大規模 Web クリックデータのためのイベント予測

松原　靖子[†a)]　櫻井　保志[†]　Christos Faloutsos[††]　岩田　具治[†††]
吉川　正俊[††††]

Fast Mining and Forecasting of Web-Click Events

Yasuko MATSUBARA[†a)], Yasushi SAKURAI[†], Christos FALOUTSOS[††],
Tomoharu IWATA[†††], and Masatoshi YOSHIKAWA[††††]

あらまし　本論文では大規模イベントデータのためのパターン検出手法である *TriMine* について述べる．具体的には Web クリックを対象とし，{*URL, userID, timestamp*} の三つ組で構成されるイベントシーケンスに対し潜在的なトレンドを発見すると同時に，将来のイベント予測を行う．実データを用いた実験では *TriMine* が Web クリックの中から有用なトレンドを発見し，長期的な将来予測を高精度に行うことを確認した．更に既存手法との比較を行い提案手法が大幅な性能向上を達成していることを明らかにした．

キーワード　複合イベントデータ，テンソル解析，トピックモデル，時系列予測

1. まえがき

多くの Web アプリケーションにおいて，時系列ログデータは高速かつ大量に生成され続けている．例えば，Web ホスティングサービスでは，ユーザと URL の情報を伴う何百万ものアクセスログが毎時刻生成される．このような大規模な生成ログ，すなわちビッグデータを効率的かつ効果的に分析することは重要な課題となっている．

本研究では，主に Web クリックデータを対象とし，イベント情報のトレンド検出と将来予測を高精度かつ高速に行うことを目的とする．Web クリックデータは，{*URL, user ID, timestamp, access devices, http/document referrer*} のような複数の属性から構成される．このようなログデータを本論文では**複合イベント**と定義する．それぞれのイベントはタイムスタンプと複数の属性で構成され，例えば Web クリックイベントの例では，URL を**オブジェクト** *(object)*，user ID を**アクター** *(actor)* と呼ぶ．複合イベントは様々なドメインにおいて生成されており，例えばウェブサイトにおけるアクセス履歴 [1] やソーシャルネットワークサービス，位置情報に基づくサービス [14] は代表的な例である．

本論文で扱う問題は以下のとおりである．

問題： 三つ組 $(object, actor, time)$ で構成されるイベントシーケンス群が与えられたとき，(a) 潜在的なトピックとトレンドを発見し，(b) 将来のイベントを高速に予測する．

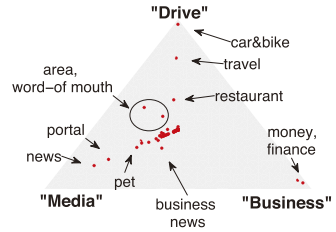
本論文では，大量に発生する複合イベント集合から主要パターンを発見する手法である *TriMine* [12](注1) について述べる．*TriMine* は Web クリックデータを $(object, actor, time)$ のそれぞれの角度から捉え，共通する潜在的トピックを発見する．なお，提案手法は上述の三つ組以外にも任意の個数の属性値をもつイベントを扱うことができるが，論述の簡略化の為に本論文では主に三つ組のイベントのみについて言及する．

1.1 具体例

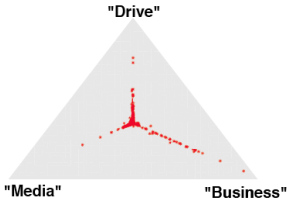
一般に，各 Web サイトには一つ以上の潜在的なト

† 熊本大学，熊本市
Kumamoto University, Kumamoto-shi, 860–8555 Japan
†† カーネギーメロン大学，米国
Carnegie Mellon University, USA
††† 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，京都府
NTT Communication Science Laboratories, NTT Corporation, Kyoto-fu, 619–0237 Japan
†††† 京都大学，京都市
Kyoto University, Kyoto-shi, 606–8501 Japan
a) E-mail: yasuko@cs.kumamoto-u.ac.jp

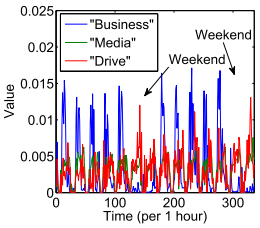
(注1)：ソースコード：http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html



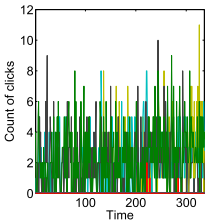
(a) トピック空間における URL の分布 (オブジェクト行列 **O**)



(b) トピック空間におけるユーザの分布 (アクター行列 **A**)



(c) トピックの時間推移 (時間行列 **C**)



(d) オリジナルのシーケンス (100 名のランダムユーザ)

図 1　Web クリックデータ (URL, user ID, time) における *TriMine* のパターン発見
Fig. 1　Results for web-click data (URL, userID, timestamp), with three hidden topics.

ピックが存在している．同様に，各ユーザも幾つかのトピックと関連性がある．例えば経済ニュースに関するサイトと株価に関するサイトは，それぞれ共通のユーザが利用する．更に，これらのユーザは同じような時間帯（平日の日中）にアクセスが偏る傾向にある．この場合，これらの Web サイト及びユーザは business トピックをもつ集合としてグループ化することができる．*TriMine* は，このような潜在的なトピックを自動的に発見する．

図 1 (a)-(c) は Web クリックデータにおける *TriMine* のトピック発見の様子である．本論文では図 1 (a)-(c) を *TriMine*-plot と呼ぶ．*TriMine*-plot は二つの散布図 (a), (b) と一つの時系列シーケンス (c) から構成される．これらの三つの図は，それぞれの潜在的トピックに対し $(object, actor, time)$ の三つの要素がどのように分布しているかを示す．図 1 は最も頻出する三つのトピックについて可視化した．この例では，分布の傾向からそれぞれのトピックに対し business, media, drive とラベルを付けている．詳細は以下のとおりである．

(a) トピック空間における URL の分布 (オブジェクト行列 **O**): 図は，三次元トピック空間上における各 URL の分布状況を表現している．ここで，各点はそれぞれの URL を示し，URL が頂点に近いほど，その頂点のトピックの特徴を強くもっていることを示す．例えば car&bike サイトは drive トピックに関連性が高く，money サイト，finance サイトは business トピックをもっている．

(b) トピック空間におけるユーザの分布 (アクター行列 **A**): 各点は個人ユーザを示す．点が頂点に近いほど，そのトピックに関連性の高いユーザであることを意味する．

(c) 潜在的トピックの時間推移 (時間行列 **C**): x 軸は時間，y 軸はその時間帯におけるトピックの強さを示す．この例では期間は 2 週間であり，ウィンドウサイズは 1 時間毎とする．全てのトピックにおいて 1 日単位の周期性が見られる．加えて，business トピック (青線) は平日に頻出し，週末に現れにくい．一方 drive トピック (赤線) は週末に高い値をもつ．

図 1 (d) はオリジナルのシーケンス例である．ここでは，100 人の任意のユーザを選び money サイトにおけるクリックの数を可視化している．*TriMine*-plot と異なり，オリジナルデータはノイズが多く，周期性やユーザのクラスタも発見できず，明確な特徴を全く把握できない．

以上のように，*TriMine* は $(object, actor, time)$ の三つの要素に対し，非常に少ない情報量で，明確な特徴を捉えることができる．これにより既存手法では困

表 1　既存手法との比較
Table 1　Capabilities of approaches.

	DWT	HOSVD /ALS	LDA	*AR* /*PLiF*	*TriMine* /*TriMine-F*
多重時間スケール	√				√
テンソル解析		√			√
離散データ			√		√
短期予測				√	√
長期予測					√

難とされる複合イベントの将来予測問題を解決することができる．

1.2　関連研究と本研究の位置づけ

表 1 は既存研究と *TriMine* の能力の比較である．

- ウェーブレット変換 (DWT: discrete wavelet transform) は単一のシーケンスにおいて多重時間スケールのトレンドを発見することができる．しかし，複数のオブジェクト，アクターから構成されるイベントデータの中から共通パターンを発見できない．
- 複合イベントはテンソルとして扱うことができる．高次特異値分解 (HOSVD: higher-order singular value decomposition) [10] と交互最小 2 乗法 (ALS: alternating least squares) [21] は，テンソル中の潜在的なコンポーネントを発見することができる．一方，イベント集合のような非ガウス性をもつカウントデータを扱うことができない．
- トピックモデル (LDA: latent Dirichlet allocation) [4] はスパースなカウントデータ集合を扱う確率モデルである．トピックモデルは潜在的なトピックを発見することでクラスタリングを行うことができるが，周期的な時系列パターンを発見することはできず，将来のデータ予測を行うこともできない．
- 自己回帰モデル (AR: autoregressive model) や *PLiF* [11] に代表される時系列モデルは，シーケンスの予測をする能力をもつ．しかし，多重スケールのトレンドを扱えず，従って長期的な時系列予測に向いていない．

1.3　本論文の貢献

提案手法は以下のような特長がある．

- *TriMine* は大規模複合イベントを効率的かつ効果的に要約し，$(objects, actors, time)$ の 3 要素に対しパターンを発見する．これによりクラスタリング，外れ値検出，そして予測問題を解決する．
- 時系列イベントシーケンスの予測を高い精度で行うことができ，計算コストは入力データの長さに対して線形である．

2.　関 連 研 究

混合分布モデルの学習については，機械学習などの分野において様々な研究がすすめられている．潜在的ディリクレ配分法 (LDA: latent Dirichlet allocation) [4] と確率的潜在意味解析 (PLSA: probabilistic latent semantic analysis) [5] は，テキストデータのための bag-of-words や画像データのための bag-of-features など，離散データ集合を分析するための潜在変数モデルとして幅広い分野で用いられている．時刻付き文書における時間発展の分析については，DTM (dynamic topic model) [3] や TOT (topics over time) [22]，その他様々な手法が提案されている [2], [6]〜[8], [23], [25]．例えば，DTM は固定長のウィンドウサイズを用いて時刻ごとの潜在的トピックを抽出する．同様に，TOT は潜在トピックの時間的推移をベータ分布に基づき表現するモデルである．Hong らは文書中に出現する各単語の総数を推定し，潜在トレンドを検出，追跡するための新たなトピックモデルを提案している．しかしこれらのモデルは多重スケールの周期トレンドに着目していないため，複雑な時系列トレンドをもつ複合イベントの長期予測を行うことは困難である．提案手法は **1.2** で述べたとおり複数の時間スケール上における周期的なパターンを発見し，将来の長期的な予測を効果的かつ効率的に行うことができる．

Web クリック分析については，Agarwal らは情報推薦を行うために，ガンマ-ポアソンモデルを用いてクリックスルーレート (click-through rate) の推定を行っている [1]．しかし，この取り組みは時刻付きイベントのトレンド発見に焦点を合わせたものではない．

テンソル分析も本研究と関連している．Kolda らは Web リンク構造を解析するためのテンソル解析手法を提案している [9]．Rendle らはテンソル分解に基づいたタグ推薦のための手法を提案している [18]．本論文での提案手法と異なり，これらの手法は将来イベントの予測を行うものではない．

大規模時系列マイニングも関連する研究テーマの一つである．時系列データにおける類似探索やパターン発見は様々な手法が提案されている [13], [19]．Papadimitriou らは時系列ストリームの主要局所パターンを見つけるためのアルゴリズムを提案している [16]．文献 [20] では，データストリーム間の遅延相関を検出するためのアルゴリズムである BRAID が提案されて

いる．AWSOM は情報予測のためのストリーム処理アルゴリズムであり，単一の数値シーケンスから周期性を発見し，将来の数値予測を行っている [15]．

3. 問題設定

本論文では，(*object*, *actor*, *time*) の三つ組で構成される複合イベントを扱う．ここで，**オブジェクト** *(object)* と**アクター** *(actor)* の総数をそれぞれ u と v とする．続いて，ウィンドウサイズ l （例えば $l = 1$ 時間）の間隔が与えられ，長さ n のイベントシーケンスを構成する場合を考える．するとこのイベントシーケンスは，3 階のテンソル $\mathcal{X} \in \mathbb{N}^{u \times v \times n}$ として表現することができる．

[**定義 1**]（**イベントテンソル**） $\mathcal{X} \in \mathbb{N}^{u \times v \times n}$ を 3 階のイベントテンソルとする．$\mathcal{X}$ の要素 $x_{i,j,t}$ は時刻 t において i 番目のオブジェクトに j 番目のアクターが出現した頻度を示す．

各要素は (*object*, *actor*, *time*; *count*) の形式で表現される．例えば ('cnn.com', 'Smith', '3pm June 1, 2003'; 23) であった場合，ユーザ Smith が cnn.com へ 2003 年 6 月 1 日の午後 3 時から 4 時の間に 23 回アクセスしたことを表す．

本論文では各イベントエントリに対し特定の潜在的トピックが存在すると仮定する．これにより，*TriMine* は (*object*, *actor*, *time*) の 3 要素に対し潜在的なトピックを発見し，テンソル $\mathcal{X}$ を三つの行列 ($\mathbf{O}$, $\mathbf{A}$, $\mathbf{C}$) に分解する．

[**定義 2**]（**オブジェクト行列 O** $(u \times k)$） 要素 $o_{i,j}$ はオブジェクト i におけるトピック j との関連度の強さを示す．

このとき要素 $o_{i,j}$ は正の実数とし，各要素の合計値を 1 とする ($o_{i,j} \geq 0$, $\sum_j o_{i,j} = 1$)．アクター行列 $\mathbf{A}$ と時間行列 $\mathbf{C}$ の定義も上記と同様であるが，簡略化のため省略する．行列 $\mathbf{O}$, $\mathbf{A}$, $\mathbf{C}$ はそれぞれ，(*actor*, *object*, *time*) の各要素において，トピック#1, #2, ..., #k に対する関連度の強さを表現する．**1.** における図 1 は実データを用いたこれらの三つの行列の可視化の例である．

なお，提案手法は三つ以上の属性 ($M > 3$) をもつイベントを扱うこともできる．この場合イベントシーケンスを M 階テンソルに変換し，M 個の行列に分解することができる ($\mathbf{O}$, $\mathbf{A}^{(1)}$, ..., $\mathbf{A}^{(M-2)}$, $\mathbf{C}$)．本論文では，簡略化のため以降 3 階テンソルについてのみ言及する．表 2 は本論文で扱う記号の定義である．

表 2 主な記号と定義
Table 2 Symbols and definitions.

記号	定義
u	オブジェクト (*object*) の総数
v	アクター (*actor*) の総数
n	イベントシーケンスの長さ
$\mathcal{X}$	3 階イベントテンソル ($\mathcal{X} \in \mathbb{N}^{u \times v \times n}$)
k	潜在的トピックの総数
$\mathbf{O}$	オブジェクト行列，$u \times k$
$\mathbf{A}$	アクター行列，$k \times v$
$\mathbf{C}$	時間行列，$k \times n$

3.1 問題定義

本論文で取り組む問題は以下のとおりである．

[**問題 1**]（複合イベント集合からのパターン発見）
三つ組 (*actor*, *object*, *time*) で構成されるイベントテンソル $\mathcal{X}$ が与えられたとき，$\mathcal{X}$ の潜在的トピックを発見し，(*actor*, *object*, *time*) 各要素に対しグループを発見する．

[**問題 2**]（複合イベントの将来予測） イベントテンソル $\mathcal{X}$ が与えられたとき，イベントの将来予測を行う．

より具体的には，例えば「スミスが明日'www.cnn.com' に何度アクセスするか」という特定の状況を予測することを目的とする．提案手法は，重要なトレンドを発見し，高い予測精度を実現すると同時に，大量のイベントデータを扱うためにスケーラブルであることが求められる．

4. 提案手法

本章では，イベントデータのためのパターン発見問題（問題 1）の解決方法として，*TriMine* について述べる．イベントの予測問題（問題 2）については次章で述べる．

4.1 提案手法の概要

TriMine は以下の二つのアイディアから構成される．

- M **次元配列分析**: まず単一のウィンドウサイズを定め（例えば $l_0 = 1$ 時間），M 方向にトピック分析を行う．具体的には，k 個の潜在的トピックを発見し，M 個の行列を生成する．例えば 3 方向の場合には，*objects* (オブジェクト行列 $\mathbf{O}$, $u \times k$), *actors* (アクター行列 $\mathbf{A}$, $k \times v$), time (時間行列 $\mathbf{C}$, $k \times n$) の 3 要素に対しそれぞれ行列を生成する．
- **多重時間スケールを用いたトピック分析**: 高い精度で予測を行うには，単一のウィンドウサイズではなく，複数の時間粒度でトピック分析を行う必要がある．そこで *TriMine* は，複数の時間粒度の行列

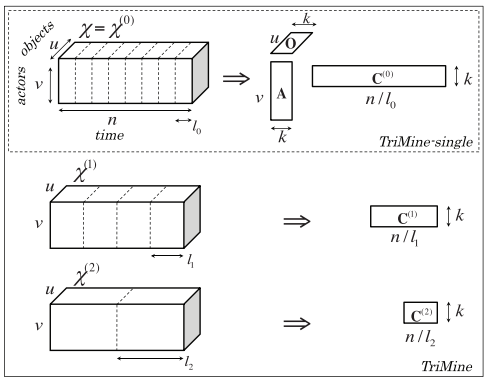


図 2　*TriMine* の概要．*TriMine-single* (点線内) は単一の時間スケール上でのトピック分析，*TriMine* (実線内) は多重時間スケール (l_0, l_1, l_2, ...) における分析の様子

Fig. 2　Illustration of *TriMine*. We perform a 3-way analysis for multiple window sizes l_0, l_1, l_2, ... to capture the multi-scale dynamics of $\mathcal{X}$.

($\{\mathbf{C}^{(0)}, \mathbf{C}^{(1)}, \ldots\}$, 例えば，分，時，日，週) を生成する．このときオブジェクトとアクターについては共通の行列 $\mathbf{O}, \mathbf{A}$ を利用する．

単一の時間スケールにおける分析 (*TriMine-single*)：図 2 (点線内) は単一の時間スケールにおけるトピック分析の様子である．オブジェクト行列 $\mathbf{O}$ は全ての時間範囲におけるオブジェクトとトピック間の関連性の強さを示す．アクター行列 $\mathbf{A}$ は，i 番目のトピック ($i = 1, \ldots, k$) に対する各アクターの頻度確率を示す．時間行列 $\mathbf{C}$ は i 番目のトピックにおける時間的な動きを表す．

多重時間スケールにおける分析 (*TriMine*)：図 2 (実線内) は，複数のウィンドウサイズを用いた場合である．*TriMine* はまずレベル $h = 0$ においてウィンドウサイズ l_0 の時間行列 $\mathbf{C}^{(0)}$ を計算する．その後，他のウィンドウサイズ l_h ($h = 0, 1, \ldots,$) に対し行列 $\mathbf{C}^{(h)}$ を得る．

4.2 TriMine

4.2.1 単一の時間スケールにおけるトピック推定

第一の課題は，イベント集合 $\mathcal{X}$ が与えられたとき，$\mathcal{X}$ を表現する k 個の潜在的トピックを発見し，これらのトピックに基づく M 個の行列 ($\mathbf{O}$, $\mathbf{A}^{(1)}$, ..., $\mathbf{A}^{(M-2)}$, $\mathbf{C}$) を推定することである．

提案モデル． 本手法ではそれぞれのイベントエントリに対し一つの潜在的トピックを割り当てる．イベント集合における生成モデルは以下のとおりである．

(1) For each topic $r = 1, \ldots, k$:
(a) For each tensor mode $m = 1, \ldots, M-2$:
i. Draw $\mathbf{A}_r^{(m)} \sim \text{Dirichlet}(\beta^{(m)})$.
(b) Draw $\mathbf{C}_r \sim \text{Dirichlet}(\gamma)$.
(2) For each object $i = 1, \ldots, u$:
(a) Draw $\mathbf{O}_i \sim \text{Dirichlet}(\alpha)$.
(b) For each entry $j = 1, \ldots, N_i$:
i. Draw a latent variable $z_{i,j} \sim \text{Multinomial}(\mathbf{O}_i)$.
ii. For each tensor mode $m = 1, \ldots, M-2$:
A. Draw an actor $e_{i,j}^{(m)} \sim \text{Multinomial}(\mathbf{A}_{z_{i,j}}^{(m)})$.
iii. Draw a timestamp $t_{i,j} \sim \text{Multinomial}(\mathbf{C}_{z_{i,j}})$.

ここで M 個から構成される行列 ($\mathbf{O}$, $\mathbf{A}^{(1)}$, ..., $\mathbf{A}^{(M-2)}$, $\mathbf{C}$) はそれぞれ，多項分布 (Multinomial) として表現され，これらはディリクレ分布 (Dirichlet) から生成されると仮定し，α, $\beta^{(m)}$, γ はそれぞれ，$\mathbf{O}$, $\mathbf{A}^{(m)}$, $\mathbf{C}$ のためのハイパーパラメータとする(注2)．

パラメータ推定． 次にトピック推定のための具体的方法について述べる．なお，ここからは簡略化のため 3 階テンソル ($M = 3$) についてのみ言及するが，以下の推定法はより高次元 ($M > 3$) についても適用可能である．本研究では，ギブスサンプリング [17] を用いてトピックの推定を行う．テンソル $\mathcal{X}$ 内における非ゼロの要素 $x_{i,j,t}$ に対し，確率 p で $x_{i,j,t}$ 個の潜在的トピックを割り振る．潜在的トピック $z_{i,j,t}$ は以下の確率によって決定される．

$$p(z_{i,j,t} = r | \mathcal{X}, \mathbf{O}', \mathbf{A}', \mathbf{C}', \alpha, \beta, \gamma) \propto \frac{o'_{i,r} + \alpha}{\sum_r o'_{i,r} + \alpha k} \cdot \frac{a'_{r,j} + \beta}{\sum_j a'_{r,j} + \beta v} \cdot \frac{c'_{r,t} + \gamma}{\sum_t c'_{r,t} + \gamma n} \quad (1)$$

ここで，$o'_{i,r}$, $a'_{r,j}$, $c'_{r,t}$ は r 番目のトピックに i 番目のオブジェクト，j 番目のアクター，時刻 t が割り振られた回数を示す．$o'_{i,r}$ 等のプライム符号は，i 番目のオブジェクト，j 番目のアクター，時刻 t について割り振られた値が除かれていることを示す．行列 $\tilde{\mathbf{O}}$, $\tilde{\mathbf{A}}$, $\tilde{\mathbf{C}}$ は次の式で計算される：$\tilde{o}_{i,r} \propto \frac{o_{i,r}+\alpha}{\sum_r o_{i,r}+\alpha k}$, $\tilde{a}_{r,j} \propto \frac{a_{r,j}+\beta}{\sum_j a_{r,j}+\beta v}$, $\tilde{c}_{r,t} \propto \frac{c_{r,t}+\gamma}{\sum_t c_{r,t}+\gamma n}$.

計算量． モデル推定の計算コストは，テンソル $\mathcal{X}$ 内のエントリの総数を N ($= \sum_{i,j,t} x_{i,j,t}$) とすると，

(注2)：オブジェクトとアクターについては，理論的にはどちらを URL (若しくはユーザ) に当てはめても同じ結果をもたらす．

N に対し線形，つまり $O(N)$ である．より具体的には，トピック数を k，学習の反復数を $iter$ とすると，潜在トピックの推定コストが $O(iter \cdot kN)$，各行列の更新コストが $O(iter \cdot k(u+v+n))$ である．ここで，$iter, k, u, v, n$ は，エントリ数 N と比較し小さい定数であるためここでは省略し，まとめると，結果として全体の計算コストは N となる．

4.2.2 多重時間スケールにおけるトピック推定

ここまでウィンドウサイズ l は固定である場合について考えた．しかし実利用のためにはデータに応じた時間スケールを選ぶ必要がある．そこで本論文では，複数のウィンドウサイズを用いることで，多重スケールにおけるパターン分析を行う．最も典型的なウィンドウサイズの選出法としては，分，時，日，週といった時間単位でスケールを扱うことが考えられる．他の方法として等比級数を用いることも可能である．この場合，各階層 $h = 0, 1, 2, ..., \lceil \log n \rceil$ において，$l_h := l_0 \cdot L^h$（例えば $L = 2$）のウィンドウサイズを利用する．

続いて，複数の時間スケール上でどのようにトピック推定を行うかについて述べる．最も単純な方法は，全ての時間スケールに対しテンソル $\{\mathcal{X}^{(0)}, \mathcal{X}^{(1)} \dots\}$ を作成し，それぞれのテンソルに対して *TriMine-single* を用いてトピック推定を行うことである．この方法を仮に *TriMine* (naive) と呼ぶ．しかしこの方法では，各時間スケールにおいて独立にトピック推定が必要となるため計算コストが非常に高い．そこで本手法では，最も短いスケール ($h = 0$) の推定結果を利用することで，他のスケールにおけるトピック推定の近似計算を行う．図 2 を用いて処理概要を示す．まずレベル $h = 0$ において，テンソル $\mathcal{X}^{(0)}$ $(= \mathcal{X})$ に対し行列 $\mathbf{O}$, $\mathbf{A}$, $\mathbf{C}^{(0)}$ を推定する．続いて，他のレベル ($h \geq 1$) に対し，$h = 0$ でのサンプリング結果を再利用しそれぞれの行列を計算する．具体的には，(a) オブジェクト行列 $\mathbf{O}$ とアクター行列 $\mathbf{A}$ を全てのレベルで共有利用し，(b) 時間行列 $\mathbf{C}^{(h)}$ については，次式を用いて計算する．

$$c^{(h)}_{r,t} \propto \sum_{i=1}^{l_h} c^{(0)}_{r,t-l_h+i}. \qquad (2)$$

ここで l_h はレベル h におけるウィンドウサイズを表す．

TriMine (naive) は全てのレベルにおいてパラメータの更新が必要となり，$O(N \log n)$ の計算時間を要するが，提案手法 *TriMine* は，データの入力サイズ N に対し線形時間 $O(N)$ である．

Algorithm 1 TriMine($\mathcal{X}^{(0)}$)

```
/* compute the triplet matrices at level h = 0 */
for each iteration do
  for each non-zero element x in X^(0) do
    for each entry for x do
      Draw hidden variable z by Equation (1)
    end for
  end for
end for
Compute O, A, C^(0)
/* compute the multi-scale matrices */
for h = 1 to ⌈log n⌉ do
  Compute C^(h) by Equation (2)
end for
return O, A, {C^(0), ..., C^(h)}
```

アルゴリズム 1 は *TriMine* の処理の流れである．イベントテンソル $\mathcal{X}^{(0)}$ 内の各エントリに対し，式 (1) を用いて隠れトピック z を割り当て，行列 $\mathbf{O}$, $\mathbf{A}$, $\mathbf{C}^{(0)}$ を推定する．その後全ての時間スケールにおいて，$\mathcal{X}^{(0)}$ の結果を用いて行列を近似計算する．

5. イベントデータの将来予測: TriMine-F

本章ではイベントデータの予測問題（問題 2）について述べる．以下では提案する予測手法を *TriMine-F* と呼ぶ．

既存手法を用いたイベント予測とその問題点: イベント集合を多次元の時系列シーケンスとみなすことで，従来の時系列解析手法を使用することができる．ウィンドウサイズ l を固定すると，イベント集合は，$u \times v$ 個 ($actor \times object$) のシーケンスに変換できる．その後それぞれのシーケンスに対し予測を行うことが可能となる．しかしこの方法では，(a) 少なくとも $O(u\ v)$ のメモリ空間と $O(u\ v\ n)$ の計算時間が必要になり，更に，(b) 各シーケンスは非常にスパースであり，例えば，$\{0, 0, 0, 1, 0, 0, 2, 0, 0, 0, 1, \dots\}$ のように一見するとただのノイズのようであるため，単体のシーケンスからの予測は非常に困難である．そこで本手法は，潜在的トピックを用いることで上記の問題を回避し，高い精度でイベントの予測を高速に行う．

5.1 トピックのダイナミックスの予測

TriMine-F は，**4.** のトピック分析で得られた行列を利用し，イベントの将来予測をする．より具体的には，(a) それぞれのトピック r $(r = 1, \dots k)$ のダイナミックス（時間行列 $\mathbf{C}$）を予測し，続いて (b) その結果と行列 $\mathbf{O}$, $\mathbf{A}$ を掛け合わせることで将来のイベント集合を生成する．

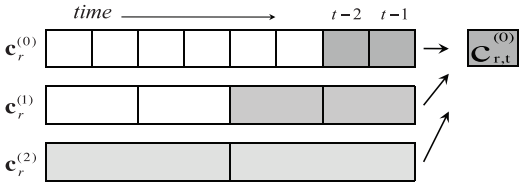


図 3 多重時間スケールにおける時間行列 $\mathbf{C}$ の予測 ($l_0 = 1, w = 2$ の場合)．各レベルの色のついたセルを用いて $c^{(0)}_{r,t}$ を予測する

Fig. 3 Illustration of multi-scale forecasting (here, $l_0 = 1, l = 2$). The gray cells indicate the variables we use to forecast $c^{(0)}_{r,t}$.

5.1.1 単一の時間スケールによる時間行列 $\mathbf{C}$ の予測

単一のウィンドウサイズ l_0 を用いる場合，AR を用いて時間行列の各要素 $c_{r,t}$ の予測を行うことができる．具体的には，w 個の係数を使い $c_{r,t-1}, \ldots c_{r,t-w}$ の関数として表現する．

$$c_{r,t} = \lambda_1 c_{r,t-1} + \cdots + \lambda_w c_{r,t-w} + \epsilon_t, \qquad (3)$$

ここで λ は回帰係数，ϵ_t はノイズとする．

5.1.2 多重時間スケールによる時間行列 $\mathbf{C}$ の予測

4.2.2 で述べたとおり，実際のイベントシーケンスは，ノイズやスパイク，周期性をはじめとする，複数のトレンドをもつ場合が多い．長期的なパターンは長いスケールの時間行列内に現れ，逆に短い周期やノイズ等は短いスケール内に出現する．そこで提案手法は，複数のレベルの時間行列 $(\mathbf{C}^{(0)}, \mathbf{C}^{(1)}, \ldots)$ を利用することで，これらの複雑な時系列パターンをモデル化する．図 3 は，多重時間スケールを用いた予測の概要を示している．提案手法は $\lceil \log n \rceil$ 個のウィンドウサイズを用い，次式のようにモデルを学習する．

$$c^{(0)}_{r,t} = \sum_{h=0}^{\lceil \log n \rceil} \sum_{i=1}^{w} \lambda^{(h)}_{i,r} c^{(h)}_{r,t-i} + \epsilon_t. \qquad (4)$$

5.2 イベントの将来予測

TriMine-F は以下の二つの予測問題を解決する．

- **イベント数の推定:** ユーザ j が URL i へ時刻 t に出現する回数 $x_{i,j,t}$ を推定する．
- **イベント集合の生成:** 将来のイベントエントリ $(object, actor, time)$ の集合を全て予測・生成する．

例えば，ユーザ Smith が CNN.com に明日以降 30 日間にアクセスする回数 $(x_{i,j,t}, t = 1, 2, ..., 30)$ を予測する．あるいは，より曖昧な条件として，Smith が明日どの URL でもよいので何度アクセスするかの予測 ($x_{*,j,t}$ の推定) を行うこともできる．

Algorithm 2 EventGeneration

$(\bar{x_1}, \ldots, \bar{x_u}, n, \mathbf{O}, \mathbf{A}, \hat{\mathbf{C}})$

```
/* Ê is a set of generated entries of form {object, actor, time } */
Ê ← ∅
for each object i = 1, ..., u do
  for each entry j = 1, ..., nx̄_i do
    Draw a hidden variable z_{i,j} ~ Multinomial(O_i)
    Draw an actor e ~ Multinomial(A_{z_{i,j}})
    Draw a timestamp t ~ Multinomial(Ĉ_{z_{i,j}})
    Ê = Ê_t ∪ {i, e, t}
  end for
end for
Return Ê
```

5.2.1 イベント数 $x_{i,j,t}$ の推定

TriMine-F は，三つの行列を用いることで，イベント数を推定することができる．具体的には，(a) 時間行列の要素 $c_{r,t}$ $(r = 1, \ldots, k)$ を予測し，行列 $\hat{\mathbf{C}}$ を得る．次に，(b) 行列 $\mathbf{O}$, $\mathbf{A}$，と予測した $\hat{\mathbf{C}}$ の積の総和を計算し，各要素 $x_{i,j,t}$ の時刻 t におけるイベント数を推定する．

$$\hat{x}_{i,j,t} = n\bar{x}_i \sum_{r=1}^{k} o_{i,r} \cdot a_{r,j} \cdot \hat{c}_{r,t}, \qquad (5)$$

ここで n は予測したいイベントの長さを示し，$\bar{x}_i$ は i 番目のオブジェクト中に含まれるイベント数の単位時間あたりの平均値とする．

5.2.2 イベント集合の生成

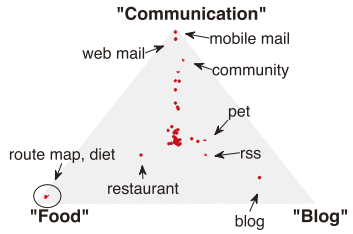
ここでは別のアプローチとして，サンプリングを用いた将来のイベント集合の生成方法について述べる．アルゴリズム 2 はイベント生成の流れである．まず時間行列の各要素 $\hat{c}_{t,r}$ を予測する．次に，$\mathbf{O}$, $\mathbf{A}$, $\hat{\mathbf{C}}$ の三つのトピック行列を用いてサンプリングを行い，{*object*, *actor*, *time*} の三つ組のエントリを生成する．最終的にこれらのエントリを全て集めて $\hat{\mathcal{E}}$ を将来のイベントエントリ集合とする．

6. 評価実験

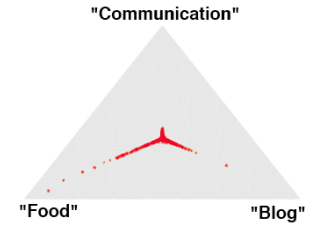
TriMine の有効性を検証するため，実データを用いた実験を行った．実験は 4GB のメモリ，Intel Core 2 Duo 1.86GHz の CPU を搭載した Linux のマシン上で実施した．本実験は，以下の諸問題に取り組む．

（1） 複合イベント集合におけるパターン発見

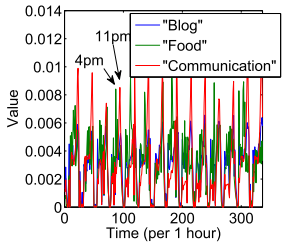
（2） イベントシーケンスに対する予測精度の検証



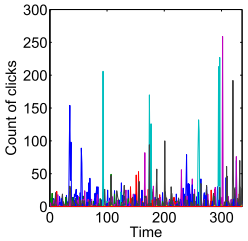
(a) トピック空間における URL の分布



(b) トピック空間におけるユーザの分布



(c) トピックの時間推移



(d) オリジナルのシーケンス

図 4　(a)-(c) *WebClick* データにおける *TriMine*-plot と (d) オリジナルデータ
Fig. 4　*TriMine* finds patterns (see (a-c)), when raw data seem noisy (part (d)).

（3） イベント予測に対する計算時間の検証

本論文では以下の二つの実データを用いて検証を行った．

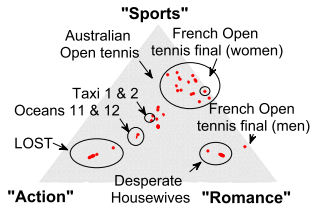
- *WebClick*： このデータセットは，1ヶ月間 (2007/4/1-4/30) のウェブアクセス履歴のデータである．このデータは URL ID (1,797 URLs), user ID (10,000 heavy users), time の三つの属性から構成される．URL には，blog, news, money を始めとする様々な種類のウェブサイトが含まれる．
- *Ondemand TV*：このデータセットは，オンラインの TV 配信サービスの視聴に関するデータである．このデータは 6ヶ月間 (2007/5/14-2007/11/15) の番組視聴履歴であり，ランダムに選出された 100,000 名の匿名ユーザそれぞれに対し，どの番組をいつ視聴したかの情報が蓄積されている．代表的な TV のジャンルはスポーツや映画等である．各レコードは，channel ID (*object*), user/viewer ID (*actor*), time の三つの属性値をもつ．本研究では，視聴回数頻度の高い 100 件の TV 番組を用いて学習を行った．

6.1 時系列イベントにおけるパターン発見

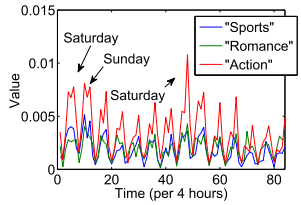
6.1.1 *WebClick*

WebClick データセットの実験結果の一部は **1.** に示しており（図 1，*TriMine*-plot），*TriMine* は効率的かつ効果的に 3 方向のパターンを検出している．図 4 は *WebClick* データセットにおける出現数の多い主要なトピックのうち，図 1 に挙げていない三つのトピック (communication, blog, food) について *TriMine*-plot を示したものである．

- **メンバーシップクラスタ**：図 4 (a) は，communication, blog, food の 3 次元トピック空間上でのオブジェクト (URL) の分布を示している．頂点に近い場所に位置するオブジェクトは，そのトピックの性質を強くもち，一方で，どの頂点からも遠く，中央に位置するオブジェクトは，これらの三つのトピックの特徴をもたないことを示す．図において，route map と restaurant のサイトは food トピックに関連しており，diet のサイトも同じトピックに含まれる．ユーザはレストラン情報と彼らの地域のルートマップをチェックし，更にその食事に関するカロリーを調べる — そのような行動を図から読み取ることができる．一方，pet サイトは communicaiton トピックと blog トピックの両方の特徴をもつ．このことから，pet に興味があるユーザは，blog あるいはメールのやり取りを通じて，pet に興味がある他のユーザと交流していることが読み取れる．図 4 (b) は，トピック空間上でのアクター（ユーザ）の分布を示している．オブジェクトの分布と異なり，単一のトピックの特徴を強くもつ点が多いことがわかる．
- **時系列トレンド**：トピック food に関連する URL は，ユーザが外出する直前の夕方頃にアクセスが増え



(a) URL の分布図（行列 $\mathbf{O}$）



(b) トピックの時間推移（行列 $\mathbf{C}$）

図 5　*Ondemand TV* データにおける *TriMine* の結果
Fig. 5　Effectiveness of *TriMine* on the *Ondemand TV* dataset.

る傾向がある（図 4 (c)）．また Web メールや SNS など，communication に関するサイトはプライベートな目的のために深夜によく使われているようである．図 4 (d) は blog サイトにおけるランダムに選択したユーザによるオリジナルの時系列データを示している．図 4 (c) と異なり，図 4 (d) からは毎日の周期性やユーザごとの関係性など，有用な情報を得ることは難しい．

6.1.2　*Ondemand TV*

図 5 は *Ondemand TV* データセットにおける出現数の多い三つの主要なトピック (sports, action, romance) について示している．

- **例外**：図 5 (a) の URL のプロットは明確なクラスタを示しているが，一つ例外が見られる．2007 年全仏オープンテニスの男子決勝戦である．'Desperate Housewives' のような romance（若しくは soap opera）に関連する番組は一般的に女性の視聴者に興味をもたれることが多く，一方でそのような視聴者はスポーツに興味をもつとは考えにくい．しかし彼女らにとって，その決勝戦だけは特別のようである．恐らくその試合の選手に彼女らは興味をもっているのではないかと考えられる．
- **時系列トレンド**：トピックの時間発展パターンは我々の直観にしたがったものとなっている．一日単位の周期は全てのトピックに見られ，action と sports については週末に高いピークが見られる．

6.2　予 測 精 度

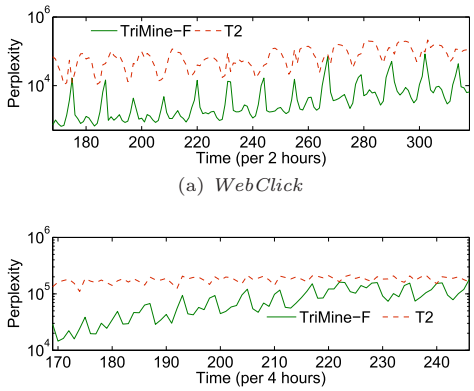
時系列イベントデータの予測は非常に挑戦的な課題である．本節では提案手法である *TriMine-F* の予測精度について，*WebClick* データを用いて検証する．予測に関する研究では短期予測の精度を検証するのが一般的であり，例えば文献 [24] では 1 時刻先の予測，すなわち時刻 t の値を得ると時刻 $t+1$ の値を推定している．これに対して本論文の目的は長期的な変動を捉えることであり，提案手法がこれを実現していることを示す．

実験では，最初の 2 週間のクリックイベントを用いてモデルを学習し，その後の 2 週間のイベントを予測することによって精度評価を行う．ウィンドウサイズを $l_0 = 2$ 時間とする．すなわち，学習データ $\mathcal{X}^{(0)}$ の長さは $n = 168$ である．潜在変数として $k = 30$，そして予測のために合計 40 個の係数値を用いる．ここでは，次の三つの手法との比較を行った．(a) **AR**: 予測問題に対する最も基本的な手法である．まずイベントテンソル $\mathcal{X}$ を $u \times v$ のシーケンスに分解し，そして *AR* モデルを各 URL と各ユーザに対して個別に適用する．公正な評価のために 40 個の回帰係数を用いる．(b) **PLiF**: 本研究ではデータベース分野において提案された *PLiF* [11] と比較実験を行う．*PLiF* は線形動的システム (Linear Dynamical System) 若しくはカルマンフィルタに基づく手法である．*PLiF* は複数のシーケンスの相関を捉え，その情報に基づきシーケンスの予測を効果的に行う優れた手法である．実データは非常にバースト性が高いため，文献 [11] に従い，実験データに対して対数計算を施す．(c) **T2**: データマイニング分野において Hong らはトレンドを検出，追跡 (tracking trends) するための新たなトピックモデルを提案している [6]．本論文ではこれを *T2* と呼ぶ．*T2* はデータ集合からトピックの時間発展を捉えることが可能な洗練された手法であり，本研究では *T2* とも比較を行う．実験では長さ n の学習データを用いてモデルパラメータを推定し，時刻 $t = n$ におけるモデルパラメータを用いて将来のイベント予測を行う．

6.2.1　予 測 精 度

パープレキシティ (perplexity) に基づく評価． 本研究ではまず *TriMine-F* と *T2* に対しパープレキシティ (perplexity) を用いた評価実験を行った．パープレキシティは次の式を用いて計算した．

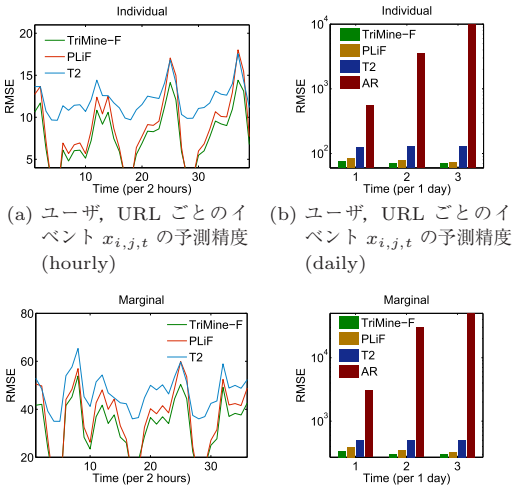
$$perplexity(t) = \exp\Big(-\frac{\sum_{ij} \log p(\boldsymbol{x}_{i,j,t} | \mathcal{X}, \mathbf{O}, \mathbf{A}, \hat{\mathbf{C}})}{\sum_{ij} \boldsymbol{x}_{i,j,t}}\Big)$$



(a) *WebClick*

(b) *Ondemand TV*

図 6　各時刻におけるパープレキシティ(perplexity)
Fig. 6　Perplexity at each time-tick.



(a) ユーザ，URL ごとのイベント $x_{i,j,t}$ の予測精度 (hourly)

(b) ユーザ，URL ごとのイベント $x_{i,j,t}$ の予測精度 (daily)

(c) ユーザごとのイベント数 $x_{*,j,t}$ の予測精度 (hourly)

(d) ユーザごとのイベント数 $x_{*,j,t}$ の予測精度 (daily)

図 7　上段：各ユーザ，URL の個別イベント $x_{i,j,t}$ に対する予測精度．下段：ユーザごとのイベントの総計 $x_{*,j,t}$ に対する予測精度
Fig. 7　Forecasting accuracy for individual $x_{i,j,t}$ (top) and marginal $x_{*,j,t}$ (bottom).

図 6 (a), (b) はそれぞれ，*WebClick*, *Ondemand TV* データセットにおける，各時刻のパープレキシティの変化を示している．低いパープレキシティは高いモデル精度を意味する．*TriMine-F* は各データセット (*WebClick*, *Ondemand TV*) の周期的なトレンドのみならず，イベントの急激なパターン変化を捉えており，適切にイベントの将来の傾向を推測している．一方で，*T2* は長期予測を得意とする手法ではないため高精度の予測が難しい．

イベント予測の精度．次に，将来のイベント数の予測に関して，提案手法と既存手法である *AR*, *PLiF*, *T2* を *WebClick* データセットを用いて比較する．図 7 (a), (b) は，全ての URL とユーザの組み合わせ ($x_{i,j,t}$) に対して，オリジナルデータの値と予測値との最小 2 乗誤差 (*RMSE*) を示している．図 7 (c), (d) は，各時刻 t においてユーザ j に関するイベント数の総計 ($x_{*,j,t}$) を予測した結果である．図 7 (a), (c) には周期的に急激な数値の下落が見られる．これは深夜にクリック回数が減少するためである．*T2* は部分的に将来のクリックイベントの生成に成功しているものの，頻繁に情報予測に失敗している．他の既存手法である *AR* と *PLiF* についても予測に失敗している．これはイベントシーケンスは非常にスパースであるため，データのトレンドや周期性を捉えることがこれらの手法にとって難しいためである．これらの手法と異なり，提案手法は全ての時刻において優れた予測結果を示している．

6.2.2　多重スケールアプローチの効果

本節では，*TriMine-F* がどのようにダイナミックスを捉えるているかについて議論する．多重時間スケールのアプローチの効果を調べるため，本研究では提案手法である *TriMine-F* から，多重時間スケールに関する機能を取り除いた手法を実装し，比較した．すなわち，これは 1 段の再帰係数のみを用いて予測するものであり，ここで *TriMine-F* (single) と呼ぶ．公正な評価のために，この手法についても同じく 40 個の係数値を用いる．図 8 は，*WebClick* データにおける二つの主要トピックの時間発展を示したものである．この実験でも，2 週間のクリックイベントを使ってモデル学習し（図における点線），その後 2 週間の予測を行う．図 8 における上段，中段，下段はそれぞれ *TriMine*, *TriMine-F* (single), *TriMine-F* の出力結果である．上段の *TriMine* の結果は予測結果ではなく，単に各時刻のトピックの重みを示したものである．下段，*TriMine-F* の結果はイベント予測に関する我々の完全な提案手法であり，多重スケール分析を含んでいる．図は *TriMine-F* (single) がダイナミックスを捉えることに失敗し，収束しているのに対し，提案手法である *TriMine-F* は数週間の予測に成功している様子を示している．多重スケール分析が有効に機能していることを示しており，その結果，長期的なトレンドと周期的なパターンを捉え，高い予測精度につながっている．

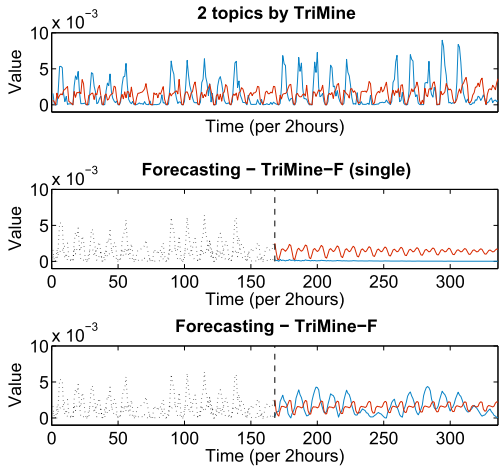


図 8　多重時間スケールの効果．上段：*WebClick* データセットにおける二つの主要トレンド，business（青線）と media（赤線）．中段と下段：*TriMine-F* は *TriMine-F* (single) より優れた予測能力をもつ
Fig. 8　Benefit of multiple time scales.

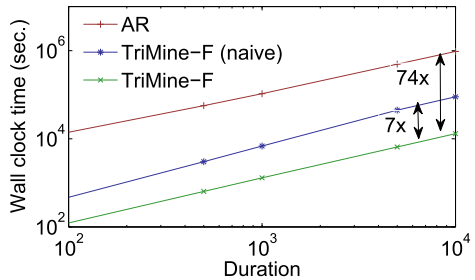


図 9　データセットのサイズ（イベントデータの長さ n）を変化させた際のイベント予測の計算コスト
Fig. 9　Scalability of event forecasting: Wall clock time vs. dataset size (= duration n).

6.3　計算コスト

本節では，イベント予測における *TriMine-F* の計算コストについて評価する．図 9 は，イベントデータの長さ n を変化させたときの AR 手法と提案手法との計算時間の比較を示している．この実験では *WebClick* データセットを用いており，URL 数は $u = 1{,}000$，ユーザ数は $v = 10{,}000$ である．ここでの計算時間は，統計値と係数値の計算と予測結果の出力に要する時間を示している．*T2* と *PLiF* はカルマンフィルタに基づく手法であり，大規模データに対しては計算コストが非常に高い．長さ $n = 100$ の場合であっても 10^6 秒以上の時間を必要とするため結果から除外した．本論文では **4.2.2** において，多重スケール分析の計算コストを削減するための高速化アプローチを提案した．このアプローチの効果を明らかにするため，提案手法から高速化アプローチに関する機能を取り除いた手法を実装し，計算コストを評価した．ここでその手法を *TriMine-F* (naive) と呼ぶ．図 9 の実験結果は *TriMine-F*（すなわち提案手法の完全版）の優位性を示している．*TriMine-F* は大幅に計算コストを低減化させており，*TriMine-F* (naive) と比較して 7 倍，AR と比べて 74 倍の高速化を達成している．

7.　む　す　び

本論文では，三つ組 (*object*, *actor*, time) の形で示される複合イベントのためのトレンド検出の問題を扱い，提案手法である *TriMine* について述べた．*TriMine* は実データから有意なパターンを発見するとともに，可視化，外れ値検出，データ要約，意味づけ (sense-making) を行うことができる．イベント予測のための手法である *TriMine-F* は効率的にイベントの予測を行うことができ，その計算コストはデータベースサイズに線形である．更に，実データを用いた実験により，*TriMine* が最新の予測手法と比べてより高い精度と性能を達成していることを示した．今後の課題として，Twitter をはじめとする，周期性を伴わず突発的なふるまいが多く見られるソーシャルネットワークデータ等に対してもイベント予測を行うために，提案手法を改良していく予定である．

文　献

[1] D. Agarwal, B.-C. Chen, and P. Elango, “Spatio-temporal models for estimating click-through rate,” WWW Conference, pp.21–30, Madrid, Spain, April 2009.
[2] L. AlSumait, D. Barbará, and C. Domeniconi, “Online lda: Adaptive topic models for mining text streams with applications to topic detection and tracking,” ICDM, pp.3–12, 2008.
[3] D.M. Blei and J.D. Lafferty, “Dynamic topic models,” ICML, pp.113–120, 2006.
[4] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.
[5] T. Hofmann, “Probabilistic latent semantic indexing,” SIGIR, pp.50–57, 1999.
[6] L. Hong, D. Yin, J. Guo, and B.D. Davison, “Tracking trends: Incorporating term volume into temporal topic models,” KDD, pp.484–492, 2011.
[7] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, “Topic tracking model for analyzing consumer purchase behavior,” IJCAI, pp.1427–1432, 2009.
[8] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, “On-

line multiscale dynamic topic models," KDD, pp.663–672, 2010.
[9] T.G. Kolda, B.W. Bader, and J.P. Kenny, "Higher-order web link analysis using multilinear algebra," ICDM, pp.242–249, 2005.
[10] L.D. Lathauwer, B.D. Moor, and J. Vandewalle, "A multilinear singular value decomposition," SIAM J. Matrix Anal. Appl., vol.21, no.4, pp.1253–1278, 2000.
[11] L. Li, B.A. Prakash, and C. Faloutsos, "Parsimonious linear fingerprinting for time series," PVLDB, vol.3, no.1, pp.385–396, 2010.
[12] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa, "Fast mining and forecasting of complex time-stamped events," KDD, pp.271–279, 2012.
[13] Y. Matsubara, Y. Sakurai, and M. Yoshikawa, "Scalable algorithms for distribution search," ICDM, pp.347–356, 2009.
[14] R.V. Nehme, E.A. Rundensteiner, and E. Bertino, "Tagging stream data for rich real-time services," PVLDB, vol.2, no.1, pp.73–84, 2009.
[15] S. Papadimitriou, A. Brockwell, and C. Faloutsos, "Adaptive, hands-off stream mining," Proc. VLDB, pp.560–571, Berlin, Germany, Sept. 2003.
[16] S. Papadimitriou and P.S. Yu, "Optimal multi-scale patterns in time series streams," SIGMOD Conference, pp.647–658, 2006.
[17] I. Porteous, D. Newman, A.T. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," KDD, pp.569–577, 2008.
[18] S. Rendle, L.B. Marinho, A. Nanopoulos, and L. Schmidt-Thieme, "Learning optimal ranking with tensor factorization for tag recommendation," KDD, pp.727–736, 2009.
[19] Y. Sakurai, C. Faloutsos, and M. Yamamuro, "Stream monitoring under the time warping distance," Proc. ICDE, pp.1046–1055, Istanbul, Turkey, April 2007.
[20] Y. Sakurai, S. Papadimitriou, and C. Faloutsos, "Braid: Stream mining through group lag correlations," Proc. ACM SIGMOD, pp.599–610, Baltimore, Maryland, June 2005.
[21] G. Tomasi and R. Bro, "Parafac and missing values," Chemometrics and Intelligent Laboratory Systems, vol.75, no.2, pp.163–180, 2005.
[22] X. Wang and A. McCallum, "Topics over time: A non-markov continuous-time model of topical trends," KDD, pp.424–433, 2006.
[23] X. Wei, J. Sun, and X. Wang, "Dynamic mixture models for multiple time-series," IJCAI, pp.2909–2914, 2007.
[24] A. Weigend and N. Gershenfeld, Time Series Prediction: Forecasting the Future and Understanding the Past, Addison-Wesley, 1993.
[25] L. Yao, D.M. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," KDD, pp.937–946, 2009.

（平成 25 年 7 月 2 日受付，10 月 29 日再受付）

松原　靖子

2006 年お茶の水女子大学理学部情報科学科卒業．2009 年同大学院人間文化創成科学研究科理学専攻博士前期課程修了．2012 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（工学）．2011～2012 年カーネギーメロン大学客員研究員．データストリーム処理，大規模データマイニングに関する研究に従事．日本データベース学会会員．

櫻井　保志　（正員）

1991 年同志社大学工学部電気工学科卒業．1991 年日本電信電話（株）入社．1999 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．2004～2005 年カーネギーメロン大学客員研究員．2013 年熊本大学大学院自然科学研究科教授．本会平成 19 年度論文賞，情報処理学会平成 18 年度長尾真記念特別賞，平成 16 年度及び平成 19 年度論文賞，日本データベース学会上林奨励賞，ACM KDD best paper awards（2008，2010）等受賞．データマイニング，データストリーム処理，センサーデータ処理，Web 情報解析技術の研究に従事．ACM，情報処理学会，日本データベース学会各会員．

Christos Faloutsos

Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation (1989), the Research Contributions Award in ICDM 2006, the SIGKDD Innovations Award (2010), seventeen best paper awards (including two 'test of time'), and four teaching awards. He has served as a member of the executive committee of SIGKDD; he is an ACM Fellow; he has published over 200 refereed articles, 11 book chapters, and one monograph. He holds five patents and he has given over 30 tutorials and over 10 invited distinguished lectures. His research interests include data mining for graphs and streams, fractals, database performance, and indexing for multimedia and bioinformatics data.

岩田　具治　（正員）

2001 年慶應義塾大学環境情報学部卒業．2003 年東京大学大学院総合文化研究科修士課程修了．同年日本電信電話 (株) 入社．2008 年京都大学大学院情報学研究科博士課程修了．博士（情報学）．2012 年から 2013 年にかけケンブリッジ大学客員研究員．現在，NTT コミュニケーション科学基礎研究所所属．機械学習，データマイニング，情報可視化の研究に従事．船井ベストペーパー賞，情報処理学会論文賞等受賞．電子情報通信学会，情報処理学会各会員．

吉川　正俊　（正員：フェロー）

京都大学大学院工学研究科博士後期課程修了．工学博士．京都産業大学，奈良先端科学技術大学院大学，名古屋大学を経て 2006 年より京都大学大学院情報学研究科教授．この間，南カリフォルニア大学客員研究員，ウォータルー大学客員准教授．The VLDB Journal 及び Information Systems (Elsevier/Pergamon) の編集委員．XML データベース，異種情報源の統合などの研究に従事．情報処理学会フェロー，ACM，IEEE Computer Society 各会員．