Mining Big Time-series Data on the Web

Yasushi Sakurai (Kumamoto University)
Yasuko Matsubara (Kumamoto University)
Christos Faloutsos (Carnegie Mellon University)

Roadmap

- Motivation
- Similarity search, pattern discovery and summarization
- Non-linear modeling and forecasting
- Extension of time-series data: tensor analysis

Part 2

Problem
- Why: “non-linear” modeling

Fundamentals
- Non-linear ("gray-box") models

Applications
- Epidemics
- Information diffusion
- (Online) competition

Non-linear mining and forecasting

Q. What are “non-linear phenomena”? Example: logistic parabola

Models population of flies [R. May/1976]

\[ x_{t+1} = a x_t \cdot (1 - x_t) \]

How to forecast?

Solution 1

Linear equations, e.g., AR, ARIMA, …
How to forecast?

Solution 1
Linear equations, e.g., AR, ARIMA, ...

Details @ part 1

\[ x_{t+1} = ax_t \cdot (1 - x_t) \]

Solution 2
“Delayed Coordinate Embedding”

= Lag Plots [Sauer92]
- Based on k-nearest neighbor search

General Intuition
(Lag Plot)

Solution 2

Lag fit: fails

\[ x_{t+1} = ax_t \]

Interpolate these…

To get the final prediction

\[ x_{t+1} = ax_t + \epsilon \]

Forecasting results
(Lag Plot)

Solution 2

Logistic parabola

LORENZ

Laser

Forecast

Forecasted \( x_{t+1} \) (green)

Original \( x_t \) (red)

Solution 2

“Delayed Coordinate Embedding”

= Lag Plots [Sauer92]
- Based on k-nearest neighbor search
- Non-linear Forecasting!
How to forecast?

**Solution 2**

“Delayed Coordinate Embedding”

“Black-box” mining (we don't know the equations)

But, still,... Hard to interpret

\[ x_{t+1} = ax_t \cdot (1 - x_t) \]

How to forecast?

**Solution 3**

“Gray-box” mining (if we know the equations)

Non-linear modeling!

\[ x_{t+1} = ax_t \cdot (1 - x_t) \]

How to forecast?

**Solution 3**

Non-linear equations

Big Time series

Part 2 Roadmap

**Problem**

Why: “non-linear” modeling

**Fundamentals**

- Non-linear (grey-box) models

**Applications**

- Epidemics
- Information diffusion
- (Online) competition

Part 2 Roadmap

**Problem**

Why: “non-linear” modeling

**Fundamentals**

- Non-linear (grey-box) models

- Logistic function
- Lotka-Volterra (prey-predator, competition)
- SI, SIR models, etc.
- Lorenz equations, etc.
Grey-box mining and non-linear equations

Information diffusion
Convection
Population growth
Competition
Epidemics

Big Time series

Logistic function
So-called “Verhulst” model (=sigmoid, =Bass)
- Population expansion with limited resources

Species
Foods
t=0
t=1
t=2

Prey (H)
Predator (P)

Lotka-Volterra equations
So-called “prey-predator” model

H : count of prey (e.g., hare)
• P : count of predators (e.g., lynx)
Lotka-Volterra equations

So-called “prey-predator” model

\[
\frac{dH}{dt} = rH - aHP
\]

\[
\frac{dP}{dt} = bHP - mP
\]

- \(H\): count of prey (e.g., hare)
- \(P\): count of predators (e.g., lynx)

Solution to the Lotka-Volterra equations.

\[
\frac{dP_i}{dt} = r_i P_i \left( 1 - \sum_{j=1}^{d} a_{ij} P_j \right) \frac{K_i}{P_i} (i = 1, \ldots, d)
\]

- \(a_{ij}\): Interaction coefficient
  - i.e., effect rate of species \(j\) on \(i\)

Extension: “Competitive” Lotka-Volterra equations

Competition between multiple (\(d\)) species

- Squirrel monkeys
- Spider monkeys
- Macaws
- Capybaras
- Fruits
- Nuts
- Grass

“Competition” in the Jungle

“Competitive” Lotka-Volterra equations

- Biological interaction
  - Table: Type of interaction
    - 0: no effect
    - -: detrimental
    - +: beneficial

<table>
<thead>
<tr>
<th>Species B</th>
<th>+</th>
<th>0</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>Mutualism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>Commensalism</td>
<td>Neutralism</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Antagonism</td>
<td>Amensalism</td>
<td>Competition</td>
</tr>
</tbody>
</table>
Grey-box mining and non-linear equations

Epidemics: Susceptible-Infected (SI) model
Each node is in one of two states

\[ S \quad \text{Susceptible (healthy)} \]
\[ I \quad \text{Infected} \]

Epidemics: Susceptible-Infected (SI) model
Each node is in one of two states

\[ S \quad \text{Susceptible (healthy)} \]
\[ I \quad \text{Infected} \]

Epidemics: Susceptible-Infected (SI) model
Each node is in one of two states

\[ S \quad \text{Susceptible (healthy)} \]
\[ I \quad \text{Infected} \]

Epidemics: Susceptible-Infected (SI) model
Each node is in one of two states

\[ S \quad \text{Susceptible (healthy)} \]
\[ I \quad \text{Infected} \]

\[ \frac{dS}{dt} = -\beta SI \]
\[ \frac{dI}{dt} = +\beta SI \]
\[ N = S(t) + I(t) \]
\[ \beta : \text{Infection strength} \]
\[ N : \text{Population size} \]
i.e., \[ \frac{dI}{dt} = \beta(N - I)I \]
Epidemics: Susceptible-Infected (SI) model

Each node is in one of two states

\[
\frac{dI}{dt} = \beta N \cdot I \cdot (1 - \frac{I}{N})
\]

i.e.,

\[
\frac{dI}{dt} = \beta (N - I)I
\]

Susceptible-Infected-recovered (SIR) model

Recovered with immunity

\[ S - \text{Susceptible (healthy)} \]
\[ I - \text{Infected} \]
\[ R - \text{Recovered (immune)} \]

\[ \beta \] : Infection rate
\[ \delta \] : Recovery rate

Details

Susceptible-Infected-recovered (SIR) model

Recovered with immunity

Susceptible-Infected-recovered (SIR) model

Recovered with immunity

Susceptible-Infected-recovered (SIR) model

Recovered with immunity

Details

Susceptible-Infected-recovered (SIR) model

Recovered with immunity

Details

Susceptible-Infected-recovered (SIR) model

Recovered with immunity

Details

Susceptible-Infected-recovered (SIR) model

Recovered with immunity

Details
Susceptible-Infected-recovered (SIR) model

Recovered with immunity

\[ S(t) \quad I(t) \quad R(t) \]

\[ \frac{dS}{dt} = -\frac{\beta SI}{N} \]
\[ \frac{dI}{dt} = \frac{\beta SI}{N} - \delta I \]
\[ \frac{dR}{dt} = \delta I \]

\[ S(t) + I(t) + R(t) = N \]

\( \beta \) : Infection rate
\( \delta \) : Recovery rate

Phase plane: \( S(t) \) vs. \( I(t) \)

Other epidemic models

Other virus propagation models ("VPM")

- SIS : susceptible-infected-susceptible, flu-like
- SIRS : temporary immunity, like pertussis
- SEIR : mumps-like, with virus incubation (E = Exposed)
- SEIR-birth/death: with birth/death rate

Underlying contact-network
- 'who-can-infect-whom'

Grey-box mining and non-linear equations

Information diffusion
Convection
Population growth
Competition
Big Time series
Epidemics
Other non-linear models

LORENZ: eqs. for atmospheric convection

\[
\begin{align*}
\frac{dx}{dt} &= \sigma(y - x) \\
\frac{dy}{dt} &= x(\rho - z) - y \\
\frac{dz}{dt} &= xy - \beta z
\end{align*}
\]

- \(x\): convective intensity
- \(y\): temperature difference between ascending and descending currents
- \(z\): difference in vertical temperature profile from linearity

Lorenz attractor

Butterfly effect (chaos)

Van del Pol oscillator
- Electric circuits, heart-beats, neurons

FitzHugh-Nagumo model
- An excitable system (e.g., a neuron)

Excitatory-inhibitory (EI) model
- Neuronal oscillations in the visual cortex
- Epilepsy

Part 2 Roadmap

Problem
- Why: “non-linear” modeling

Fundamentals
- Non-linear (“gray-box”) models

Applications
- Epidemics (skips, competition, “shocks”)
- Information diffusion
- Online competition

Mining and forecasting of co-evolving epidemics

Flu
Measles
Mumps

Q. Can we forecast future epidemics?

Future
Time (years)
Real-time monitoring of co-evolving epidemics

- Influenza (ILI) prediction using search engine query data [Ginsberg+, Nature’09]

- CDC-reported ILI percentages
- Model estimates

CDC: Centers for Disease Control and Prevention
ILI: influenza-like illness


Epidemics - roadmap

A. Non-linear (gray-box) modeling!

- Outbreak vs. Skips [Stone+ Nature’07]
- Interaction between diseases [Rohani+ Nature’03]
- FUNNEL [Matsubara+ KDD’14]

Recurrent epidemics: Outbreak or skip?

- Time series of reported measles cases

New York

London


© 2016 Sakurai, Matsubara & Faloutsos
Recurrent epidemics: Outbreak or skip?

- Time series of reported measles cases

Q. Outbreak vs. skip?

- Conditions for predicting “outbreak vs. skip”
  - SIR model with high/low seasons

Phase plane diagram (S vs. log(I))

Contact rate
- \( \beta^+ \): high season
- \( \beta^- \): low season

Details

[Stone+ Nature '07]

Recurrent epidemics: Outbreak or skip? [Stone+ Nature’07]

- Conditions for predicting “outbreak vs. skip”
  - SIR model with high/low seasons

Phase plane diagram (S vs. log(I))

Threshold $S_c$: “Outbreak vs. Skip”

$S_0 > S_c = \frac{\gamma + \mu}{\beta_0} - \frac{\mu}{2} \quad \Rightarrow \quad \text{epidemic}$

if $S_0 < S_c$ there is a skip in the following year.

Epidemics - roadmap

A. Non-linear (gray-box) modeling!

Solutions

- Outbreak vs. Skips [Stone+ Nature’07]
- Interaction between diseases [Rohani+ Nature’03]
- FUNNEL [Matsubara+ KDD’14]

Ecological interference between fatal diseases

Q. Any relationship (i.e., interaction) between two different diseases (e.g., measles vs. whooping cough)?

A. Yes. There are “competing” diseases!

Ecological interference between fatal diseases

Weekly case fatality reports for two diseases

Measles   VS   Whooping cough

Birmingham Glasgow

Berlin Liverpool

Biennial (opposite) cycles

© 2016 Sakurai, Matsubara & Faloutsos

© 2016 Sakurai, Matsubara & Faloutsos
Ecological interference between fatal diseases

Extension of SIR model [Rohani’98]

Equations for 3 disease model

\[
\begin{align*}
\frac{dS}{dt} &= \mu S - \beta S \frac{I}{N} - \gamma S \\
\frac{dI}{dt} &= \beta S \frac{I}{N} - \gamma I - \delta I \\
\frac{dV}{dt} &= \delta I - \gamma V
\end{align*}
\]

[Stone+ Nature’07]

E2. Interaction between diseases

Rohani+Nature’03)

E3. FUNNEL [Matsubara+ KDD’14]

Non-linear (gray-box) modeling!

Solutions

- E1. Outbreak vs. Skips
- E2. Interaction between diseases
- E3. FUNNEL

FUNNEL with a single epidemic

With a single epidemic: Funnel-RE

People of 3 classes
- S: Susceptible
- I: Infected
- V: Vigilant/vaccinated

© 2016 Sakurai, Matsubara & Faloutsos
With a single epidemic: Funnel-RE

\[ S(t+1) = S(t) - \beta(t)e(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \]
\[ I(t+1) = I(t) + \beta(t)e(t)S(t)I(t) - \delta I(t) \]
\[ V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \]

\[ S(t) \text{: susceptible} \]
\[ I(t) \text{: Infected} \]
\[ V(t) \text{: Vigilant/Vaccinated} \]

\[ \beta(t) \text{: strength of infection} \]
\[ \gamma \text{: yearly periodic func} \]
\[ \theta(t) \text{: healing rate} \]
\[ \delta(t) \text{: disease reduction effect} \]

\[ \epsilon(t) \text{: temporal susceptible rate} \]
Information diffusion in social networks

MemeTracker [Leskovec+ KDD’09]
- Short phrases sourced from U.S. politics in 2008
  “you can put lipstick on a pig” (# of mentions in blogs)

“yes we can”

Q. How news/rumors spread in social media?

News spread in social media

MemeTracker [Leskovec+ KDD’09]
- Short phrases sourced from U.S. politics in 2008
  “you can put lipstick on a pig” (# of mentions in blogs)

“yes we can”

Q. How many patterns are there?
– Four classes on YouTube, etc.
  [Crane et al. PNAS’08]
– Six classes on Social media
  [Yang et al. WSDM’11]

News spread in social media

• Twitter (# of hashtags per hour)

• Google trend (# of queries per week)
  “tsunami” (in 2005)
  “harry potter” (2010 - 2011)
News spread in social media

- The volume of Google searches

[Tutorial@WWW'16](http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/)

© 2016 Sakurai, Matsubara & Faloutsos

News spread in social media

- Based on self-excited Hawkes Poisson process*

\[
\frac{dB(t)}{dt} = S(t) + \sum_{t_i \leq t} \mu_i \cdot \phi(t - t_i)
\]

[Tutorial@WWW'16](http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/)

© 2016 Sakurai, Matsubara & Faloutsos

News spread in social media

- Based on self-excited Hawkes Poisson process*

\[
\frac{dB(t)}{dt} = S(t) + \sum_{t_i \leq t} \mu_i \cdot \phi(t - t_i)
\]

 Rate of spread of infection/propagation
 Exogenous/External source
 # of Potential viewers
 Decaying virus/news strength

Exogenous
Endogenous

[Tutorial@WWW'16](http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/)

© 2016 Sakurai, Matsubara & Faloutsos

News spread in social media

- Four classes on YouTube* [Crane et al. PNAS'08]

[Tutorial@WWW'16](http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/)

© 2016 Sakurai, Matsubara & Faloutsos

© 2016 Sakurai, Matsubara & Faloutsos 16
News spread in social media

- Four classes on YouTube [Crane et al. PNAS’08]
- News spread
- Six classes of information diffusion patterns on social media [Yang et al. WSDM’11]

News spread in social media

- Four classes on YouTube [Crane et al. PNAS’08]
- News spread

Q. How many patterns are there, after all?

A. Our answer is “ONE”!
A single non-linear model!

“SpikeM”

Rise and Fall Patterns of Information Diffusion:
Model and Implications
Yasuko Matsubara (Kyoto University),
Yasushi Sakurai (NTT),
B. Aditya Prakash (CMU),
Lei Li (UCB), Christos Faloutsos (CMU)
Rise and fall patterns in social media

SpikeM captures 3 properties of real spike

1. periodicities

2. avoid infinity

3. power-law fall

SpikeM can capture behavior of real spikes using few parameters

Main idea (details)

- 1. Un-informed bloggers (clique of N bloggers/nodes)

- 2. External shock at time \( n_s \) (e.g., breaking news)
Main idea (details)
- 1. Un-informed bloggers (clique of N bloggers/nodes)
- 2. External shock at time t0 (e.g., breaking news)
- 3. Infection (word-of-mouth effects)

Infectiveness of a blog-post
\( \beta \) = Strength of infection (quality of news)
\( f(n) \) = Decay function (how infective a blog posting is)

SpikeM-base (details)
Equations of SpikeM (base)
\[
\Delta B(n+1) = U(n) \cdot \sum_{i=1}^{n} (\Delta B(i) + S(i)) \cdot f(n+1-i) + \varepsilon
\]
\[
U(n+1) = U(n) - \Delta B(n+1)
\]

- \( N \) = Total population of available bloggers
- \( \beta \) = Strength of infection/news
- \( n_s, S_s \) = External shock \( S_s \) at birth (time \( n_s \))
- \( \varepsilon \) = Background noise

Model fitting (Details)
- SpikeM consists of 7 parameters
\( \theta = \{N, \beta, n_s, S_s, \varepsilon, p_s, P_s\} \)

Learning parameters
- Given a real time sequence
  \( X = \{X(1), ..., X(n_1), ..., X(n_2)\} \)
- Minimize the error
  (Levenberg-Marquardt (LM) fitting)
  \[
  D(X, \theta) = \sum_{n=0}^{n} \frac{(X(n) - \Delta B(n))^2}{\varepsilon}
  \]

Analysis
SpikeM matches reality
exponential rise and power-law fall

SpikeM vs. SI model (susceptible infected model)
Q1-1 Explaining K-SC clusters

- Six patterns of K-SC [Yang et al. WSDM’11]

  - SpikeM can generate all patterns in K-SC

Q1-2 Matching MemeTracker patterns

- MemeTracker (memes in blogs) [Leskovec et al. KDD’09]

  - SpikeM matches reality

Q1-3 Matching Twitter data

- Twitter data (hashtags)

  - It can generate various patterns in social media

Q1-4 Matching Google trend data

- Volume of searches for queries on Google

  - SpikeM can capture various patterns
Q2 Tail-part forecasts
- Given a first part of the spike
- Forecast the tail part

SpikeM can capture tail part (AR: fail)

A1. “What-if” forecasting
Forecast not only tail-part, but also rise-part!
- (1) First spike
- (2) Release date of two sequel movies
- (3) Access volume before the release date

SpikeM can forecast upcoming spikes!

A2. Outlier detection
- Fitting result of “tsunami (Google trend)”
- in log-log scale

Another earthquake

Indian Ocean earthquake

One year after Indian Ocean earthquake

A3. Reverse engineering
SpikeM provide an intuitive explanation
PDF of parameters over 1,000 memes/hashtags

Meme

Twitter
A3. Reverse engineering

**SpikeM** provide an intuitive explanation PDF of parameters over 1,000 memes/hashtags

Observation 2
Strength of first burst (news) is $\beta N = 1.0$

Observation 3
Daily periodicity with phase shift $P_s = 0$
Every meme has the same periodicity without lag

(Twitter)
Daily periodicity with more spread in $P_s$ (i.e., Multiple timezone)

Part 2
Roadmap

- Problem
  - Why: “non-linear” modeling
- Fundamentals
  - Non-linear (grey-box) models
- Applications
  - Epidemics
  - Information diffusion

Online competition

- Roadmap

  - A. Non-linear (gray-box) modeling!
    - Winner-Takes-All [Prakash+ WWW’12]
    - Co-existence of the two viruses [Beutel+ KDD’12]
    - The Web as a Jungle [Matsubara+ WWW’15]

Q. How can we describe “virtual competition”?
Online competition - roadmap

A. Non-linear (gray-box) modeling!

Solutions
- Winner-Takes-All [Prakash+ WWW’12]
- Co-existence of the two viruses [Beutel+ KDD’12]
- The Web as a Jungle [Matsubara+ WWW’15]

Competing contagions
Contagions: viruses, online activities

iPhone v Android  Blu-ray v HD-DVD

Q. What happen when two viruses compete?

Competing contagions

ASSUME: Virus 1 is stronger than Virus 2

Q: What happens in the end?

Answer: Winner-Takes-All!

ASSUME: Virus 1 is stronger than Virus 2

Footprint @ Steady State
Footprint @ Steady State

© 2016 Sakurai, Matsubara & Faloutsos

© 2016 Sakurai, Matsubara & Faloutsos
A simple model

- Modified flu-like (SIS) model
- Mutual Immunity ("pick one of the two")
- Susceptible-Infected1-Infected2-Susceptible

Result: Winner-Takes-All

Given this model, and any graph, the weaker virus always dies-out, completely

1. The stronger survives only if it is above threshold
2. Virus 1 is stronger than Virus 2, if:
   \[ \text{strength}(\text{Virus 1}) > \text{strength}(\text{Virus 2}) \]
3. Strength(Virus) = \( \frac{\lambda \beta}{\delta} \) \( \rightarrow \) same as before!

Real Examples of “WTA”

Online competition in social networks

A. Non-linear (gray-box) modeling!

Solutions
- Winner-Takes-All [Prakash+ WWW’12]
- Co-existence of the two viruses [Beutel+ KDD’12]
- The Web as a Jungle [Matsubara+ WWW’15]

Interacting Viruses:
Can Both Survive?

Real example of “co-existence”

Interacting Viruses:
Can Both Survive?

Real example of “co-existence”
A simple model: $SI_{1/2}S$

- Modified flu-like (SIS)
- Susceptible-Infected, or 2°Susceptible
- Interaction Factor $\varepsilon$
  - Full Mutual Immunity: $\varepsilon = 0$
  - Partial Mutual Immunity (competition): $\varepsilon < 0$
  - Cooperation: $\varepsilon > 0$

Answer: Yes!
There is a phase transition

ASSUME: Virus 1 is stronger than Virus 2

Question:
What happens in the end?

$\varepsilon = 0$
Winner takes all

$\varepsilon = 1$
Co-exist independently

$\varepsilon = 2$
Viruses cooperate

What about for $0 < \varepsilon < 1$?
Is there a point at which both viruses can co-exist?

Answer: Yes!
There is a phase transition

Assume: Virus 1 is stronger than Virus 2

Result:
Viruses can Co-exist

Given this model and a fully connected graph, there exists an $\varepsilon_{\text{critical}}$ such that for $\varepsilon \geq \varepsilon_{\text{critical}}$, there is a fixed point where both viruses survive.

1. The stronger survives only if it is above threshold
2. Virus 1 is stronger than Virus 2, if:
   \[ \text{strength}(\text{Virus 1}) > \text{strength}(\text{Virus 2}) \]
3. Strength(\text{Virus}) $\sigma = N \beta / 8$
Online competition in social networks

A. Non-linear (gray-box) modeling!

Solutions
- Winner-Takes-All [Prakash+ WWW’12]
- Co-existence of the two viruses [Beutel+ KDD’12]
- The Web as a Jungle [Matsubara+ WWW’15]

The Web as a Jungle: Non-Linear Dynamical Systems for Co-evolving Online Activities

Yasuko Matsubara (Kumamoto University)
Yasushi Sakurai (Kumamoto University)
Christos Faloutsos (CMU)

Given: online user activities

e.g., Google search volumes for

Xbox, PlayStation, Wii, Android

Volume @ time

Time (weekly)

Volume @ time

Time (weekly)

1. Exponential growth

2. (Hidden) interaction between keywords

Given: online user activities

e.g., Google search volumes for

Xbox, PlayStation, Wii, Android

Volume @ time

Time (weekly)

Volume @ time

Time (weekly)

Q. Any trends?
Given: online user activities
e.g., Google search volumes for

3. Seasonality

Goal: find patterns and rules “fully-automatically”

Problem definition

Given: Co-evolving online activities
X (activity x time)

Find: Compact description of X

EcoWeb

NO magic numbers!
Parameter-free!

Modeling power of EcoWeb
Wii vs. Android!

© 2016 Sakurai, Matsubara & Faloutsos
Modeling power of EcoWeb

EcoWeb: seasonal component

EcoWeb: seasonal component

EcoWeb: seasonal component

EcoWeb: seasonal component
Problem definition

Given: Co-evolving online activities 
X (activity * time)

Find: Compact description of X

EcoWeb: Main idea

Q. How can we describe the evolutions of X?

A. The Web as a jungle!
- “Virtual species” living on the Web
- Interacting with other species (activities)

EcoWeb: Main idea

Q. How can we describe the evolutions of X?

A. Web as a jungle:

Ecosystem on the Web

G1: EcoWeb-individual

Popularity size increases over time
G1: EcoWeb-individual

Non-linear evolution of a single keyword

\[
P(t+1) = P(t) \left[ 1 + r \left( 1 - \frac{P(t)}{K} \right) \right],
\]

- \( p \) = Initial condition (i.e., \( P(0) = p \))
- \( r \) = Growth rate, attractiveness
- \( K \) = Carrying capacity (= available user resources)


EcoWeb: Main idea

Q. How can we describe the evolutions of \( X \) ?

A. Web as a jungle!

G1 G2 G3


G2: EcoWeb-interaction

Interaction between multiple keywords

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

- \( a_{ij} \) = Interaction coefficient
  - i.e., effect rate of keyword \( j \) on \( i \)

EcoWeb: Main idea

Q. How can we describe the evolutions of X?

A. Web as a jungle!

Non-linear evolution
Interaction/competition
Seasonality

G3: EcoWeb-seasonality

“Hidden” seasonal activities

Users change their behavior according to seasonal events!

Climate
Events

Estimated volume of keyword $i$

\[
C_i(t) = P_i(t) \left[ 1 + \frac{c_i(t)}{e_i(t)} \right] \quad (i = 1, \ldots, d),
\]

where

- $P_i(t)$: latent popularity
- $C_i(t)$: volume
- $E_i(t)$: seasonality

Seasonality matrix $B$
Participation (weight) matrix $W$

Seasonal events
Season/Climate
EcoWeb: Main idea

Q. How can we describe the evolutions of $X$?

$$X = \{p, r, K, A, W, B\}$$

Full parameters

$$S = \{p, r, K, A, W, B\}$$

Algorithms

Q1. How can we automatically find “seasonal components”?

Idea (1): Seasonal component analysis

Q2. How can we efficiently estimate full-parameters?

Idea (2): Multi-step fitting

Idea (1): Seasonal component analysis

Q1. How can we automatically find “$k$-seasonal components”?

Idea (1):

a. Seasonal component detection
b. Automatic component analysis
Idea (1): Seasonal component analysis

Idea (1-a) Seasonal component detection

\[
E \quad d = 2
\]

Time (1,... n)

\[
\hat{E} \quad d \times n / n_p
\]

ICA

Independent components

E: seasonality

\[
W \times B
\]

\[
B \times n_p
\]

Find optimal number \( k \) (1 \( \leq \) \( k \) \( \leq \) \( d \))

\( d \) : dimension

\[
k = ? \quad \text{opt } k = ?
\]

Cost \( M \) (\( S \)) + Cost \( C \) (\( X \mid S \))

Model cost

Coding cost

Good compression

Good description

Cost \( T \) (\( X \mid S \)) = \log^* (d) + \log^* (n) + Cost \( M \) (\( p, r, K \))

+ Cost \( M \) (\( A \)) + Cost \( M \) (\( k, W, B \)) + Cost \( C \) (\( X \mid S \))

\( k_{opt} = \arg \min_k \text{Cost}_T (X \mid S) \)
Idea (1): Seasonal component analysis

Idea (1-b) Automatic component analysis

Find optimal number $k$ ($1 \leq k \leq d$)

- $d$: dimension

- Optimal $k$

$I = \{1, 2, 3\}$

Cost(1) = $$

Cost(2) = $

Cost(3) = $$$

- $k = 1$

- $k = 2$

- $k = 3$

$Idea (2): EcoWeb-Fit$

Q2. How can we efficiently estimate model parameters?

- Idea (2-a): StepFit: Update parameters alternately

  a. StepFit (sub)
  b. EcoWeb-Fit (full)

$X = \{A, B, C, D\}$

Step A

Step B

Idea (2-b): EcoWeb-Fit: full algorithm

- 1. Individual-Fit
- 2. Pair-Fit
- 3. Full-Fit

Experiments

We answer the following questions...

Q1. Effectiveness
   How successful is it in spotting patterns?

Q2. Accuracy
   How well does it match the data?

Q3. Scalability
   How does it scale in terms of computational time?

© 2016 Sakurai, Matsubara & Faloutsos
Q1. Effectiveness

(#1) Video games

Seasonality

Interactions between keywords


Q1. Effectiveness

(#1) Video games


Q1. Effectiveness

(#2) Programming language


Q1. Effectiveness

(#3) Social media


Q1. Effectiveness

(#4) Apparel companies


Q1. Effectiveness

(#5) Retail companies

Q1. Effectiveness

(#5) Retail companies

Amazon, Walmart, Home Depot, BestBuy, Lowes, Costco

Fitting result - RMSE=0.065/73

Seasonality

Q2. Accuracy

RMSE between original and fitted volume

(Lower is better)

EcoWeb consistently wins!

Q3. Scalability

Wall clock time vs. dataset size (years)

EcoWeb-Fit scales linearly, i.e., \( O(n) \)

7x faster than LV, 20x faster than EcoWeb-Plain

EcoWeb at work - forecasting

Forecasting future activities

Train: 2/3 sequences

Forecast: 1/3 following years

EcoWeb-Fit scales linearly, i.e., \( O(n) \)

7x faster than LV, 20x faster than EcoWeb-Plain

EcoWeb at work - forecasting

Forecasting future activities

Train: 2/3 sequences

Forecast: 1/3 following years

EcoWeb can capture future patterns!
Part 2
Roadmap

**Problem**
- Why: “non-linear” modeling

**Fundamentals**
- Non-linear (grey-box) models

**Applications**
- Epidemics
- Information diffusion
- Online competition

Goal!

References (1)

- Non-linear forecasting

- Non-linear equations and modeling

- Others

References (2)

- Epidemics

- Information diffusion

- Online activities and competition

- Others

Non-linear mining and forecasting

Yasuhi Sakurai
(Yokohama City University)

Yusako Matsubara
(Kumamoto University)

Christos Faloutsos
(Carnegie Mellon University)