大規模時系列データのためのモデル学習とイベント予測

松原　靖子†

† NTT コミュニケーション科学基礎研究所
E-mail: †matsubara.yasuko@lab.ntt.co.jp

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．さら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

Yasuko MATSUBARA†

† NTT Communication Science Laboratories, NTT Corporation
E-mail: †matsubara.yasuko@lab.ntt.co.jp

Abstract　Given huge collections of time-evolving events such as web-click logs, which consist of multiple attributes (e.g., URL, userID, timestamp), how do we find patterns and trends? How do we go about capturing daily patterns and forecasting future events? We introduce *TriMine*, which performs three-way mining for all three attributes, namely, URLs, users, and time. Specifically *TriMine* discovers hidden topics, groups of URLs, and groups of users, simultaneously. Thanks to its concise but effective summarization, it makes it possible to accomplish the most challenging task, namely, to forecast future events.

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発生しており，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 [13] は代表的な例である．

本論文で扱う問題は以下の通りである．

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．

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持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．

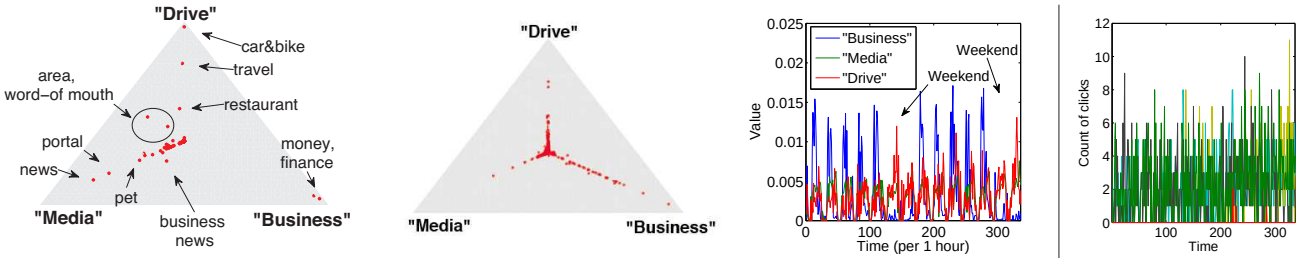
本論文では，大量に発生する複合イベント集合から主要パターンを発見する手法である *TriMine* [9] (注1) について述べる．*TriMine* は Web クリックデータを $(object, actor, time)$ のそれぞれの角度から捉え，共通する潜在的トピックを発見する．

1.1 具　体　例

一般に，各 Web サイトには 1 つ以上の潜在的なトピックが存在している．同様に，各ユーザもいく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．さらに，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 トピックを持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．ここでは図 1 (a)-(c) を *TriMine*-plot と呼

(注1)：ソースコード：http://www.kecl.ntt.co.jp/csl/sirg/people/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 $\mathbf{O}$)　(b) トピック空間におけるユーザの分布 (アクター行列 $\mathbf{A}$)　(c) トピックの時間推移 (時間行列 $\mathbf{C}$)　(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.

ぶ．*TriMine*-plot は 2 つの三角プロットと 1 つの時系列シーケンスから構成される．図 1 (a)-(c) はそれぞれの潜在的トピックに対して，$(object, actor, time)$ の 3 つの要素がどのように分布しているかを示す．図は最も頻出する 3 つのトピックについて可視化している．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．詳細は以下の通りである．

(a) トピック空間における URL の分布：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持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 トピックを持っている．

(b) トピック空間におけるユーザの分布：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の様子： 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．トピック全体において 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) は週末に高い値を持つ．

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 3 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 1.2 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) [6] と交互最小二乗法 (ALS: alternating least squares) [19] 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持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) [3] 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* [7] 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持つ．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（図 7，8 ）．

	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					√

表 1　既存手法との比較.

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

大規模時系列マイニングは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 [8], [10]〜[12], [16]〜[18], [20]．潜在的ディリクレ配分法 (LDA: latent Dirichlet allocation) [3] はテキストデータのための bag-of-words などの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) [2] や TOT (topics over time) [21]，その他様々な手

表 2　主な記号と定義.

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，$k \times n$

法が提案されている.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 [5]．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 [15]．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．

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 3 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 $\mathbf{O}$ ($u \times k$)） 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 ($o_{i,j} \geqq 0, \sum_j o_{i,j} = 1$)．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. 章における図 1 は実データを用いたこれらの 3 つの行列の可視化の例である．

なお，提案手法は 3 つ以上の属性 ($M > 3$) を持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．

3.1 問題定義

本論文で取り組む問題は以下の通りである．

[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 2 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 ($\{\mathbf{C}^{(0)}, \mathbf{C}^{(1)}, \ldots\}$, 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 $\mathbf{O}$ はすべ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．本手法ではそれぞれのイベントエントリに対し 1 つの潜在的トピックを割り当てる．ここでは，ギブスサンプリング [14]

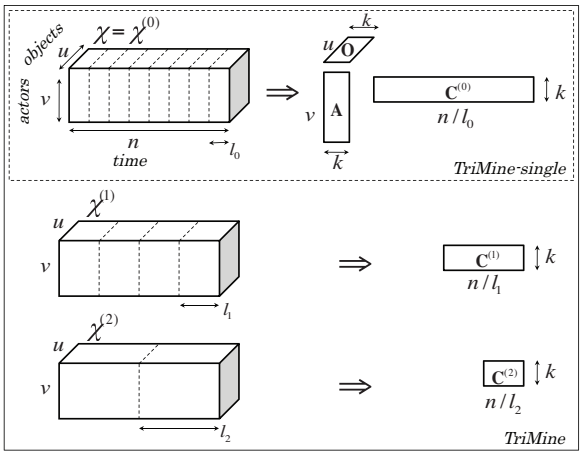


図 2　*TriMine* の概要．*TriMine-single* (点線内) は単一の時間スケール上でのトピック分析，*TriMine* (実線内) は多重時間スケール ($l_0, l_1, l_2, \ldots$) における分析の様子．

を用いてトピックの推定を行う．イベント集合における生成モデルは以下の通りである．

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

ここで $\alpha, \beta^{(m)}, \gamma$ 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．

次にトピック推定のための具体的方法について述べる．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．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 $\tilde{\mathbf{O}}$, $\tilde{\mathbf{A}}$, $\tilde{\mathbf{C}}$ は以下の式で計算される．

$$\tilde{o}_{i,r} \propto \frac{o_{i,r} + \alpha}{\sum_r o_{i,r} + \alpha k}, \tilde{a}_{r,j} \propto \frac{a_{r,j} + \beta}{\sum_j a_{r,j} + \beta v}, \tilde{c}_{r,t} \propto \frac{c_{r,t} + \gamma}{\sum_t c_{r,t} + \gamma n}.$$

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，サンプリングの計算コストは $O(N)$ である．

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．

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)}
```

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，すべての時間スケールに対しテンソル $\{\mathcal{X}^{(0)}, \mathcal{X}^{(1)} \ldots\}$ 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 ($h \geqq 1$) 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 $\mathbf{A}$ をすべ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．

$$c_{r,t}^{(h)} \propto \sum_{i=1}^{l_h} c_{r,t-l_h+i}^{(0)} \quad (2)$$

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) はすべ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．

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．その後すべ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

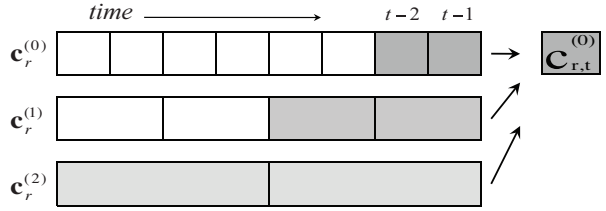


図 3　多重時間スケールにおける時間行列 $\mathbf{C}$ の予測 ($l_0 = 1, w = 2$ の場合)．各レベルの色のついたセルを用いて $c_{r,t}^{(0)}$ を予測する．

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見なすことで，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クスの予測

TriMine-F は，4. 章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．

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 $c_{r,t-1}, \dots c_{r,t-w}$ の関数として表現する．

$$c_{r,t} = \lambda_1 c_{r,t-1} + \dots + \lambda_w c_{r,t-w} + \epsilon_t, \tag{3}$$

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 節で述べた通り，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，複数のトレンドを持つ場合が多い．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 ($\mathbf{C}^{(0)}, \mathbf{C}^{(1)}, \dots$) 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．

$$c_{r,t}^{(0)} = \sum_{h=0}^{\lceil \log n \rceil} \sum_{i=1}^{w} \lambda_{i,r}^{(h)} c_{r,t-i}^{(h)} + \epsilon_t. \tag{4}$$

5.2　イベントの将来予測

TriMine-F は以下の 2 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 ($object, actor, time$) の集合をすべて予測・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．

5.2.1　イベント数 $x_{i,j,t}$ の推定

TriMine-F は，3 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 $c_{r,t}$ ($r = 1, \dots, k$) 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．

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

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 3 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すべて集めて $\hat{\mathcal{E}}$ を将来のイベントエントリ集合とする．

Algorithm 2 EventGeneration ($\bar{x_1}, \dots, \bar{x_u}, n, \mathbf{O}, \mathbf{A}, \hat{\mathbf{C}}$)

```
/* 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, ..., n x̄_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
```

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．

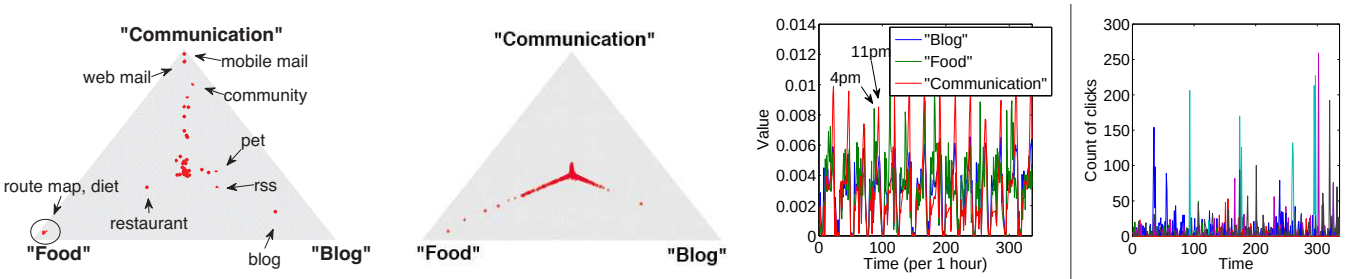
（1） 複合イベント集合におけるパターン発見

（2） イベントシーケンスに対する予測精度の検証

（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 2 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)



(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.

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 3 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 3 つの属性値を持つ．

6.1　時系列イベントにおけるパターン発見

6.1.1　*WebClick*

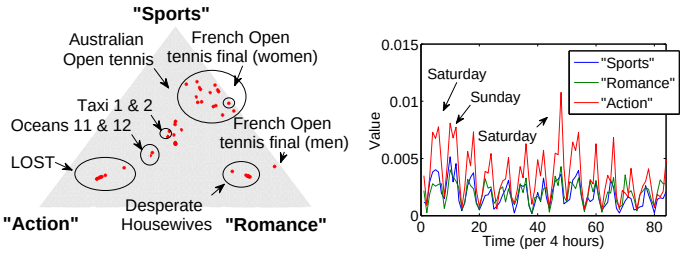
WebClick データセットの実験結果の一部は 1. 章に示しており（図 1，*TriMine*-plot），*TriMine* は効率的かつ効果的に 3 方向のパターンを検出している．図 4 は *WebClick* データセットにおける communication, blog, food という 3 つの異なるトピックの *TriMine*-plot を示している．

- メンバーシップクラスタ：ほとんどのオブジェクト (URL) は中央から頂点を結ぶスポーク状の線に沿って分布している（図 4 (a)）．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，さら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．
- 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毎の関係性など，有用な情報を得ることは難しい．

6.1.2　*Ondemand TV*

図 5 は主要な 3 つのトピック sports, action, romance について示している．

- 外れ値：URL のプロットは各 URL の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（もしくは soap opera）に関連する番組は一般的に女性の視聴者に興味を持たれることが多く，一方でそのような視聴者はスポーツに興味を持つとは考えにくい．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．おそらくその試合の選手に彼女らは興味を持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．一日単位の周期はすべ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．



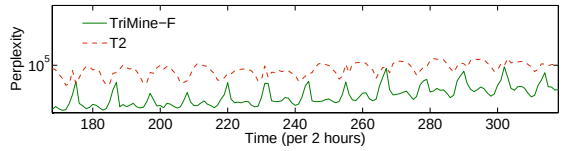
(a) URL の分布図（行列 $\mathbf{O}$）　(b) トピックの時間推移（行列 $\mathbf{C}$）

図 5　*Ondemand TV* データにおける *TriMine* の結果.

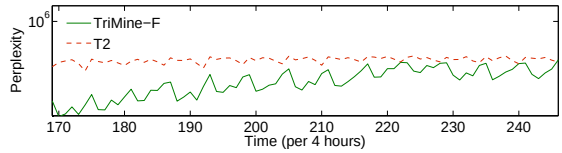
6.2　予 測 精 度

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 [22] 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 3 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* [7] と比較実験を行う．*PLiF* は線形動的システム (Linear Dynamical System) も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

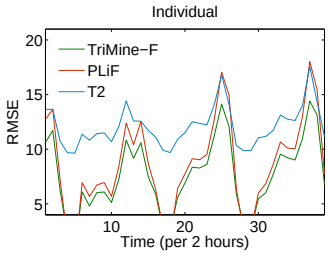


(a) *WebClick*

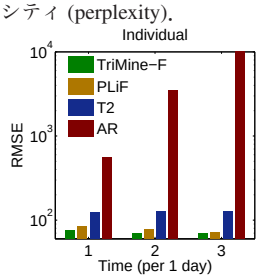


(b) *Ondemand TV*

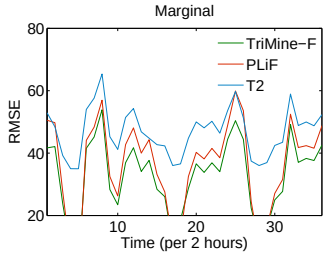
図 6　各時刻におけるパープレキシティ (perplexity).



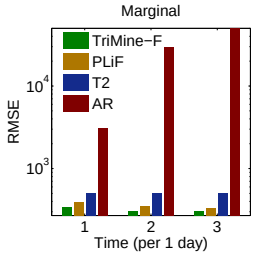
(a) ユーザ，URL 毎のイベント $x_{i,j,t}$ の予測精度 (hourly)



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



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



(d) ユーザ毎のイベント数 $x_{*,j,t}$ 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：ユーザ毎のイベントの総計 $x_{*,j,t}$ に対する予測精度.

をとらえ，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 [7] 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 [4]．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．図 6 は各時刻におけるパープレキシティを示している．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，すべての 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，提案手法はすべ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 2 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．図では *TriMine-F* (single) がダイナミ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．

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 節において，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

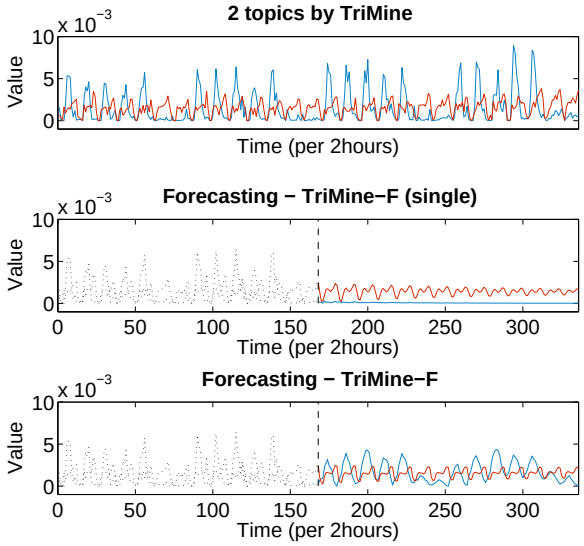


図 8　多重時間スケールの効果．上段：*WebClick* データセットにおける 2 つの主要トレンド，business（青線）と media（赤線）．中段と下段：*TriMine-F* は *TriMine-F* (single) より優れた予測能力を持つ．両手法ともに時刻 $t = 168$（2 週間後）より長期予測を開始している．多重スケールアプローチを採り入れた *TriMine-F* は実データの特徴を捉えている．

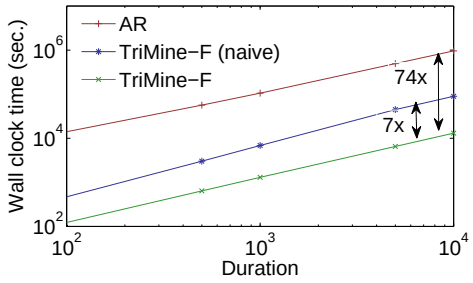


図 9　データセットのサイズ（イベントデータの長さ n）を変化させた際のイベント予測の計算コスト．

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．さらに，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．

謝　辞

本研究の一部は，内閣府最先端研究開発支援プログラム「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」の助成により行われた．

文　献

[1] D. Agarwal, B.-C. Chen, and P. Elango. Spatio-temporal models for estimating click-through rate. In *WWW Conference*, pages 21–30, Madrid, Spain, April 2009.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] L. Hong, D. Yin, J. Guo, and B. D. Davison. Tracking trends: incorporating term volume into temporal topic models. In *KDD*, pages 484–492, 2011.

[5] T. G. Kolda, B. W. Bader, and J. P. Kenny. Higher-order web link analysis using multilinear algebra. In *ICDM*, pages 242–249, 2005.

[6] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[7] L. Li, B. A. Prakash, and C. Faloutsos. Parsimonious linear fingerprinting for time series. *PVLDB*, 3(1):385–396, 2010.

[8] Y. Matsubara, L. Li, E. Papalexakis, D. Lo, Y. Sakurai, and C. Faloutsos. F-trail: Finding patterns in taxi trajectories. In *PAKDD*, 2013 (to appear).

[9] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa. Fast mining and forecasting of complex time-stamped events. In *KDD*, pages 271–279, 2012.

[10] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *KDD*, pages 6–14, 2012.

[11] Y. Matsubara, Y. Sakurai, and M. Yoshikawa. Scalable algorithms for distribution search. In *ICDM*, pages 347–356, 2009.

[12] Y. Matsubara, Y. Sakurai, and M. Yoshikawa. *D-Search*: an efficient and exact search algorithm for large distribution sets. *Knowl. Inf. Syst.*, 29(1):131–157, 2011.

[13] R. V. Nehme, E. A. Rundensteiner, and E. Bertino. Tagging stream data for rich real-time services. *PVLDB*, 2(1):73–84, 2009.

[14] I. Porteous, D. Newman, A. T. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *KDD*, pages 569–577, 2008.

[15] S. Rendle, L. B. Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *KDD*, pages 727–736, 2009.

[16] Y. Sakurai, C. Faloutsos, and M. Yamamuro. Stream monitoring under the time warping distance. In *Proceedings of ICDE*, pages 1046–1055, Istanbul, Turkey, April 2007.

[17] Y. Sakurai, L. Li, Y. Matsubara, and C. Faloutsos. Windmine: Fast and effective mining of web-click sequences. In *SDM*, pages 759–770, 2011.

[18] Y. Sakurai, S. Papadimitriou, and C. Faloutsos. Braid: Stream mining through group lag correlations. In *Proceedings of ACM SIGMOD*, pages 599–610, Baltimore, Maryland, June 2005.

[19] G. Tomasi and R. Bro. Parafac and missing values. *Chemometrics and Intelligent Laboratory Systems*, 75(2):163–180, 2005.

[20] M. Toyoda, Y. Sakurai, and Y. Ishikawa. Pattern discovery in data streams under the time warping distance. In *VLDB Journal*, 2013 (to appear).

[21] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, pages 424–433, 2006.

[22] A. Weigend and N. Gershenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley, 1993.