SpringerBriefs in Computer Science

Series Editors

Stan Zdonik, Brown University, Providence, RI, USA

Shashi Shekhar, University of Minnesota, Minneapolis, MN, USA

Xindong Wu, University of Vermont, Burlington, VT, USA

Lakhmi C. Jain, University of South Australia, Adelaide, SA, Australia

David Padua, University of Illinois Urbana-Champaign, Urbana, IL, USA

Xuemin Sherman Shen, University of Waterloo, Waterloo, ON, Canada

Borko Furht, Florida Atlantic University, Boca Raton, FL, USA

V. S. Subrahmanian, University of Maryland, College Park, MD, USA

Martial Hebert, Carnegie Mellon University, Pittsburgh, PA, USA

Katsushi Ikeuchi, University of Tokyo, Tokyo, Japan

Bruno Siciliano, Università di Napoli Federico II, Napoli, Italy

Sushil Jajodia, George Mason University, Fairfax, VA, USA

Newton Lee, Institute for Education, Research and Scholarships, Los Angeles, CA, USA

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic.

Typical topics might include:

- A timely report of state-of-the art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that students must understand in order to make independent contributions

Briefs allow authors to present their ideas and readers to absorb them with minimal time investment. Briefs will be published as part of Springer's eBook collection, with millions of users worldwide. In addition, Briefs will be available for individual print and electronic purchase. Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, easy-to-use manuscript preparation and formatting guidelines, and expedited production schedules. We aim for publication 8–12 weeks after acceptance. Both solicited and unsolicited manuscripts are considered for publication in this series.

**Indexing: This series is indexed in Scopus, Ei-Compendex, and zbMATH **

Chen Ye · Hongzhi Wang · Guojun Dai

Knowledge Discovery from Multi-Sourced Data

Springer

Chen Ye
Computer and Software Department
Hangzhou Dianzi University
Hangzhou, China

Guojun Dai
Computer and Software Department
Hangzhou Dianzi University
Hangzhou, China

Hongzhi Wang
Computer Science and Technology
Harbin Institute of Technology
Harbin, China

ISSN 2191-5768 ISSN 2191-5776 (electronic)
SpringerBriefs in Computer Science
ISBN 978-981-19-1878-0 ISBN 978-981-19-1879-7 (eBook)
https://doi.org/10.1007/978-981-19-1879-7

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

This book is dedicated to all contributors in this field.

Preface

With the rapid development of information technology, all areas have ushered in the era of big data. One big challenge in analyzing the overwhelming generated data is the veracity of the data. Data, even describing the same object or event, can come from a variety of sources such as crowd workers and social media users. However, noisy pieces of data or information are unavoidable. Facing the daunting scale of data, it is unrealistic to expect humans to "label" or tell which data source is more reliable. Hence, it is crucial to identify trustworthy information from multiple noisy information sources, referring to the task of knowledge discovery.

At present, the knowledge discovery research for multi-sourced data mainly faces two challenges. On the structural level, it is essential to consider the different characteristics of data composition and application scenarios and define the knowledge discovery problem on different occasions. On the algorithm level, the knowledge discovery task needs to consider different levels of information conflicts and design efficient algorithms to mine more valuable information using multiple clues. Existing knowledge discovery methods have defects on both the structural level and the algorithm level, making the knowledge discovery problem far from totally solved.

In this book, the theories, techniques, and methods in data cleaning, data mining, and natural language processing are synthetically used to study the knowledge discovery problem on multi-source data. This book mainly focuses on three data models: the first is multi-source isomorphic data, which has a clear and significant entity-attribute-source structure; the second is multi-source heterogeneous data, where the entities and attributes from different sources may have various representations; and the third is text data, which does not intuitively reflect the entity-attribute-source structure and contains a lot of irrelevant words. On the basis of three data models, this book studies the knowledge discovery problems including truth discovery, pattern discovery, and fact discovery on multi-source data from four important properties: relevance, inconsistency, sparseness, and heterogeneity.

We hope the proposed ideas in this book can inspire researchers in both academics and industry, and further prompt them to join the field of knowledge discovery.

Hangzhou, P.R. China
December 2021

Chen Ye
Hongzhi Wang
Guojun Dai

Acknowledgements This book was partially supported by the Fundamental Research Funds for the Provincial Universities of Zhejiang (No. GK219909299001-011), the Natural Science Foundation of Zhejiang Province (No. KYZ054122042CZ), and the National Key Research and Development Program of China (No. 2017YFE0118200).

Contents

1 Introduction 1
1.1 Why Knowledge Discovery from Multi-source Noisy Data? 1
1.2 Summary of Existing Work 3
1.2.1 Knowledge Discovery on Multi-source Isomorphic Data 3
1.2.2 Knowledge Discovery on Multi-source Heterogeneous Data 6
1.2.3 Knowledge Discovery on Text Data 6
1.2.4 The Drawbacks of Existing Work 7
1.3 Overview of the Book 8
References 9

2 Functional-Dependency-Based Truth Discovery for Isomorphic Data 13
2.1 Motivations 13
2.2 Problem Definition 15
2.3 AutoRepair Algorithm 17
2.3.1 Framework Overview 17
2.3.2 Handling Independent Constraints 18
2.3.3 Handling Inter-Related Constraints 22
2.3.4 Handling Inter-Source Conflicts 25
2.3.5 Updating Source Weights 26
2.3.6 Summary of the Algorithm 27
2.3.7 Performance Evaluation 27
2.4 Summary 29
References 30

3 Denial-Constraint-Based Truth Discovery for Isomorphic Data 33
3.1 Motivations 33
3.2 Problem Definition 35
3.3 CTD Algorithm 37
3.3.1 The Framework Overview 38
3.3.2 Partition and Reduction 38

3.3.3 Summary of the Algorithm 42
3.3.4 Discussions and Extensions 44
3.3.5 Performance Evaluation 45
3.4 Summary 48
References 50

4 Pattern Discovery for Heterogeneous Data 53
4.1 Motivations 53
4.2 Problem Definition 55
4.3 PatternFinder Algorithm 57
4.3.1 Solution Overview 57
4.3.2 Optimization Framework 58
4.3.3 Summary of the Algorithm 59
4.3.4 Optimized Grouping Strategy 62
4.3.5 Pattern and Truth Generation 63
4.3.6 Performance Evaluation 64
4.4 Summary 66
References 67

5 Fact Discovery for Text Data 69
5.1 Motivations 69
5.2 Pattern-Based Fact Discovery 71
5.2.1 Preprocessing 72
5.2.2 Problem Definition 73
5.3 The CNN-LSTM Architecture 73
5.3.1 EAV-Tuple Encoder 74
5.3.2 Pattern Embedding 76
5.3.3 Training and Inference 78
5.3.4 Performance Evaluation 79
5.4 Summary 82
References 83

About the Authors

Chen Ye is currently an Associate Researcher at the School of Computer Science and Technology, Hangzhou Dianzi University, China. She received the Ph.D. degree in Computer Software and Theory from Harbin Institute of Technology, China. Her current research interests include data repairing, truth discovery, and crowdsourcing. She has won the ACM SIGMOD China Doctoral Dissertation Award in 2020.

Hongzhi Wang is a Professor and Doctoral Supervisor at the School of Computer Science and Technology, Harbin Institute of Technology, China. His research interests include big data management and analysis, data quality, graph data management, and web data management. He has published more than 150 papers, and he is the Primary Investigator of more than 10 projects including three NSFC projects, and co-PI of 973, 863, and NSFC key projects. He was awarded as Microsoft fellowship, China Excellent Database Engineer, and IBM Ph.D. fellowship.

Guojun Dai is now working in the School of Computer Science and Technology of Hangzhou Dianzi University, as the Head of the National Brain-Computer Collaborative Intelligent Technology International Joint Research Center, the director of the Institute of Computer Application Technology. His research interests include Internet of Things, industrial big data, network collaborative manufacturing, edge computing, brain-computer interface, cognitive computing, artificial intelligence. He has published over 50 research papers in top-quality international conferences and journals, particularly, INFOCOM, IEEE Transactions on Industrial Informatics, and IEEE Transactions on Mobile Computing.