社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ICS－136　(21)
2004／8／5

Webのトピックダイナミックスモデル

木村 昌弘　斉藤 和巳　上田 修功
NTT コミュニケーション科学基礎研究所

概要: Web上で日々流通する大量の情報に基づいて，米国同時多発テロのようなある種の例外的な社会現象を検出する問題を考察する．本稿では特に，あるカテゴリーに属する文書群の時系列データに基づいて，その中でアウトブレークしたトピック（ホットトピック）と，その存在期間を抽出する手法を提案する．社会ニュースの実データおよび，国際ニュースの実データを用いた実験により，提案法の有効性を検証する．

Extracting hot topics from the Web

Masahiro KIMURA　Kazumi SAITO　Naonori UEDA
NTT Communication Science Laboratories

Abstract: We consider the problem of detecting anomalous social phenomena from the Web. In this report, we propose the method of extracting the hot topics from time-series data of documents that belong to some category in the Web. Moreover, we experimentally examine the effectiveness of our method by using real data of news articles.

1　はじめに

近年，World Wide Webをはじめ様々な複雑ネットワークに対して，その構造やダイナミックスに対する関心が高まっている[12, 6]．それに応じて，社会ネットワーク上での病気の蔓延についての数理研究[9]や，blog空間におけるトピック伝搬モデルの研究[4]等，情報の伝搬（流通）に関する研究も注目されるようになってきた．ところで，Webは，新たなコミュニケーションメディアとして発展し続けており，今や人間社会の縮図とも考えられうる．したがって，Web上で日々流通（伝搬）する大量の情報（文書情報，リンク情報，ユーザ情報等）に基づいて，ある種の異常な社会現象を抽出することや説明すること，さらにはその予兆を発見することは，極めて興味深い研究課題と考えられる．本稿では，このような研究目標への第一歩として，ある種の異常（例外的）現象と考えられる，トピックのアウトブレーク現象の抽出について探求する．すなわち，あるカテゴリーに属する文書群の時系列データに基づいて，その中でアウトブレークしたトピック（**ホットトピック**）と，その存在期間を抽出するという問題を考察する．

ところで，トピック検出と追跡(TDT)の研究[1, 13]において，「ニュース記事の時系列から，新トピックの出現を検出したり，既知トピックについての記事を追跡する．」という問題が提起され，そのための手法が研究されている．また，TDT研究に引き続いてSwanとAllan[10]は，ニュース記事の時系列データ集合における，主要なトピックとその存在期間を抽出するための統計的手法を開発している．さらに，Klienberg[7]は，隠れマルコフモデルを用いた手法により，Swanらの手法において一つのトピックの存在期間が短く切れるという問題の解決を試みるとともに，トピックの階層構造を構築することも試みている．

さて，Swanらの手法においては，まず前処理として自然言語処理による特徴語抽出処理が必要であり，主要トピックはそれら特徴語の組み合わせで表現されている．また，Kleinbergの手法においても，主要トピックは，タイトルに現れる単語のようななんらかの特徴語で表現されている．しかしながら，我々の研究では，ホットトピックの抽出にとどまらず，その予兆発見という予測問題への発展をも念頭においているので，抽出するホットトピックは，これら既存研究のような単なる主要な出来事というより，もっと広義の概念であるべきだと考えられる．そこで本稿では，文書表現としてBag-of-Words(BOW)表現という超高次元データ表現に基づき，潜在トピック（ナイーブベイズ）モデルを導入し，文書の大まかな内容や概念に基づいた微妙なトピックを，ホットトピックとして抽出する手法を探求する．

ところで，そのような広義トピック抽出法としては，主成分分析(PCA)法[3]や潜在的意味解析

(LSA) 法 [2] がよく知られている．しかしながら，これらの手法は，文書データの（確率）モデルを考慮していない．これに対して Hofmann[5] は，文書データの確率モデルに基づく，確率的潜在意味解析 (PLSA) 法を提案し，その LSA 法等に対する有効性を実証している．しかしながら，PLSA 法は，超大規模データに対しては不安定となり，効率的でないという問題がある．よって本稿では，文書データの確率モデルに基づく，ホットトピックの生成消滅現象のモデルを提案するとともに，効率的なホットトピック抽出法を提案する．そして，Web 上で配信されている最近の約 2ヶ月間の社会ニュースの実データ，および，10 年間の国際ニュースの実データを用いた実験により，提案法の有効性を検証する．

2 ホットトピックのモデル

あるカテゴリーにおけるホットトピックの生成と消滅の現象を，このカテゴリーに属する文書群時系列の確率的生成モデルとして捕らえることを考える．

2.1 文書群時系列の確率的生成モデル

このカテゴリーの第 t 日目の文書群データを，

$$D(t) = \{d(t,n);\ n = 1, \cdots, N(t)\} \tag{1}$$

とする．ここに，$N(t)$ は第 t 日目の文書の総数である．我々は，文書データの表現方法として，文書中にどのような単語がどのような頻度で出現するかという情報のみに着目した，Bag-of-Words(BOW) 表現[1]を用いる．すなわち，各文書 $d(t,n)$ を，単語頻度ベクトル

$$\boldsymbol{x}(t,n) = (x_1(t,n), \cdots, x_V(t,n)) \tag{2}$$

で表現する．ここに，V はコーパス全体における異なる単語の総数[2]（語彙総数）であり，$x_i(t,n)$ は文書 $d(t,n)$ 中で第 i 単語（語彙）が出現した回数である．

まず，式 (1) で与えられる第 t 日目の文書群データ $D(t)$ を，その日の全文書を単純にマージすることにより構成された大きな文書と考え，BOW 表現を用いて，単語頻度ベクトル

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

で表現する．このとき，式 (2) より明らかに，

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

である．次に，BOW 表現に基づくナイーブベイズ (naive Bayes) モデル[3]を仮定して，このカテゴリー—における第 t 日目の文書群データ $D(t)$ の生成確率を，語彙に関する多項分布

$$P(D(t)) \propto \prod_{i=1}^{V} \{\psi_i(t)\}^{X_i(t)} \tag{3}$$

によりモデル化する．ここに，$\psi_i(t)$ は，このカテゴリ—の第 t 日目の文書群で第 i 単語（語彙）が生起する確率であり，

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

を満している．このとき，

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

は，このカテゴリーの第 t 日目の文書群における**単語生成確率ベクトル**と呼ばれる．

以上より，単語生成確率ベクトル $\boldsymbol{\psi}(t)$ のダイナミックスをモデル化することにより，このカテゴリーの文書群時系列の確率的生成モデルが構築されることになる．

2.2 ホットトピックと単語生成確率ベクトル

まず，このカテゴリーにおける一般的な文書群というものが存在すると仮定し，ナイーブベイズモデルに基づいて，そのような文書群の単語生成確率ベクトルを，

$$\bar{\boldsymbol{\psi}} = (\bar{\psi}_1, \cdots, \bar{\psi}_V)$$

とする．例えば，第 t 日目がホットトピックなど存在しない通常日ならば，$\boldsymbol{\psi}(t) = \bar{\boldsymbol{\psi}}$ なる単語生成確率ベクトルにしたがって，その日の文書群データが生成されると考えるのである．

我々は，ホットトピックとしては，我々が興味ある期間全体で起るようなものを考えるのではなく，（全期間に比べてそれほど長くない）ある期間においてのみ起こり，それ以外の期間では起らないようなものを考える．さて，我々が興味がある期間には，L 個のホットトピックが起ったとしよう．そし

[1]文書の BOW 表現は，大規模文書の分類では，より複雑な文書表現法よりも分類性能が良いと報告されており，現在の文書分類において共通に用いられる表現法となっている [8].

[2]正確には，有意な単語の総数と言うべきである．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 人称単数形や過去形などで変化した語は同一視されている．

[3]ナイーブベイズモデルは，BOW 表現に基づいた大規模文書データの分類等において，ロバストな手法として広く用いられている [3].

て，第 ℓ ホットトピックが起ったとき，ナイーブベイズモデルに基づき，それに関連する文書群の単語生成確率ベクトルを，

$$\boldsymbol{\phi}_\ell = (\phi_{\ell,1}, \cdots, \phi_{\ell,V}),\ (\ell = 1, \cdots, L)$$

とする．すなわち，第 ℓ ホットトピックが起った日には，それに関する文書群は，単語生成確率ベクトル $\boldsymbol{\phi}_\ell$ にしたがって生成されると考える．また，第 ℓ ホットトピックは，期間 $[T_{\ell,0}, T_{\ell,1}]$ においてのみ起こり，その他の期間では起らないとする．

以上に基づいて，我々は，このカテゴリーにおけるホットトピックの生成と消滅の現象を，このカテゴリーに属する文書群時系列の確率的生成モデルにより，次のようにモデル化する．すなわち，このカテゴリーの第 t 日目の文書群における単語生成確率ベクトル $\boldsymbol{\psi}(t)$ は，一般的文書群の単語生成確率ベクトル $\bar{\boldsymbol{\psi}}$ と，ホットトピックに関連する文書群の単語生成確率ベクトル $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$ の（パラメトリック）混合 [11]

$$\boldsymbol{\psi}(t) = \left(1 - \sum_{\ell=1}^{L} h_\ell(t)\right)\bar{\boldsymbol{\psi}} + \sum_{\ell=1}^{L} h_\ell(t)\boldsymbol{\phi}_\ell \qquad (4)$$

であるとモデル化する．ここに，

$$h_\ell(t) = \begin{cases} c_\ell, & t \in [T_{\ell,0}, T_{\ell,1}], \\ 0, & \text{otherwise}, \end{cases}$$

であり，c_ℓ は $0 \le c_\ell \le 1$ で $\sum_{\ell=1}^{L} c_\ell \le 1$ なる定数である．

3 ホットトピックの抽出法

このカテゴリーに属する文書群の T 日間のデータ

$$\mathcal{D}_T = \{D(1), \cdots, D(T)\}$$

が与えられたとき，そこでのホットトピックとその存在期間を抽出したい．ここに，第 t 日目の文書群 $D(t)$ は式 (1) で与えられているとする．

このとき，式 (3),(4) で与えられている，このカテゴリーに属する文書群時系列の確率的生成モデルを，訓練データ $\mathcal{D}_T$ に基づいて学習することにより，そのパラメータ $(\bar{\boldsymbol{\psi}},\ \{\boldsymbol{\phi}_\ell,\ [T_{\ell,0},\ T_{\ell,1}];\ \ell = 1, \cdots, L\})$ を決定することができれば，我々の問題は解決できる．しかしながら，我々の実験においては，T は数十や数千で語彙数 V は数万にもなるので，訓練データからそれらのパラメータを決定するのは，極めて困難と考えられる．したがって，我々は，それとは異なるアプローチ（射影法）を取ることにする．

3.1 射影法の考え方

第 t 日目の文書群 $D(t)$ は，その日の全文書を単純にマージすることにより構成された大きな文書と考える．そして，BOW 表現において用いる V 個の語彙以外の単語は抜きさり，$D(t)$ を単語 ID の羅列として表現しよう．このとき，$D(t)$ の第 m 番目にでてくる単語 ID を $w_m(t)$ とする．すなわち，文書 $D(t)$ を，

$$D(t) = \langle w_1(t), \cdots, w_{M(t)}(t)\rangle$$

として表現しよう．ここに，$M(t)$ は $D(t)$ の総単語数であり，明らかに，

$$M(t) = \sum_{i=1}^{V} X_i(t)$$

となる．このとき，$\{1, \cdots, V\}$ 上の確率過程 $m \mapsto w_m(t)$ は，確率分布 $\boldsymbol{\psi}(t)$ に従う，独立同分布な確率過程である．よって，任意の V 次元ベクトル

$$\boldsymbol{u} = (u_1, \cdots, u_V)$$

に対して，実軸上の確率過程 $m \mapsto u_{w_m(t)}$ もまた，確率分布 $\boldsymbol{\psi}(t)$ に従う，独立同分布な確率過程となる．

さて，実軸上の確率変数

$$A(t) = \frac{1}{M(t)} \sum_{m=1}^{M(t)} u_{w_m(t)}$$

を考えよう．今 $M(t)$ は十分大きいので，中心極限定理より，$A(t)$ は，平均が，

$$\mu(t) = \sum_{i=1}^{V} \psi_i(t) u_i \qquad (5)$$

で，分散が

$$\sigma(t)^2 = \frac{1}{M(t)} \left\{ \sum_{i=1}^{V} \psi_i(t) {u_i}^2 - \mu(t)^2 \right\} \qquad (6)$$

であるガウス分布により近似できる．したがって，式 (4) に注意すれば，ホットトピックの存在に応じて，$A(t)$ のガウス分布として振る舞いが変わることがわかる．特に，ホットトピックが 1 つしかない場合，すなわち $L = 1$ の場合を考えると，ホットトピックがある期間とない期間で，$A(t)$ はガウス分布として異なるので，$A(t)$ の時系列を観察することにより，ホットトピック期間が抽出できると考えられる．

ところで，

$$A(t) = \boldsymbol{\theta}(t) \cdot \boldsymbol{u}$$

であることに注意しよう．ここに，

$$\boldsymbol{\theta}(t) = (\theta_1(t), \cdots, \theta_V(t)) = \frac{1}{M(t)}\boldsymbol{X}(t)$$

であり，“·” はベクトルの内積を表している．すなわち，$\boldsymbol{u}$ が大きさ 1 のベクトル ($\|\boldsymbol{u}\| = 1$) ならば，$A(t)$ は第 t 日目の規格化された文書群サンプル $\boldsymbol{\theta}(t)$ の $\boldsymbol{u}$ 軸方向への射影に他ならない．我々は，射影軸 $\boldsymbol{u}$ をうまく選ぶことによって，$A(t)$ の時系列を観察することから，様々なホットトピックを抽出することを考える．

3.2 ホットトピック軸

ホットトピックに関連する文書は，このカテゴリーにおける通常の文書とは大きく違うと考えられるので，単語生成確率ベクトル $\boldsymbol{\phi}_\ell$ と $\bar{\boldsymbol{\psi}}$ は大きく異なると考えられる．また，通常の文書を中心に考えたとき，異なるホットトピックの文書どうしは独立であると思われるので，異なるホットトピック ℓ_1 と ℓ_2 においては，$\boldsymbol{\phi}_{\ell_1} - \bar{\boldsymbol{\psi}}$ と $\boldsymbol{\phi}_{\ell_2} - \bar{\boldsymbol{\psi}}$ がほぼ直交していると考える．すなわち，

$$(\boldsymbol{\phi}_{\ell_1} - \bar{\boldsymbol{\psi}}) \cdot (\boldsymbol{\phi}_{\ell_2} - \bar{\boldsymbol{\psi}}) = 0, \text{ 但し}, \ell_1 \neq \ell_2. \quad (7)$$

このカテゴリーにおける通常文書群の単語生成確率ベクトル $\bar{\boldsymbol{\psi}}$ は，ナイーブベイズモデルを仮定したときの観測データ $\mathcal{D}_T$ に基づく最尤推定量 $\bar{\boldsymbol{\theta}} = (\bar{\theta}_1, \cdots, \bar{\theta}_V)$,

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

により近似的に求められる．また，第 t 日目の規格化された文書群データ $\boldsymbol{\theta}(t)$ を，第 t 日目の真の単語生成確率ベクトル $\boldsymbol{\psi}(t)$（すなわち，式 (3) の多項分布の平均値）で近似する．このとき，式 (4) より，$\boldsymbol{\theta}(t)$ は，

$$\boldsymbol{\theta}(t) = \bar{\boldsymbol{\theta}} + \sum_{\ell=1}^{L} h_\ell(t)(\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}}) \quad (8)$$

と近似される．

さて，$\boldsymbol{u}_\ell$ を $\|\boldsymbol{u}_\ell\| = 1$ で $\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}}$ に平行な V 次元ベクトルとし，

$$A_\ell(t) = \boldsymbol{\theta}(t) \cdot \boldsymbol{u}_\ell$$

とする．このとき，式 (7),(8) より，$A_\ell(t)$ は，

$$A_\ell(t) = \left\{(1 - h_\ell(t))\bar{\boldsymbol{\theta}} + h_\ell(t)\boldsymbol{\phi}_\ell\right\} \cdot \boldsymbol{u}_\ell$$

で近似される．よって，式 (5),(6) より，「確率変数 $A_\ell(t)$ は，第 ℓ ホットトピックが存在する期間 $[T_{\ell,0}, T_{\ell,1}]$ においては，平均が μ_ℓ で分散が ${\sigma_\ell}^2/M(t)$ のガウス分布にしたがって生成され，第 ℓ ホットトピックが存在しない期間においては，平均が f_ℓ で分散が ${g_\ell}^2/M(t)$ のガウス分布にしたがって生成される．」と近似される．すなわち，

$$A_\ell(t) \sim \begin{cases} \mathcal{N}\left(\mu_\ell; {\sigma_\ell}^2/M(t)\right), & t \in [T_{\ell,0}, T_{\ell,1}], \\ \mathcal{N}\left(f_\ell; {g_\ell}^2/M(t)\right), & \text{otherwise}, \end{cases} \quad (9)$$

と考えられる．したがって，時系列 $A_\ell(t)$ を調べることにより，第 ℓ ホットトピックの存在期間 $[T_{\ell,0}, T_{\ell,1}]$ を推定することが可能になる．

3.3 ホットトピック期間の推定

規格化された文書群データの集合 $\{\boldsymbol{\theta}(t); t = 1, \cdots, T\}$ は，$V-1$ 次元標準単体 Δ^{V-1} 内に $\bar{\boldsymbol{\theta}}$ を中心に分布するが，ホットトピックの存在によりその分布の分散は変化していると考えられる．すなわち，データ点は，通常は $\bar{\boldsymbol{\theta}}$ の周りに分布するが，第 ℓ ホットトピックが存在すると $\boldsymbol{\phi}_\ell$ の方向に大きく引っ張られるので，データ集合の分散が大きくなると考えられる．したがってデータ集合の分散の大きい方向を調べることで，ホットトピック軸が抽出できる可能性がある．

我々は，$\bar{\boldsymbol{\theta}}$ を中心としてデータ集合 $\{\boldsymbol{\theta}(t);\ t = 1, \cdots, T\}$ の分布を考えるとき，その分散が最大となる軸に第 1 ホットトピックがあり，それに直交する軸で次にその分散が最大となる軸に第 2 ホットトピックがある，というようにホットトピック軸が存在すると考える．ここで，各データ点 $\boldsymbol{\theta}(t)$ は，規格化されていることに注意しよう．すなわち，本来観測されるデータは，$\boldsymbol{\theta}(t)$ ではなく $\boldsymbol{X}(t)$ であり，それには総単語数 $M(t)$ が非常に多い場合（$D(t)$ が多くの文書，または長い文書を含む場合）もあれば，それほどでもない場合も存在する．当然，$M(t)$ が大きい日のデータほど信頼すべきである．したがって，$\|\boldsymbol{u}\| = 1$ のもとで，

$$E(\boldsymbol{u}) = \sum_{t=1}^{T} M(t)\left\{(\boldsymbol{\theta}(t) - \bar{\boldsymbol{\theta}}) \cdot \boldsymbol{u}\right\}^2 \quad (10)$$

なる目的関数の最大化問題を考えることにより，ホットトピック軸を抽出することを考える．これは，$V \times V$ 行列 $B = (b_{ij})$ の固有値問題として求められる．ここに，

$$b_{ij} = \sum_{t=1}^{T} M(t)(\theta_i(t) - \bar{\theta}_i)(\theta_j(t) - \bar{\theta}_j)$$

である．すなわち，B の第 ℓ 固有値の固有ベクトル $\boldsymbol{u}_\ell$ が，第 ℓ ホットトピック軸である．

上記のようにして第 ℓ ホットトピック軸 $\boldsymbol{u}_\ell$ が求められたならば，$A_\ell(t)$ を計算し，式 (9) に基づいてホットトピック ℓ の存在期間 $[T_{\ell,0},\ T_{\ell,1}]$，および，$A_\ell(t)$ を生成するガウス分布のパラメータを推定する．ただし，推定の安定化のために，我々は $g_\ell = \sigma_\ell$ と仮定する．さて，$T_{\ell,0}$ と $T_{\ell,1}$ の値を指定すると，時系列データ $\{A_\ell(t);\ \ell = 1, \cdots, T\}$ に基づいた最尤推定法により，式 (9) における $\mu_\ell,\ \sigma_\ell,\ f_\ell,\ g_\ell$ の推定値が一意に求まり，さらに尤度も計算できる．したがって，尤度を最大化する $[T_{\ell,0},\ T_{\ell,1}]$ を求めることにより，第 ℓ ホットトピックの存在期間が推定できる．

3.4 ホットトピック文書の抽出

次に，存在する期間が推定された各ホットトピック ℓ に対して，それがどのようなトピックであるかを調べたい．そのために，各文書 $d(t,n)$ の第 ℓ ホットトピック度 $r_\ell(t,n)$ を定義し，その値に基づき文書をランキングすることを考える．

$\boldsymbol{u}_\ell = (u_{\ell,1},\cdots,u_{\ell,V})$ を推定された第 ℓ ホットトピック軸とし，$\mu_\ell,\ f_\ell$ をそれぞれ，ホットトピック ℓ が存在する期間と存在しない期間における式 (9) のガウス分布の平均値の推定値とする．このとき，式 (4),(5) より，

$$\begin{aligned} \mu_\ell &= \left\{\bar{\boldsymbol{\theta}} + c_\ell(\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}})\right\} \cdot \boldsymbol{u}_\ell \\ f_\ell &= \bar{\boldsymbol{\theta}} \cdot \boldsymbol{u}_\ell \end{aligned}$$

が成り立つと考えられる．したがって，$\mu_\ell > f_\ell$ ならば，$(\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}}) \cdot \boldsymbol{u}_\ell > 0$ なので，$\bar{\boldsymbol{\theta}}$ から $\boldsymbol{u}_\ell$ の正の方向にホットトピック $\boldsymbol{\phi}_\ell$ があることになる．一方，$\mu_\ell < f_\ell$ ならば，$(\boldsymbol{\phi}_\ell - \bar{\boldsymbol{\theta}}) \cdot \boldsymbol{u}_\ell < 0$ なので，$\bar{\boldsymbol{\theta}}$ から $\boldsymbol{u}_\ell$ の負の方向にホットトピック $\boldsymbol{\phi}_\ell$ があることになる．以上を踏まえた上で式 (10) に基づき，文書 $d(t,n)$ の第 ℓ ホットトピック度 $r_\ell(t,n)$ を，

$$r_\ell(t,n) = \varepsilon\sqrt{M(t,n)}\left\{\left(\frac{1}{M(t,n)}\boldsymbol{x}(t,n) - \bar{\boldsymbol{\theta}}\right)\cdot\boldsymbol{u}_\ell\right\}$$

で定義する．ここに，$M(t,n)$ は文書 $d(t,n)$ の総有意単語数，すなわち，

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

であり，

$$\varepsilon = \begin{cases} 1, & \text{if } \mu_\ell > f_\ell, \\ -1, & \text{if } \mu_\ell < f_\ell, \end{cases}$$

である．

さて，文書 $d(t,n)$ の第 ℓ ホットトピック度 $r_\ell(t,n)$ の統計的意味を考えよう．文書群 $D(t)$ に対する第 ℓ ホットトピック軸への射影 $A_\ell(t)$ と同様に，文書 $d(t,n)$ に対する射影 $A_\ell(t,n)$ を考える．すなわち，

$$A_\ell(t,n) = \frac{1}{M(t,n)}\boldsymbol{x}(t,n)\cdot\boldsymbol{u}_\ell$$

を考える．このとき，$A_\ell(t,n)$ に対しても $A_\ell(t)$ と同様の結果，すなわち，式 (9) と類似した結果が成立していると考えられる．$f_\ell(t,n)$ と $g_\ell(t,n)^2$ をそれぞれ，第 ℓ ホットトピックが存在しない期間におけるガウス分布の平均と分散としよう．式 (5),(6) より，$f_\ell(t,n) = f_\ell$ であり，

$$g_\ell(t,n)^2 = \frac{\sum_{i=1}^{V}\bar{\theta}_i {u_{\ell,i}}^2 - {f_\ell}^2}{M(t,n)}$$

である．ここで，確率変数 $\{A_\ell(t,n) - f_\ell\}\ /\ g_\ell(t,n)$ を考えると，第 ℓ ホットトピックが存在しない期間においてそれは，平均 0 で分散 1 のガウス分布に従うことに注意しておく．また，

$$r_\ell(t,n) = \varepsilon\frac{A_\ell(t,n) - f_\ell}{g_\ell(t,n)}\sqrt{\sum_{i=1}^{V}\bar{\theta}_i {u_{\ell,i}}^2 - {f_\ell}^2}$$

が成り立つことが容易に確かめられる．ここで，右辺第 3 因子は，文書 $d(t,n)$ に依存しない定数であることに注意．したがって，文書 $d(t,n)$ の第 ℓ ホットトピック度 $r_\ell(t,n)$ は，統計的には，通常文書の生成に対応するガウス分布（第 ℓ ホットトピックがない場合に対応するガウス分布）を基準にして，文書 $d(t,n)$ の異常度（分散）を測定したと考えられる．

4 実験評価

4.1 社会ニュースデータ

まず，2004 年 3 月 31 日から 2004 年 6 月 15 日までの期間において収集した，“asahi.com” の社会カテゴリーの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．

図 1，2，3，4 は，それぞれ，$\ell = 1,\ 2,\ 3,\ 4$ に対する時系列 $A_\ell(t)$ を表示している．ここに，時間ステップは 1 日であり，総記事数は 2,421，総単語数は 12,640 であった．これらの図より，時系列 $A_\ell(t)$ を観察することから，ホットトピック期間が抽出できそうなことが，視覚的に見て取れる．提案法により抽出したホットトピック期間は図上では実線と丸印で記され，それ以外の期間は点線で記されて

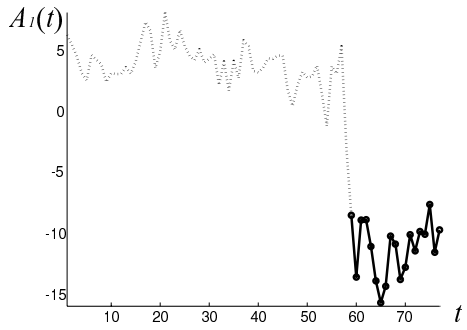

図 1: 社会第 1 ホットトピック (期間: 0528 - 0615)

1：女児の精神鑑定決める　佐世保小 6 事件で第 1 回審判 (0614)
2：次第に孤立深めた女児、ＨＰで心情記す　佐世保事件 (0608)
3：女児、付添人面会に初の涙　週内に精神鑑定を申請 (0609)
4：小 6 女児、同級生女児に切られ死亡　長崎・佐世保 (0601)

表 1: 社会第 1 ホットトピック

いる．したがって，これらの図より，提案法は妥当な期間を抽出していると考えられる．

表 1, 2, 3, 4 は，それぞれ，$\ell = 1, 2, 3, 4$ に対して，第 ℓ ホットトピックがどのようなトピックであるかを，ホットトピック ℓ 度に関し，その存在期間で上位にランクされた文書のタイトルにより表示している．これらの表より大雑把に言えば，ホットトピック 1 は「佐世保の女児殺人事件を中心とするトピック」，ホットトピック 2 は「北朝鮮拉致被害者の家族の帰国問題を中心とするトピック」，ホットトピック 3 は「人身事故関連のトピック」，ホットトピック 4 は「イラクにおける日本人人質事件を中心とするトピック」と考えられる．

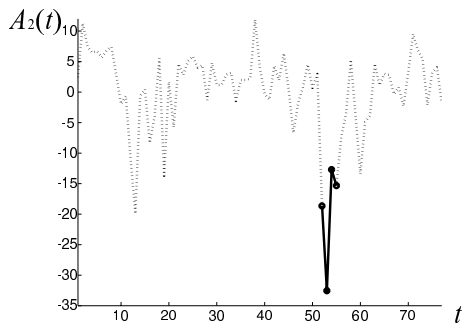

図 2: 社会第 2 ホットトピック (期間: 0521 - 0524)

1：地村さん「複雑な気持ち」　曽我さんへの気遣い満ちる (0522)
2：「再調査」１０人の家族、首相に不満・落胆 (0522)
3：「首相を信じて待つ」　拉致被害者 5 人が都内で会見 (0521)
4：「4 人で一緒に一晩寝たい」　家族を待つ拉致被害者 5 人 (0522)

表 2: 社会第 2 ホットトピック

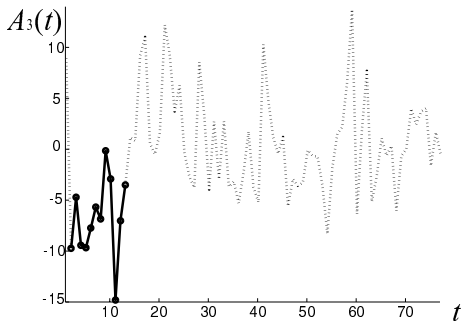

図 3: 社会第 3 ホットトピック (期間: 0401 - 0412)

4.2　国際ニュースデータ

次に，1993 年から 2002 年までの 10 年間の毎日新聞における，国際面の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．

図 5，6，7，8 は，それぞれ，$\ell = 1, 2, 3, 4$ に対する時系列 $A_\ell(t)$ を表示している．ここに，時間ステップは 1 日であり，総記事数は 76,765，総単語数は 72,156 であった．提案法により抽出したホットトピック期間は，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．また，表 5, 6, 7, 8 は，それぞれ，$\ell = 1, 2, 3, 4$ に対して，第 ℓ ホットトピックがどのようなトピックであるかを，第 ℓ ホットトピック度に関し，その存在期間で上位にランクされた文書のタイトルにより表示している．これらの表より大雑把に言えば，ホットトピック 1 は「米国同時多発テロを中心とするトピック」，ホットトピック 2 は「中国の台湾・香港問題関連のトピック」，ホットトピック 3 は「小泉首相の初訪朝を中心とする北朝鮮問題関連のトピッ

1：ヒルズ回転ドア、事故ごとの速報押収　社内回覧状況調査 (0405)
2：同じ公園遊具で 2 児童が指切断　大阪、午前と午後 (0403)
3：事故と同型の回転遊具、和歌山県の小学校でもボルト異状 (0405)
4：東海道線が人身事故で上下線で一時運転見合わせ (0407)

表 3: 社会第 3 ホットトピック

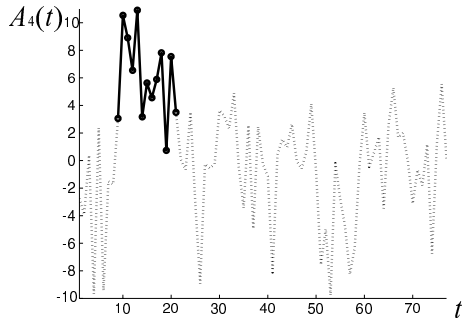

図 4: 社会第 4 ホットトピック (期間: 0408 - 0420)

1：イラク市民、人質事件の賛否二分　「仕方ない」の声も (0415)
2：イラク人質の３人と家族がドバイ出発　夕には関空に到着 (0418)
3：元東大生ら２人にも実刑判決　スーパーフリー事件 (0409)
4：「本人たちの会見は義務」と高遠さんの母 (0419)

表 4: 社会第 4 ホットトピック

ク」，ホットトピック 4 は「シャロン・イスラエル政権よるアラファト議長の幽閉を中心とする，イスラエル・パレスチナ紛争関連のトピック」と考えられる．以上より，提案法は，妥当なホットトピックを抽出していると考えられる．

5　おわりに

ホットトピックの生成消滅現象の確率モデルを提案し，あるカテゴリーに属する文書群の時系列データから，ホットトピックとその存在期間を抽出する，効率的手法を提案した．また，抽出されたホットトピックがどのようなトピックかを知るために，ホットトピック度に関する文書のランキング法も

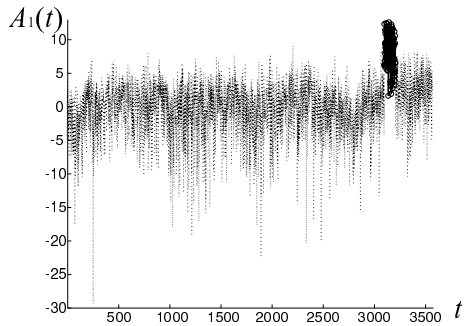

図　5:　国際第 1 ホットトピック (期間:　010912 - 011229)

1：米国同時多発テロ　ブッシュ米大統領・議会演説（全文） (010922)
2：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 (010930)
3：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 (011003)
4：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 (010917)

表 5: 国際第 1 ホットトピック

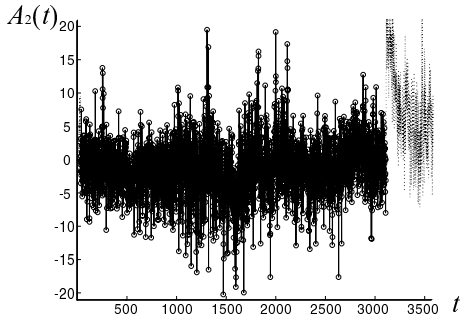

図　6:　国際第 2 ホットトピック (期間:　930126 - 010908)

提案した．そして，Web 上で配信されている社会ニュース記事の約 2ヶ月間の実データ，および，10 年間の国際面の新聞記事の実データを用いた実験により，提案法の有効性を実証した．

ところで，どのようなホットトピックが抽出されたかについては，極めて上位にランクされている記事だけで判断したが，当然，もっと下位の記事まで詳細に調べる必要がある．これについては，今後検討していきたい．また，従来法との詳細比較は，今後の重要な検討課題である．さらに，ハイパーリンク情報の積極的利用についても，今後検討していく予定である．

1：［探眼複眼］中台統一の構図に異変 – 台湾、「独立派」の勢力台頭 (940806)
2：関係転換へ期待込め – 台湾総統選、中国の視点 (000219)
3：ダライ・ラマ台湾訪問　同床異夢の「歴史的和解」(970401)
4：「９７香港返還」着々と進む中国化 – いよいよ、あと１年 (960626)

表 6: 国際第 2 ホットトピック

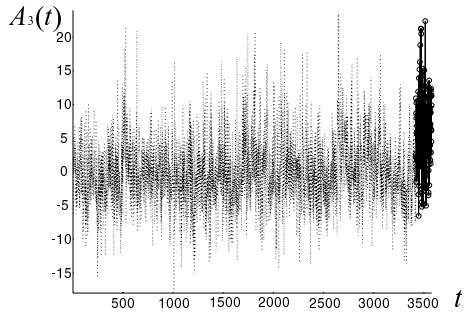

図 7: 国際第 3 ホットトピック (期間: 020726 - 021231)

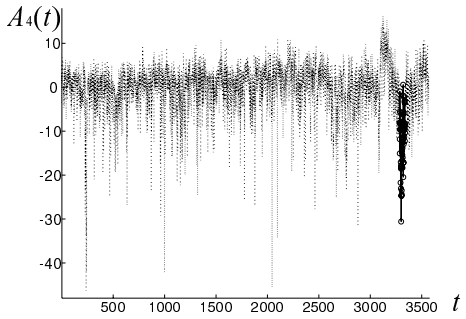

図 8: 国際第 4 ホットトピック (期間: 020321 - 020514)

1 : 北朝鮮・核開発計画、放棄を要求
食い違う関係国の思惑 (021028)
2 : 韓国与党、金正日総書記の訪韓を切望
– 大統領選の逆転を狙う (020831)
3 : 北朝鮮、何が変わったか – 日米韓との対話に積極姿勢 (020817)
4 : 小泉首相訪朝　首脳会談、各国が注視 (020916)

表 7: 国際第 3 ホットトピック

1 : ［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 (020407)
2 : 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 (020330)
3 : ［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」 (020410)
4 : 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 (020426)

表 8: 国際第 4 ホットトピック

謝辞：データの収集から評価実験までを行って頂いた，NTT コムウェア株式会社の 飯野斉 氏 に感謝致します.

参考文献

[1] Allan, J., Papka, R., & Lavrenko, V. (1998). Online new event detection and tracking, In *Proc. SIGIR'98* (pp. 37-45).

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

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

[4] Gruhl, D., Guha, R, Liben-Nowell, D, & Tomkins, A. (2004). Information diffusion through blogspace, In *Proc. WWW'04* (pp. 491-501).

[5] Hofmann, T. (1999). Probabilistic latent semantic indexing, In *Proc. SIGIR'99* (pp. 50-57).

[6] Kimura, M., Saito, K., & Ueda, N. (in press). Modeling of growing networks with directional attachment and communities, *Neural Networks*.

[7] Kleinberg, J. (2002). Bursty and hierarchical structure in streams, In *Proc. SIGKDD'02* (pp. 91-101).

[8] Manning, C. D. & Schtze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

[9] Newman, M. E. J. (2002). Spread of epidemic disease on networks, *Physical Review E*, *66*, 016128.

[10] Swan, R. & Allan, J. (2000). Automatic generation of overview timelines, In *Proc. SIGIR'00* (pp. 49-56).

[11] Ueda, N. & Saito, K. (2002). Single-shot detection of multiple topics using parametric mixture models, In *Proc. SIGKDD'02* (pp. 626-631).

[12] Storgatz, S. H. (2001). Exploring complex networks, *Nature*, *410*, 268-276.

[13] Yang, Y., Pierce, T., & Carbonell, J. (1998). A study on retrospective and on-line event detection, In *Proc. SIGIR'98* (pp. 28-36).