Naveen Kumar
Vasudha Bhatnagar (Eds.)

LNCS 9498

Big Data Analytics

4th International Conference, BDA 2015
Hyderabad, India, December 15–18, 2015
Proceedings

Springer

Lecture Notes in Computer Science 9498

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

Editorial Board

More information about this series at http://www.springer.com/series/7409

Naveen Kumar · Vasudha Bhatnagar (Eds.)

Big Data Analytics

4th International Conference, BDA 2015 Hyderabad, India, December 15–18, 2015 Proceedings

Springer

Editors
Naveen Kumar
Department of Computer Science
University of Delhi
Delhi
India

Vasudha Bhatnagar
Department of Computer Science
University of Delhi
Delhi
India

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-27056-2 ISBN 978-3-319-27057-9 (eBook)
DOI 10.1007/978-3-319-27057-9

Library of Congress Control Number: 2015955359

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Big data is no longer just a buzz word. It is a serious analytics area, requiring a rethinking of platforms, computing paradigms, and architectures. Big data analytics is now quintessential for a wide variety of traditional as well as modern applications, which has orchestrated intense research efforts in recent years.

This volume contains the papers presented at the 4th International Conference on Big Data Analytics (BDA 2015) held during December 15–18, 2015, in Hyderabad, India. The aim of the conference was to highlight recent advancements in the area while stimulating fresh research. There were 61 submissions in the research track. Each submission was reviewed by on average 2.5 Program Committee members. The committee decided to accept nine papers, which are included in this volume. The volume also includes nine invited papers. The volume is divided into four sections: *Security and Privacy, Commerce, Models and Algorithms, and Medicine.*

The section "Big Data: Security and Privacy" includes two papers. Barker describes the current state of the art and makes a call to open a dialog between the analytics and privacy communities. Agarwal et al. discuss online radicalization and civil unrest as two important applications of open source social media intelligence. The authors also discuss open research problems, important papers, publication venues, research results, and future directions.

The section "Big Data in Commerce" includes four papers. Singh discusses information exploration in e-commerce databases and identifies limitations in the query and result panel that deter exploratory search. Add-on extensions are proposed to address these limitations. Reddy proposes a framework to harvest the page views of the Web by forming clusters of similar websites. Anand et al. propose the utility-based fuzzy miner (UBFM) algorithm to efficiently mine a process model driven by a utility threshold. The utility can be measured in terms of profit, value, quantity, or other expressions of user's preference. The work incorporates statistical and semantic aspects while driving a process model. Bansal et al. integrate the notion of discount in state-of-the-art utility-mining algorithms and propose an algorithm for efficiently mining high-utility itemsets.

The section "Models and Algorithms" includes papers on a wide range of algorithms. Bhatnagar emphasizes that a completely new rethink of the solution from the perspective of the powers of the Map-Reduce paradigm can provide substantial gains. He presents an example of the design of a model-learning solution from high-volume monitoring data from a manufacturing environment. Toyoda addresses the requirement to increase the resilience and safety of transportation systems in view of the 2020 Olympic games in Tokyo, amidst predictions of big earthquakes. Based on large-scale smart card data from the Tokyo Metro subway system and vehicle recorder data, the author presents methods to estimate passenger flows, changes in the flows after accidents, and visualizes traffic with social media streams. Kiran and Kitsuregawa review issues involved in finding periodic patterns in the context of time series and

transactional databases. They describe the basic model of finding periodic patterns, and its limitations, and suggest approaches to address them. Bhardwaj and Dash propose a novel density-based clustering algorithm for distributed environments using the MapReduce programming paradigm. Goel and Chaudhary present an incremental approach for mining formal concepts from a context that is assumed to be present in the form of object-attribute pairs. The approach utilizes the Apache Spark framework to discover and eliminate the redundant concepts in every iteration. Sachdev et al. present an implementation of the alpha-miner algorithm in MySQL and Cassandra using SQL and CQL with the aim of conducting a performance benchmarking and comparison of the alpha-miner algorithm on row-oriented and NoSQL column-oriented databases. Nagpal and Gaur propose a novel algorithm to filter out both irrelevant and redundant features from the data set using a greedy technique with gain ratio.

The last section, "Big Data in Medicine," addresses the applications of big data in the area of medicine. Talukder reviews the science and technology of big data translational genomics and precision medicine with respect to formal database systems. The paper presents two big-data platforms iOMICS (deployed at Google cloud) for translational genomics and DiscoverX (deployed at Amazon Web Services) for precision medicine. Adhil et al. present a clinical expert system (CES) that uses the clinical and genomic marker of the patient combined with a knowledge-base created from distributed, dissimilar, diverse big data. The system predicts the prognosis using a cancer registry compiled between 1997 and 2012 in the USA. Agarwal et al. present an integrative analytics approach for cancer genomics that takes the multiscale biological interactions as key considerations for model development. Sharma et al. propose a class-aware exemplar discovery algorithm, which assigns preference value to data points based on their ability to differentiate samples of one class from others. Goel et al. suggest a system for predicting the risk of cardiovascular disease based on electrocardiogram tests. The authors also discuss a recommendation system for suggesting nearby relevant hospitals based on the prediction.

In closing, we wish to thank the supporting institutes: NIT-Warangal, the University of Delhi, University of Aizu, and the Indian Institute of Technology Delhi. Thanks are also due to our sponsors: KDNuggets, NIT Warangal, Tata Consultancy Services, and Springer. We take this opportunity to thank Niladri Chatterjee (Student Symposium Chair), and Ramesh Agrawal (Tutorial Chair) for helping us in organizing the respective tracks. We gratefully acknowledge the Program Committee members and external reviewers for their time and diligent reviews. Thanks are due to Steering Committee members for their guidance and advisory role. The Organizing Committee and student volunteers of BDA 2015 deserve special mention for their support.

Last but not the least, we appreciate and acknowledge the EasyChair conference management system, which transformed the tedious task of managing submissions into an enjoyable one.

December 2015

Naveen Kumar
Vasudha Bhatnagar

Organization

Steering Committee Chair

S.K. Gupta	IIT Delhi, India

Proceedings Chair

Subhash Bhalla	University of Aizu, Japan

General Chair

D.V.L.N. Somayajulu	NIT Warangal, India

Program Chairs

Naveen Kumar	University of Delhi, India
Vasudha Bhatnagar	University of Delhi, India

Tutorial Chair

R.K. Agrawal	JNU, New Delhi

Student Symposium Chair

Niladri Chatterjee	IIT Delhi, India

Organizing Chairs

P. Krishna Reddy	IIIT Hyderabad, India
R.B.V. Subramanyam	NIT Warangal, India

Finance Chairs

D.V.L.N. Somayajulu	NIT Warangal, India
Vikram Goyal	IIIT Delhi, India

Sponsorship Chair

T. Ramesh	NIT Warangal, India

Website Chair

R.B.V. Subramanyam	NIT Warangal, India

Publicity Chair

Vikram Goyal	IIIT Delhi, India

Steering Committee

Alexander Vazhenin	University of Aizu, Japan
D. Jankiram	IIT Madras, India
Jaijit Bhattacharyya	KPMG, India
Manish Gupta	Xerox Research Centre, India
Mukesh Mohania	IBM Research, India
N. Vijayaditya	Formerly at NIC, India
Nirmala Datla	HCL Technologies, India
R.K. Arora	Formerly at IIT Delhi, India
R.K. Datta	Formerly at IMD, India
S.K. Gupta (Chair)	IIT Delhi, India
Subhash Bhalla	University of Aizu, Japan
T. Srinivasa Rao	NIT Warangal, India

Program Committee

Muhammad Abulaish	Jamia Millia Islamia, India
Vijay Aggarwal	Jagan Institute of Management Studies, India
Ramesh Agrawal	Jawaharlal Nehru University, India
Avishek Anand	L3S Research Center, Germany
Rema Ananthanarayanan	IBM India Research Lab, New Delhi
Zeyar Aung	Masdar Institute of Science and Technology, United Arab Emirates
Amitabha Bagchi	Indian Institute of Technology, Delhi, India
Amit Banerjee	South Asian University, New Delhi, India
Srikanta Bedathur	IBM India Research Lab, New Delhi, India
Subhash Bhalla	University of Aizu, Japan
Raj Bhatnagar	University of Cincinnati, USA
Vasudha Bhatnagar	University of Delhi, India
Arnab Bhattacharya	Indian Institute of Technology, Kanpur, India
Xin Cao	Queen's University Belfast, UK
Niladri Chatterjee	Indian Institute of Technology, Delhi, India
Debashish Dash	Institute of Genomics and Integrative Biology, India
Prasad Deshpande	IBM India Research Lab, India
Dejing Dou	University of Oregon, USA
Shady Elbassuoni	American University of Beirut, Lebanon

Anita Goel	Dyal Singh College, University of Delhi, India
Vikram Goyal	Indraprastha Institute of Information Technology, Delhi, India
Rajeev Gupta	IBM India Research Lab, New Delhi
S.C. Gupta	Indian Institute of Technology, Delhi, India
Renu Jain	CSJM University, India
Kalapriya Kannan	HP Research, India
Sharanjit Kaur	AND College, University of Delhi, India
Akhil Kumar	Penn State University, USA
Naveen Kumar	University of Delhi, India
Ulf Leser	Institut für Informatik, Humboldt-Universität zu Berlin, Germany
Aastha Madan	International Institute of Information Technology, Banglore, India
Ravi Madipadaga	Carl Zeiss
Sameep Mehta	IBM India Research Lab, Delhi, India
Yasuhiko Morimoto	Hiroshima University, Japan
Saikat Mukherjee	Siemens Medical Solutions, India
Mandar Mutalikdesai	Siemens Technology and Services Pvt. Ltd., India
S.K. Muttoo	University of Delhi, India
Ankur Narang	Data Science Mobileum Inc., India
Anjaneyulu Pasala	Infosys Limited, India
Dhaval Patel	IIT Roorkee, India
Nishith Pathak	Ninja Metrics, India
Lukas Pichl	International Christian University, Japan
Krishna Reddy Polepalli	IIIT, Hyderabad, India
Vikram Pudi	IIIT, Hyderabad, India
Mangsuli Purnaprajna	Honeywell Technology Solutions, India
Santhanagopalan R.	International Institute of Information Technology, Bangalore, India
Ramakrishnan Raman	Honeywell Technology Solutions, India
Maya Ramanath	Indian Institute of Technology, Delhi, India
Sreenivasan Sengamedu	Yahoo Labs, India
Mark Sifer	University of Wollongong, Australia
Alok Singh	University of Hyderabad, India
Manish Singh	Indian Institute of Technology, Hyderabad, India
Srinath Srinivasa	International Institute of Information Technology, Bangalore, India
Shamik Sural	Indian Institute of Technology, Kharagpur, India
Ashish Sureka	Software Analytics Research Lab, India
Asoke Talukedar	InterpretOmics, Bangalore, India
Durga Toshniwal	Indian Institute of Technology, Roorkee, India
Chenthamarakshan Vijil	IBM India Research Lab, India

Additional Reviewers

de Silva, N.H.N.D.
Gupta, Anamika
Gupta, Shikha
Kafle, Sabin
Kundu, Sonia
Puri, Charu
Saxena, Rakhi

Contents

Big Data: Security and Privacy

Privacy Protection or Data Value: Can We Have Both?

Ken Barker(✉)

University of Calgary, Calgary, AB, Canada
kbarker@ucalgary.ca

Abstract. Efforts to derive maximum value from data have led to an expectation that this is "just the cost of living in the modern world." Ultimately this form of data exploitation will not be sustainable either due to customer dissatisfaction or government intervention to ensure private information is treated with the same level of protection that we currently find in paper-based systems. Legal, technical, and moral boundaries need to be placed on how personal information is used and how it can be combined to create inferences that are often highly accurate but not guaranteed to be correct. Agrawal's initial call-to-arms in 2002 has generated a large volume of work but the analytics and privacy communities are not truly communicating with the goal of providing high utility from the data collected but in such a way that it does not violate the intended purpose for which it was initially collected [2]. This paper describes the current state of the art and makes a call to open a true dialog between these two communities. Ultimately, this may be the only way current analytics will be allowed to continue without severe government intervention and/or without severe actions on behalf of the people from whom the data is being collected and analyzed by either refusing to work with exploitative corporations or litigation to address the harms arising from the current practices.

1 Introduction

Users provide information about themselves to either receive a value of some kind or because they are compelled to do so for legal or moral reasons. For example, a patient needing health care due to an illness or injury must disclose to their physician personal information that is considered in almost all cultures to be highly private so they can receive the health care they need. The reason the information is being disclosed is to receive the medical care they need to maintain or regain their health. There is an expectation, and indeed a high ethical standard (Hippocratic Oath: Article 8) that compels this information be kept private and not used for any other purpose.[1] This example is often cited in the literature

[1] There are societally and individually acceptable deviations from this high standard such as informing an insurance company about the costs of the service so the physician can be paid. However, this is done with the explicit informed consent of the patient and there is an expectation that this information will be kept private and will not be used for any other purpose.

N. Kumar and V. Bhatnagar (Eds.): BDA 2015, LNCS 9498, pp. 3–20, 2015.
DOI: 10.1007/978-3-319-27057-9_1

to motivate the need for privacy and the expectations implied here are often applied to other interactions that occur. However, there are interactions that occur where there is no expectation of privacy or that a "private" communication will be kept confidential. For example, a person interviewed by a press reporter does not have the same expectation that the communication will be protected. In fact, the exact opposite is the expectation and both parties entering into such a communication understand that the reason for the conversation is to capture the discussion with the purpose of writing a very public story. Analytic advocates argue that data collection is often done with both parties understanding that the information collected will be used for many forms of analysis well beyond its apparent initial purpose. For example, search query data submitted to a search engine is not considered "private" by the search engine provider and in many cases, the collection of such data is the basis of their business model. However, police investigations now regularly include searches of suspects browser search histories for evidence that might link the suspect to the crime. It is a fairly obvious claim that this was not the intention of the suspect and as a result will likely feel privacy is being violated.

A commonly held myth is that privacy cannot be protected; that people no longer care about their privacy because any interactions in the modern world, by definition, requires that we forfeit our privacy; and that we should simply accept it and live without concern about it only because we gave up our right to privacy a long time ago. Clearly this myth is easily revealed as "false" by virtue of the large number of things each individual does in their daily lives to protect their privacy. Very few people would consider cameras in the bathroom a reasonable privacy of violation even if there might be substantial public good that might accrue from such cameras. The ubiquity of public survivance cameras does not imply the public accepts being monitored everywhere so there are societal limits that restrict the extent to which cameras can be used. Why should there these limits not also be placed on other forms of survivance?

Analytics, is effectively survivance. Click stream data describes how a user interacts on a website and some would argue it reveals their intent. Mining of such data allows an analyst to predict a particular users interests and to tailor the way information is presented to the user. This tailoring, it is argued, is a "user value" because it allows the webpage to provides "opportunities of *specific* interest" to the user. The argument continues that this is a "win-win-win" because the webpage vendor can provide a marketing opportunity to third-parties by promising to deliver advertising content to users with the maximum likelihood of purchasing their products. This, in effect, creates a "third" winner and this the third-party product provider who is able to more effectively sell their product or service.

Unfortunately there are serious issues associated with this particular three way "win". First, even if we accept that analytics for the purpose of "adding value to the customer's experience" is an acceptable tradeoff, the data collected may be used in any number of unanticipated ways. A person who is interested in understanding how a disease impacts a friend could be denied insurance coverage if there was a large number of searches undertaken about a particular disease,

even if the person does not have such a disease. A teenage person searching for information about date rape drugs so they can better understand how a friend may have been subjected to such a criminal act may be brought under suspicion as a result of the police investigation into who might have committed the crime.

Second, the analytics will combine data in ways that will produce inferences that are inaccurate. Data mining will fail to produce useful inferences if accuracy is considered sacrosanct. The *accuracy paradox* [9] demonstrate that accuracy can even improve when known errors are introduced into the system. This has led researchers to develop terms and definitions that are more appropriate for analytical work such as *precision* and *recall* when evaluating the effectiveness of their algorithms. This is not problematic when data is considered in aggregate but introduces significant issues if the analytics purports to identify instances that may introduce inaccurate descriptions of individuals. The primary goal of classification systems is to place individuals within each class so they can be considered largely homogeneous. When an instance is misclassified the potential harm is substantial and it is unlikely that anyone would agree to a privacy policy statement that explicitly allowed for such errors.

Third, the analytics will lead to unintended consequences. The famous case of Target^{TM} identifying a pregnant teenager and "helpfully" providing product information is often touted as how precise and successful data analytics is [5]. However, it does not address the harm that comes from such "helpful" marketing campaigns. Although the father in this particular situation was quite polite in his response when he says, "[I]t turns out there's been some activities in my house I haven't been complete aware of," is exceptionally muted, it is clear that not all parents would react quite so calmly. If the father's trust in such analytics was so great that he assumed Target^{TM}'s analysis was indeed correct but that daughter had chosen to seek an abortion without her parent's consent, the consequences could have been even more significant. Target argues that this is providing value for their customer and admits that the goal is to capture a high value market segment but acknowledges that not everyone receives such "helpful" information in the same way [5] so they have taken steps to disguise how specific this one-to-one marketing is for individuals.[2]

2 Big Data

Big Data is a term that has been used since the inception of modern computer systems and in particular since the development of database management systems (DBMS). Unfortunately, it is a term that can only be defined relative to something and even when care is taken in its definition, gaps readily form that lead to disagreements. Absolute size does not work as demonstrated by the date of the inaugural *Very Large Data Base* conference that occurred in 1975 to

[2] One-to-One marketing "... means being willing and able to change your behaviour toward an individual customer based on what the customer tells you and what else you know about that customer." [8].

help deal with "large" databases which would be consider miniscule by today's systems. Gudiveda *et al.* [6] define big data as "data too large and complex to capture, process and analyze using current computing infrastructure" and then point to the popular characterization using five "V"s, which may not be applicable in all big data environments. The "V"s are:

- Volume - currently defined as terabytes (2^{40}) or even as large as exabytes (2^{60}).
- Velocity - data is produced at extremely high rates so streaming processes are required to help limit the volume.
- Variety - often data captures are very heterogeneous and may include structured, unstructured or semi-structured elements.
- Veracity - provenance is often crucial in determining if the source is valid, trustworthy, and of high quality.
- Value - the data should be of value either in itself or via subsequent analytics.

Each of these characteristics has a different impact on how big data is managed but since our focus here is on privacy, we consider briefly how each impacts on analytics and the nature of the data that is privacy sensitive. Volume, in and of itself, is not necessarily a privacy sensitive characteristic. For example, large volumes of radio telescope data does not contain data that is private. The challenges here are found in stream processing through sparse data which is not relevant to this paper. The amount of data collected is largely orthogonal to privacy issues but as the volume increases relative to the available compute power, there is a risk that privacy (or for that matter security) considerations will be displaced by the need for efficiency. Thus, volume introduces a risk but is not in itself a privacy threat.

Similarly the velocity with which data is received is not inherently problematic for privacy. The challenge becomes one of ensuring that any privacy requirements for the quickly arriving data is managed in a timely way. Much like volume, the strong temptation would be to sacrifice proper meta-data management (including privacy annotations) in the interest of simply collecting the data. Later we argue that data must be appropriately annotated with privacy provenance and if the collection mechanisms are too slow, this critical foundational step could be compromised.

A key aspect of big data is variety. The analytics that can be undertaken when large volumes of heterogeneous data is collected provides both an opportunities for tremendous insight and to invade privacy. Research into how to integrate and validate heterogeneous data sources has been underway for at least three decades. The new issues introduced by privacy meta-data that will likely be in conflict because they come from different sources, in different formats, will provide a rich set of research opportunities.

One of the critical features of any big data project is ensuring the veracity of the data collected. This will be critical in privacy as one of the key privacy obligations is to ensure that data collected can be checked for correctness. In addition, any data found to be incorrect must be deleted and possibly replaced with correct values. Thus, the veracity characteristic will be challenging because

the data must be validated upon collection, maintained correctly throughout its lifecyle and deleted (really deleted) at the end of its life [7].

Data analytics is fundamentally motivated by the desire to derive value from data. In privacy research, this is often expressed as the privacy-utility tradeoff. Privacy is trivially guaranteed by not providing any access to the data collected, but this sacrifices all utility. Conversely, utility is maximized by eliminating any data access but this provides no data privacy at all. Thus, the definition of "value" in big data needs to be tempered by the responsibility to facilitate appropriate privacy for the data provider.

3 Privacy

Data ownership is likely a central issue in how people think about data privacy. If you, as an individual, believe that information about yourself is ultimately your "property" and should be controlled and disseminated by you, your perspective about privacy will likely hold that the individual owns that data and should have the right to determine how, or even if, it is used by others. If you, possibly thinking as a corporation, believe that information about those with whom you interact belongs to the corporation, then you will view the artifacts arising from interactions with individuals as your property. Some would argue that these artifacts are "shared" property so both parties have the right to claim ownership but this position is difficult to realize pragmatically when considering privacy. In other words, once a piece of private data is released, it cannot be easily retracted and protected from further disclosure.[3] This suggests that there are at least three distinct schools of thought when it comes to data ownership and we consider each below.

3.1 Individuals Own Data About Themselves

Medical data, religious beliefs, sexual orientation and political viewpoints are often considered private by individuals and most jurisdictions have laws that prevent, for example, an employer asking questions about these beliefs or viewpoints. However, if a person seeks treatment for a particular disease, they must reveal detailed medical histories to ensure their physician will treat them appropriately. This highly personal data is often believed to be owned by the patient and they have the right to control how it is used and to who it might be disseminated. However, hospitals collect this data and often in the interest of the greater good argue that this data, appropriately anonymized, should be used for medical research. The Hippocratic Oath (Article 8) requires that this not be done without explicit permission so patients are often asked to sign documents

[3] This is often the argument made by those who believe that "nothing is private" any longer and we should just accept this as a reality. However, the argument is self-evidently specious since the argument's proponents are often quite protective of some aspects of themselves as discussed earlier.

that allow their data to be used for this purpose.[4] Thus, the hospital is acknowledging that ownership of the data resides with the patient and the research to be done is with explicit permission from the patient (i.e. data owner).

3.2 Corporations Own Data Collected

Users interact with websites for many reasons but their primary purpose is the acquisition of some knowledge or service. The way users interact is often a high value commodity that the corporation can use for analytics for its own operation or to resell this behavioural information to other interested parties. For example, a search engine may collect the queries posed by users to help identify what is trending or of interest to its user community. The value may be substantial for organizations that want to communicate with the public about things that may be of interest. A political party that makes a policy statement would likely want to see if their announcement is generating interest and if it is being received favourably. This kind of data is often considered the property of the website and as such can be used to derive benefits for themselves. The corporation will attempt to protect the individual users by anonymizing the data through aggregation or using some other technique but it strongly believes this data is owned by them and it is likely that most people would agree with their claim.

An online shopping site will also collect similar search queries and use it in the same way as a search engine. However, the shopping site may have an opportunity to collect additional data in the event that a user makes a purchase. The user must enter data relevant to the purchase itself including name, address, contact information, credit card information and other details directly applicable the purchase. The shopping site will use this information in the first instance for the purpose of selling and delivering the purchased product but is now able to link very specific identifying information to the individual doing the search. Often these website will also install, often with permission, an artifact on the users computer so future trips to the site will also be readily tracked. The shopping site would claim that this data is owned by them. Few people would likely agree that information such as name, address, and credit card data is now "owned" by the website but this is the current state of the art and there is not typically a need for the company to delete the information after its initial intended use.

3.3 Shared Ownership of Data

One sharing mechanism suggested is that the data remains the property of the data owner but because they have willing shared it with the corporation, they

[4] It is unclear if this permission is collected in a completely non-coercive way. A patient might feel that by failing to sign such a document they may not receive the best possible care. Clearly this would not be the case but the perception may be a critical factor in providing such permission and this would likely be considered coercive in some way by a reasonable person. However, and much more likely, the patient is simply overwhelmed with the amount of documents required as they seek treatment so they may simply sign the documents presented to them with due consideration as they seek care.