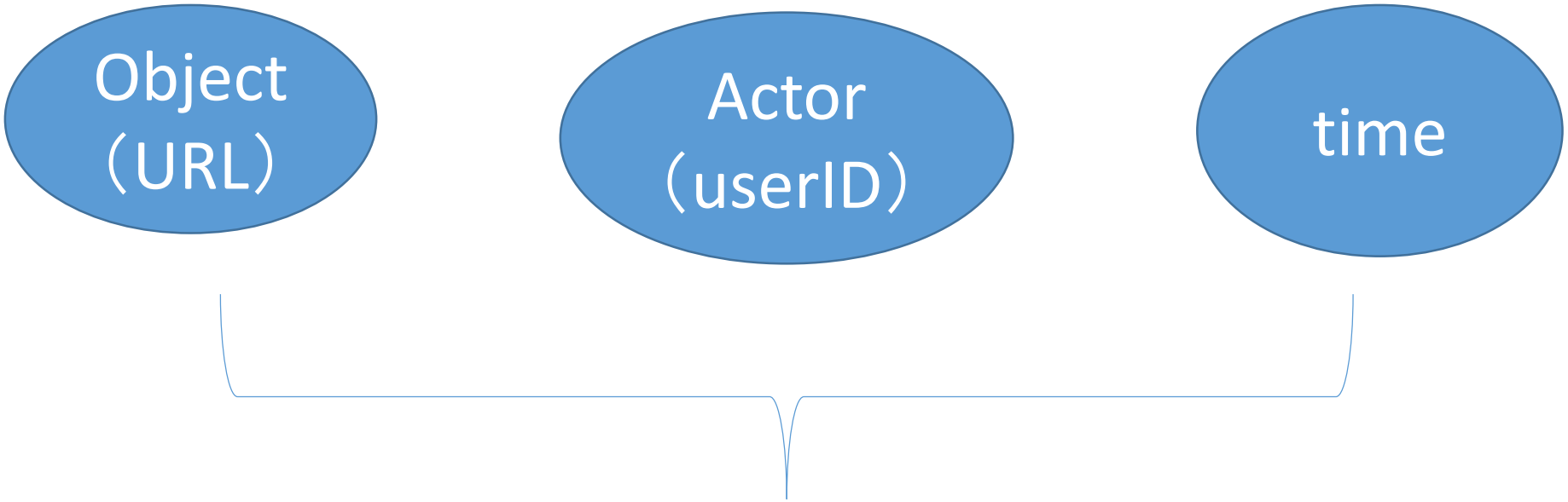
大規模Webクリックデータのためのイベント予測

1515050　山本聖也

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.
- 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

本研究で扱う問題



この3つで構成されるイベントシーケンス群が与えられたとき

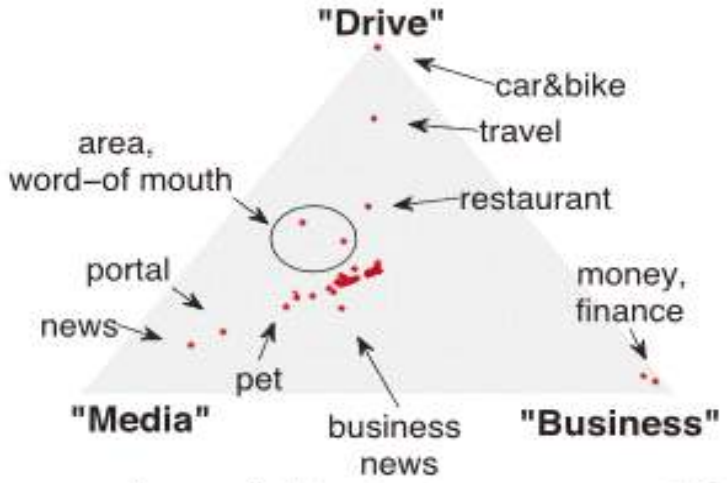
a 潜在的なトピックとトレンドを発見

b 将来のイベントを高速に予測

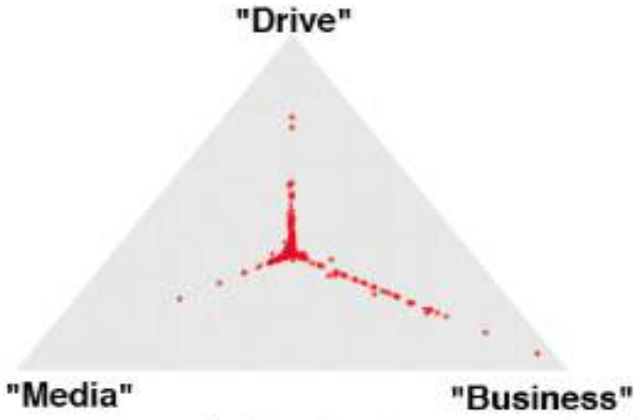
TriMine

- Webサイトには一つ以上の潜在的なトピックが存在
- 例えば経済ニュースに関するサイトは,それぞれ共通のユーザが利用．同じような時間帯にアクセスが偏る傾向
- TriMineはWebクリックデータを（object,actor,time）の潜在的トピックを発見

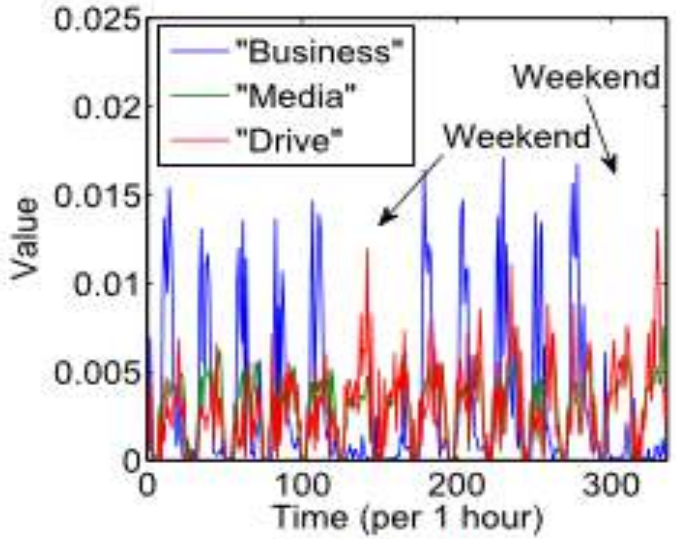
- 非常に少ない情報量から複合イベントの将来予想問題を解決する



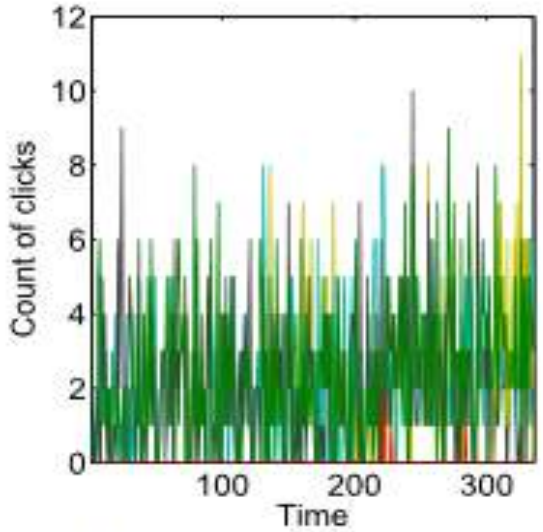
(a) トピック空間における URL の分布
(オブジェクト行列 $\mathbf{O}$)



(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)

2. 問題設定

- 定義1（イベントテンソル）
- $X \in \mathbb{N}^{u \times v \times n}$を3階のイベントテンソルとする．u,v,はそれぞれオブジェクトとアクターの総数，nはイベントシーケンスの長さを表している．Xの要素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表している．

（object,actor,time;count）

↓

（’cnn.com’,’Smith’,’3pm June 1’,’2003’;23）

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している．

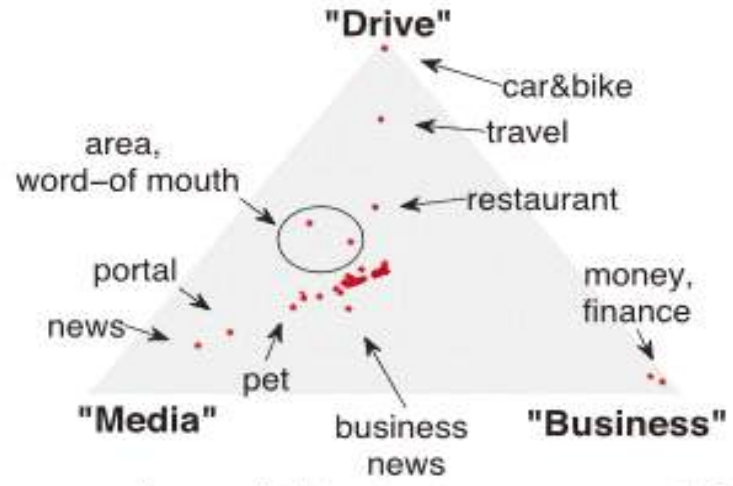
• 定義2（オブジェクト行列O（$u \times k$））

要素$o_{i,j}$ は正の実数とし，各要素の合計を1とする.

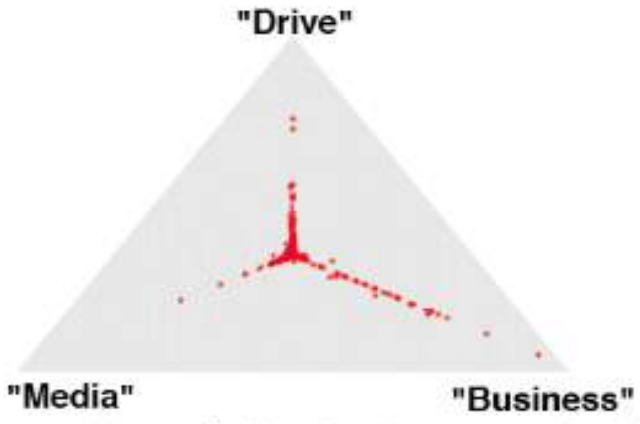
アクター行列A,Cも同様なものとする.

それぞれの行列は（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.



(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**)

2.1問題定義

- 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Xが与えられたとき，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

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.

3提案手法

- 概要
- M次元配列分析（本研究ではM=3）

まず単一のウィンドウサイズを定める（例　l_0 =1時間）．M方向にトピック分析を行う．objects(オブジェクト行列O,u×k),actors(アクター行列A,k×v),time(時間行列C,k×n)の3要素に対しそれぞれ行列を生成

- 多重時間スケールを用いたトピック分析

　複数の時間粒度の行列（$\{C^{(0)}, C^{(1)}, \ldots\ldots\ldots\}$,例えば,分,時,日,週）を生成する．このときOとAも同様である．

単一の時間スケールにおける分析

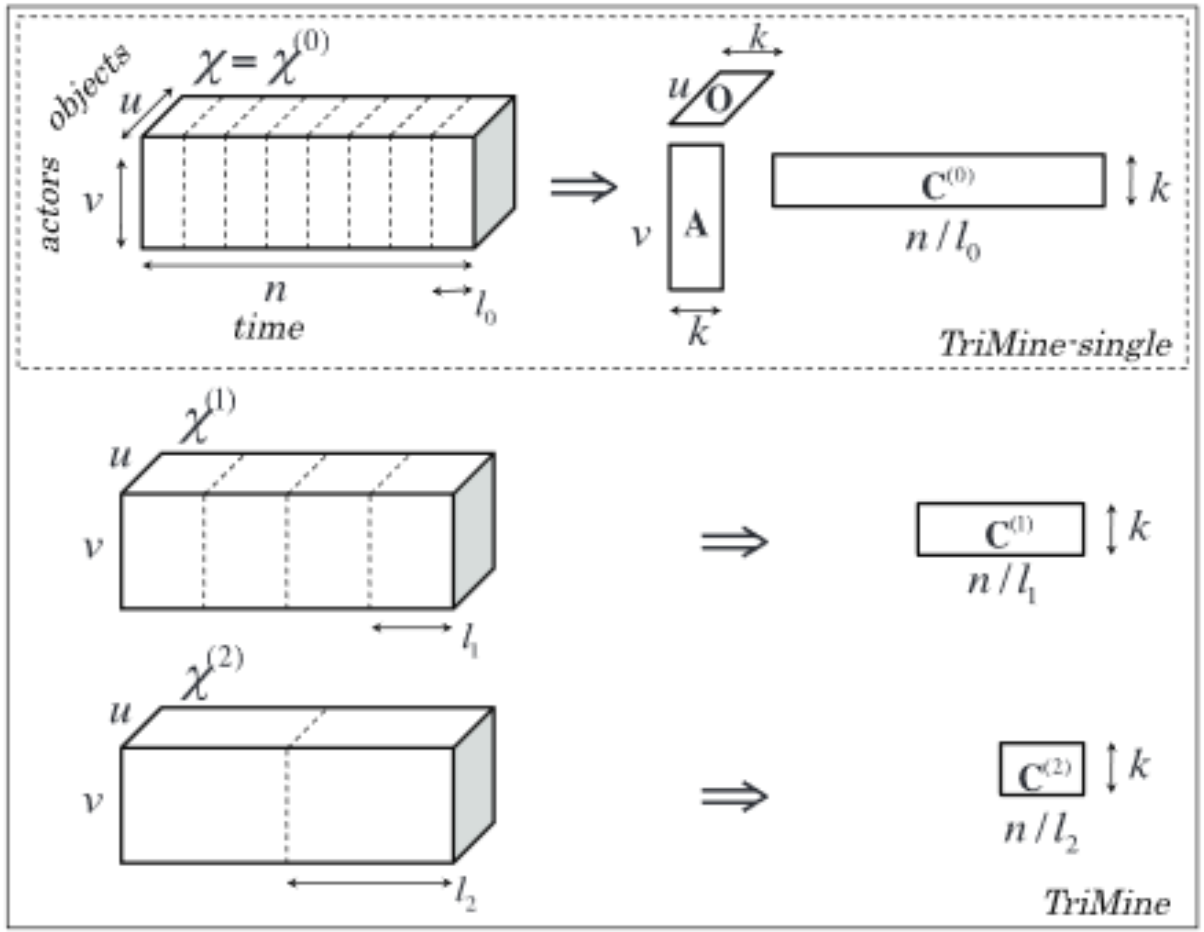
O　オブジェクトとトピック間の関連性の強さ

A　各アクターの頻度確立

C　i番目のトピックにおける時間的な動き

多重時間スケールにおける分析

h=0において$C^{(0)}$について計算したのち，他のウィンドウサイズl_h (h=0,1,…)に対し行列$C^{(h)}$を得る．



TriMineの詳細
パラメータ推定

- テンソルX内における非ゼロ要素$x_{i,j,t}$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．

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

- 行列O，A，C,は以下の式で計算される．

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

- 計算量

テンソルX内のエントリの総数をN（$=\sum_{i,j,t} x_{i,j,t}$）とすると，Nに対し線形，つまりO（N）である.

トピック数　k　学習の反復数 iter

例　　推定コスト　O（iter・kN）　　更新コスト　O（iter・k(u+v+k)）

iter,k,u,v,nは小さい定数のため省略し、まとめると全体の計算コストはNとなる.

- 多重時間スケールにおけるトピック推定

　ウィンドウサイズがlで固定である場合だけでなく複数のウィンドウサイズを用いる.

　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スケールにおけるトピックの近似計算を行う.

- h=0においてテンソル$X^{(0)}$に対し行列O，A，$C^{(0)}$を推定する．続いてほかのレベルに対してO，Aを共有利用し，$C^{(h)}$については以下の式で計算.

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

TriMineの処理の流れ

イベントテンソルX^0内の各エントリに対し，式(1)で潜在的トピックzを割り当て，行列O,A,C^0を推定.

その後すべての時間スケールにおいて，X^0の結果から行列を近似計算する.

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)}
```

4. イベントデータの将来予測
TriMine-F

- TriMine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行う.
- 単一の時間スケールによる時間行列Cの予測

単一のウィンドウサイズの場合ARを用いて時間行列Cの予測を行う.

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

w:係数の数　λ:回帰係数　$\in_t$:ノイズ

- 多重時間スケールによる時間行列Cの予測

　複数のレベルの時間行列($C^{(0)}$, $C^{(1)}$ …)を利用することで複雑な時系列パターンをモデル化する

　以下の式ではlogn個のウィンドウサイズを用いる

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

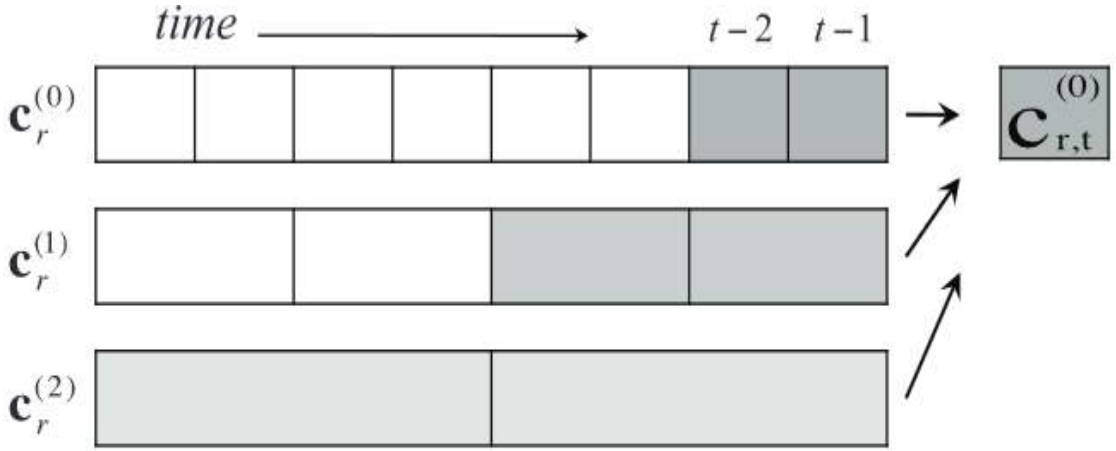


図 3　多重時間スケールにおける時間行列 C の予測 ($l_0 = 1, w = 2$ の場合)．各レベルの色のついたセルを用いて $c_{r,t}^{(0)}$ を予測する

• 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.

時間行列の要素$c_{r,t}$（r=1,….,k）を予測し行列Cを得る

↓

行列O,Aと予測したCの積の総和を計算しイベント数を計算

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

nはイベントの長さ，$\bar{x}$はイベント数mp単位時間当たりの平均値

TriMine-Fの処理の流れ
時間行列の各要素$\hat{c}_{t,r}$を予測
O,A,$\hat{\mathbf{C}}$の行列を用いて
{object,actor,time}の3つ組のエントリ
を生成する

Algorithm 2 EventGeneration

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

/* $\hat{\mathcal{E}}$ is a set of generated entries of form {*object*, *actor*, *time*} */
$\hat{\mathcal{E}} \leftarrow \emptyset$
for each object $i = 1, \ldots, u$ **do**
  for each entry $j = 1, \ldots, n\bar{x_i}$ **do**
    Draw a hidden variable $z_{i,j} \sim Multinomial(\mathbf{O}_i)$
    Draw an actor $e \sim Multinomial(\mathbf{A}_{z_{i,j}})$
    Draw a timestamp $t \sim Multinomial(\hat{\mathbf{C}}_{z_{i,j}})$
    $\hat{\mathcal{E}} = \hat{\mathcal{E}}_t \cup \{i, e, t\}$
  end for
end for
Return $\hat{\mathcal{E}}$

5 評価実験

- 本実験では以下の諸問題について取り組む

(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

- 以下のデータを用いて検証を行う

Webclick: ウェブアクセス履歴のデータ

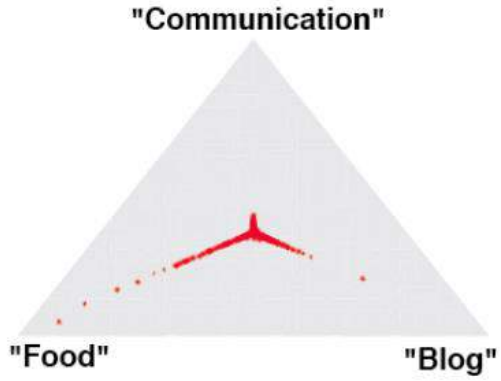
URLID,userID,timeの3つの属性から構成

Ondemand TV: オンラインのTV配信サービスの視聴に関するデータ

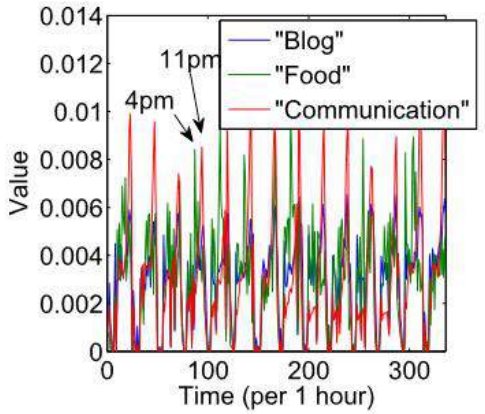
channelID,user/viewerID,timeの3つの属性から構成

Webclick

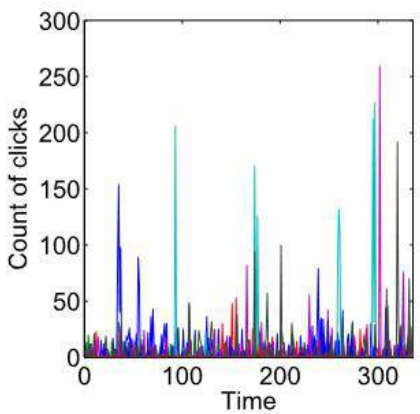
(a) トピック空間における URL の分布



(b) トピック空間におけるユーザの分布

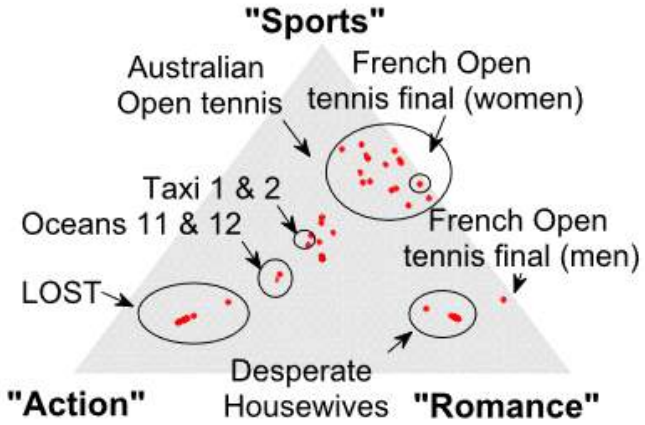


(c) トピックの時間推移

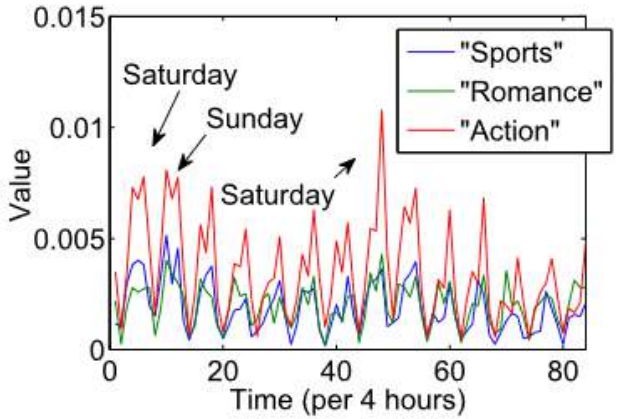


(d) オリジナルのシーケンス

OndemandTV



(a) URL の分布図（行列 $\mathbf{O}$）



(b) トピックの時間推移（行列 $\mathbf{C}$）

予測精度

- TriMine-Fの予測精度についてWebClickデータを用いて検証

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

↓

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l_0 ＝2時間

$X^{(0)}$の長さ　n=168

潜在変数　k=30

予測のために合計40個の係数値を用いる

既存手法との比較

- 本研究では以下の3つの既存手法との比較を行う

(a)AR　予測問題に対する最も基本的な手法

(b)PLiF　線形動的システムに基づく手法

(c)T2　Hongらが提案した新たな提案モデル

パープレキシティ（perplexity）

- TriMine-FとT2に対しパープレキシティを用いた評価実験を行った.

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

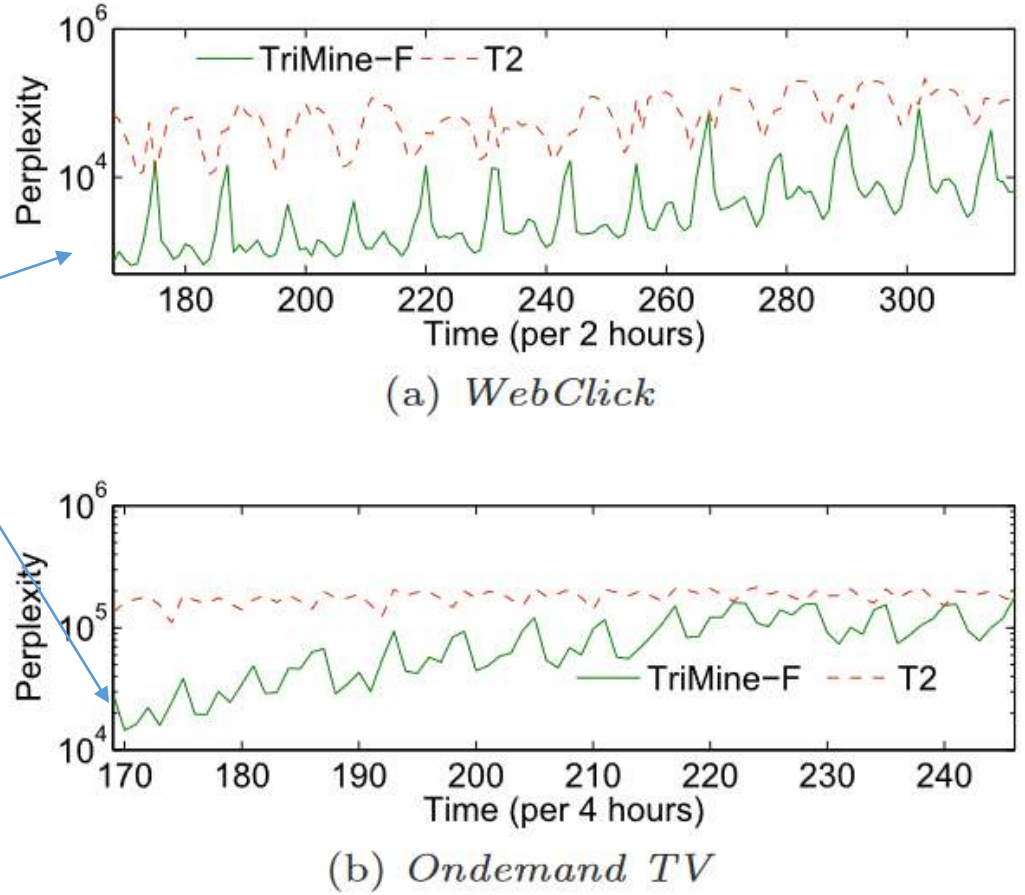
- パープレキシティ

  確率の逆数で定義されており，分岐数，または選択肢の数

→モデルの予測性能を表している

実行結果

低い値は高い
モデル精度を
意味している



TriMine-Fの方がT2よりも周期的なトレンドのみではなく，イベントの急激なパターン変化を捉え，イベントの将来の傾向を予測している

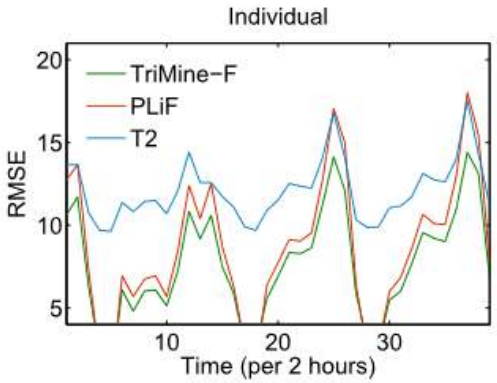
- イベント数の予測に関してAR，PLiF，T2とWebClickデータセットを用いて比較する

a,bは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.

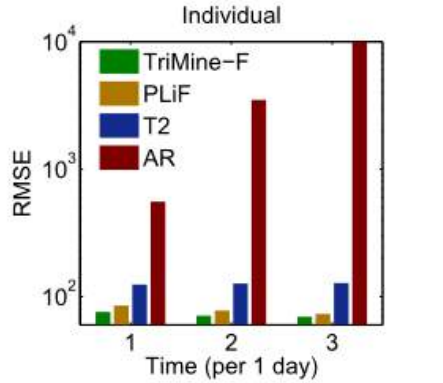
c,dはイベント数の総計を予測した結果

a,cの値の下落は深夜にクリック数が減少するため

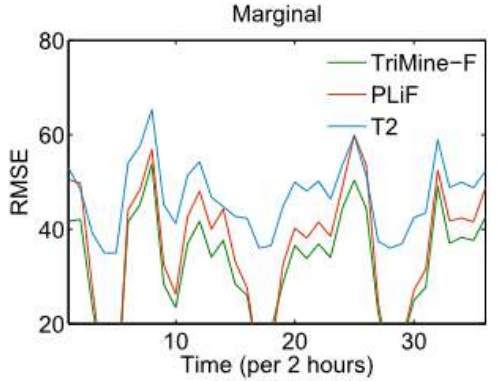
TriMine-Fはどの時間帯においても既存手法より優れた結果となっている.



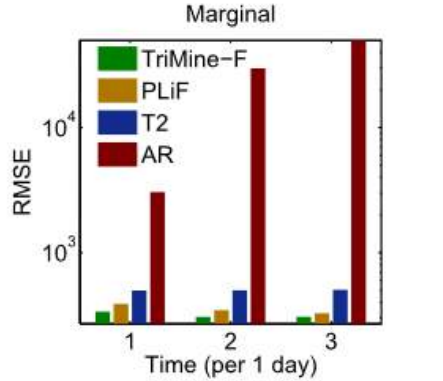
(a) ユーザ，URL ごとのイベント $x_{i,j,t}$ の予測精度 (hourly)



(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)



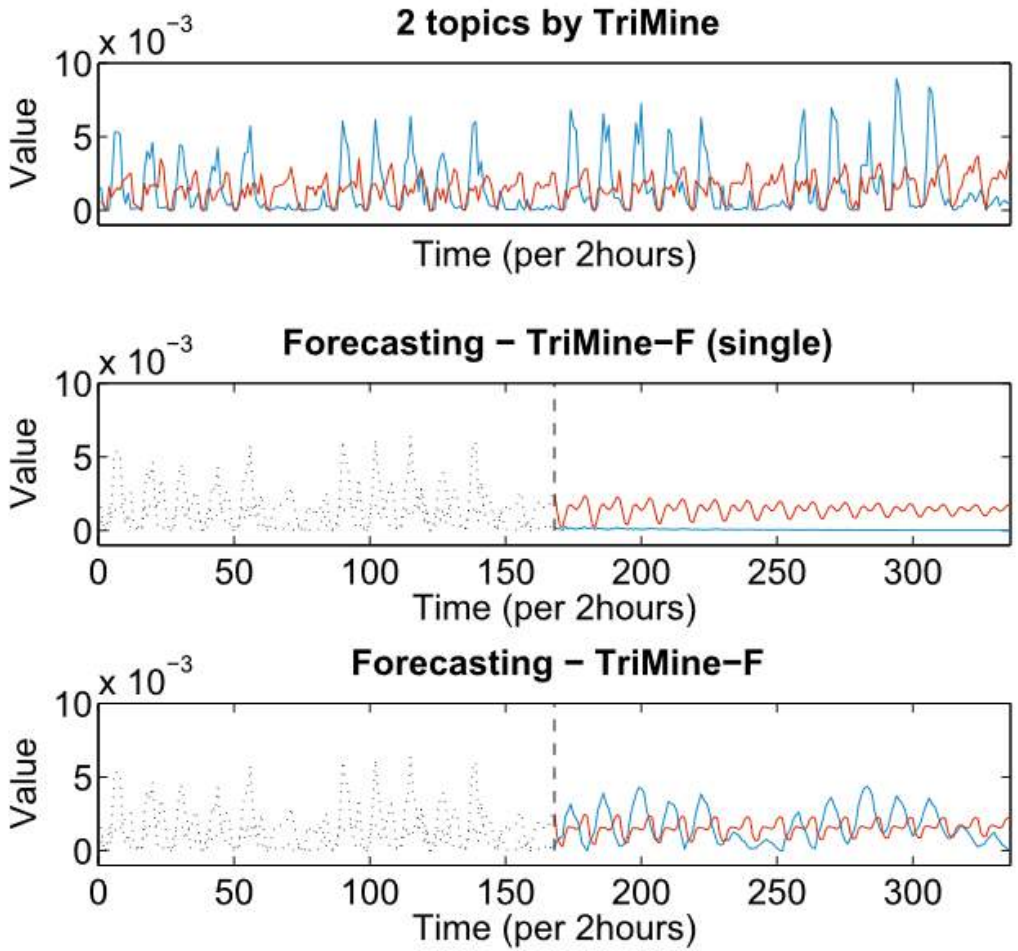
(d) ユーザごとのイベント数 $x_{*,j,t}$ の予測精度 (daily)

- 多重時間スケールのアプローチ効果を調べるためTriMine-FとTriMine-F（single）（多重時間スケールを取り除いたもの）とTriMineの3つを用いて比較する．この実験でもWebClickデータを前実験と同様に用いる．

TriMine-F（single）はイベント予測に失敗し，

収束している

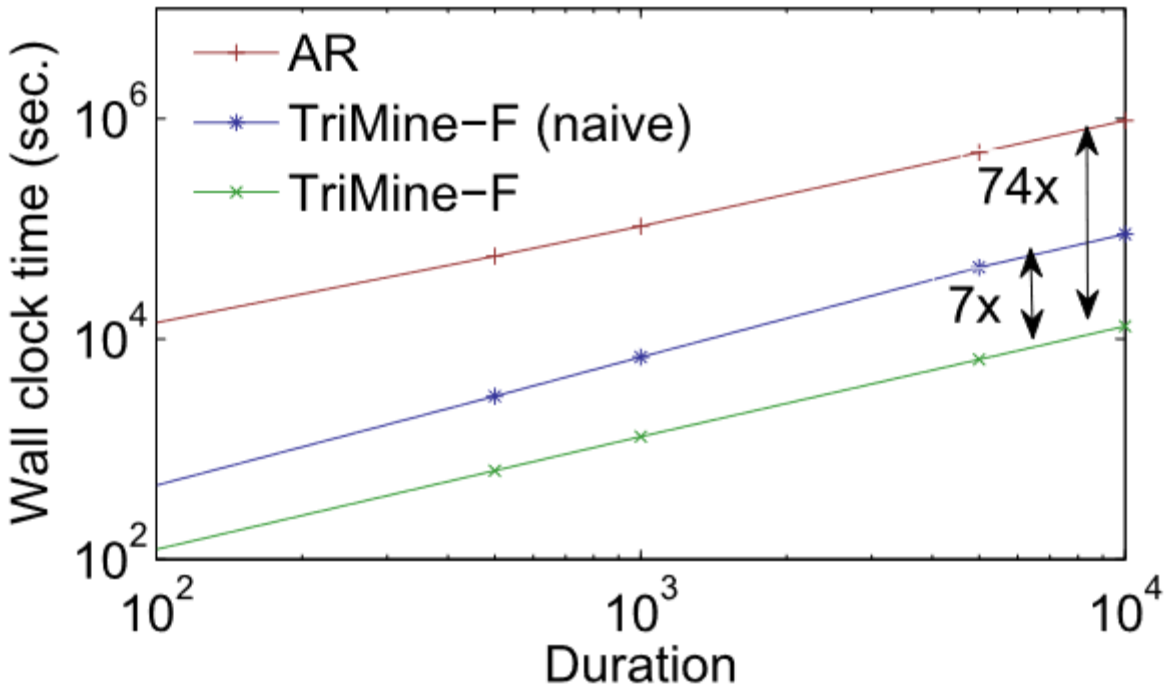
TriMine-Fは数週間の予測に成功している．



- 計算コスト

TriMine-Fの計算コスト（計算時間）に関してARとTriMine-F（native）(高速化アプローチを取り除いたもの)を比較対象として実験する．本実験もWebClickデータを用いており，URL数　u=1000,ユーザ数 v=10000となっている．

図より，TriMine-FはARの７４倍，TriMine-F(naïve)の７倍の高速化を実現している．



6 むすび

- 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，意味づけを行うことができる.
- TriMine-Fは効率的にイベントの予測を行うことができ，計算コストはデータベースサイズに線形である.
- 今後の課題は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みられるSNSデータ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.