Mining Big Time-series Data on the Web

Yasushi Sakurai
Kumamoto University
yasushi@cs.kumamoto-u.ac.jp

Yasuko Matsubara
Kumamoto University
yasuko@cs.kumamoto-u.ac.jp

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

ABSTRACT

Online news, blogs, SNS and many other Web-based services has been attracting considerable interest for business and marketing purposes. Given a large collection of time series, such as web-click logs, online search queries, blog and review entries, how can we efficiently and effectively find typical time-series patterns? What are the major tools for mining, forecasting and outlier detection? Time-series data analysis is becoming of increasingly high importance, thanks to the decreasing cost of hardware and the increasing on-line processing capability.

The objective of this tutorial is to provide a concise and intuitive overview of the most important tools that can help us find meaningful patterns in large-scale time-series data. Specifically we review the state of the art in three related fields: (1) similarity search, pattern discovery and summarization, (2) non-linear modeling and forecasting, and (3) the extension of time-series mining and tensor analysis. We also introduce case studies that illustrate their practical use for social media and Web-based services.

1. INTRODUCTION

The increasing volume of online, time-stamped activity represents a vital new opportunity for data scientists and analysts to measure the collective behavior of social, economic, and other important evolutions on the Web. Time-series data occur naturally in many online applications, and the logging rate has increased greatly with the progress made on hardware and storage technology. One big challenge for Web mining is to handle and analyze such large volumes of data (i.e., "big" time-series data) at a very high logging rate.

Time-series data comes in various types of formats including co-evolving numerical sequences (e.g., IoT device data, video, audio), complex time-stamped events (e.g., web-click logs of the form <user-ID, URL, time>), and time-evolving graph (e.g., social networks over time). Data variety imposes new requirements to data mining, therefore recent studies has revealed some new directions for research on time-series analysis, which include:

- **Large-scale tensor analysis:** Given huge collections of time-evolving online activities such as Google search queries, which consist of multiple attributes (e.g., keywords, locations, time), how can we analyze temporal patterns and relationships among all these activities and find location-specific trends? Time-evolving online activities and many other time-series data and can be modeled as tensors, and tensor analysis is an important data mining tool that has various applications including web-click logs for multiple users/URLs, IoT data streams, hyperlinks and social networks over time.
- **Non-linear modeling:** Non-linear models are widely used in a variety of areas, such as epidemiology, biology, physics and economics, since the nature of real data sets suggests that non-linear models are appropriate for describing their dynamics. In the Web mining field, analyses of social media and online user activities have attracted considerable interest, and recent studies have focused on non-linear time-series analysis to understand the dynamic behavior of social networks (e.g., information diffusion, influence propagation).
- **Automatic mining:** We also emphasize the importance of fully-automatic mining. Most of the existing time-series tools require parameter tuning, and they are very sensitive to these parameters. In fact, faced with "big data on the Web", fully automatic mining is even more important: otherwise, the data scientists and analysts would have to try several parameter tuning steps, each of which would take too long (e.g., hours, or days). Namely, as regards real big data analysis, we *cannot afford* human intervention.

This tutorial provides a concise and intuitive overview of the most important tools that we can use to help us understand and find patterns in large-scale time evolving sequences. We will provide a comprehensive overview and the above new directions for time-series analysis, and deal specifically with the following key topics: (1) similarity search, pattern discovery and summarization, (2) non-linear modeling and forecasting, and (3) the extension of time-series mining and tensor analysis.

Who should attend. The target audience is researchers and advanced professionals of Web, social media, IoT data mining, who wish to get up to speed with the major tools used in time sequence analysis. Also, practitioners who want a concise, intuitive overview of the state of the art.

Prerequisites. None. The emphasis is on the intuition behind all these mathematical tools.

2. CONTENT AND OUTLINE

Our tutorial is structured as follows:

1. Similarity search, pattern discovery and summarization (60 minutes)

WWW 2016 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2891061.

(a) Why we need similarity search (indexing, fast searching, similarity measure)
(b) Feature extraction (discrete Fourier transform, wavelets, singular value decomposition, independent component analysis)
(c) Linear modeling and forecasting (main idea behind linear forecasting, AR methodology and multivariate regression, recursive least square)
(d) Streaming pattern discovery (component analysis, correlation monitoring)
(e) Sequence summarization (linear dynamic systems, probabilistic models, automatic mining of co-evolving sequences)

2. Non-linear modeling and forecasting (60 minutes)
   (a) Non-linear forecasting (lag-plots, fractal dimension and power-law)
   (b) Non-linear dynamical systems (main idea behind non-linear equations, non-linear epidemic models, gray-box non-linear mining, non-linear dynamical systems for online activities, information diffusion in social networks)
3. Extension of time-series mining - tensor analysis (60 minutes)
   (a) Tensor decomposition (basic approaches, decompositions of higher-order tensors)
   (b) Mining of complex time-stamped tensors (complex time-stamped events and big sparse tensors, feature extraction from sparse tensors, forecasting of complex time-stamped events)
   (c) New directions of tensor analysis (non-linear modeling for tensors, automatic non-linear analysis)

2.1 Similarity search, pattern discovery and summarization

In this first part of the tutorial, we explain the most common and fundamental tools of time series data mining. More specifically, we demonstrate some traditional approaches applied to time series data mining including similarity search (e.g., Euclidean distance and dynamic time warping [5, 17, 24, 19, 16, 53, 50, 49, 39]), feature extraction (singular value decomposition (SVD), independent component analysis (ICA) [43, 51]), segmentation [18], multi-dimensional scaling [11]. We also introduce several mining algorithms for online data streams, including component analysis [45, 46], correlation monitoring [64, 52] and time warping over streams [50, 57]. For linear modeling, auto regression and moving averaging models have been studied for many years in statistics and finance [5], and have been applied to time-series data mining [7, 15, 28]. We introduce AR methodology and several important tools, including MUSCLES [63] and AWSOM [44]. We also introduce linear dynamical systems (LDS), Kalman filters (KF) and their variants [14, 30, 29, 56]. For probabilistic modeling, hidden Markov models (HMMs) have been used in various applications including speech recognition [60] and sensor monitoring [27, 37, 13, 59]. As regards probabilistic time-series analysis, [32] developed AutoPlait, a fully-automatic mining algorithm for co-evolving time sequences. Given a large collection of co-evolving multiple sequences, which contains an unknown number of patterns of different durations, AutoPlait automatically identifies all distinct patterns and spots the time position of each variation. We introduce case studies that illustrate their practical use for social media and Web-based services.

2.2 Non-linear modeling and forecasting

In this part, we introduce several advanced techniques, and focus specifically on non-linear time series analysis. We start by explaining non-linear forecasting e.g., lag-plots [8], which is based on nearest-neighbor search. We also explain some fundamental concepts such as fractal dimension and power law [54, 41, 2, 31]. We then review the most common non-linear equations, including the logistic function (LF) [6], the susceptible-infected (SI) model [1], the independent cascade (IC) model [10], the so-called "Bass" model [3], the Lotka-Volterra (LV) model [40] and other non-linear equations [42]. We explain the importance of non-linear equations and the concept of gray-box non-linear mining. In this part, we also review recent work on understanding the non-linear time evolution of online user activities. Analyses of epidemics, blogs, social media, propagation and the cascades they create have attracted much interest [25, 61, 48, 22, 26, 47, 4]. We answer several important topics such as how popularity of "memes" changes over time [25]; how to find temporal patterns in information diffusion process through online media, e.g., blogs, hashtags [62, 61], and YouTube [9, 12]; how to describe rising and falling patterns of information propagation (e.g., memes, hashtags and keyword search volume) using non-linear dynamical systems [36, 33].

2.3 Extension of time-series mining - tensor analysis

The goal in this part is to present large-scale studies of complex time-stamped events and big sparse tensors. We first introduce some basic approaches including Tucker, PARAFAC, and higher-order SVD (HOSVD) [20, 21, 55, 23]. Complex time-stamped events can be represented as a tensor with several dimensions. For example, given a set of time-stamped event entries of the form {object, actor, timestamp} (e.g., web-clicks: {URL, userID, timestamp}), we can treat them as a 3rd order tensor. Here, one subtle, but important issue is that the complex time-stamped tensor is *very sparse*, which derails all typical time-series mining and forecasting tools. We introduce a scalable algorithm, TriMine [35] to deal with this issue. TriMine has the ability to find meaningful patterns in complex time-stamped tensors, and forecast future events, e.g., estimate the number of clicks from user "Smith" to URL "CNN.com" for the next 30 days.

Finally, we show new directions for tensor analysis, namely, automatic and non-linear analysis for big time-series tensors. Specifically, we introduce a unifying analytical model, FUNNEL [38], for mining and forecasting large-scale epidemiological data (e.g., the Project Tycho [58]) as well as an efficient fitting algorithm, which solves the problem. As regards online activities on the Web, [34] developed CompCube, which identifies the most probable competitor for each keyword among all possible keywords. It operates on large collections of co-evolving activities and summarizes them succinctly with respect to multiple aspects (i.e., activity/keyword, location, time).

We also discuss the importance of fully-automatic mining for time-series tensor analysis. There are many fascinating and useful tools for time-series analysis. However, most existing methods require parameter settings and fine tuning, such as the number of coefficients, and the reconstruction error thresholds, and they are very sensitive to these parameters. The ideal method should look for arbitrary patterns and require no initial human intervention to guide it.

3. PRESENTERS - BIOS

Yasushi Sakurai is a Professor at Kumamoto University. He obtained his B.E. degree from Doshisha University in 1991, and his M.E. and Ph.D. degrees from Nara Institute of Science and Technology in 1996 and 1999, respectively. In 1998, he joined NTT Laboratories, and became a Senior Research Scientist in 2005. He was a Visiting Researcher at Carnegie Mellon University during 2004-2005. He received two KDD best paper awards in 2008 and 2010. His research interests include time-series analysis, web mining, and sensor data processing.

Yasuko Matsubara is an Assistant Professor in the Department of Computer Science and Electrical Engineering at Kumamoto University, Japan. She obtained her BS and MS degrees from Ochanomizu University in 2007 and 2009 respectively, and her PhD from Kyoto University in 2012. She was a Visiting Researcher at Carnegie Mellon University during 2011-2012 and 2013-2014. Her research interests include time-series data mining and non-linear dynamic systems.

Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation (1989), the Research Contributions Award in ICDM 2006, the SIGKDD Innovations Award (2010), twenty "best paper" awards (including two "test of time" awards), and four teaching awards. Five of his advisees have attracted KDD or SCS dissertation awards. He is an ACM Fellow, he has served as a member of the executive committee of SIGKDD; he has published over 300 refereed articles, 17 book chapters and two monographs. He holds eight patents and he has given over 35 tutorials and over 15 invited distinguished lectures. His research interests include data mining for graphs and streams, fractals, database performance, and indexing for multimedia and bio-informatics data.

Acknowledgement

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research Number 15H02705, 26730060, 26280112. This material is based upon work supported by the National Science Foundation under Grants No. CNS-1314632 and IIS-1408924; and by the Army Research Laboratory (ARL) under Cooperative Agreement Number W911NF-09-2-0053; and by a Google Focused Research Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, ARL, or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

4. REFERENCES

[1] R. M. Anderson and R. M. May. *Infectious Diseases of Humans Dynamics and Control*. Oxford University Press, 1992.

[2] A. L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435, 2005.

[3] F. M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.

[4] A. Beutel, B. A. Prakash, R. Rosenfeld, and C. Faloutsos. Interacting viruses in networks: can both survive? In *KDD*, pages 426–434, 2012.

[5] G. E. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliffs, NJ, 3rd edition, 1994.

[6] F. Brauer and C. Castillo-Chavez. *Mathematical models in population biology and epidemiology*, volume 40. Springer Verlag, New York, 2001.

[7] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag New York, Inc., New York, NY, USA, 1987.

[8] D. Chakrabarti and C. Faloutsos. F4: large-scale automated forecasting using fractals. In *CIKM*, pages 2–9, 2002.

[9] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. In *PNAS*, 2008.

[10] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[11] C. Faloutsos and K.-I. Lin. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *SIGMOD*, pages 163–174, 1995.

[12] F. Figueiredo, J. M. Almeida, Y. Matsubara, B. Ribeiro, and C. Faloutsos. Revisit behavior in social media: The phoenix-r model and discoveries. In *PKDD*, pages 386–401, 2014.

[13] Y. Fujiwara, Y. Sakurai, and M. Yamamuro. Spiral: efficient and exact model identification for hidden markov models. In *KDD*, pages 247–255, 2008.

[14] A. Jain, E. Y. Chang, and Y.-F. Wang. Adaptive stream resource management using kalman filters. In *SIGMOD*, pages 11–22, 2004.

[15] K. Kalpakis, D. Gada, and V. Puttagunta. Distance measures for effective clustering of arima time-series. In *ICDM 2001: Proceeding of 2001 IEEE International Conference on Data Mining*, pages 273–280, 2001.

[16] E. J. Keogh. Exact indexing of dynamic time warping. In *Proceedings of VLDB*, pages 406–417, Hong Kong, China, August 2002.

[17] E. J. Keogh, K. Chakrabarti, S. Mehrotra, and M. J. Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. In *Proceedings of ACM SIGMOD*, pages 151–162, May 2001.

[18] E. J. Keogh, S. Chu, D. Hart, and M. J. Pazzani. An online algorithm for segmenting time series. In *ICDM*, pages 289–296, 2001.

[19] E. J. Keogh, T. Palpanas, V. B. Zordan, D. Gunopulos, and M. Cardle. Indexing large human-motion databases. In *VLDB*, pages 780–791, 2004.

[20] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[21] T. G. Kolda, B. W. Bader, and J. P. Kenny. Higher-order web link analysis using multilinear algebra. In *ICDM*, pages 242–249, 2005.

[22] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *KDD*, pages 553–562, 2010.

[23] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[24] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: a partition-and-group framework. In *SIGMOD*, pages 593–604, 2007.

[25] J. Leskovec, L. Backstrom, and J. M. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.

[26] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD*, pages 462–470, 2008.

[27] J. Letchner, C. Ré, M. Balazinska, and M. Philipose. Access methods for markovian streams. In *ICDE*, pages 246–257, 2009.

[28] L. Li, C.-J. M. Liang, J. Liu, S. Nath, A. Terzis, and C. Faloutsos. Thermocast: A cyber-physical forecasting model for data centers. In *KDD*, 2011.

[29] L. Li, J. McCann, N. Pollard, and C. Faloutsos. Dynammo: Mining and summarization of coevolving sequences with missing values. In *KDD*, 2009.

[30] L. Li, B. A. Prakash, and C. Faloutsos. Parsimonious linear fingerprinting for time series. *PVLDB*, 3(1):385–396, 2010.

[31] Y. Matsubara, L. Li, E. E. Papalexakis, D. Lo, Y. Sakurai, and C. Faloutsos. F-trail: Finding patterns in taxi trajectories. In *PAKDD*, pages 86–98, 2013.

[32] Y. Matsubara, Y. Sakurai, and C. Faloutsos. Autoplait: automatic mining of co-evolving time sequences. In *SIGMOD*, pages 193–204, 2014.

[33] Y. Matsubara, Y. Sakurai, and C. Faloutsos. The web as a jungle: Non-linear dynamical systems for co-evolving online activities. In *WWW*, 2015.

[34] Y. Matsubara, Y. Sakurai, and C. Faloutsos. Non-linear mining of competing local activities. In *WWW*, 2016.

[35] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa. Fast mining and forecasting of complex time-stamped events. In *KDD*, pages 271–279, 2012.

[36] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *KDD*, pages 6–14, 2012.

[37] Y. Matsubara, Y. Sakurai, N. Ueda, and M. Yoshikawa. Fast and exact monitoring of co-evolving data streams. In *ICDM*, 2014.

[38] Y. Matsubara, Y. Sakurai, W. G. van Panhuis, and C. Faloutsos. FUNNEL: automatic mining of spatially coevolving epidemics. In *KDD*, pages 105–114, 2014.

[39] Y. Matsubara, Y. Sakurai, and M. Yoshikawa. Scalable algorithms for distribution search. In *ICDM*, pages 347–356, 2009.

[40] R. M. May. Qualitative stability in model ecosystems. *Ecology*, 54(3):638–641, 1973.

[41] M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, and N. Glance. Finding patterns in blog shapes and blog evolution. In *International Conference on Weblogs and Social Media*, Boulder, Colo., March 2007.

[42] M. Nowak. *Evolutionary Dynamics*. Harvard University Press, 2006.

[43] J.-Y. Pan, H. Kitagawa, C. Faloutsos, and M. Hamamoto. Autosplit: Fast and scalable discovery of hidden variables in stream and multimedia databases. In *PAKDD*, May 26-28 2004.

[44] S. Papadimitriou, A. Brockwell, and C. Faloutsos. Adaptive, hands-off stream mining. In *VLDB*, pages 560–571, 2003.

[45] S. Papadimitriou, J. Sun, and C. Faloutsos. Streaming pattern discovery in multiple time-series. In *VLDB*, pages 697–708, 2005.

[46] S. Papadimitriou and P. S. Yu. Optimal multi-scale patterns in time series streams. In *SIGMOD*, pages 647–658, 2006.

[47] B. A. Prakash, A. Beutel, R. Rosenfeld, and C. Faloutsos. Winner takes all: competing viruses or ideas on fair-play networks. In *WWW*, pages 1037–1046, 2012.

[48] B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, and C. Faloutsos. Threshold conditions for arbitrary cascade models on arbitrary networks. In *ICDM*, pages 537–546, 2011.

[49] T. Rakthanmanon, B. J. L. Campana, A. Mueen, G. E. A. P. A. Batista, M. B. Westover, Q. Zhu, J. Zakaria, and E. J. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *KDD*, pages 262–270, 2012.

[50] Y. Sakurai, C. Faloutsos, and M. Yamamuro. Stream monitoring under the time warping distance. In *ICDE*, pages 1046–1055, Istanbul, Turkey, April 2007.

[51] Y. Sakurai, L. Li, Y. Matsubara, and C. Faloutsos. Windmine: Fast and effective mining of web-click sequences. In *SDM*, pages 759–770, 2011.

[52] Y. Sakurai, S. Papadimitriou, and C. Faloutsos. Braid: Stream mining through group lag correlations. In *SIGMOD*, pages 599–610, 2005.

[53] Y. Sakurai, M. Yoshikawa, and C. Faloutsos. Ftw: Fast similarity search under the time warping distance. In *PODS*, pages 326–337, Baltimore, Maryland, June 2005.

[54] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise.* W. H. Freeman, 1991.

[55] J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *KDD*, pages 374–383, 2006.

[56] Y. Tao, C. Faloutsos, D. Papadias, and B. Liu. Prediction and indexing of moving objects with unknown motion patterns. In *SIGMOD*, pages 611–622, 2004.

[57] M. Toyoda, Y. Sakurai, and Y. Ishikawa. Pattern discovery in data streams under the time warping distance. *VLDB J.*, 22(3):295–318, 2013.

[58] W. G. van Panhuis, J. Grefenstette, S. Y. Jung, N. S. Chok, A. Cross, H. Eng, B. Y. Lee, V. Zadorozhny, S. Brown, D. Cummings, and D. S. Burke. Contagious diseases in the united states from 1888 to the present. *NEJM*, 369(22):2152–2158, 2013.

[59] P. Wang, H. Wang, and W. Wang. Finding semantics in time series. In *SIGMOD Conference*, pages 385–396, 2011.

[60] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11):1870–1878, 1990.

[61] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM*, pages 599–608, 2010.

[62] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186, 2011.

[63] B.-K. Yi, N. Sidiropoulos, T. Johnson, H. Jagadish, C. Faloutsos, and A. Biliris. Online data mining for co-evolving time sequences. *ICDE*, pages 13–22, 2000.

[64] Y. Zhu and D. Shasha. Statstream: Statistical monitoring of thousands of data streams in real time. In *VLDB*, pages 358–369, 2002.