KDD 2012 BEIJING
THE 18TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING
Beijing, China
August 12-16, 2012

Fast Mining and Forecasting of Complex Time-Stamped Events

Yasuko Matsubara
Kyoto University
y.matsubara@db.soc.i.kyoto-u.ac.jp

Yasushi Sakurai
NTT Communication Science Labs
yasushi.sakurai@acm.org

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Tomoharu Iwata
NTT Communication Science Labs
iwata.tomoharu@lab.ntt.co.jp

Masatoshi Yoshikawa
Kyoto University
yoshikawa@i.kyoto-u.ac.jp

Motivation

Complex time-stamped events

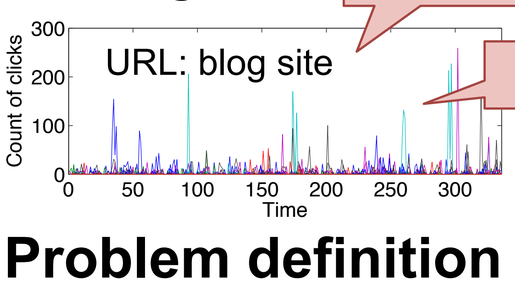
Web click events: *{**time**, url, user, access device, http referrer,…}*

Timestamp	URL	User	Device
2012-08-01-12:00	CNN.com	Smith	iphone
2012-08-02-15:00	YouTube.com	Brown	iphone
2012-08-02-19:00	CNET.com	Smith	mac
…	…	…	…

Any trends?

Can we forecast future events?

Challenges



Original access counts (100 users, 1 week)

☹ We cannot see any trends!!

Problem definition

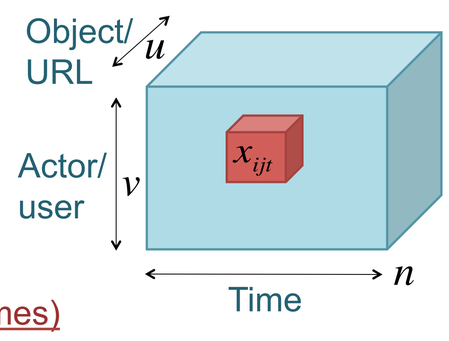
Given: A sequence of complex time-stamped events
Goal: (1) Find major topics (2) Forecast future events

Proposed method: TriMine

Main idea (1): M-way analysis

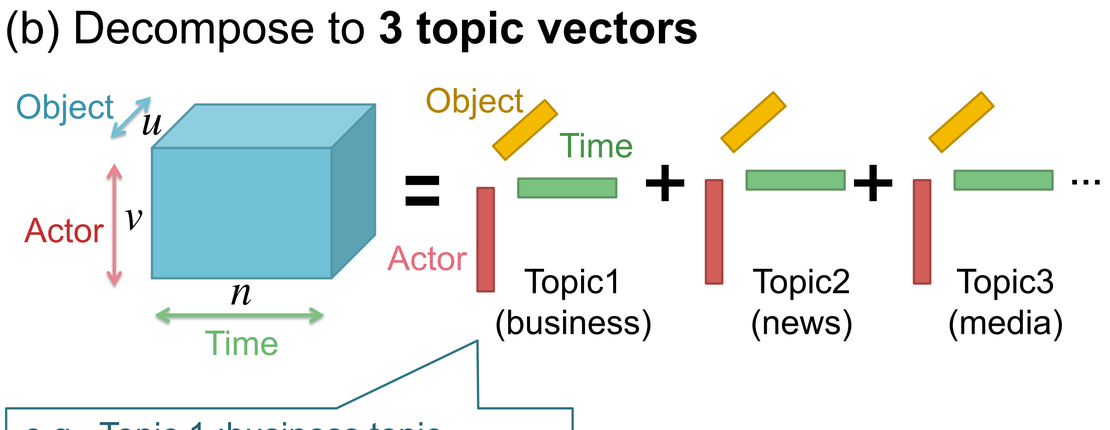
(a) Complex time-stamped tensor

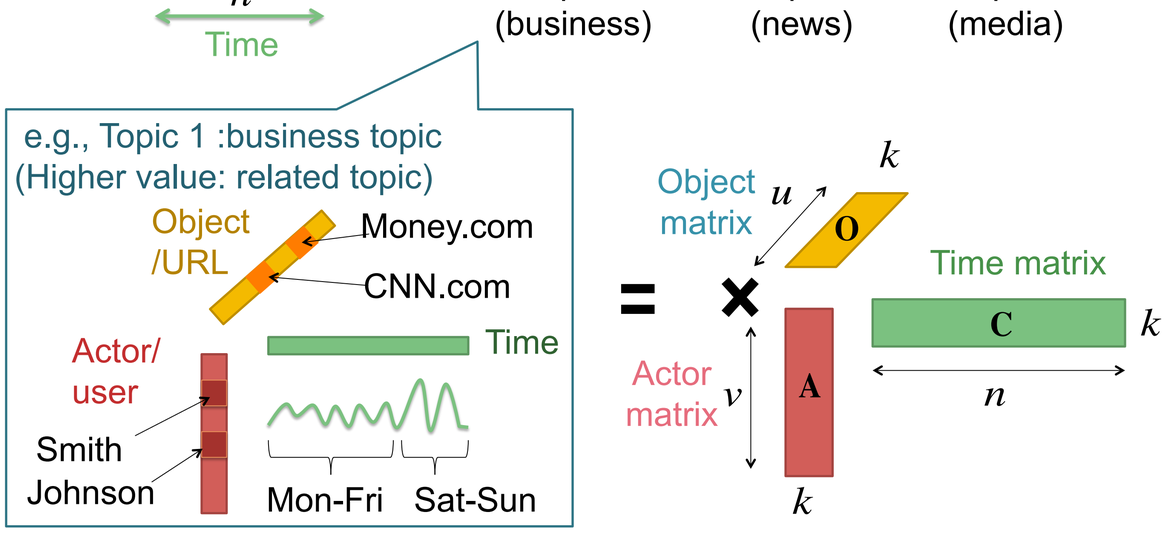
M[th] order tensor (M=3)
cube {object x actor x time}
x_{ijt} : event count of object i, actor j at time t



e.g., (‘Smith’, ‘CNN.com’, ‘Aug1 12pm’; 21 times)

(b) Decompose to **3 topic vectors**



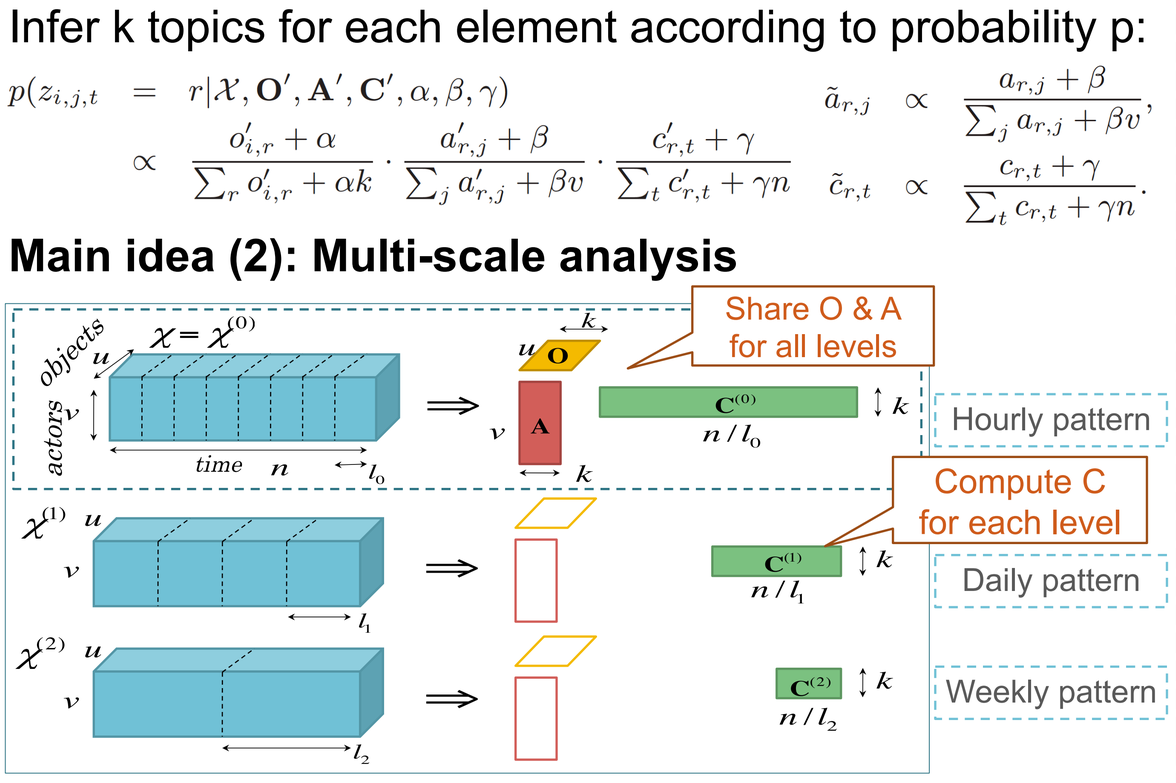


(c) Inference (Gibbs sampling)
Infer k topics for each element according to probability p:

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

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

Main idea (2): Multi-scale analysis

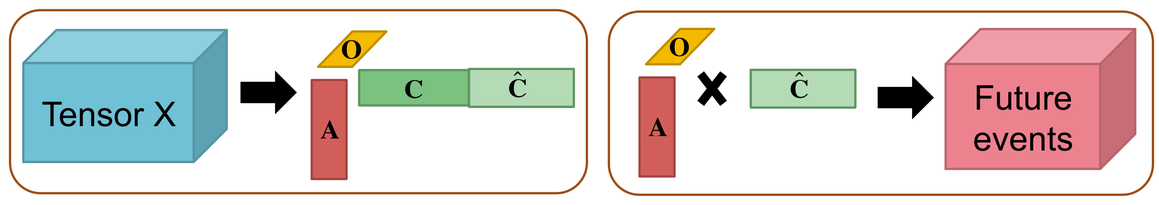


TriMine is linear on the input size N, i.e., $O(N \log n) \rightarrow O(N)$
(N: counts of events in X, n: duration of X)

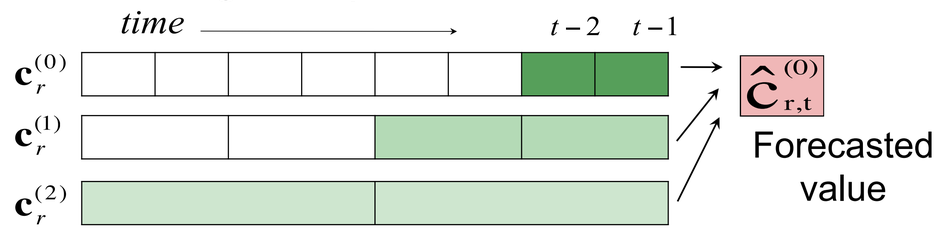
TriMine-F – forecasts

Use topic matrices O, A, C

[Step 1] [Step 2]



[step1] Forecast time-topic matrix C’
Forecast C’ using multiple levels of matrices

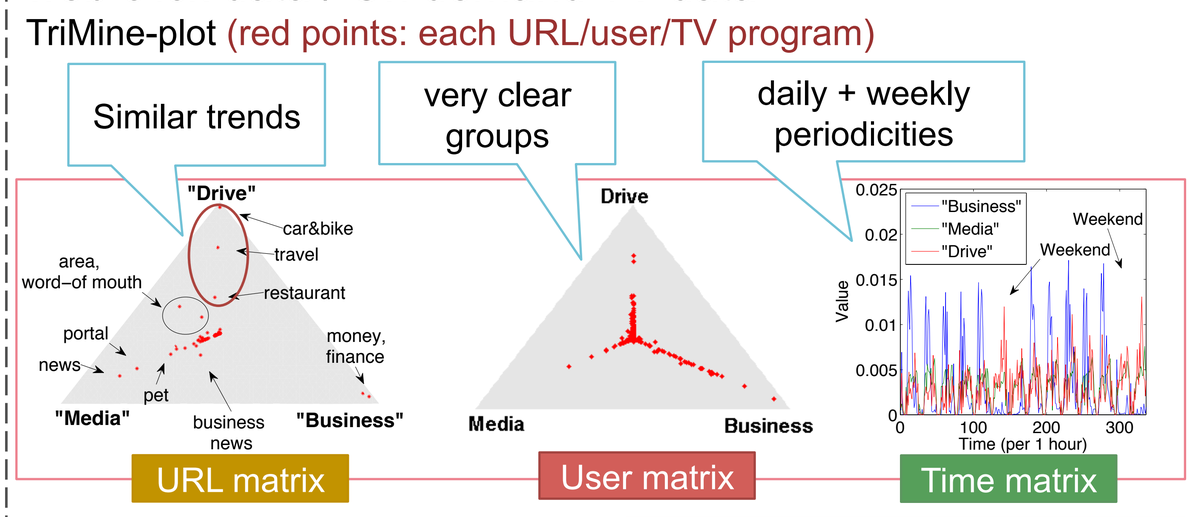


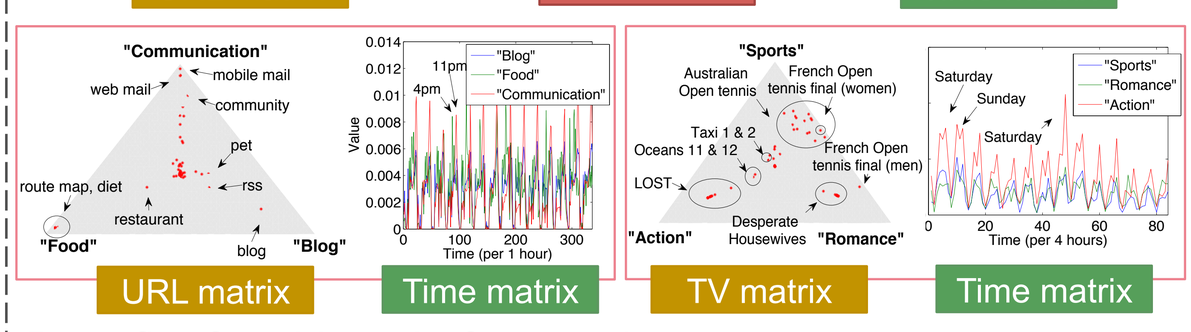
[step2] Generate events using three matrices O, A, C’
(a) Count estimation – X’ = O * A * C’
(b) Complex event generation – sampling approach

Experiments

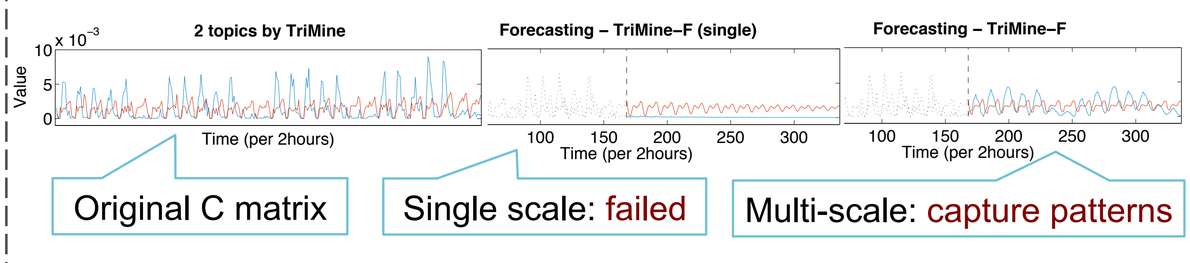
web click data / On demand TV data

TriMine-plot (red points: each URL/user/TV program)





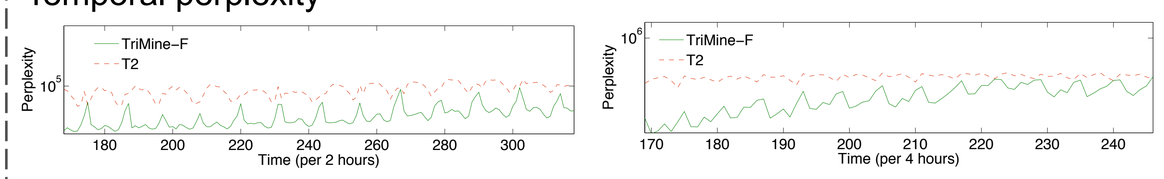
Benefit of multi-scale forecasting



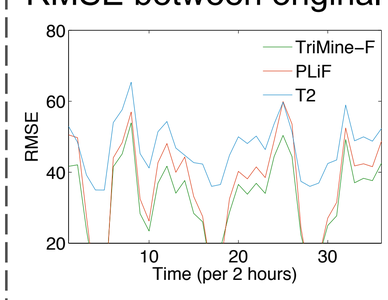
Forecasting accuracy

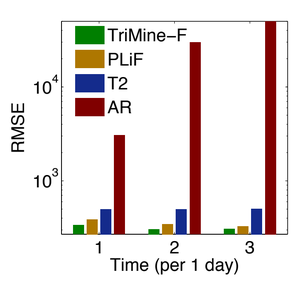
Temporal perplexity

T2: [Hong et al.’11], PLiF [Li et al.’10]

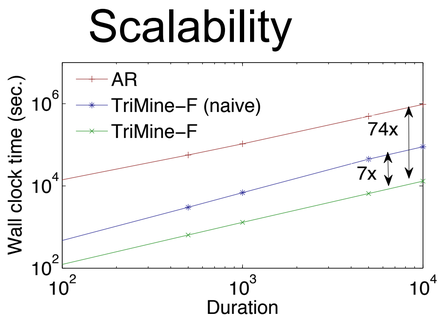


RMSE between original & forecasted events





Scalability



Conclusions

We addressed the problem of **complex event mining**
TriMine has following properties:

- Effective: It finds meaningful patterns in real datasets
- Accurate: It enables forecasting
- Scalable: It is linear on the database size

Code: http://www.kecl.ntt.co.jp/csl/sirg/people/yasuko/software.html