Fast Mining and Forecasting of Complex Time-Stamped Events

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

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 referer, ...]

Goals:
1. Find major topics
2. Forecast future events

Challenge:
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
Mth order tensor (M=3) cube (object x actor x time)
Tijl - event count of object i, actor j at time l

(b) Decompose to 3 topic vectors
Object/URL = Time + Topic1 + Topic2 + Topic3

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

\[ p(a_{ij}, k) = \frac{1}{Z} \exp \left( \sum_{i=1}^{M} \alpha_i T_{ijl} + \sum_{j=1}^{N} \beta_j A_{ij} + \sum_{l=1}^{Q} \gamma_l \right) \]

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] Forecast time-topic matrix C' using multiple levels of matrices
(a) Count estimation \(- X = O \times A \times C'\)
(b) Complex event generation – sampling approach

[Step 2] Generate events using three matrices O, A, C'

Experiments
web click data / On demand TV data
TriMine-plot (red points: each URL/user/TV program)

Similar trends
very clear groups
daily + weekly periodicities

Benefits of multi-scale forecasting
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

Notes: