KDD-2006 Workshop Report
Theory and Practice of Temporal Data Mining

Tao Li
School of Computer Science, Florida International University
taoli@cs.fiu.edu

Chang-shing Perng
IBM T.J. Watson Research Center
perng@us.ibm.com

ABSTRACT

In this paper we provide a summary of the workshop on Theory and Practice of Temporal Data Mining held in conjunction with ACM SIGKDD 2006, on August 20st in Philadelphia, Pennsylvania, USA. We report in detail about the research issues addressed in the talks at the workshop.

1. INTRODUCTION

Many real-world applications deal with huge amounts of temporal data. Examples include alarms/events and performance measurements generated by distributed computer systems and by telecommunication networks, the web server logs, online transaction logs, financial data, workflow process logs, and sensor data collected from sensor networks. Conventionally, temporal data is classified to either categorical event streams or numerical time series and both types have been intensively studied in data mining and statistics. However, several previously less emphasized aspects of temporal data have proven their importance in emerging applications and posed several challenges calling for more research. In addition, the applications of temporal data analysis, such as web services, information navigation, system management, adaptive workflow management, program behavior analysis, security management and bioinformatics, are enjoying a growing amount of attention. This workshop aims to gather researchers and practitioners to tackle these challenges and attempts to study the common tasks that need to be addressed in practical applications.

The setting of traditional temporal data analysis is to apply one algorithm on a static, regular and relatively small temporal data set. Many practitioners found existing analysis methods inadequate for their real-world data. Many struggle to transform the data in order to apply existing methods or even to reduce the original problems to better studied ones; either ways induce in more preprocessing effort, more artificial parameters and less interpretable results. We believe these new aspects of temporal data deserve theories and algorithms of their own. Some of these new aspects are:

- Irregularity: Many types of numerical temporal data are not equally paced.
- Asynchronism: In distributed computing environments like sensor networks, data from different sources tend to be not aligned and hence can not apply synchronous methods.
- Distributed analysis: A trend in temporal data analysis is to perform data filtering, transformation and analysis as close as possible to the data sources to avoid the prohibitive amount of data being transmitted and analyzed. This new computing paradigm calls for a new theoretical foundation.
- Streaming Data: Some temporal data is stored only temporally and requires near real-time analysis.
- Heterogeneous data types: It is very common that temporal data is partly categorical events and partly numerical time series. It remains to be an interesting challenging to best analyze all possible data in a uniform way.
- Huge Volume: The stream of data can be huge for a long, continuous observation period. Many types of measurements can be obtained from a large number of data sources. This requires designing scalable solutions in analyzing a large volume of temporal data, in terms of both the large number of data points and the large number of types of measurements.

Driven by the new aspects of temporal data, several fundamental problems need to be revisited. Just to name a few of them: *Prediction*, *Correlation*, *Regression*, *Benchmarking*, *Periodic Pattern Mining*, *Temporal Association Finding Causality Analysis*, *Sequential Event Patterns*, *Threshold selection Frequency Analysis*, and *Anomaly Detection*.

2. WORKSHOP OVERVIEW

We wish to re-establish the platform for exchanging ideas in the field of temporal data mining by organizing workshops. We have successfully organized two temporal data mining workshops in 2004 and 2005 with IEEE International Conference on Data Mining (ICDM). To continue this tradition of fostering interactions among researchers sharing the same interest, we organized the *KDD 2006 workshop on Theory and Practice of Temporal Data Mining (TPTDM 2006)*. In particular, the workshop aims to bring together researchers from both industry and academia with diverse backgrounds: data mining, machine learning, database, statistical analysis, and application knowledge to foster interactions, to propose new ideas, to identify promising technologies, to create a forum for discussing recent advances, to better understand the practical challenges in applications, and to inspire new research directions.

The program of the workshop included an invited talk by Prof. Jian Pei from Simon Fraser University and six research paper presentations. The on-line proceedings of the workshop is available at
http://temporaldatamining.com/TDM-2006.html.

3. INVITED TALK

In his invited talk titled "Mining Temporal Orders from Event Sequences", Jian Pei (Simon Fraser University) discussed the problem of finding partial order of events. He asked two questions: i) How to mine ordering information? and ii) How to represent ordering information succinctly?
He answered the first question by a two-step algorithm. Step 1 of the algorithm is to find all non-trivial closed sequential patterns, and step 2 is to ensemble frequent closed partial orders using closed sequential patterns. He addressed the second question by using the *Frecpo Framework* which represents a partial order uniquely as the set of edges in its transitive reduction. All edges in a set can be sorted in the alphabetical order or any global order. The Frecpo Framework has the following advantages

1. Mining in transitive reduction to avoid substantial space and I/O overhead.

2. Directly extracting frequent closed partial orders in transitive reduction.

3. Aggressively and progressively pruning futile branches in the recursive depth-first search

In summary, Pei concluded that mining frequent ordering information from sequences is interesting with many applications; Mining frequent closed partial orders is also a challenging modeling problem. The method presented is an adequate solution to the problem. The talk is joint work with H. Wang, J. Liu, K. Wang, J. Wang, and P. S. Yu.

4. OVERVIEW OF THE RESEARCH PAPERS

Nan Jiang and Le Gruenwald (University of Oklahoma) presented their work on online data stream mining. Comparing to the more studied problem of mining frequent closed itemsets which are mainly intended for traditional transaction databases and thus do not take data stream characteristics into consideration. They proposed a novel approach for mining closed frequent itemsets over data streams. It computes and maintains closed itemsets online and incrementally and can output the current closed frequent itemsets in real time based on users specified thresholds. Experimental results show that the proposed method is both time and space efficient with good scalability.
The following talk was by Fabian Morchen from Siemens Corporate Research. Allen's 13 relations has been the de facto representation of temporal intervals. Morchen pointed out several drawbacks of the model including i) not robust against noise, ii) ambiguous for similar patterns, and iii) require lengthy descriptions for the relationship among multiple intervals. Morchen proposed an alternative representation TSKR (Time Series Knowledge Representation) using *Chords*, a notion borrowed from music theory. TSKR is not robust and unambiguous, experiments shows it is very terse for describing relationship among multiple intervals.
Tamraparni Dasu, Deborah F. Swayne and David Poole (AT&T Labs Research) presented their work on grouping multivariate time series. Much work has been done on time series clustering. The goal of this paper is very specific: grouping massive multivariate time series. A methodology containing preprocessing, nonparametric summaries, and clustering is discussed. The performance of the algorithm is particularly impressive.
Wolfgang Jank, Galit Shmueli, Shanshan Wang (University of Maryland) delivered a talk drew the most attention titled *Dynamic, Realtime Forecasting of Online Auctions via Functional Models*. The study is to predict the 7th day final auction price on eBay by observing the bidding history on the first 6 days, then applying functional data analysis. The mean absolute percentage error of the proposed dynamic model is about 5% (with a standard deviation of 12%) at the auction-end, in comparison to 40%using exponential smoothing (standard deviation of 19%). Most audiences have since been busy using the method to make a killing on eBay.
The last talk by Jan Ramon (Dept. of Computer Science, K.U.Leuven) described his work on applying temporal data mining in the health-care Industry. He discussed several issues and possible approaches for predicting evolution of critically ill patients. The goal of the study is to better plan and allocate medical resources to meet patients' need.

5. WORKSHOP ORGANIZATION

Work Co-chairs
Tao Li, Florida International University
Charles Perng, IBM Research
Haixun Wang, IBM Research
Carlotta Domeniconi, George Mason University

Committee Members
Daniel Barbara, George Mason University
Christos Faloustos, Carnegie Mellon University
Johannes Gehrke, Cornell University
Dimitrios Gunopulos, UC Riverside
John Handley, Xerox Research
Oscar Kipersztok, Boeing Research
Feng Liang, Duke University
Mitsunori Ogihara, Univ. of Rochester
Srinivasan Parthasarathy, Ohio State University
Tong Sun, Xerox Research
Michalis Vlachos, IBM Research
Hui Xiong, Rutgers University
Philip S. Yu, IBM Research Center
Mohammed Zaki, Rensselaer Polytechnic Institute
Shenghuo Zhu, NEC Labs America

Most submissions were reviewed and discussed by two reviewers and workshop co-chairs. We are very indebted to all program committee members who helped us organize the workshop and reviewed the papers very carefully. We would also like to thank all the authors who submitted their papers to the workshop; they provided us with an excellent workshop program. More information about the workshop can be found at: http://temporaldatamining.com.