Data Analytics

Series editors

Longbing Cao, Advanced Analytics Institute, University of Technology, Sydney, Broadway, NSW, Australia
Philip S. Yu, University of Illinois at Chicago, Chicago, IL, USA

Aims and Goals:

Building and promoting the field of data science and analytics in terms of publishing work on theoretical foundations, algorithms and models, evaluation and experiments, applications and systems, case studies, and applied analytics in specific domains or on specific issues.

Specific Topics:

This series encourages proposals on cutting-edge science, technology and best practices in the following topics (but not limited to):

Data analytics, data science, knowledge discovery, machine learning, big data, statistical and mathematical methods for data and applied analytics,

New scientific findings and progress ranging from data capture, creation, storage, search, sharing, analysis, and visualization,

Integration methods, best practices and typical examples across heterogeneous, interdependent complex resources and modals for real-time decision-making, collaboration, and value creation.

More information about this series at http://www.springer.com/series/15063

Chuan Shi · Philip S. Yu

Heterogeneous Information Network Analysis and Applications

Springer

Chuan Shi
Beijing University of Posts and Telecommunications
Beijing
China

Philip S. Yu
University of Illinois at Chicago
Chicago, IL
USA

Data Analytics
ISBN 978-3-319-56211-7 ISBN 978-3-319-56212-4 (eBook)
DOI 10.1007/978-3-319-56212-4

Library of Congress Control Number: 2017936890

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The interacting and multi-typed components in the real-world environment constitute interconnected networks, which can be called information networks. These ubiquitous information networks form a critical component of modern information infrastructure. In recent years, the information network analysis has gained extremely wide attentions from researchers in many disciplines, such as computer science, social science, physics. Particularly, the information network analysis has become a mainstream direction in data mining, database and information retrieval fields in the past decades. The basic paradigm is to mine hidden patterns through mining linkage relations from networked data. The information network analysis is also related to the works in social network analysis, link mining, graph mining and network science.

Contemporary information network analyses are usually based on homogeneous information networks, where there is only one type of objects or links in the network. An example is the author collaboration network which only contains the author object and the co-author relation. These homogeneous information networks usually are the simplification of real interacting systems by simply ignoring the heterogeneity of objects and links or only considering one type of links among one type of objects. However, most real interacting systems contain multi-typed interacting components which can be modeled as heterogeneous information networks which include different types of objects and links. For example, the bibliographic database, like DBLP, can be organized as a heterogeneous information network which includes multiple types of objects (e.g., papers, authors, and venues) and links (e.g., written by/writing relations between papers and authors, published/publishing relations between papers and venues). Obviously, the author collaboration network is implicitly contained in the heterogeneous information network, which can be derived from the written by/writing relation between papers and authors.

Compared to homogeneous information network, the heterogeneous information network can effectively fuse more information and contain richer semantics in objects and links, and thus it forms a new development of data mining. Since the concept of heterogeneous information network is first proposed in 2009, it rapidly

became a hot research topic in data mining, and many innovative data mining tasks have been exploited in this kind of networks. In addition, some unique analysis techniques (e.g., meta-path-based mining) are developed to demonstrate the benefits of heterogeneous information networks. Particularly, with the arrival of the era of big data, heterogeneous information networks offer the potential to be an effective way to model and analyze complex objects and their relations in big data.

This book first provides a comprehensive survey of current developments of heterogeneous information network analysis, as well as some novel data mining tasks in this field. This book includes two parts. In the first part, it deeply and comprehensively summarizes the newest developments of this field in Chaps. 1, 2, and 9. This book introduces in-depth understanding of heterogeneous information network in Chap. 1 and investigates the research developments in most data mining tasks in Chap. 2. Furthermore, based on the newest developments and trends, we point out the future research directions in Chap. 9. In the second part, it illustrates the traits of heterogeneous information network analysis through several data mining tasks in Chaps. 3–8. This book presents relevance measure in Chap. 3, ranking and clustering in Chap. 4, recommendation in Chap. 5, fusion learning in Chap. 6, and schema-rich heterogeneous network mining in Chap. 7. Moreover, some interesting prototype systems are discussed in Chap. 8.

The readers of this book are engineers and researchers in the field of data mining, especially social network analysis. It is also suitable for engineers and researchers in artificial intelligences and informatics. More broadly, readers also include those who are interesting in social network analysis in other disciplines, such as statistics, social sciences, physical, and biology. This book can be used in those courses, such as data mining, social network analysis, complex network, advanced artificial intelligences. These courses are suitable for advanced undergraduates or graduate students specializing in computer sciences and related fields. The readers are suggested to quickly understand this field through the first part and deeply study data mining tasks with the second part.

We would like to express our sincere thanks to all those who work with us on this project. First of all, we appreciate Dr. Jiawei Zhang for his contribution in Chap. 6, which makes this book more integrated. Then, we are grateful to our co-authors in the work of heterogeneous information network. They are Xiangnan Kong, Yizhou Sun, Bin Wu, Yitong Li, Zhiqiang Zhang, Jian Liu, Ran Wang, Yuyan Zheng, Jing Zheng, Xiaohuan Cao, Jiawei Hu, Xiaofeng Meng, Chong Zhou, et al. We also wish to thank supporters during writing this book. They are Xin Wan, Xiaoji Chen, Yugang Ji, Houye Ji, Yiding Zhang, Yang Xiao, Binbin Hu, Xiaotian Han, Pudi Chen, Li Song, Govardhana K., Melissa Fearon, Jennifer Malat, et al. In addition, this work is supported by the National Key Basic Research and Department (973) Program of China (No. 2013CB329600), the National Natural Science Foundation of China (No. 61375058 and 61672313), and US National Science Foundation through grant III-1526499. We also thank the supports of these grants. Finally, we thank our families for their wholehearted support throughout this project.

Contents