Lecture Notes in Computer Science 7678

Srinath Srinivasa Vasudha Bhatnagar (Eds.)

Big Data Analytics

First International Conference, BDA 2012
New Delhi, India, December 24-26, 2012
Proceedings

Springer

Volume Editors

Srinath Srinivasa
International Institute of Information Technology
26/C, Electronics City, Hosur Road, Bangalore 560100, India
E-mail: sri@iiitb.ac.in

Vasudha Bhatnagar
University of Delhi, Faculty of Mathematical Sciences
107, Department of Computer Science
Delhi 110007, India
E-mail: vbhatnagar@cs.du.ac.in

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-35541-7 e-ISBN 978-3-642-35542-4
DOI 10.1007/978-3-642-35542-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012953578

CR Subject Classification (1998): H.3, I.2, H.4, H.2.8, I.4, H.5

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

The arrival of the so-called *Petabyte Age* has compelled the analytics community to pay serious attention to development of scalable algorithms for intelligent data analysis. In June 2008, *Wired* magazine featured a special section on "The Petabyte Age" and stated that "..our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology." The recent explosion in social computing has added to the vastly growing amounts of data from which insights can be mined. The term "Big Data" is now emerging as a catch-all phrase to denote the vast amounts of data at a scale that requires a rethink of conventional notions of data management.

There is a saying among data researchers that "more data beats better algorithms." Big Data provide ample opportunities to discern hitherto inconceivable insights from data sets. This, however, comes with significant challenges in terms of both computational and storage expense, of the type never addressed before. Volume, velocity, and variability in Big Data repositories necessitate advancing analytics beyond operational reporting and dashboards. Early attempts to address the issue of scalability were handled by development of incremental data mining algorithms. Other traditional approaches to solve scalability problems included sampling, processing data in batches, and development of parallel algorithms. However, it did not take long to realize that all of these approaches, except perhaps parallelization, have limited utility.

The International Conference on Big Data Analytics (BDA 2012) was conceived against this backdrop, and is envisaged to provide a platform to expose researchers and practitioners to ground-breaking opportunities that arise during analysis and processing of massive volumes of distributed data stored across clusters of networked computers. The conference attracted a total of 42 papers, of which 37 were research track submissions. From these, five regular papers and five short papers were selected, leading to an acceptance rate of 27%.

Four tutorials were also selected and two tutorials were included in the proceedings. The first tutorial entitled "Scalable Analytics: Algorithms and Systems" addresses implementation of three popular machine learning algorithms in a Map-Reduce environment. The second tutorial, "Big-Data: Theoretical, Engineering and Analytics Perspectives," gives a bird's eye view of the Big Data landscape, including technology, funding, and the emerging focus areas. It also deliberates on the analytical and theoretical perspectives of the ecosystem.

The accepted research papers address several aspects of data analytics. These papers have been logically grouped into three broad sections: Data Analytics Applications, *Knowledge Discovery Through Information Extraction*, and *Data Models in Analytics.*

In the first section, Basil et al. compare several statistical machine learning techniques over electro-cardiogram (ECG) datasets. Based on this study, they make recommendations on features, sampling rate, and the choice of classifiers in a realistic setting. Yasir et al. present an approach for information requirements elicitation (IRE), which is an interactive approach for building queries, by asking a user his/her information needs. Gupta et al. look at Big Data from the perspective of database management. They divide analytics over Big Data into two broad classes: data in rest and data in motion, and propose separate database solutions for both of them. Reddy et al. describe their efforts in imparting practical education in the area of agriculture by means of a virtual lab. A virtual "crop lab" designed by the authors contains large amounts of practical data about crops that are indexed and summarized. The authors speculate on pedagogic methodologies necessary for imparting practical education using such crop data.

In the second section, Yasir et al. address the problem of schema summarization from relational databases. Schema summarization poses significant challenges of semantically correlating different elements of the database schema. The authors propose a novel technique of looking into the database documentation for semantic cues to aid in schema summarization. Nambiar et al. present a faceted model for computing various aspects of a topic from social media datasets. Their approach is based on a model called the *colon classification scheme* that views social media data along five dimensions: Personality, Matter, Energy, Space, and Time. Gupta et al. address text segmentation problems and propose a technique called Analog Text Entailment that assigns an entailment score to extracted text segments from a body of text in the range [0,1], denoting the relative importance of the segment based on its constituent sentences. Kumar et al. study the price movements of gold and present a model explaining the price movements using three feature sets: macro-economic factors, investor fear features, and investor behavior features.

Finally, in the last section, Ryu et al. propose a method for dynamically generating classifiers to build an ensemble of classifiers for handling variances in streaming datasets. Mohanty and Sajith address the problem of eigenvalue computations for very large nonsymmetric matrix datasets and propose an I/O efficient algorithm for reduction of the matrix dataset into Hessenberg form, which is an important step in the eigenvalue computation. Gupta et al. address the problem of information security in organizational settings and propose a notion of "context honeypot" – a mechanism for luring suspected individuals with malicious intent into false areas of the dataset. They analyze the elements required for luring conditions and propose a mathematical model of luring. Kim et al. address recommender systems for smart TV contents, which are characterized by large sparse matrix datasets. They propose a variant of collaborative filtering for building efficient recommender systems. Kumar and Kumar address the problem of selecting optimal materialized views over high-dimensional datasets, and propose simulated annealing as a solution approach.

We would like to extend our gratitude to the supporting institutes: University of Delhi, University of Aizu, Indian Institute of Technology Delhi, and ACM Delhi-NCR Chapter. Thanks are also due to our sponsors: Microsoft Corporation, Hewlett Packard India, and IBM India Research Lab. And last but not the least, we extend our hearty thanks to all the Program, Organizing, and Steering Committee members, external reviewers, and student volunteers of BDA 2012.

December 2012

Srinath Srinivasa
Vasudha Bhatnagar

Organization

BDA 2012 was organized by the Department of Computer Science, Univeristy of Delhi, India, in collaboration with University of Aizu, Japan.

Steering Committee

N. Vijyaditya	Ex-Director General (NIC), Government of India
Ajay Kumar	Dean (Research), University of Delhi, India
R.K. Arora	Ex-Professor, IIT Delhi, Delhi, India
Rattan Datta	Ex-Director, Indian Meteorological Department, Delhi, India
Pankaj Jalote	Director, IIIT Delhi, India
Jaijit Bhattacharya	Director, Government Affairs, HP India

Executive Committee

Conference Chair

S.K. Gupta	IIT, Delhi, India

Program Co-chairs

Srinath Srinivasa	IIIT, Banglore, India
Vasudha Bhatnagar	University of Delhi, India

Organizing Co-chairs

Naveen Kumar	University of Delhi, India
Neelima Gupta	University of Delhi, India

Publicity Chair

Vikram Goyal	IIIT, Delhi, India

Publications Chair

Subhash Bhalla	University of Aizu, Japan

Sponsorship Chair

DVLN Somayajulu	NIT Warangal, India

Local Organizing Committee

Ajay Gupta	University of Delhi, India
A.K. Saini	GGSIP University, Delhi, India
Rajiv Ranjan Singh	SLCE, University of Delhi, India
R.K. Agrawal	JNU, New Delhi, India
Sanjay Goel	JIIT, University, Noida, India
Vasudha Bhatnagar	University of Delhi, India
V.B. Aggarwal	Jagannath Institute of Management Sciences, New Delhi, India
Vikram Goyal	IIIT Delhi, India

Program Committee

V.S. Agneeswaran	I-Labs, Impetus, Bangalore, India
R.K. Agrawal	Jawaharlal Nehru University, New Delhi, India
Srikanta Bedathur	Indraprastha Institute of Information Technology Delhi, India
Subhash Bhalla	University of Aizu, Japan
Vasudha Bhatnagar	University of Delhi, India
Vikram Goyal	Indraprastha Institute of Information Technology Delhi, India
S.K. Gupta	Indian Institute of Technology, Delhi, India
S.C. Gupta	National Informatics Center, Delhi, India
Ajay Gupta	University of Delhi, India
Sharanjit Kaur	University of Delhi, India
Akhil Kumar	Pennsylvania State University, USA
Naveen Kumar	University of Delhi, India
C. Lakshminarayan	Hewlett-Packard Labs, USA
Yasuhiko Morimoto	Hiroshima University, Japan
Saikat Mukherjee	Siemens India
Peter Neubauer	Neo Technologies, Sweden
Sanket Patil	Siemens India
Jyoti Pawar	University of Goa, India
Lukas Pichl	International Christian University, Japan
Maya Ramanath	Indian Institute of Technology Delhi, India
B. Ravindran	Indian Institute of Technology Madras, India
Pollepalli Krishna Reddy	International Institute of Information Technology Hyderabad, India
S.H. Sengamedu	Komli Labs, India
Myra Spiliopoulou	University of Magdeburg, Germany
Srinath Srinivasa	International Institute of Information Technology Bangalore, India

Ashish Sureka	Indraprastha Institute of Information Technology, Delhi, India
Shamik Sural	Indian Institute of Technology Kharagpur, India
Srikanta Tirthapura	University of Iowa, USA

Sponsors

University of Delhi, India
University of Aizu, Japan
IBM India Research Institute
Hewlett-Packard India
Indian Institute of Technology Delhi, India
ACM Delhi-NCR Chapter, India
Microsoft Corp.

Table of Contents

Perspectives on Big Data Analytics

Data Analytics Applications

Knowledge Discovery through Information Extraction

Data Models in Analytics