Mining and Forecasting of Big Time-series Data

Yasushi Sakurai

Sakurai Lab.

@ Kumamoto University
Big time-series data

Social/natural phenomena

climate

Big Data

IoT, sensors

Web

Economy

Epidemic
Big time-series data

Social/natural phenomena

climate

Web

Economy

Environmental sensors
- Temperature,
- Air pressure, etc.

Application
- Agriculture

IoT, sensors
Big time-series data

Social/natural phenomena

climate

Online user activities
- Web clicks,
- Online reviews

Application
- Market analysis

Economy

Epidemic
Big time-series data

Social/natural phenomena
- Climate
- Web
- Economy

Sensor streams
- Vibration, Acceleration

Application
- Self-driving car
- Structural health monitoring
Motivation

• **Given:** Big time-series data
  – Sensor data
  – Web, online activities
  – Medical data

• **Goal:** Real-time forecasting

• **At-work:**
  – Online marketing
  – Sensor monitoring
  – Forecasting future epidemics
New challenges

Time-series data analysis

Indexing, Similarity search
- ED, DTW, SAM

Feature extraction
- DFT, DWT, SVD, ICA

Linear-modeling
- AR, LDS, ARIMA

Data stream
- Correlation monitoring etc.

Data (X)  Model (M)

Sakurai Lab. © 2016 Yasushi Sakurai
New challenges

They cannot handle “Big” data, i.e.,

Non-linear phenomena

Complex events

Real-time forecasting

Automatic parameter-tuning

Model (M)

Data (X)

Parameter-tuning

Correlation monitoring

Feature extraction

Data stream

Linear modeling

They cannot handle “Big” data, i.e.,

Y

X

PCA

ICA

DTW

ED,

SAM

SVD,

ICA

AR,

LDS,

ARIMA

DFT,

DWT,

Indexing,

Similarity search

Sakurai Lab. © 2016 Yasushi Sakurai
They cannot handle “Big” data, i.e.,

Non-linear phenomena

Complex events

Real-time forecasting

Automatic parameter-tuning

Automatic parameter-tuning

Model (M)

Data (X)

StatStream

Correlation

AR, ARIMA, LDS

ED, DTW

DFT, DWT, SVD, ICA

PCA

ICA

DTW

Non-linear phenomena

Complex events

Real-time forecasting

Automatic parameter-tuning

Model (M)

Data (X)

StatStream

Correlation

AR, ARIMA, LDS

ED, DTW

DFT, DWT, SVD, ICA

PCA

ICA

DTW
New research directions

- Time-series data analysis
  - Indexing and fast searching
  - Sequence matching
  - Clustering
  - etc.

- New research directions
  1. Automatic mining
  2. Non-linear modeling
  3. Large-scale tensor analysis

NO magic #
(R1) Automatic mining

No magic numbers! ... because,

Manual
- sensitive to the parameter tuning
- long tuning steps (hours, days, ...)

Automatic (no magic numbers)
- no expert tuning required

Big data mining:

-> we cannot afford human intervention!!
“Automatic” mining algorithm

Find: compact description of data X

Given:
- Chicken dance
- Left/right legs
- Left/right arms
- Tail feathers
- Claps
- Beaks
- Wings

Find:
1. Beaks
2. Claps
3. Wings
4. Tail feathers
(R1) Automatic mining

“Automatic” mining algorithm

Idea (1): Multi-level chain model

- HMM-based probabilistic model
- with “across-regime” transitions
(R1) Automatic mining

Idea(2): Minimize encoding cost!

$$\min \left( \text{Cost}_M(M) + \text{Cost}_C(X|M) \right)$$

Model cost

Coding cost

Good compression

Good description

CostM

CostC

CostT

(# of r, m)
(R1) Automatic mining

"Automatic" mining algorithm

AutoPlait (NO magic numbers)
(R2) Non-linear (gray-box) modeling

- Gray-box mining
  - If we know the equations
- Non-linear (differential) equations
  - Epidemic
  - Biology
  - Physics, Economics, etc.
- Modeling non-linear social phenomena
  - Non-linear analysis for Web online activities (e.g., information diffusion, interaction)
(R2) Non-linear (gray-box) modeling

News spread in social media

Number of mentions in blogs/Twitter

Breaking news

News spread

Decay (power law)

(per hour, 1 week)
(R2) Non-linear (gray-box) modeling

News spread in social media

1. Un-informed bloggers
2. External shock
3. Infection (word-of-mouth)

Time $n=0$ \rightarrow Time $n=n_b$ \rightarrow Time $n=n_b+1$

$\beta$
(R2) Non-linear (gray-box) modeling

News spread in social media

1. Un-informed bloggers
2. External shock

Decay function:

\[ f(n) = \beta \cdot n^{-1.5} \]

Power law

- Linear
- Log
Non-linear (gray-box) modeling

News spread in social media

Memes in social media/blogs

Sakurai Lab. © 2016 Yasushi Sakurai
Non-linear (gray-box) modeling

Forecast tail-part and rise-part!

Given (1) first spike,
(2) release date of two sequel movies
(3) access volume before the release date
(R2) Non-linear (gray-box) modeling

Online competition

Interactions between keywords

Fitting result - RMSE=0.058817
Online competition

(Google Search volume)

Fitting result - RMSE=0.058817

Volume @ time

Time (weekly)


PlayStation  Wii  Android

Xbox
(R2) Non-linear (gray-box) modeling

The Web as a Jungle!

Ecosystem on the Web

Ecosystem in the Jungle
(R2) Non-linear (gray-box) modeling

The Web as a Jungle!

--- Non-Linear equations ---

\[ P_i(t + 1) = P_i(t) \left[ 1 + r_i \left( 1 - \frac{\sum_{j=1}^{d} a_{ij}P_j(t)}{K_i} \right) \right] \]
Ecosystem on the Web

Jungle

- Biological species
- Food resources
- Population
- Climate /season

Web

- Online activities
- User resources
- Popularity
- Annual events (e.g., Xmas)

EcoWeb WWW’15

Sakurai Lab. © 2016 Yasushi Sakurai
(R2) Non-linear (gray-box) modeling

The Web as a Jungle!

Interactions between keywords

Fitting result - RMSE=0.058817

Sakurai Lab. © 2016 Yasushi Sakurai
(R2) Non-linear (gray-box) modeling

The Web as a Jungle!

Seasonality

Fitting result - RMSE=0.058817

Time (weekly)

Time (weekly)

PlayStation  Wii  Android

Xbox

Sakurai Lab.  © 2016 Yasushi Sakurai
(R3) Large-scale tensor analysis

• Time-stamped events
  – *e.g.*, *web clicks*

<table>
<thead>
<tr>
<th>Time</th>
<th>URL</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>08-01-12:00</td>
<td>CNN.com</td>
<td>Smith</td>
</tr>
<tr>
<td>08-02-15:00</td>
<td>YouTube.com</td>
<td>Brown</td>
</tr>
<tr>
<td>08-02-19:00</td>
<td>CNET.com</td>
<td>Smith</td>
</tr>
<tr>
<td>08-03-11:00</td>
<td>CNN.com</td>
<td>Johnson</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Represent as $M^\text{th}$ order tensor ($M=3$)

\[
\mathbf{X} \in \mathbb{N}^{u \times v \times n}
\]

Element $x$: # of events

- *e.g.*, ‘Smith’, ‘CNN.com’, ‘Aug 1, 10pm’; 21 times
(R3) Large-scale tensor analysis

Complex time-stamped events
\{time, URL, user ID, access devices, http referrer,...\}

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>URL</th>
<th>User</th>
<th>Device</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012-08-01-12:00</td>
<td>CNN.com</td>
<td>Smith</td>
<td>iphone</td>
</tr>
<tr>
<td>2012-08-02-15:00</td>
<td>YouTube.com</td>
<td>Brown</td>
<td>iphone</td>
</tr>
<tr>
<td>2012-08-02-19:00</td>
<td>CNET.com</td>
<td>Smith</td>
<td>mac</td>
</tr>
<tr>
<td>2012-08-03-11:00</td>
<td>CNN.com</td>
<td>Johnson</td>
<td>ipad</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2012-08-05-12:00</td>
<td>CNN.com</td>
<td>Smith</td>
<td>iphone</td>
</tr>
<tr>
<td>2012-08-05-19:00</td>
<td>CNET.com</td>
<td>Smith</td>
<td>iphone</td>
</tr>
</tbody>
</table>
(R3) Large-scale tensor analysis

Complex time-stamped events

Object/
URL

Actor/
user

Web clicks

= 

Object

Time

Actor

(business) +

(news) +

(media)...

TriMine
KDD’12

© 2016 Yasushi Sakurai
Large-scale tensor analysis

Complex time-stamped events

Higher value: Highly related topic

e.g., business topic vectors

Object/URL

Actor/user

Smith Johnson

Money.com

CNN.com

Time

Mon-Fri

Sat-Sun

(business)

(news)

(media)
(R3) Large-scale tensor analysis

Complex time-stamped events

Noisy 😞  (Original data)  Sparse 😞

URL matrix
User matrix
Time matrix

TriMine KDD’12
New research directions

1. Automatic mining
2. Non-linear (gray-box) modeling
3. Large-scale tensor analysis

New challenge: MANT analysis

Multi-Aspect Non-linear Time-series
New challenge: MANT analysis

Non-Linear tensor analysis

P1 Seasonality
P4 External shocks
P5 Mistakes, errors
P3 Local patterns
P2 Vaccination
New challenge: MANT analysis

Non-Linear tensor analysis

Given:
Tensor $\mathcal{X}$ (disease x location x time)

Find:
Compact description of $\mathcal{X}$, "automatically"
New challenge: MANT analysis

Non-Linear tensor analysis

Given: Tensor $X$ (disease x location x time)

Find: Compact description of $X$, "automatically"

- Seasonality
- Discontinuities

$X = \begin{pmatrix} B & R & N \end{pmatrix} \begin{pmatrix} P1 & P2 & P3 \end{pmatrix}$

© 2016 Yasushi Sakurai
New challenge: MANT analysis

Non-Linear mining & forecasting of local competition

(a) Local competition strength between Kindle and Nexus

(b) Original/fitted search volumes for Kindle and Nexus

(c) Seasonal/annual patterns for each location/country
New challenge: MANT analysis

Time-stamped events:
\{activity, location, time\}

CompCube WWW’16

© 2016 Yasushi Sakurai
Sakurai Lab.
New challenge: MANT analysis

Given: Tensor $\mathbf{X}$
(activity x location x time)

Find: Compact description of $\mathbf{X}$

$\mathbf{X} = \text{CompCube} \mathbf{B} \mathbf{C} \mathbf{S} \mathbf{D}$
New challenge: MANT analysis

Given: Tensor $\chi = \text{CompCube}$

Find: Compact description of $\chi$

$\chi$ = CompCube

Basics

Competition

Seasonality

Deltas

© 2016 Yasushi Sakurai

Sakurai Lab.
Modeling power of CompCube

CompCube WWW’16

Google search for Kindle, Nexus

Local Competition strength

© 2016 Yasushi Sakurai
Modeling power of CompCube

Google search for Kindle, Nexus

Local Competition strength

© 2016 Yasushi Sakurai
Modeling power of CompCube

Local seasonality for iPod

Component #1

Component #2

© 2016 Yasushi Sakurai
Modeling power of CompCube

Local seasonality for iPod

Dec.

Xmas

Feb.

Chinese New Year

Component #1

Component #2

© 2016 Yasushi Sakurai
Modeling power of CompCube

Fitting for News resources

Fitting result – RMSE = 0.056

Detected!

Wikipedia
Modeling power of CompCube

Fitting for News resources

Weak/Strong

Local attention to US election

US election Nov. 2008

Sakurai Lab. © 2016 Yasushi Sakurai
New challenge: MANT analysis

Non-Linear mining & forecasting of local competition

3. Forecast future activities

Future

Time

CompCube (RMSE=0.2600)

FUNNEL (RMSE=0.4525)

SARIMA+ (RMSE=0.4369)

TBATS (RMSE=0.5839)
Putting it all together

New research directions
1. Automatic mining
2. Non-linear (gray-box) modeling
3. Large-scale tensor analysis

New challenge: MANT analysis

Multi-Aspect Non-linear Time-series

Sakurai Lab. © 2016 Yasushi Sakurai
• Bridge structural health monitoring
  – Structural monitoring using vibration/shock sensors
  – Keep track of lag correlations for sensor data streams
• **Bridge structural health monitoring**
  – Goal: real-time anomaly detection for disaster prevention
  – Several thousands readings (per sec) from several hundreds sensor nodes

- Uses BRAID
- Metropolitan Expressway (Tokyo, Japan)
Industrial contribution

• Bridge structural health monitoring with BRAID

Metropolitan Expressway (Tokyo, Japan)

Tokyo Gate Bridge (Tokyo, Japan)

Can Tho Bridge (Vietnam)
Industrial contribution

– Automobile sensor data
  \{ \text{trip ID}, \text{zone ID}, \alpha \}

TrailMarker

Original tensor $X$

Object (sensors)

velocity

Longitudinal/
lateral acceleration
Industrial contribution

– Automobile sensor data

\{ \text{trip ID}, \text{zone ID}, a \}

TrailMarker

Original tensor $X$

Trip 1

Trip 2

Trip 3

zone

zone

zone
Industrial contribution

— Automobile sensor data

\{ trip ID, zone ID, object \}

TrailMarker
Thank you

Code/Data/Tutorial: Sakurai Lab.
http://www.cs.kumamoto-u.ac.jp/~yasushi/

Fitting result - RMSE=0.058817

NO magic!