることに注意する．我々は，一般に，長い文書の方が短い文書よりも確かな情報を提供すると考える．よって，$r_\ell(d(t,n))$ は，文書 $d(t,n)$ がどのくらい第 ℓ 主要潜在トピックに特徴的で重要な文書かを測定していると考えられる．

3.6 PMM-PCA と PCA の差異

主軸抽出アプローチという意味において，PMM-PCA と PCA は類似している．しかしながら，PMM-PCA ではデータ $\{\boldsymbol{X}(t); t = 1, \cdots, T\}$ が多項分布に準拠した多重トピック文書の数理モデルである PMM に従うという仮定に基づいているが，PCA ではそれがガウス分布に従うという仮定に基づいている．したがって，PMM-PCA ではトピック軸を，目的関数 $E(\boldsymbol{u})$ の $\boldsymbol{u} \cdot \boldsymbol{u} = 1$ の下での最大化問題として見つけることになるが（式 (3.11) を参照），PCA ではそれを，目的関数 $E_0(\boldsymbol{u})$ の $\boldsymbol{u} \cdot \boldsymbol{u} = 1$ の下での最大化問題として見つけることになる．ここに，

$$E_0(\boldsymbol{u}) = \sum_{t=1}^{T} \left\{ \left(\boldsymbol{X}(t) - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{X}(t) \right) \cdot \boldsymbol{u} \right\}^2 \tag{3.18}$$

である．

ところで，式 (3.2), (3.11), (3.18) より，任意の t に対して $M(t) = M$（M はある正の定数）ならば $E(\boldsymbol{u}) = E_0(\boldsymbol{u})$ が成り立つことが容易にわかる．すなわち，$D(t)$ における総単語数 $M(t)$ が t に依存しない場合，PMM-PCA はデータ点群における主軸抽出という意味では PCA と一致することになる．よって，PMM-PCA は PCA の一つの拡張と考えられる．

我々は，モデルに $M(t)$ や $M(t,n)$ を組み込むこと，すなわち，文書の長さを考慮に入れることは，重要な拡張であると考えている．実際，潜在トピックを推論する場合，一般に長い文書の方が短い文書よりも確かな情報を提供すると考えるからである．

4. 実験評価

Web 空間内の実際の英語および日本語の大規模文書ストリームデータを用いて，PMM-PCA 法の有効性を検証した．

Table 1. Top 10 documents for the first major-latent-topic extracted from the NSF dataset (Proposed method).

文書タイトル	年月
An Event-driven Programmable Network Architecture for the Next Generation Internet	October 1998
Secure Communications for Ad Hoc Networking	June 2000
CAREER: Towards an Efficient Ubiquitous Computing Infrastructure	June 2001
Collaborative Research: Design and Restoration Techniques for Fault-Tolerant Wireless Access Networks	October 2000
CAREER: Flexible, Large-Scale Best-Effort Quality of Service in the Internet	July 2002
ITR: Protocol Coordination for Multistream Applications	October 2002
ITR/SI - A Networking Protocol for Underwater Acoustic Networks	September 2001
SGER: Exploratory Research on A New Survivable, Scalable, and Self-Adapting Network Architecture Based on Biological Concepts	September 1999
ITR: The Bio-Networking Architecture: A Biologically Inspired Approach to the Design of Scalable, Adaptive, and Survivable/Available Network Applications	September 2000
CAREER: Programmable Mobile Networking	May 1999

4.1 NSF データ

まず，文書ストリームデータ $\mathcal{D}$ として，1989 年 8 月から 2003 年 9 月までの "NSF(National Science Foundation) Research Award Abstracts" データ[*5] を用いて，提案法の有効性を検証した．ただし，本実験では，その中から "Computer and Information Science and Engineering" という分野のデータのみを利用した．本データセットにおける総文書数は 10,595 であり，それらの形態素解析後の語彙総数 V は $V = 21,036$ であった．Fig. 2 に，時系列 $M(t), (t = 1, \cdots, T)$ を示す．ここに，時間ステップは 1 ヶ月とし，$T = 170$ であった．Fig. 2 より，$M(t)$ は月ごとに激しく変動することが見て取れるので，PMM-PCA を適用する必要があると考える．

Figs. 3, 4 は，それぞれ，抽出した第 1 および第 2 トピック軸への単体 Δ^{V-1} 上でのサンプルデータ $\{\boldsymbol{X}(t)/M(t); t = 1, \cdots, T\}$ の正規化射影値

$$\bar{A}_0(t; \boldsymbol{u}_\ell) = \frac{A_0(t; \boldsymbol{u}_\ell)}{\bar{\sigma}(\boldsymbol{u}_\ell)/\sqrt{M(t)}}, \quad (1 \le t \le T), \quad (\ell = 1, 2), \tag{4.1}$$

[*5] http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.data.html を参照.

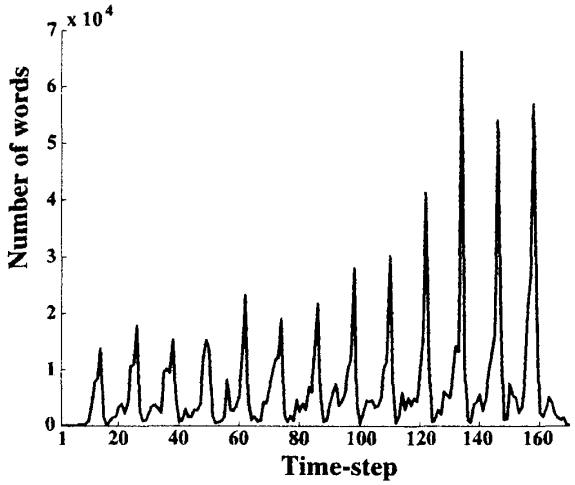

Fig. 2. Fluctuation in the number $M(t)$ of words for the NSF dataset.

を表している（3.3 節を参照）．推定した各主要潜在トピックの主要アクティブ期間は，図上で実線と丸印で記されており，それ以外の期間は実線だけで記されている．ここに，第 1 および第 2 主要潜在トピックの主要アクティブ期間は，それぞれ，$[s_1, e_1] = [110, 170]$ で 1998 年 9 月から 2003 年 9 月，$[s_2, e_2] = [39, 68]$ で 1992 年 10 月から 1995 年 3 月であった．ところで，$A(t, \boldsymbol{u}_\ell)$ は，$t \in [s_\ell, e_\ell]$ ではガウス分布 $\mathcal{N}(\mu_\ell, \sigma_\ell{}^2/M(t))$ に従い，$t \notin$

Table 2. Top 10 documents for the second major-latent-topic extracted from the NSF dataset (Proposed method).

文書タイトル	年月
High Performance Connection to the Internet	March 1998
Connections to netILLINOIS	July 1993
HPNC: HPNC for Science Research at Loyola University Chicago	October 2002
Connections to NetIllinois — Phase V	June 1994
Connections to netILLINOIS — Phase IV	June 1994
High-Performance Network Connectivity for the University of Nebraska-Lincoln	March 1998
Connections to netILLINOIS — Phase III	September 1993
Connections to NSFNET for North Carolina Institutions	September 1993
Proposal to Create an Academic / Research Network for the State of Kentucky	September 1993
Connection of School to OARnet and NSFNET — Washington State Community College	July 1995

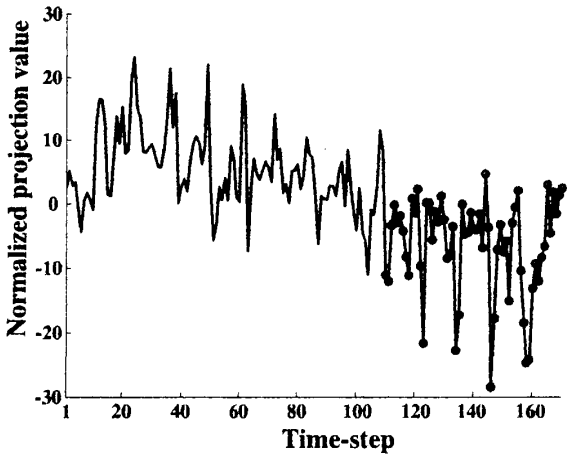

Fig. 3. Fluctuation in the normalized projection value $\bar{A}_0(t; \boldsymbol{u}_1)$ of sample data $\boldsymbol{X}(t)/M(t)$ to the first topic axis for the NSF dataset.

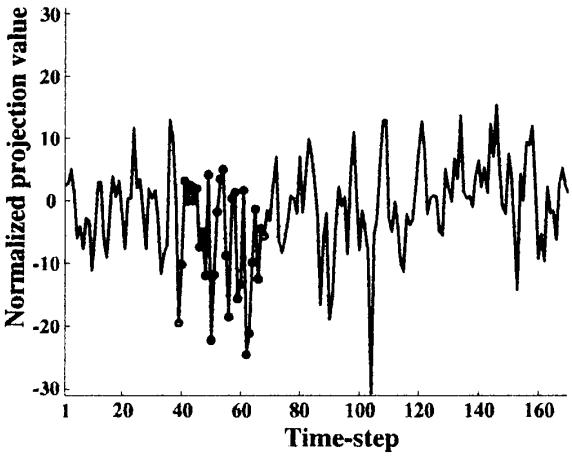

Fig. 4. Fluctuation in the normalized projection value $\bar{A}_0(t; \boldsymbol{u}_2)$ of sample data $\boldsymbol{X}(t)/M(t)$ to the second topic axis for the NSF dataset.

$[s_\ell, e_\ell]$ ではガウス分布 $\mathcal{N}(f_\ell, g_\ell{}^2/M(t))$ に従うとモデル化した（3.4 節参照）．よって，正規化射影値 $\bar{A}_0(t, \boldsymbol{u}_\ell)$ は，$t \in [s_\ell, e_\ell]$ では一つのガウス分布に従い，$t \notin [s_\ell, e_\ell]$ ではそれとは分散は同じだが平均値が異なるあるガウス分布に従うとモデル化していることになる．Figs. 3, 4 より，NSF データでは，このようなモデル化に基づいた妥当な期間が，主要アクティブ期間 $[s_\ell, e_\ell]$, $(\ell = 1, 2)$ として抽出されていることが見て取れる．

Tables 1, 2 は，それぞれ，抽出した第 1 および第 2 主要潜在トピックを表す，上位 10 位の文書タイトルとその年月を表示している．第 1 および第 2 主要潜在トピックのタイトルを，それぞれの上位にランキングされた文書群を人手で調べることにより作成した．第 1 主要潜在トピックは「ネットワークアーキテクチャーなどのネットワーキングに関連する研究」，第 2 主要潜在トピックは「高性能ネットワークへの接続などのネットワークインフラに関連する研究」というタイトルがそれぞれつけられた．よって，20 世紀後半か

ら 21 世紀初頭の "Computer and Information Science and Engineering" 分野における研究企画ということを考慮すれば，PMM-PCA 法は妥当なトピックを主要潜在トピックとして抽出していると言える．また，第 3 主要潜在トピックも，主要アクティブ期間が 1993 年 7 月から 1995 年 1 月で「並列分散計算に関連する研究」というタイトルがつけられる妥当なトピックであった．

4.2 国際ニュースデータ

次に，文書ストリームデータ $\mathcal{D}$ として，1993 年から 2002 年までの 10 年間の毎日新聞における，国際面ニュース記事データを用いて，提案法の有効性を検証した．本データセットにおける総文書数は 72,765 であり，それらの形態素解析後の語彙総数 V は $V = 72,156$ であった．Fig. 5 に，時系列 $M(t), (t = 1, \cdots, T)$ を示す．ここに，時間ステップは 1 日とし，$T = 3,577$ であった．Fig.5 より，$M(t)$ は日ごとに変動することが見て取

Table 3. Top 10 documents for the first major-latent-topic extracted from the Mainichi-Shinbun dataset (Proposed method).

文書タイトル	年月日
米国同時多発テロ　ブッシュ米大統領・議会演説（全文）	2001 年 9 月 22 日
米国同時多発テロを契機、「テロ支援国」に激変 ――対米関係、改善の動き	2001 年 9 月 30 日
イスラム諸国、「国家テロ」で米国批判 パレスチナ問題念頭――国連総会	2001 年 10 月 3 日
米国同時多発テロ　「報復戦争」準備 ――民間シンクタンクの軍事専門家２人に聞く	2001 年 9 月 17 日
［人と世界］２００１　新世代のイスラム指導者、 ハーリド・アルジェンディ師に聞く	2001 年 10 月 29 日
イスラエル連続爆破テロ　報復攻撃、過激派一掃を宣言 ――「反テロ」追い風に	2001 年 12 月 4 日
イスラエル軍事報復　米国は支持し共同歩調 ――「反テロ包囲網」に亀裂	2001 年 12 月 6 日
パレスチナの連続爆弾テロ、米仲介を妨害か ――アラファト議長、苦境に	2001 年 12 月 3 日
米国同時多発テロ　周到なテロ、憶測交錯 ――資金力誇るウサマ氏、米国内にも足場	2001 年 9 月 13 日
米国同時多発テロ　米高官「シリアも攻撃対象に」発言、 反発強める中東諸国	2001 年 10 月 18 日

Table 4. Top 10 documents for the second major-latent-topic extracted from the Mainichi-Shinbun dataset (Proposed method).

文書タイトル	年月日
［探眼複眼］中台統一の構図に異変――台湾、「独立派」の勢力台頭	1994 年 8 月 6 日
関係転換へ期待込め――台湾総統選、中国の視点	2000 年 2 月 19 日
ダライ・ラマ台湾訪問　同床異夢の「歴史的和解」	1997 年 4 月 1 日
「９７香港返還」着々と進む中国化――いよいよ、あと１年	1996 年 6 月 26 日
マカオの中国返還、「統一の基地化」警戒、１国２制度に反発――台湾	1999 年 12 月 21 日
［フロント・ライン］返還まで、あと３年　香港に広がる「中国の影」	1994 年 7 月 2 日
中台、交流機関のトップ会談　議題あいまい、非公式のまま ――あすから５年半ぶり	1998 年 10 月 13 日
中国、硬軟織り交ぜ「圧力」　有権者心理、巧みに突く――台湾総統選、きょう投開票	2000 年 3 月 18 日
［探眼複眼］決着か決裂か、中英メンツの戦い　香港民主化交渉、来月にもヤマ場	1993 年 10 月 20 日
下関条約 100 周年で揺れる台湾　独立論が絡み評価は二分	1995 年 4 月 20 日

れ，PMM-PCA を適用する必要があると考えられる．

Figs. 6, 7 は，それぞれ，抽出した第 1 および第 2 トピック軸への単体 Δ^{V-1} 上でのサンプルデータ $\{\boldsymbol{X}(t)/M(t);\ t = 1, \cdots, T\}$ の正規化射影値 $\{\bar{A}_0(t; \boldsymbol{u}_1);\ 1 \le t \le T\}$, $\{\bar{A}_0(t; \boldsymbol{u}_2);\ 1 \le t \le T\}$ を表している（式 (4.1) を参照）．推定した各主要潜在トピックの主要アクティブ期間は，図上で実線と丸印で記されており，それ以外の期間は点線で記されている．ここに，第 1 および第 2 主要潜在トピックの主要アクティブ期間は，それぞれ，$[s_1, e_1] = [3104, 3212]$ で 2001 年 9 月 12 日から 2001 年 12 月 29 日，$[s_2, e_2] = [25, 3100]$ で 1993 年 1 月 26 日から 2001 年 9 月 8 日であった．4.1 節でも述べたように，正規化射影値 $\bar{A}_0(t, \boldsymbol{u}_\ell)$ は，$t \in [s_\ell, e_\ell]$ では一つのガウス分布に従い，$t \notin [s_\ell, e_\ell]$ ではそれとは分散は同じだが平均値が異なるあるガウス分布に従うとモデル化している．Figs. 6, 7 より，毎日新聞データでは，このようなモデル化に基づいた妥当な期間が，主要アクティブ期間 $[s_\ell, e_\ell]$ $(\ell = 1, 2)$ として抽出されていることが見て取れる．

Tables 3, 4 は，それぞれ，抽出した第 1 および第 2 主要潜在トピックを表す，上位 10 位の文書タイトルとその年月日を表示している．第 1 および第 2 主要潜在トピックのタ

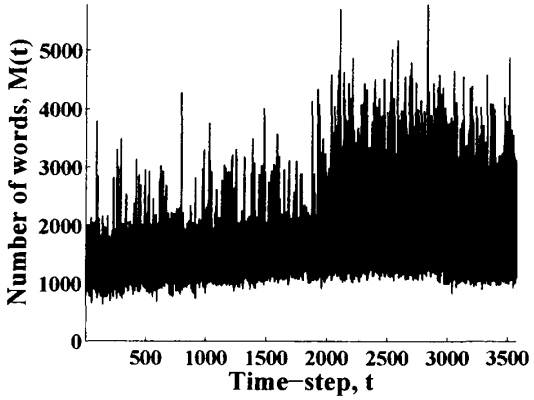

Fig. 5. Fluctuation in the number $M(t)$ of words for the Mainichi-Shinbun dataset.

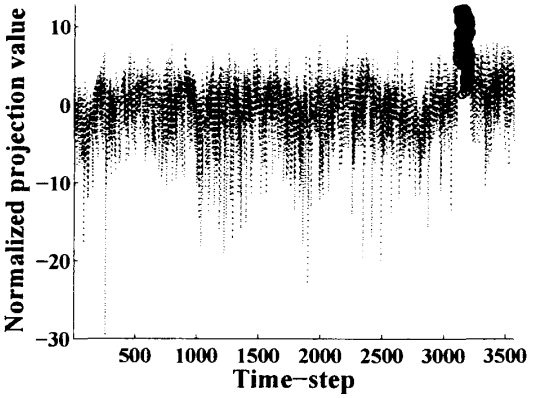

Fig. 6. Fluctuation in the normalized projection value $\bar{A}_0(t; \boldsymbol{u}_1)$ of sample data $\boldsymbol{X}(t)/M(t)$ to the first topic axis for the Mainichi-Shinbun dataset.

イトルを，それぞれの上位にランキングされた文書群を人手で調べることにより作成した．第 1 主要潜在トピックは「米国同時多発テロに関連する中東問題」，第 2 主要潜在トピックは「中国の台湾および香港問題」というタイトルがそれぞれつけられた．よって，PMM-PCA 法は，1993 年から 2002 年の国際ニュースという観点から，妥当なトピックを主要潜在トピックとして抽出していると言える．

4.3 詳細評価

本節では，毎日新聞データを用いて，PMM-PCA 法の有効性をより詳細に検証する．

大規模文書ストリームデータ（BOW 表現された文書群の時系列）と抽出する主要潜在トピック数 L が与えられたならば，PMM-PCA 法を適用することにより，任意の ℓ, $(1 \leq \ell \leq L)$ に対して，第 ℓ 主要潜在トピックの主要アクティブ期間とそのトピックを代表する文書のランキング結果を得ることができる．よって，大規模文書ストリームデータ

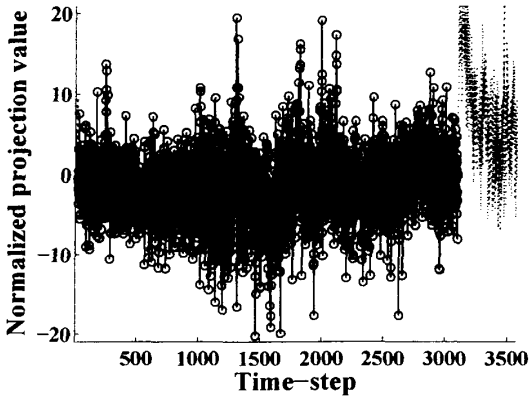

Fig. 7. Fluctuation in the normalized projection value $\bar{A}_0(t; \boldsymbol{u}_2)$ of sample data $\boldsymbol{X}(t)/M(t)$ to the second topic axis for the Mainichi-Shinbun dataset.

の理解促進の一助として，次のような提案手法のアプリケーションが考えられる．まず，抽出した L 個の主要潜在トピックのタイトルと主要アクティブ期間を組にした表を，ユーザに提示する[*6]．この表により，ユーザに，大規模文書ストリームをすべて読むことなく，どのような主要潜在トピックがどの期間に顕著であったかを大まかに把握してもらう．次に，その中からユーザが指定した主要潜在トピックに対しては，文書ランキング結果を提示（ランキングされた文書タイトルと発行日時を表示）するとともに，ユーザが望めば各文書の内容を読めるようにする．これにより，ユーザに，興味ある主要潜在トピックに対

Table 5. Five major-latent-topics extracted from the Mainichi-Shinbun dataset (Proposed method).

	トピック	期間
1	米国同時多発テロに関連する中東問題	2001 年 9 月 12 日から 2001 年 12 月 29 日
2	中国の台湾および香港問題	1993 年 1 月 26 日から 2001 年 9 月 8 日
3	小泉首相の初訪朝に関連する北朝鮮問題	2002 年 7 月 26 日から 2002 年 12 月 31 日
4	シャロン・イスラエル政権によるアラファト議長幽閉事件に関連する，イスラエル・パレスチナ紛争問題	2002 年 3 月 21 日から 2002 年 5 月 14 日
5	米国同時多発テロを契機とした米中関係の新展開（ブッシュ米大統領の中国訪問）	2001 年 9 月 12 日から 2002 年 2 月 24 日

[*6] ここでは，抽出した各主要潜在トピックのタイトルは，その上位にランキングされた文書群から人手で作成した．文書ランキング結果からそのトピックのタイトルを自動的に付与する手法に関しては，今後の課題としたい．

Table 6. Top 10 documents for the third major-latent-topic extracted from the Mainichi-Shinbun dataset (Proposed method).

文書タイトル	年月日
北朝鮮・核開発計画、放棄を要求　食い違う関係国の思惑	2002 年 10 月 28 日
韓国与党、金正日総書記の訪韓を切望	2002 年 8 月 31 日
――大統領選の逆転を狙う	
北朝鮮、何が変わったか――日米韓との対話に積極姿勢	2002 年 8 月 17 日
小泉首相訪朝　首脳会談、各国が注視	2002 年 9 月 16 日
露朝首脳会談　金正日総書記、ロシアと信頼関係強調	2002 年 8 月 24 日
――日米韓と対話を模索	
ロシア大統領、外交力をアピール　北朝鮮と首脳会談	2002 年 8 月 19 日
――極東の経済発展も期待	
韓国、「核」を議題に　北朝鮮の真意探る	2002 年 10 月 19 日
――きょう南北閣僚級会談	
韓国次期大統領に盧武鉉氏　「太陽政策継承」に思惑交錯	2002 年 12 月 21 日
北朝鮮、対話路線を印象付け　「主導権得た」の見方も	2002 年 8 月 1 日
――米朝外相会談実現	
北朝鮮が核兵器開発認める　体制維持へ苦渋の選択	2002 年 10 月 17 日
――米政権の態度厳しく	

しては，上位にランキングされた文書群を実際に読むことでより詳しくトピックの内容を理解してもらうようにする．

Table 5 に，$L = 5$ の場合に構築した「主要潜在トピックのタイトルと主要アクティブ期間を組にした表」を示す．また，Tables 6, 7 および 8 に，それぞれ，第 3, 第 4 および第 5 主要潜在トピックを表す上位 10 位の文書タイトルとその年月日を示す．日本では，米国，中国，北朝鮮および中東に関わる問題は最も関心が高い国際問題の一部であるが，Table 5 に提示された主要潜在トピックは，「米国同時多発テロ（2001 年 9 月 11 日）」，「香港の中国への返還（1997 年 7 月 1 日）とそれに関連した中国の台湾との統一問題」，「小泉首相初訪朝（2002 年 9 月 17 日）」，「アラファト議長幽閉（2002 年 3 月 29 日）」および「ブッシュ大統領による 30 年ぶりの米大統領の訪中（2002 年 2 月 21 日）」という，1993 年から 2002 年までの期間の国際ニュースにおいて特筆する出来事や問題に関連したものである（新聞に掲載されるのは，通常 1 日遅れることに注意）．すなわち，トピックによっては主要アクティブ期間は長いものもあれば短いものもあるが，妥当なトピックが主要潜在トピックとして抽出されていると考えられる．よって，提案法は有効であると考えられる．

次に，ベースラインシステムとして，PMM-PCA 法における主軸抽出とトピック文書

Table 7. Top 10 documents for the fourth major-latent-topic extracted from the Mainichi-Shinbun dataset (Proposed method).

文書タイトル	年月日
［憎悪の連鎖］パレスチナ衝突　仲介成果は不透明 ――米国務長官、あす中東へ出発	2002 年 4 月 7 日
屈辱、幽閉アラファト氏　イスラエル、面目つぶす狙い ――トイレ行くにも許可	2002 年 3 月 30 日
［憎悪の連鎖］パレスチナ衝突 「長官、まずイスラエルに行くべきだった」	2002 年 4 月 10 日
イスラエル・ゼエビ観光相殺害事件　過激派に実刑判決 ――パレスチナ軍事裁判	2002 年 4 月 26 日
ヒズボラ、パレスチナと連携強化　レバノン南部では英雄視	2002 年 4 月 2 日
自治区侵攻 1 カ月（その 1）　硬軟両面の外交展開 ――イスラエル	2002 年 4 月 29 日
シャロン・イスラエル首相に決断迫る　軍事作戦に歯止めか ――米大統領、撤退要求	2002 年 4 月 5 日
［憎悪の連鎖］パレスチナ衝突　イスラエル、作戦を継続 米国の要求に応じず	2002 年 4 月 6 日
［流血の行方］混迷のイスラエル社会／下　衝突防げぬ指導者	2002 年 3 月 23 日
パレスチナ衝突　米仲介、入り口で難航　停戦条件かみあわず 和平交渉再開にらむが	2002 年 4 月 15 日

ランキングを通常の PCA 法で置き換えた手法を調べた．すなわち，PCA 法による主要潜在トピック抽出では，一つの主軸につき正方向と負方向の二つの主要潜在トピックが抽出されるが，主要アクティブ期間抽出を組み込みことにより，一つの主軸につき一つの主要潜在トピックのみを抽出する手法を調べた．Table 9 に，本手法により抽出された第 1 主要潜在トピックを表す上位 10 位の文書タイトルとその年月日を示す．ここに，抽出された第 1 主要潜在トピックの主要アクティブ期間は 1998 年 10 月 1 日から 2001 年 12 月 28 日であった．Table 9 からもわかるように，本手法により上位にランキングされた文書群を実際に読んでみても，国際ニュースにおけるどのようなトピックが主要潜在トピックなのかを理解するのは極めて困難であった．すなわち，文書ストリームデータ（BOW 表現された文書群の時系列）からの主要潜在トピック抽出においては，通常の PCA 法に基づいた手法よりも PMM-PCA 法がより有効と考えられる．このように，通常の PCA 法の結果が PMM-PCA 法の結果と著しく異なったのは，第 t 日目の文書群の総単語数 $M(t)$ が t に関して大きく変動していたこと (Fig. 5 参照) に原因があると考えられる (3.6 節を参照).

Table 8. Top 10 documents for the fifth major-latent-topic extracted from the Mainichi-Shinbun dataset (Proposed method).

文書タイトル	年月日
ブッシュ大統領発言「米国が防衛」、台湾政権は高く評価	2002 年 2 月 23 日
――「対中国」に支援不可欠	
ニクソン訪中から３０年　米中「記念日」に協調演出	2002 年 2 月 1 日
――首脳会談、対立避ける	
米中首脳会談（要旨）	2001 年 10 月 20 日
中国駐在の台湾人１８１人、金門島経由で初の帰郷	2002 年 2 月 10 日
米中首脳会談　対立点を棚上げ――協調態勢、維持へ	2001 年 10 月 19 日
［東論西談］「テロ絶対悪」論　中国の主張、米保守派と酷似	2001 年 10 月 29 日
台湾立法委員選　民進党主導、連立枠組み	2001 年 12 月 2 日
――国民党は分裂危機、政界再編の可能性も	
中国、台湾に「一つの中国」迫る	2001 年 12 月 6 日
公開書簡で民主活動家・徐文立氏の釈放求める――中国	2002 年 2 月 20 日
米中首脳会談　共同会見（要旨）	2002 年 2 月 22 日

そして，通常の PCA 法で主要潜在トピック抽出がうまくいかなかったのは，BOW 表現した文書群データ $\{\boldsymbol{X}(t); t = 1, \cdots, T\}$ をガウス分布に従うと近似したことが不適切であったためと考えられる．

5. 関連研究

本論文のように BOW 表現に基づき時間を考慮するトピックモデルに関しても，いくつかの研究がなされている．Blei と Lafferty [2] は，Latent Dirichlet Allocation (LDA) という多重トピック文書生成モデルに基づいた，マルコフ型の離散時間のダイナミックトピックモデルを提案している．また，Wang と McCallum [16] は，LDA に基づいた非マルコフ型の連続時間のダイナミックトピックモデルを提案している．前者は，トピックの意味（単語連想）が時間的に変化するという考え方に基づいたものであるが，後者は本論文と同様，トピックの意味が時間的に変化しないという考え方に基づいている．LDA は，各文書に対して各単語ごとにトピックが変化するという考え方に基づいた，多重トピック文書生成モデルである．一方 PMM は，各文書ごとに一組の多重トピックが割り当てられるという考え方に基づいたモデルである．本論文では，PMM に基づいて，文書ストリームデータにおける主要潜在トピック抽出とその主要アクティブ期間同定を行う手法を提案した．上記の研究のような LDA に基づいた手法との定性的および定量的な比較評価は，今後の重要な研究課題の一つである．

Table 9. Top 10 documents for the first major-latent-topic extracted from the Mainichi-Shinbun dataset (PCA method).

文書タイトル	年月日
米国同時多発テロ　報復、２段階か　拠点攻撃後、長期態勢	2001 年 9 月 17 日
――国際手続きは後回し？	
クリントン米大統領、訪朝断念　韓国に不安と懸念	2000 年 12 月 30 日
――米国に協調堅持要請へ	
［人と世界］２００１　在日米軍と基地	2001 年 4 月 30 日
――ハワイの国際関係研究者スミスさんに聞く	
「ロシア首相解任」繰り返す　失政の火の粉避け、保身	1999 年 8 月 10 日
――エリツィン大統領	
戦略核の削減で条約　米露首脳が調印へ	2002 年 5 月 14 日
――備蓄問題、なお流動的	
米露首脳会談、１３日から　なるか戦略核弾頭の大幅削減	2001 年 11 月 10 日
――ＡＢＭ条約の改廃がカギ	
北朝鮮・核開発計画、放棄を要求　食い違う関係国の思惑	2002 年 10 月 28 日
成果見えぬ「4 者会談」（米・中・南北朝鮮）	1999 年 8 月 12 日
思惑絡み、つばぜり合い	
どうみる米中首脳会談――識者に聞く	1998 年 6 月 24 日
米中首脳会談　対立点を棚上げ――協調態勢、維持へ	2001 年 10 月 19 日

一方，社会ネットワーク分析において，主要潜在トピックの時間変動を，友人関係の時間変動や人と単語の共起関係の時間変動などから説明するダイナミックモデルが研究されている．毎年開催される国際会議論文集のアーカイブは，そのような研究で対象とされる文書ストリームの代表例である．Sarkar と Moore [11] は，このような国際会議論文アーカイブにおいて時間的に変動する論文共著関係を説明する，マルコフ型の離散時間のダイナミックモデル（状態空間モデル）を提案し，それに基づいて著者群を低次元空間上に可視化している．さらに，Sarkar ら [12] は，このような国際会議論文アーカイブにおいて時間変動する著者と単語の論文における共起関係を説明する，マルコフ型の離散時間のダイナミックモデル（状態空間モデル）を提案し，それに基づいて著者群と単語群を同時に低次元空間上に可視化している．このような文書ストリームに関しては，論文共著関係や著者とその論文の単語情報の関係を利用することで，主要潜在トピック抽出の性能向上が期待される．提案法に対して共著情報や著者と単語の共起情報を組み込むことは，今後の重要な研究課題の一つである．

しかしながら，実データを用いた実験より，提案した PMM-PCA 法は，大規模な文書ス

トリームデータにおける主要潜在トピックを抽出しその主要アクティブ期間を同定する，有効でシンプルな手法の一つと考えられる．我々は，大規模文書ストリームからの効率的な主要潜在トピック抽出という目標に向けて，大きく前進できたと考えている．

6. おわりに

文書ストリームデータ $\mathcal{D}$ における主要潜在トピック抽出とその主要アクティブ期間同定を，文書の BOW 表現に基づいて効率よく行うために，PMM-PCA 法と呼ぶ新たな教師なし学習法を提案した．PMM-PCA 法は，V 次元単語頻度ベクトル群の時系列データとして表現された文書ストリームデータに対して，その適切な確率的生成モデルに基づいているという性質を有している．PMM-PCA 法では，まず，Δ^{V-1} 上でサンプルデータ $\boldsymbol{D} = \{\boldsymbol{X}(t)/M(t)\}$ の射影値の分散を最大にする軸として第 1 トピック軸を推定し，Δ^{V-1} 上で第 1 トピック軸に直交する軸でかつ $\boldsymbol{D}$ の射影値の分散を最大にする軸として第 2 トピック軸を推定し，これを続けて一般に第 ℓ トピック軸を推定する．次に，サンプルデータ $\boldsymbol{D}$ の第 ℓ トピック軸上への射影値の分布の時間変化に基づいて，第 ℓ 主要潜在トピックの主要アクティブ期間 $[s_\ell, e_\ell]$ を推定する．さらに，期間 $[s_\ell, e_\ell]$ における文書群を第 ℓ トピック度に基づいてランキングすることにより，第 ℓ 主要潜在トピックを表す文書群を抽出する．ただし，“単体 Δ^{V-1} 上でのサンプルデータ $\boldsymbol{D}$ の軸上への射影値の分散” という概念や文書の “第 ℓ トピック度” に関しては，中心極限定理に基づいて $\mathcal{D}$ の確率的生成モデルをガウス分布近似することを通じて定義した．

我々は，約 5 年間の NSF Research Awards Abstract データおよび 10 年間の毎日新聞国際面記事データを用いた実験で，PMM-PCA 法の有効性を実証した．また，PMM-PCA 法は，データ点群における主軸抽出という観点において，頻度データに適用するための PCA 法の一つの拡張と考えられることを示した．

謝辞 本研究は，科学研究費補助金基盤研究 (C)(No. 20500147) の補助を受けた．

参考文献

[1] Baldi, P., Fransconi, P., and Smyth, P., *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, Wiley, Chichester, 2003.

[2] Blei, D. M. and Lafferty, J. D., Dynamic topic models, *Proceedings of the 23rd International Conference on Machine Learning* (2006), 113–120.

[3] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R, Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, **41** (1990), 391–407.

[4] Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern Classification*, Wiley, New York, 2000.

[5] Kimura, M., Saito, K., and Ueda, N., Modeling of growing networks with directional attachment and communities, *Neural Networks*, **17** (2004), 975–988.

[6] Kimura, M., Saito, K., and Ueda, N., Modeling share dynamics by extracting competition structure, *Physica D*, **198** (2004), 51–73.

[7] Kleinberg, J. and Lawrence, S., The structure of the web, *Science*, **294** (2001), 1849–1850.

[8] Kleinberg, J., Bursty and hierarchical structure in streams, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002), 91–101.

[9] Manning, C. D. and Schütze, H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, 1999.

[10] Pal, S. K., Talwar, V., and Mitra, P., Web mining in soft computing framework: Relevance, state of the art and future directions, *IEEE Transactions on Neural Networks*, **13** (2002), 1163–1177.

[11] Sarkar, P. and Moore, A. W., Dynamic social network analysis using latent space models, *Advances in Neural Information Processing Systems*, **18** (2006), 1145–1152.

[12] Sarkar, P., Siddiqi, S. M., and Gordon, G. J., A latent space approach to dynamic embedding of co-occurrence data, *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics* (2007).

[13] Swan, R. and Allan, J., Automatic generation of overview timelines, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2000), 49–56.

[14] Ueda, N. and Saito, K., Single-shot detection of multiple topics using parametric mixture models, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002), 626–631.

[15] Ueda, N. and Saito, K., Parametric mixture models for mult-labeled text, *Advances in Neural Information Processing Systems*, **15** (2003), 737–744.

[16] Wang, X. and McCallum, A., Topics over time: A non-Markov continuous-time model of topical trends, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006), 424–433.

木村 昌弘 (正会員) 〒520-2194 滋賀県大津市瀬田大江町横谷 1-5

1989 年大阪大学大学院理学研究科数学専攻修士課程修了．博士（理学）．NTT コミュニケーション科学基礎研究所を経て，現在，龍谷大学理工学部電子情報学科准教授．知能情報学の研究に従事．日本数学会，人工知能学会，電子情報通信学会各会員．

斉藤 和巳 (非会員) 〒422-8526 静岡県静岡市駿河区谷田 52-1

1985 年慶応義塾大学理工学部数理科学科卒業．同年 NTT 入社．2007 年より，静岡県立大学経営情報学部教授．工学博士．知能情報学, 発見科学, ネットワークの科学等の研究に従事．情報処理学会, 電子情報通信学会, 人工知能学会, 日本神経回路学会各会員．

（2007年11月16日受付）

（2008年 8 月 5 日最終稿受付）