Advanced Information and Knowledge Processing

Springer-Verlag London Ltd.

Also in this series

Gregoris Mentzas, Dimitris Apostolou, Andreas Abecker and Ron Young
Knowledge Asset Management
1-85233-583-1

Michalis Vazirgiannis, Maria Halkidi and Dimitrios Gunopulos
Uncertainty Handling and Quality Assessment in Data Mining
1-85233-655-2

Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho
Ontological Engineering
1-85233-551-3

Arno Scharl
Environmental Online Communication
1-85233-783-4

Shichao Zhang, Chengqi Zhang and
Xindong Wu

Knowledge Discovery in Multiple Databases

With 21 Figures

Springer

Shichao Zhang, PhD, MSc
Chengqi Zhang, PhD, MSc, BSc, DSc
FIT, University of Technology Sydney, Australia

Xindong Wu, PhD, MSc
Department of Computer Science, University of Vermont, USA

Series Editors
Xindong Wu
Lakhmi Jain

British Library Cataloguing in Publication Data
Zhang, Shichao
Knowledge discovery in multiple databases. - (Advanced information and knowledge processing)
1. Data mining 2. Distributed databases
I. Title II. Zhang, Chengqi III. Wu, Xindong
005.7'58

ISBN 978-1-4471-1050-7

Library of Congress Cataloging-in-Publication Data
Zhang, Shichao.
Knowledge discovery in multiple databases / Shichao Zhang, Chengqi Zhang, Xindong Wu.
p. cm.
Includes bibliographic references and index.
ISBN 978-1-4471-1050-7 ISBN 978-0-85729-388-6 (eBook)
DOI 10.1007/978-0-85729-388-6
1. Database management. 2. Database searching. I. Zhang, Chengqi, 1957- II. Wu, Xindong
III. Title

QA76.9.D3Z54 2004
005.74-dc22 2004048100

AI&KP ISSN 1610-3947

ISBN 978-1-4471-1050-7

springeronline.com

Originally published by Springer-Verlag London Limited in 2004

Softcover reprint of the hardcover 1st edition 2004

The use of registered names, trademarks etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Typesetting: Electronic text files prepared by authors

34/3830-543210 Printed on acid-free paper SPIN 10894388

Preface

Many organizations have an urgent need of mining their multiple databases inherently distributed in branches (distributed data). In particular, as the Web is rapidly becoming an information flood, individuals and organizations can take into account low-cost information and knowledge on the Internet when making decisions. How to efficiently identify quality knowledge from different data sources has become a significant challenge.

This challenge has attracted a great many researchers including the authors who have developed a local pattern analysis, a new strategy for discovering some kinds of potentially useful patterns that cannot be mined in traditional multi-database mining techniques. Local pattern analysis delivers high-performance pattern discovery from multiple databases. There has been considerable progress made on multi-database mining in such areas as hierarchical meta-learning, collective mining, database classification, and peculiarity discovery. While these techniques continue to be future topics of interest concerning multi-database mining, this book focuses on these interesting issues under the framework of local pattern analysis.

The book is intended for researchers and students in data mining, distributed data analysis, machine learning, and anyone else who is interested in multi-database mining. It is also appropriate for use as a text supplement for broader courses that might also involve knowledge discovery in databases and data mining.

The book consists of ten chapters. Chapter 1 states the multi-database mining problem and its importance. Chapter 2 lays a common foundation for subsequent material. This includes the preliminaries on data mining and multi-database mining, as well as necessary concepts, previous efforts, and applications. Chapter 3 introduces the framework of local pattern analysis. The later chapters are essentially self-contained and may be read selectively, and in any order. Chapters 4, 5, and 6 develop techniques for preprocessing the data in multi-databases. Chapters 7, 8, and 9 presents techniques for identifying interesting patterns from multi-databases based on local pattern analysis. And Chapter 10 presents a summary of the previous chapters and demonstrates some open problems.

Beginners should read Chapters 1 and 2 before selectively reading other chapters. Although the opening problems are very important, techniques in other chapters may be helpful for experienced readers who want to attack such problems.

Shichao Zhang March 2004

Chengqi Zhang

Xindong Wu

Acknowledgments

We are deeply indebted to Jenny Wolkowicki for the carefully proofreading, as well as many colleagues for the advice and support they gave during the writing of this book. We are especially grateful to Tony King for his editorial efforts when he worked with Springer.

For many suggested improvements and discussions on the material, we thank Professor Geoffrey Webb from Monash University, Mr. Zili Zhang from Deakin University, Dr. Huan Liu from Arizona State University, President Hong Liang and Ms. Yanchun Zhou from Guangxi Teachers University, Ms. Li Liu and Mr. Xiaowei Yan from the University of Technology, Sydney, Professor Xiaopei Luo from the Chinese Academy of Sciences;, and Professor Guoxi Fan from the Education Bureau of Quanzhou.

Contents