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 have 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 have 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 can and 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)
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 derail 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.

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.
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 Ochamizu 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 GoogleFocused 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
