Ming-Syan Chen
Philip S. Yu
Bing Liu (Eds.)

LNAI 2336

Advances in Knowledge Discovery and Data Mining

6th Pacific-Asia Conference, PAKDD 2002
Taipei, Taiwan, May 2002
Proceedings

Springer

Ming-Syan Chen Philip S. Yu
Bing Liu (Eds.)

Advances in Knowledge Discovery and Data Mining

6th Pacific-Asia Conference, PAKDD 2002
Taipei, Taiwan, May 6-8, 2002
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Ming-Syan Chen
National Taiwan University, EE Department
No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan, ROC
E-mail: mschen@cc.ee.ntu.edu.tw

Philip S. Yu
IBM Thomas J. Watson Research Center
30 Sawmill River Road, Hawthorne, NY 10532, USA
E-mail: psyu@us.ibm.com

Bing Liu
National University of Singapore, School of Computing
Lower Kent Ridge Road, Singapore 119260
E-mail: liub@comp.nus.edu.sg

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Advances in knowledge discovery and data mining : 6th Pacific Asia conference ; proceedings / PAKDD 2002, Taipei, Taiwan, May 6 - 8, 2002. Ming-Syan Chen ... (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; Vol. 2336 : Lecture notes in artificial intelligence)
ISBN 3-540-43704-5

CR Subject Classification (1998): I.2, H.2.8, H.3, H.5.1, G.3, J.1, K.4

ISSN 0302-9743
ISBN 3-540-43704-5 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10869781 06/3142 5 4 3 2 1 0

Preface

Knowledge discovery and data mining have become areas of growing significance because of the recent increasing demand for KDD techniques, including those used in machine learning, databases, statistics, knowledge acquisition, data visualization, and high performance computing. In view of this, and following the success of the five previous PAKDD conferences, the sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2002) aimed to provide a forum for the sharing of original research results, innovative ideas, state-of-the-art developments, and implementation experiences in knowledge discovery and data mining among researchers in academic and industrial organizations.

Much work went into preparing a program of high quality. We received 128 submissions. Every paper was reviewed by 3 program committee members, and 32 were selected as regular papers and 20 were selected as short papers, representing a 25% acceptance rate for regular papers. The PAKDD 2002 program was further enhanced by two keynote speeches, delivered by Vipin Kumar from the Univ. of Minnesota and Rajeev Rastogi from AT&T. In addition, PAKDD 2002 was complemented by three tutorials, XML and data mining (by Kyuseok Shim and Surajit Chadhuri), mining customer data across various customer touchpoints at e-commerce sites (by Jaideep Srivastava), and data clustering analysis, from simple groupings to scalable clustering with constraints (by Osmar Zaiane and Andrew Foss). Moreover, PAKDD 2002 offered four international workshops on "Knowledge Discovery in Multimedia and Complex Data", "Mining Data across Multiple Customer Touchpoints for CRM", "Toward the Foundation of Data Mining", and "Text Mining". Articles from these workshops have been published separately.

All of this work would not have been possible without the dedication and professional work of many colleagues. We would like to express our sincere appreciation to all contributors to the conference for submitting papers, offering tutorials, and organizing workshops. Special thanks go to our honorary chairs, David C. L. Liu from the National Tsing Hua University and Benjamin Wah from the University of Illinois, Urbana-Champaign, for their leadership and advice on planning this conference. We are also deeply grateful to Huan Liu for serving as the Workshop Chair, Yao-Nan Lien for being the Tutorial Chair, Chia-Hui Chang and Vincent Tseng for serving as the Industrial Chairs, Show-Jane Yen and Yue-Shi Lee for being the Publication Chairs, and Chun-Nan Hsu for acting as Local Arrangement Chair. Last but not least, we are indebted to the Steering Committee Members (Chaired by Hongjun Lu) for their timely guidance and unwavering support. All of them deserve our very sincere gratitude.

Finally, we hope you all had a very pleasant stay at PAKDD 2002 in Taipei, and we wish you great success in your KDD endeavors.

May 2002

Arbee L. P. Chen
Jiawei Han
Ming-Syan Chen
Philip S. Yu
Bing Liu

PAKDD 2002 Conference Committee

Honorary Chairs:

David C. L. Liu	National Tsing Hua University, Taiwan
Ben Wah	University of Illinois, Urbana-Champaign, USA

Conference Co-chairs:

Arbee L.P. Chen	National Tsing Hua University, Taiwan
Jiawei Han	University of Illinois, Urbana-Champaign, USA

Program Chairs/Co-chairs:

Ming-Syan Chen	National Taiwan University, Taiwan (Chair)
Philip S. Yu	IBM T.J. Watson Research Center, USA (Co-chair)
Bing Liu	National University of Singapore, Singapore (Co-chair)

Workshop Chair:

Huan Liu	Arizona State University, USA

Tutorial Chair:

Yao-Nan Lien	National Chengchi University, Taiwan

Industrial Chairs:

Chia-Hui Chang	National Central University, Taiwan
Vincent S. M. Tseng	National Cheng Kung University, Taiwan

Publication Chairs:

Show-Jane Yen	Fu Jen Catholic University, Taiwan
Yue-Shi Lee	Ming Chuan University, Taiwan

Local Arrangement Chair:

Chun-Nan Hsu	Academia Sinica, Taiwan

PAKDD Steering Committee:

Hongjun Lu	Hong Kong University of Science & Technology, Hong Kong (Chair)
Hiroshi Motoda	Osaka University, Japan (Co-chair)
David W. Cheung	The University of Hong Kong, Hong Kong
Masaru Kitsuregawa	The University of Tokyo, Japan
Rao Kotagiri	University of Melbourne, Australia
Huan Liu	Arizona State University, USA
Takao Terano	University of Tsukuba, Japan
Graham illiams	CSIRO, Australia
Ning Zhong	Maebashi Institute of Technology, Japan
Lizhu Zhou	Tsinghua University, China

PAKDD 2002 Program Committee

Roberto Bayardo	IBM Almaden Research Center, USA
Michael Berthold	U.C. Berkeley, USA
Chia-Hui Chang	National Central University, Taiwan
Edward Chang	UCSB, USA
Meng-Chang Chen	Academia Sinica, Taiwan
David W. Cheung	The University of Hong Kong
Umeshwar Dayal	Hewlett-Packard Labs, USA
Guozhu Dong	Wright State University, USA
Ada Fu	Chinese University of Hong Kong, Hong Kong
Yike Guo	Imperial College, UK
Eui-Hong Han	University of Minnesota, USA
Jiawei Han	University of Illinois, Urbana-Champaign, USA
David Hand	Imperial College, UK
Jayant Haritsa	Indian Institute of Science, India
Robert Hilderman	University of Regina, Canada
Howard Ho	IBM Almaden Research Center, USA
Se-June Hong	IBM T.J. Watson Research Center, USA
Kien Hua	University of Central Florida, USA
Ben Kao	University of Hong Kong, Hong Kong
Masaru Kitsuregawa	University of Tokyo, Japan
Willi Klosgen	GMD, Germany
Kevin Korb	Monash University, Australia
Rao Kotagiri	University of Melbourne, Australia
Vipin Kumar	University of Minnesota, USA
Wai Lam	Chinese University of Hong Kong, Hong Kong
C. Lee Giles	Penn State University, USA
Chiang Lee	Cheng-Kung University, Taiwan
Suh-Yin Lee	Chia-Tung University, Taiwan
Chung-Sheng Li	IBM T.J. Watson Research Center, USA
Qing Li	City University of Hong Kong, Hong Kong
T. Y. Lin	San Jose State University, USA
Huan Liu	Arizona State University, USA
Ling Liu	Georgia Tech Institute, USA
Xiaohui Liu	University of Brunel, UK
Hongjun Lu	Hong Kong UST, Hong Kong
Yuchang Lu	Tsing Hua University, China
Hiroshi Motoda	Osaka University, Japan
Raymond Ng	UBC, Canada
Yen-Jeng Oyang	National Taiwan University, Taiwan
Zhongzhi Shi	China
Kyuseok Shim	KAIST, Korea
Arno Siebes	Holland
Ramakrishnan Srikant	IBM Almaden Research Center, USA
Jaideep Srivastava	University of Minnesota, USA
Einoshin Suzuki	Yokohama National University, Japan

Ah-Hwee Tan	Kent Ridge Digital Labs, Singapore
Changjie Tang	Sichuan University, China
Takao Terano	University of Tsukuba, Japan
Bhavani Thurasingham	MITRE, USA
Hannu T. T. Toivonen	Nokia Research, Finland
Shin-Mu Tseng	Cheng-Kung University, Taiwan
Anthony Tung	National University of Singapore, Singapore
Jason Wang	New Jersey IT, USA
Ke Wang	Simon Fraser University, Canada
Kyu-Young Whang	KAIST, Korea
Graham Williams	CSIRO, Australia
Xingdong Wu	Colorado School of Mines, USA
Jiong Yang	IBM T.J. Watson Research Center, USA
Yiyu Yao	University of Regina, Canada
Clement Yu	University of Illinois, USA
Osmar R. Zaiane	University of Alberta, Canada
Mohammed Zaki	Rensselaer Poly Institute, USA
Zijian Zheng	Blue Martini Software, USA
Ning Zhong	Maebashi Institute of Technology, Japan
Aoying Zhou	Fudan University, China
Lizhu Zhou	Tsinghua University, China

Table of Contents

Interestingness

Sequence Mining

Clustering

Web Mining

Association Rules (II)

Semi-structure & Concept Mining

Data Warehouse and Data Cube

Bio-Data Mining

Classification (II)

Temporal Mining

Classification (III)

Outliers, Missing Data, and Causation

Network Data Mining and Analysis: The $\mathcal{NEMESIS}$ Project

Minos Garofalakis and Rajeev Rastogi

Bell Labs, Lucent Technologies

Abstract. Modern communication networks generate large amounts of operational data, including traffic and utilization statistics and alarm/fault data at various levels of detail. These massive collections of *network-management* data can grow in the order of several Terabytes per year, and typically hide "knowledge" that is crucial to some of the key tasks involved in effectively managing a communication network (e.g., capacity planning and traffic engineering). In this short paper, we provide an overview of some of our recent and ongoing work in the context of the $\mathcal{NEMESIS}$ project at Bell Laboratories that aims to develop novel data warehousing and mining technology for the effective storage, exploration, and analysis of massive network-management data sets. We first give some highlights of our work on *Model-Based Semantic Compression (MBSC)*, a novel data-compression framework that takes advantage of attribute semantics and data-mining models to perform lossy compression of massive network-data tables. We discuss the architecture and some of the key algorithms underlying $\mathcal{SPARTAN}$, a model-based semantic compression system that exploits predictive data correlations and prescribed error tolerances for individual attributes to construct concise and accurate *Classification and Regression Tree (CaRT)* models for entire columns of a table. We also summarize some of our ongoing work on warehousing and analyzing network-fault data and discuss our vision of how data-mining techniques can be employed to help automate and improve fault-management in modern communication networks. More specifically, we describe the two key components of modern fault-management architectures, namely the *event-correlation* and the *root-cause analysis* engines, and propose the use of mining ideas for the automated inference and maintenance of the models that lie at the core of these components based on warehoused network data.

1 Introduction

Besides providing easy access to people and data around the globe, modern communication networks also *generate* massive amounts of operational data throughout their lifespan. As an example, Internet Service Providers (ISPs) continuously collect traffic and utilization information over their network to enable key network-management applications. This information is typically collected through monitoring tools that gather switch- and router-level data, such as SNMP/RMON probes [13] and Cisco's NetFlow measurement tools [1]. Such tools typically collect traffic data for each network element at fine granularities (e.g., at the level of individual packets or packet flows between source-destination pairs), giving rise to massive volumes of network-management data over time [7]. Packet traces collected for traffic management in the Sprint IP backbone

M.-S. Chen, P.S. Yu, and B. Liu (Eds.): PAKDD 2002. LNAI 2336, pp. 1–12, 2002.

amount to 600 Gigabytes of data per day [7]! As another example, telecommunication providers typically generate and store records of information, termed "Call-Detail Records" (CDRs), for every phone call carried over their network. A typical CDR is a fixed-length record structure comprising several hundred bytes of data that capture information on various (categorical and numerical) attributes of each call; this includes network-level information (e.g., endpoint exchanges), time-stamp information (e.g., call start and end times), and billing information (e.g., applied tariffs), among others [4]. These CDRs are stored in tables that can grow to truly massive sizes, in the order of several Terabytes per year.

A key observation is that these massive collections of network-traffic and CDR data typically hide invaluable "knowledge" that enables several key network-management tasks, including application and user profiling, proactive and reactive resource management, traffic engineering, and capacity planning. Nevertheless, techniques for effectively managing these massive data sets and uncovering the knowledge that is so crucial to managing the underlying network are still in their infancy. Contemporary network-management tools do little more than elaborate report generation for all the data collected from the network, leaving most of the task of inferring useful knowledge and/or patterns to the human network administrator(s). As a result, effective network management is still viewed as more of an "art" known only to a few highly skilled (and highly sought-after) individuals. It is our thesis that, in the years to come, network management will provide an important application domain for very innovative, challenging and, at the same time, practically-relevant research in data mining and data warehousing.

This short abstract aims to provide an overview of our recent and ongoing research efforts in the context of $\mathcal{NEMESIS}$ (NEtwork-Management data warEhousing and analySIS) , a Bell Labs' research project that targets the development of novel data warehousing and mining technology for the effective storage, exploration, and analysis of massive network-management data sets. Our research agenda for $\mathcal{NEMESIS}$ encompasses several challenging research themes, including data reduction and approximate query processing [2,5,6], mining techniques for network-fault management, and managing and analyzing continuous data streams. In this paper, we first give some highlights of our recent work on *Model-Based Semantic Compression (MBSC)*, a novel data-compression framework that takes advantage of attribute semantics and data-mining models to perform lossy compression of massive network-data tables. We also describe the architecture and some of the key algorithms underlying $\mathcal{SPARTAN}$, a system built based on the MBSC paradigm, that exploits predictive data correlations and prescribed error tolerances for individual attributes to construct concise and accurate *Classification and Regression Tree (CaRT)* models for entire columns of a table [2]. We then turn to our ongoing work on warehousing and analyzing network-fault data and discuss our vision of how data-mining techniques can be employed to help automate and improve fault-management in modern communication networks. More specifically, we describe the two key components of modern fault-management architectures, namely the *event-correlation* and the *root-cause analysis* engines, and offer some (more speculative) proposals on how mining ideas can be exploited for the automated inference and

maintenance of the models that lie at the core of these components based on warehoused network data.

2 Model-Based Semantic Compression for Network-Data Tables

Data compression issues arise naturally in applications dealing with massive data sets, and effective solutions are crucial for optimizing the usage of critical system resources like storage space and I/O bandwidth, as well as network bandwidth (for transferring the data) [4,7]. Several statistical and dictionary-based compression methods have been proposed for text corpora and multimedia data, some of which (e.g., Lempel-Ziv or Huffman) yield provably optimal asymptotic performance in terms of certain ergodic properties of the data source. These methods, however, fail to provide adequate solutions for compressing massive data tables, such as the ones that house the operational data collected from large ISP and telecom networks. The reason is that all these methods view a table as a large byte string and do not account for the complex dependency patterns in the table. Compared to conventional compression problems, effectively compressing massive tables presents a host of novel challenges due to several distinct characteristics of table data sets and their analysis.

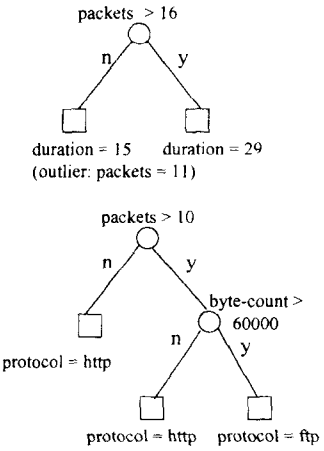
• **Semantic Compression.** Existing compression techniques are "syntactic" in the sense that they operate at the level of consecutive bytes of data. Such syntactic methods typically fail to provide adequate solutions for table-data compression, since they essentially view the data as a large byte string and do not exploit the complex dependency patterns in the table. Effective table compression mandates techniques that are *semantic* in nature, in the sense that they account for and exploit both (1) existing data dependencies and correlations among attributes in the table; and, (2) the meanings and dynamic ranges of individual attributes (e.g., by taking advantage of the specified error tolerances).

• **Approximate (Lossy) Compression.** Due to the exploratory nature of many data-analysis applications, there are several scenarios in which an exact answer may not be required, and analysts may in fact prefer a fast, approximate answer, as long as the system can guarantee an *upper bound on the error of the approximation*. For example, during a drill-down query sequence in ad-hoc data mining, initial queries in the sequence frequently have the sole purpose of determining the truly interesting queries and regions of the data table. Thus, in contrast to traditional lossless data compression, the compression of massive tables can often afford to be *lossy*, as long as some (user- or application-defined) upper bounds on the compression error are guaranteed by the compression algorithm. This is obviously a crucial differentiation, as even small error tolerances can help us achieve much better compression ratios.

In our recent work [2], we have proposed *Model-Based Semantic Compression (MBSC)*, a novel data-compression framework that takes advantage of attribute semantics and data-mining models to perform guaranteed-error, lossy compression of massive data tables. Abstractly, MBSC is based on the novel idea of exploiting data correlations and user-specified "loss"/error tolerances for individual attributes to construct concise data mining models and derive the best possible compression scheme for the data based

protocol	duration	byte-count	packets
http	12	2,000	1
http	16	24,000	5
ftp	27	100,000	24
http	15	20,000	8
ftp	32	300,000	35
http	19	40,000	11
http	26	58,000	18
ftp	18	80,000	15

(a) Tuples in Table



(b) CaRT Models

Fig. 1. Model-Based Semantic Compression.

on the constructed models. To make our discussion more concrete, we focus on the architecture and some of the key algorithms underlying *SPARTAN*[1], a system that takes advantage of attribute correlations and error tolerances to build concise and accurate *Classification and Regression Tree (CaRT)* models [3] for entire columns of a table. More precisely, *SPARTAN* selects a certain subset of attributes (referred to as *predicted* attributes) for which no values are explicitly stored in the compressed table; instead, concise CaRTs that predict these values (within the prescribed error bounds) are maintained. Thus, for a predicted attribute X that is strongly correlated with other attributes in the table, *SPARTAN* is typically able to obtain a very succinct CaRT predictor for the values of X, which can then be used to completely eliminate the column for X in the compressed table. Clearly, storing a compact CaRT model in lieu of millions or billions of actual attribute values can result in substantial savings in storage.

Example 21 *Consider the table with 4 attributes and 8 tuples shown in Figure 1(a), where each tuple represents a* data flow *in an IP network. The categorical attribute* protocol *records the application-level protocol generating the flow; the numeric attribute* duration *is the time duration of the flow; and, the numeric attributes* byte-count *and* packets *capture the total number of bytes and packets transferred. Let the acceptable errors due to compression for the numeric attributes* duration, byte-count, *and* packets *be 3, 1,000, and 1, respectively. Also, assume that the* protocol *attribute has to be compressed without error (i.e., zero tolerance). Figure 1(b) depicts a regression tree for predicting the* duration *attribute (with* packets *as the predictor attribute) and a classi-*

[1] [From Webster] **Spartan:** */'spart-*n/* (1) of or relating to Sparta in ancient Greece, (2) a: marked by strict self-discipline and avoidance of comfort and luxury, b: sparing of words : TERSE : LACONIC.

fication tree for predicting the protocol *attribute (with* packets *and* byte-count *as the predictor attributes). Observe that in the regression tree, the predicted value of duration (label value at each leaf) is almost always within 3, the specified error tolerance, of the actual tuple value. For instance, the predicted value of* duration *for the tuple with* packets = *1 is 15 while the original value is 12. The only tuple for which the predicted value violates this error bound is the tuple with* packets = *11, which is an marked as an outlier value in the regression tree. There are no outliers in the classification tree. By explicitly storing, in the compressed version of the table, each outlier value along with the CaRT models and the projection of the table onto only the predictor attributes (*packets *and* byte-count*), we can ensure that the error due to compression does not exceed the user-specified bounds. Further, storing the CaRT models (plus outliers) for* duration *and* protocol *instead of the attribute values themselves results in a reduction from 8 to 4 values for* duration *(2 labels for leaves + 1 split value at internal node + 1 outlier) and a reduction from 8 to 5 values for* protocol *(3 labels for leaves + 2 split values at internal nodes).* ∎

To build an effective CaRT-based compression plan for the input data table, $\mathcal{SPARTAN}$ employs a number of sophisticated techniques from the areas of knowledge discovery and combinatorial optimization. Below, we list some of $\mathcal{SPARTAN}$'s salient features.

• **Use of Bayesian Network to Uncover Data Dependencies.** A Bayesian network is a directed acyclic graph (DAG) whose edges reflect strong predictive correlations between nodes of the graph [12]. $\mathcal{SPARTAN}$ uses a Bayesian network on the table's attributes to dramatically reduce the search space of potential CaRT models since, for any attribute, the most promising CaRT predictors are the ones that involve attributes in its "neighborhood" in the network.

• **Novel CaRT-selection Algorithms that Minimize Storage Cost.** $\mathcal{SPARTAN}$ exploits the inferred Bayesian network structure by using it to intelligently guide the selection of CaRT models that minimize the overall storage requirement, based on the prediction and materialization costs for each attribute. We demonstrate that this model-selection problem is a strict generalization of the *Weighted Maximum Independent Set (WMIS)* problem [8], which is known to be $\mathcal{NP}$-hard. However, by employing a novel algorithm that effectively exploits the discovered Bayesian structure in conjunction with efficient, near-optimal WMIS heuristics, $\mathcal{SPARTAN}$ is able to obtain a good set of CaRT models for compressing the table.

• **Improved CaRT Construction Algorithms that Exploit Error Tolerances.** Since CaRT construction is computationally-intensive, $\mathcal{SPARTAN}$ employs the following three optimizations to reduce CaRT-building times: (1) CaRTs are built using random samples instead of the entire data set; (2) leaves are not expanded if values of tuples in them can be predicted with acceptable accuracy; (3) pruning is integrated into the tree growing phase using novel algorithms that exploit the prescribed error tolerance for the