Lecture Notes in Computer Science 8506

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Hua Wang Mohamed A. Sharaf (Eds.)

Databases Theory and Applications

25th Australasian Database Conference, ADC 2014
Brisbane, QLD, Australia, July 14-16, 2014
Proceedings

Volume Editors

Hua Wang
Victoria University
College of Engineering and Science
Centre for Applied Informatics (CAI)
Ballarat Road
Footscray, VIC 8001, Australia
E-mail: hua.wang@vu.edu.au

Mohamed A. Sharaf
The University of Queensland
Faculty of Engineering, Architecture and Information Technology
School of Information Technology and Electrical Engineering
Brisbane St. Lucia, QLD 4072, Australia
E-mail: m.sharaf@uq.edu.au

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-08607-1 e-ISBN 978-3-319-08608-8
DOI 10.1007/978-3-319-08608-8
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014941783

LNCS Sublibrary: SL 3 – Information Systems and Application,
incl. Internet/Web and HCI

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

It is our pleasure to present to you the proceedings of the 25th Australasian Database Conference (ADC2014), which took place in Brisbane, Australia. ADC2014 is an annual forum for researchers and practitioners from Australia, New Zealand and around the world to share the latest research progress and novel applications of database systems, data driven applications, and data analytics. The mission of ADC is to exchange novel research solutions to problems of today's information society that fulfill the needs of heterogeneous applications and environments, as well as to identify new issues and directions for future research and development work. ADC2014 seeks papers from academia and industry presenting research on all practical and theoretical aspects of advanced database theory and applications, as well as case studies and implementation experiences. All topics related to databases are of interest and within the scope of the conference. ADC gives researchers and practitioners a unique opportunity to share their perspectives with others interested in the various aspects of database systems.

The ADC 2014 Program Committee accepted those papers considered to be of ADC quality without setting any predefined quota, and was impressed by the quality of the submissions. The conference received 38 submissions and accepted 15 papers. The Program Committee who selected the papers consisted of 28 members from around the globe including Singapore, Bangladesh, Germany, Ireland, Japan, Switzerland, China, and the United States. The Program Committee was thorough and dedicated to the reviewing process with each paper peer reviewed in full by at least two independent reviewers, and in some cases three or four referees produced independent reviews. A conscious decision was made to select papers for which all reviews were positive and favorable. While this challenged the determination, and some high-quality papers were finally not included, we are confident that the results demonstrate a very solid program and that each paper makes a strong contribution to the proceedings.

We would like to thank all our colleagues who served on the Program Committee or acted as external reviewers. We would also like to thank all the authors who submitted papers, both accepted and rejected, as well as the conference attendees. This conference is held for you, and we hope that these proceedings provide an overview of our vibrant research community and its activities. We encourage all database researchers to contribute and make submissions to the next ADC conference.

July 2014

Hua Wang
Mohamed A. Sharaf

General Chair's Welcome Message

Welcome to the 25th Australasian Database Conference (ADC2014)! ADC is a leading Australia and New Zealand based international conference on research, development and applications of database systems and related areas. Previous ADC conferences were held as part of the Australasian Computer Science Week (ACSW). In the past 10 years, ADC was held in Adelaide (2013), Melbourne (2012), Perth (2011), Brisbane (2010), Wellington (2009), Wollongong (2008), Ballarat (2007), Hobart (2006), Newcastle (2005), and Dunedin (2004).

Australasia has an increasingly large, very active and internationally highly visible group of database researchers. Based on wide community consultation, ADC 2014 departs from its tradition as part of the Australasian Computer Science Week (ACSW) to become an independent conference with an expanded research program, a PhD School and a community-building focus. So, in that sense, ADC 2014 is the first of the new ADC conference series.

The conference this year had two eminent keynote speakers: Timos Sellis from the Royal Melbourne Institute of Technology, Australia, and Divesh Srivastava from AT&T, USA. In addition to 15 full papers and 6 short papers carefully selected by the Program Committee, we were also very fortunate to have three invited talks presented by world-leading researchers, ChengXiang Zhai from UIUC, Lei Chen from HKUST and Jeffrey Yu from CUHK. We also have a three-day PhD School program as part of this year's ADC. We wish to take this opportunity to thank all speakers, authors and organisers. I would also specially thank the Program Committee co-chairs Hua Wang and Mohamed A. Sharaf for their dedication and effort in ensuring a high quality program. I would also like to thank our PhD School convenor, Heng Tao Shen, and our key local organiser, Kath Williamson, for their contributions to making this year's new ADC a success.

As a member of the new ADC's Steering Committee, I would like to sincerely thank my fellow Committee members: Rao Kotagiri (The University of Melbourne), Timos Sellis (RMIT), Gill Dobbie (University of Auckland), Alan Fekete (The University of Sydney), Xuemin Lin (UNSW), and Yanchun Zhang (Victoria University), for their support to set a new direction for ADC and offering the opportunity to host the first new ADC conference at The University of Queensland in Brisbane.

Brisbane is a beautiful city and UQ has one of the best university campuses in Australia. All ADC2014 participants are sure to enjoy the conference, the campus and the city.

General Chair — Xiaofang Zhou (The University of Queensland)

Organization

General Chair

Xiaofang Zhou	University of Queensland, Australia

PC Co-chairs

Hua Wang	Victoria University, Australia
Mohamed A. Sharaf	University of Queensland, Australia

Steering Committee

Rao Kotagiri	University of Melbourne, Australia
Timos Sellis	RMIT University, Australia
Gill Dobbie	University of Auckland, New Zealand
Alan Fekete	University of Sydney, Australia
Xuemin Lin	University of New South Wales, Australia
Yanchun Zhang	Victoria University, Australia

Program Committee

Ying Zhang	University of New South Wales, Australia
Sebastian Maneth	University of Edinburgh, UK
Junhu Wang	Griffith University, Australia
Gang Li	Deakin University, Australia
Mohammed Eunus Ali	University of Engineering and Technology, Bangladesh
Michael E. Houle	National Institute of Informatics, Japan
Ruixuan Li	Huazhong University of Science and Technology, China
Sarana Nutanong	Johns Hopkins University, USA
Shichao Zhang	Guangxi Normal University, China
Panos Chrysanthis	University of Pittsburgh, USA
Chaoyi Pang	CSIRO, Australia
James A. Thom	RMIT University, Australia
Xue Li	University of Queensland, Australia
Yoshiharu Ishikawa	Nagoya University, Japan
Xiangmin Zhou	CSIRO, Australia

Markus Stumptner	University of South Australia, Australia
Annika Hinze	University of Waikato, New Zealand
Zahir Tari	RMIT University, Australia
Bing Tian Dai	Singapore Management University, Australia
Jinli Cao	Latrobe University, Australia
Miyuki Nakano	University of Tokyo, Japan
Evaggelia Pitoura	University of Ioannina, Greece
Bela Stantic	Griffith University, Australia
Hye-Young Paik	University of New South Wales, Australia
Ge Yu	Northeastern University, China
Laurianne Sitbon	Queensland University of Technology, Australia
Chengfei Liu	Swinburne University of Technology, Australia

Invited Talks

Analysing Big Trajectory Data: Theory, Algorithms and Applications

Kai Zheng

University of Queensland

Abstract. The prevalence of GPS sensors and mobile devices has enabled tracking the movements of almost any kind of moving objects such as vehicles, humans and animals. As a result, in the past decade we have witnessed unprecedented increase of trajectory data both in volume and variety. With some attributes such as variable lengths, uncontrolled quality, high redundancy and uncertainty and so on, trajectory data challenge the traditional methodologies and practices in many research areas including data storage and indexing, data mining and analytics, information retrieve, etc. Trajectory data management has been attracting numerous research interests from both academia and industry due to its tremendous value and benefits in a variety of critical applications like traffic analysis, fleet management, trip planning, location-based recommendation, etc. In this tutorial, we will talk about the challenges, techniques and open problems with the focus on similarity-based analytics, the foundation of trajectory management, and covering a range of topics from fundamental theory, algorithms to advanced applications.

Boosting Methods in Machine Learning

Chunhua Shen
University of Adelaide

Abstract. Many machine learning and data mining tasks favour fast and yet accurate classification methods. The classification speed is not only a matter of time-efficiency but is often crucial to achieve good accuracy. Standard kernel machines such as Support Vector Machine (SVM) are slow and methods for rapid classification have been pursued. Boosting classifiers have been so successful owing to its fast computation and yet comparable or sometimes better accuracy to kernel methods, being a standard method in many areas. Boosting as a representative ensemble learning method, which aggregates simple weak learners, can be seen as a flat tree structure when each learner is a decision-stump. When trees are used as weak learners, boosting methods learn a linearly weighted decision forest. We will overview the fundamental theory of boosting in the first part of this course.

Recently, structured learning has found many applications in text analysis and computer vision. Thus far it has not been clear how one can train a boosting model that is directly optimised for predicting multivariate or structured outputs. To bridge this gap, inspired by structured support vector machines, a boosting algorithm for structured output prediction is introduced, which we refer to as StructBoost. StructBoost supports nonlinear structured learning by combining a set of weak structured learners. As structured SVM generalises SVM, the StructBoost generalises standard boosting approaches such as AdaBoost, or LPBoost to structured learning. The resulting optimization problem of StructBoost is more challenging than Structured SVM in the sense that it may involve exponentially many variables and constraints. In contrast, for Structured SVM one usually has an exponential number of constraints and a cutting-plane method is used. In order to efficiently solve StructBoost, we formulate an equivalent 1-slack formulation and solve it using a combination of cutting planes and column generation. We show the versatility and usefulness of StructBoost on a range of problems.

Big Data Mining on SAP HANA

Hoyoung Jeung

SAP

Abstract. This talk will cover how the theories in data mining and machine learning are implemented and used in the state-of-the-art in-memory computing technology, SAP HANA. In particular, Dr.Jeung will share his extensive experience in predictive analysis on big business data, discussing about hidden insights when dealing with complex algorithms on extremely large data.

Statistical Methods for Mining Big Text Data

Chengxiang Zhai

University of Illinois Urbana-Champaign, USA

Abstract. Text data, broadly including all kinds of natural language text produced by humans (e.g., web pages, social media, email messages, news articles, government documents, and scientific literature), have been growing dramatically recently. This creates great opportunities for applying computational methods to mine large amounts of text data to discover all kinds of useful knowledge, especially knowledge about people's opinions, preferences, and behavior. Due to the difficulty in precisely understanding natural language by computers, scalable text mining algorithms tend to be based on statistical analysis and probabilistic reasoning. In this tutorial, I will systematically review the major statistical methods developed for mining text data, with a focus on covering probabilistic topic models for mining topics and topical patterns in text data, and statistical methods for integrating and analyzing scattered online opinions.

Crowdsourcing over Big Data, Are We There Yet?

Lei Chen

Hong Kong University of Science & Technology

Abstract. Recently, the popularity of crowdsourcing has brought a new opportunity to engage human intelligence into various data analysis tasks. Compared with computer systems, crowds are good at handling items with human-intrinsic values or features. Existing approaches develop sophisticated methods by utilizing the crowd as a new type of processor, a.k.a. HPU (Human Processing Unit). As a consequence, tasks executed on HPU are called HPU-based tasks. Now we are in the Big Data Era, a nature question arises: How about crowdsourcing over Big Data, are we there yet? In this talk, I will first briefly review the history of crowdsourcing and discuss the key issues related to crowdsourcing. Then, I will demonstrate the power of crowdsourcing in solving the well-known and very hard data integration problem, schema matching, and discuss how to migrate the power of crowdsourcing to a social media platform whose users can serve as a huge reservoir of workers. Finally, I will highlight some research challenges about crowdsourcing over Big Data.

Large Graph Processing

Jeffrey Yu

Chinese University of Hong Kong

Abstract. The real applications that need graph processing techniques to handle a large graph can be found from many real applications including online social networks, biological networks, ontology, transportation networks, etc. In this talk, we will discuss some selected research topics on graph mining and graph query processing over large graphs. For graph mining, we will focus on ranking nodes in a large graph. We will discuss ranking over trust networks, random-walk domination, and diversified ranking. For ranking nodes over trust network, we discuss how to take the trust score into consideration while ranking. For the random-walk domination, we discuss the techniques for handling item-placement in online social networks and ads-placement in advertisement networks. For diversified ranking, we discuss how to find top-k nodes that match the user query and are very different from each other. For graph query processing, we will discuss top-k structural diversity search, finding the maximal cliques in massive networks, and I/O efficient computing techniques that make a large directed graph small and simple. The other related topics may be also addressed in this talk.

Keynotes

Selecting Sources Wisely for Integration

Xin Luna Dong[1], Theodoros Rekatsinas[2], Barna Saha[3], and Divesh Srivastava[3]

[1]Google Inc., Mountain View, CA 94043, USA
lunadong@google.com
[2]University of Maryland, College Park, MD 20742, USA
thodrek@cs.umd.edu
[3]AT&T Labs-Research, Bedminster, NJ 07921, USA
{barna, divesh}@research.att.com

Abstract. Data integration is a challenging task due to the large numbers of autonomous data sources, which necessitates the development of techniques to reason about the costs and benefits of acquiring and integrating data. Too many sources can result in a huge integration cost, and low quality sources can be detrimental to the benefit of integration. In this talk, we present the problem of *source selection*, that is, identifying the subset of sources before integration that maximize the profit (benefit − cost) of integration, for static and dynamic sources. To address this problem, we propose techniques that, inspired by the marginalism principle in economic theory, integrate a source only if its marginal benefit is higher than its marginal cost. We quantify the integration benefit in terms of the quality of the integrated data, which is characterized using a set of data quality metrics, including coverage, freshness and accuracy, and develop statistical models for estimating these metrics. Although source selection is NP-complete, we show that for many practical cases solutions to our problem can be found in polynomial time with approximation guarantees. Finally, we empirically establish the effectiveness and scalability of our techniques on real-world and synthetic data.

Data Ecosystems: From Very Large Data Bases to Big Data Infrastructures

Timos Sellis

Computer Science & Info Tech
RMIT University
timos.sellis@rmit.edu.au

Abstract. Data ecosystems involve the coexistence of one or more data collections, typically databases, and their surrounding applications for data entry and retrieval. For decades, both data and ecosystem management have failed to address significant, costly and labor-consuming challenges which involve (a) the departure from databases focusing on alphanumeric data only, (b) their inability to be integrated and provide transparent access and composition facilities for heterogeneous data, (c) their static querying nature, which is deprived of personal, context-aware or interactive characteristics, (d) the enforcement of DBMS operation over monolithic servers, and, (e) the complete indifference to problems of evolution and adaptation over time.

In this talk we address issues around the methodologies, the theoretical and modeling foundations as well as the algorithmic techniques and the necessary software architectures that will facilitate the personalization, integration, and evolution management facilities for data ecosystems that operate over a decentralized infrastructure for a large variety of data types.

Table of Contents