

Lecture Notes in Artificial Intelligence 2307

Subseries of Lecture Notes in Computer Science

Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

**Springer**

*Berlin*

*Heidelberg*

*New York*

*Barcelona*

*Hong Kong*

*London*

*Milan*

*Paris*

*Tokyo*

Chengqi Zhang Shichao Zhang

# Association Rule Mining

Models and Algorithms



Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA  
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Authors

Chengqi Zhang  
Shichao Zhang  
University of Technology, Sydney, Faculty of Information Technology  
P.O. Box 123 Broadway, Sydney, NSW 2007 Australia  
E-mail: {chengqi,zhangsc}@it.uts.edu.au

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Zhang, Chengqi:  
Association rule mining : models and algorithms / Chengqi Zhang ;  
Shichao Zhang. - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ;  
London ; Milan ; Paris ; Tokyo : Springer, 2002  
(Lecture notes in computer science ; Vol. 2307 : Lecture notes in  
artificial intelligence)  
ISBN 3-540-43533-6

CR Subject Classification (1998): I.2.6, I.2, H.2.8, H.2, H.3, F.2.2

ISSN 0302-9743

ISBN 3-540-43533-6 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York  
a member of BertelsmannSpringer Science+Business Media GmbH

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2002  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign  
Printed on acid-free paper SPIN: 10846539 06/3142 5 4 3 2 1 0

## Preface

Association rule mining is receiving increasing attention. Its appeal is due, not only to the popularity of its parent topic ‘knowledge discovery in databases and data mining’, but also to its neat representation and understandability. The development of association rule mining has been encouraged by active discussion among communities of users and researchers. All have contributed to the formation of the technique with a fertile exchange of ideas at important forums or conferences, including SIGMOD, SIGKDD, AAAI, IJCAI, and VLDB. Thus association rule mining has advanced into a mature stage, supporting diverse applications such as data analysis and predictive decisions.

There has been considerable progress made recently on mining in such areas as quantitative association rules, causal rules, exceptional rules, negative association rules, association rules in multi-databases, and association rules in small databases. These continue to be future topics of interest concerning association rule mining. Though the association rule constitutes an important pattern within databases, to date there has been no specialized monograph produced in this area. Hence this book focuses on these interesting topics.

The book is intended for researchers and students in data mining, data analysis, machine learning, knowledge discovery in databases, and anyone else who is interested in association rule 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 eight chapters, with bibliographies after each chapter. Chapters 1 and 2 lay a common foundation for subsequent material. This includes the preliminaries on data mining and identifying association rules, as well as necessary concepts, previous efforts, and applications. The later chapters are essentially self-contained and may be read selectively, and in any order. Chapters 3, 4, and 5 develop techniques for discovering hidden patterns, including negative association rules and causal rules. Chapter 6 presents techniques for mining very large databases, based on instance selection. Chapter 7 develops a new technique for mining association rules in databases which utilizes external knowledge, and Chapter 8 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 open problems are very important, techniques in other chapters may be helpful for experienced readers who want to attack these problems.

January 2002

*Chengqi Zhang and Shichao Zhang*

## Acknowledgments

We are deeply indebted to many colleagues for the advice and support they gave during the writing of this book. We are especially grateful to Alfred Hofmann for his efforts in publishing this book with Springer-Verlag. And we thank the anonymous reviewers for their detailed constructive comments on the proposal of this work.

For many suggested improvements and discussions on the material, we thank Professor Geoffrey Webb, Mr. Zili Zhang, and Ms. Li Liu from Deakin University; Professor Huan Liu from Arizona State University, Professor Xindong Wu from Vermont University, Professor Bengchin Ooi and Dr. Kianlee Tan from the National University of Singapore, Dr. Hong Liang and Mr. Xiaowei Yan from Guangxi Normal University, Professor Xiaopei Luo from the Chinese Academy of Sciences, and Professor Guoxi Fan from the Education Bureau of Quanzhou.

# Contents

<b>1. Introduction</b> .....	1
1.1 What Is Data Mining? .....	1
1.2 Why Do We Need Data Mining? .....	2
1.3 Knowledge Discovery in Databases (KDD) .....	4
1.3.1 Processing Steps of KDD .....	4
1.3.2 Feature Selection .....	6
1.3.3 Applications of Knowledge Discovery in Databases.....	7
1.4 Data Mining Task .....	7
1.5 Data Mining Techniques.....	9
1.5.1 Clustering .....	9
1.5.2 Classification .....	10
1.5.3 Conceptual Clustering and Classification .....	14
1.5.4 Dependency Modeling .....	15
1.5.5 Summarization .....	15
1.5.6 Regression .....	16
1.5.7 Case-Based Learning .....	16
1.5.8 Mining Time-Series Data .....	17
1.6 Data Mining and Marketing .....	17
1.7 Solving Real-World Problems by Data Mining .....	18
1.8 Summary .....	21
1.8.1 Trends of Data Mining .....	21
1.8.2 Outline .....	22
<b>2. Association Rule</b> .....	25
2.1 Basic Concepts .....	25
2.2 Measurement of Association Rules .....	30
2.2.1 Support-Confidence Framework .....	30
2.2.2 Three Established Measurements .....	31
2.3 Searching Frequent Itemsets .....	33
2.3.1 The Apriori Algorithm .....	33
2.3.2 Identifying Itemsets of Interest .....	36
2.4 Research into Mining Association Rules .....	39
2.4.1 Chi-squared Test Method .....	40
2.4.2 The FP-tree Based Model .....	43



2.4.3	OPUS Based Algorithm .....	44
2.5	Summary .....	46
<b>3.</b>	<b>Negative Association Rule .....</b>	<b>47</b>
3.1	Introduction .....	47
3.2	Focusing on Itemsets of Interest .....	51
3.3	Effectiveness of Focusing on Infrequent Itemsets of Interest ..	53
3.4	Itemsets of Interest .....	55
3.4.1	Positive Itemsets of Interest .....	55
3.4.2	Negative Itemsets of Interest .....	58
3.5	Searching Interesting Itemsets .....	59
3.5.1	Procedure .....	59
3.5.2	An Example .....	62
3.5.3	A Twice-Pruning Approach .....	65
3.6	Negative Association Rules of Interest .....	66
3.6.1	Measurement .....	66
3.6.2	Examples .....	71
3.7	Algorithms Design .....	73
3.8	Identifying Reliable Exceptions .....	75
3.8.1	Confidence Based Interestingness .....	75
3.8.2	Support Based Interestingness .....	77
3.8.3	Searching Reliable Exceptions .....	78
3.9	Comparisons .....	80
3.9.1	Comparison with Support-Confidence Framework .....	80
3.9.2	Comparison with Interest Models .....	80
3.9.3	Comparison with Exception Mining Model .....	81
3.9.4	Comparison with Strong Negative Association Model ..	82
3.10	Summary .....	83
<b>4.</b>	<b>Causality in Databases .....</b>	<b>85</b>
4.1	Introduction .....	85
4.2	Basic Definitions .....	87
4.3	Data Partitioning .....	90
4.3.1	Partitioning Domains of Attributes .....	90
4.3.2	Quantitative Items .....	92
4.3.3	Decomposition and Composition of Quantitative Items	93
4.3.4	Item Variables .....	95
4.3.5	Decomposition and Composition for Item Variables ...	96
4.3.6	Procedure of Partitioning .....	98
4.4	Dependency among Variables .....	99
4.4.1	Conditional Probabilities .....	100
4.4.2	Causal Rules of Interest .....	101
4.4.3	Algorithm Design .....	103
4.5	Causality in Probabilistic Databases .....	105
4.5.1	Problem Statement .....	105

4.5.2	Required Concepts .....	108
4.5.3	Preprocess of Data .....	108
4.5.4	Probabilistic Dependency .....	110
4.5.5	Improvements .....	115
4.6	Summary .....	119
<b>5.</b>	<b>Causal Rule Analysis .....</b>	<b>121</b>
5.1	Introduction .....	121
5.2	Problem Statement .....	122
5.2.1	Related Concepts .....	124
5.3	Optimizing Causal Rules .....	126
5.3.1	Unnecessary Information .....	126
5.3.2	Merging Unnecessary Information .....	127
5.3.3	Merging Items with Identical Properties .....	130
5.4	Polynomial Function for Causality .....	131
5.4.1	Causal Relationship .....	132
5.4.2	Binary Linear Causality .....	132
5.4.3	N-ary Linear Propagating Model .....	137
5.4.4	Examples .....	139
5.5	Functions for General Causality .....	143
5.6	Approximating Causality by Fitting .....	149
5.6.1	Preprocessing of Data .....	149
5.6.2	Constructing the Polynomial Function .....	150
5.6.3	Algorithm Design .....	155
5.6.4	Examples .....	156
5.7	Summary .....	159
<b>6.</b>	<b>Association Rules in Very Large Databases .....</b>	<b>161</b>
6.1	Introduction .....	161
6.2	Instance Selection .....	164
6.2.1	Evaluating the Size of Instance Sets .....	164
6.2.2	Generating Instance Set .....	167
6.3	Estimation of Association Rules .....	169
6.3.1	Identifying Approximate Frequent Itemsets .....	169
6.3.2	Measuring Association Rules of Interest .....	171
6.3.3	Algorithm Designing .....	172
6.4	Searching True Association Rules Based on Approximations ..	173
6.5	Incremental Mining .....	179
6.5.1	Promising Itemsets .....	180
6.5.2	Searching Procedure .....	182
6.5.3	Competitive Set Method .....	187
6.5.4	Assigning Weights .....	188
6.5.5	Algorithm of Incremental Mining .....	190
6.6	Improvement of Incremental Mining .....	193
6.6.1	Conditions of Termination .....	193

6.6.2	Anytime Search Algorithm .....	194
6.7	Summary .....	197
<b>7.</b>	<b>Association Rules in Small Databases .....</b>	<b>199</b>
7.1	Introduction .....	200
7.2	Problem Statement .....	201
7.2.1	Problems Faced by Utilizing External Data .....	201
7.2.2	Our Approach .....	203
7.3	External Data Collecting .....	204
7.3.1	Available Tools .....	204
7.3.2	Indexing by a Conditional Associated Semantic .....	206
7.3.3	Procedures for Similarity .....	208
7.4	A Data Preprocessing Framework .....	209
7.4.1	Pre-analysis: Selecting Relevant and Uncontradictable Collected Data-Sources .....	209
7.4.2	Post-analysis: Summarizing Historical Data .....	212
7.4.3	Algorithm Designing .....	214
7.5	Synthesizing Selected Rules .....	217
7.5.1	Assigning Weights .....	218
7.5.2	Algorithm Design .....	221
7.6	Refining Rules Mined in Small Databases .....	222
7.7	Summary .....	223
<b>8.</b>	<b>Conclusion and Future Work .....</b>	<b>225</b>
8.1	Conclusion .....	225
8.2	Future Work .....	226
	<b>References .....</b>	<b>229</b>
	<b>Subject Index .....</b>	<b>237</b>