

Big Data Management

Editor-in-Chief

Xiaofeng Meng, School of Information, Renmin University of China, Beijing, China

Editorial Board Members

Daniel Dajun Zeng, University of Arizona, Tucson, AZ, USA

Hai Jin, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei, China

Haixun Wang, Facebook Research, USA

Huan Liu, Arizona State University, Tempe, AZ, USA

X. Sean Wang, Fudan University, Shanghai, Shanghai, China

Weiyi Meng, Binghamton University, Binghamton, NY, USA

Advisory Editors

Jiawei Han, Department of Computer Science, University Illinois at Urbana-Champaign, Urbana, IL, USA

Masaru Kitsuregawa, National Institute of Informatics, University of Tokyo, Chiyoda, Tokyo, Japan

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

Tieniu Tan, Chinese Academy of Sciences, Beijing, Beijing, China

Wen Gao, Room 2615, Science Buildings, Peking University, Beijing, Beijing, China

The big data paradigm presents a number of challenges for university curricula on big data or data science related topics. On the one hand, new research, tools and technologies are currently being developed to harness the increasingly large quantities of data being generated within our society. On the other, big data curricula at universities are still based on the computer science knowledge systems established in the 1960s and 70s. The gap between the theories and applications is becoming larger, as a result of which current education programs cannot meet the industry's demands for big data talents.

This series aims to refresh and complement the theory and knowledge framework for data management and analytics, reflect the latest research and applications in big data, and highlight key computational tools and techniques currently in development. Its goal is to publish a broad range of textbooks, research monographs, and edited volumes that will:

- Present a systematic and comprehensive knowledge structure for big data and data science research and education
- Supply lectures on big data and data science education with timely and practical reference materials to be used in courses
- Provide introductory and advanced instructional and reference material for students and professionals in computational science and big data
- Familiarize researchers with the latest discoveries and resources they need to advance the field
- Offer assistance to interdisciplinary researchers and practitioners seeking to learn more about big data

The scope of the series includes, but is not limited to, titles in the areas of database management, data mining, data analytics, search engines, data integration, NLP, knowledge graphs, information retrieval, social networks, etc. Other relevant topics will also be considered.

More information about this series at <https://link.springer.com/bookseries/15869>

Lizhen Wang • Yuan Fang • Lihua Zhou

Preference-based Spatial Co-location Pattern Mining

Lizhen Wang
School of Information Science and
Engineering
Yunnan University
Kunming, China

Yuan Fang
School of Information Science and
Engineering
Yunnan University
Kunming, China

Lihua Zhou
School of Information Science and
Engineering
Yunnan University
Kunming, China

ISSN 2522-0179
Big Data Management
ISBN 978-981-16-7565-2
<https://doi.org/10.1007/978-981-16-7566-9>

ISSN 2522-0187 (electronic)
ISBN 978-981-16-7566-9 (eBook)

Jointly published with Science Press
The print edition is not for sale in China (Mainland). Customers from China (Mainland) please order the print book from Science Press.

© China Science Publishing & Media Ltd (Science Press) 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 publishers, 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 publishers 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 publishers remain 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

Foreword

With the continuous improvement in Global Positioning System (GPS) accuracy, our location-based services have already gone deeply into all aspects of people's life, and the data containing spatial location information is increasing. In the face of the continuous growth of massive spatial data, people are finding it difficult to fully understand the data without knowledge assistance, and so spatial data mining technology has emerged. Spatial co-location pattern mining is an important branch of the field of spatial data mining. By mining the spatial co-location patterns, we can find interesting relationships between spatial features and play a positive guiding role in various location-based application domains.

Massive spatial data brings new challenges to spatial co-location pattern mining. How can we find the compressed or condensed representation of spatial patterns from a large number of mining results? How can we design appropriate preference constraints for users in spatial co-location pattern mining? How do we derive some pruning strategies from the preference constraints to improve the efficiency of the preference-based mining algorithms? This monograph answers these questions quite well.

Professor Wang's research team has systematically and continuously researched spatial co-location pattern mining since 2008. In particular, fruitful and pioneering research on preference-based spatial co-location pattern mining has been conducted. Many of their results have been published in internationally renowned journals and conferences and have been accepted by the scholars in the world, of course, including myself.

This monograph studies a series of preference-based pattern mining techniques, including: maximal prevalent co-location pattern mining, maximal sub-prevalent co-location pattern mining, SPI-closed co-location pattern mining, top- k prevalent co-location pattern mining, dominant co-location pattern mining, non-redundant co-location pattern mining, high-utility co-location pattern mining, interactive mining, the similarity measurement between spatial co-location patterns, etc. The research results are systematic and promising.

Throughout the monograph, the author gives a clear explanation of the motivation for the research, the definition of the problem, the key research, the statement and analysis of the method, and the application evaluation, etc. It is believed that this monograph may serve researchers and application developers in spatial data mining technology and related fields and may help them explore this exciting field and develop new methods and applications. It may also provide the current status of this promising latest research theme, both for graduate students and other interested readers.

I find the monograph is enjoyable to read and fully recommend this monograph to you.

Renmin University of China, Beijing,
China
July, 2021

Xiaofeng Meng

Preface

The development of information technology has enabled many different technologies to collect large amounts of spatial data every day. It is of very great significance to discover implicit, non-trivial, and potentially valuable information from this spatial data. Spatial co-location patterns expose the distribution rules of spatial features, and their discovery can be of great value to application users. This book is intended to provide commercial software developers with proven and effective algorithms for detecting and filtering these implicit patterns, and easily-implemented pseudocode is provided for all our algorithms. The book is also intended to provide a base for further research in such a promising field. We take the demand from user applications and provide a mathematical and systematic study of preference-based spatial co-location pattern mining.

Spatial co-location pattern mining has broad prospects for spatial data owners. Potential economic value has been found in mining the data from Earth science, public safety, biological information processing, location-based personalized recommendation, geo-information system (GIS), and military strategy planning. We know it has aroused strong interest amongst researchers, both at research institutions (Google, Microsoft, IBM, etc.) and at universities (Stanford University, University of Washington, University of Minnesota, Hong Kong University of Science and Technology, etc.). High-level research results have emerged in authoritative international data engineering journals such as “IEEE Transactions on Knowledge and Data Engineering (TKDE)” and at top academic conferences such as “ACM SIGSPATIAL, ACM SIGKDD, and IEEE ICDE.”

The authors’ research team began conducting research on spatial pattern mining in 2008. Since then it has been involved with 5 National Natural Science Fund projects on spatial co-location pattern mining, has trained 7 doctoral students and over 80 master students, and has published more than 90 relevant academic papers. We have been given prizes from the Yunnan Science and Technology Award. The research team led by the first author is called the innovation team of the “Yunnan Province Spatial Big Data Mining and Decision Support Research Group,” the first provincial innovation team of our college.

Similar to frequent item set mining in transaction databases, spatial co-location pattern mining often generates a huge number of prevalent co-location patterns, but only a few of them satisfy user interests. User preferences are often subjective, and a pattern preferred by one user may not be favored by another, and so cannot be measured by objective-oriented *prevalence* measures. Therefore, the following challenges and/or issues will be answered in this book:

- What is a natural criterion for ranking patterns and presenting the “best” patterns?
- How to find a condensed representation of spatial patterns from a huge number of mined results.
- How to enable a user to have proper constraints/preferences in spatial co-location pattern mining.
- How to derive pruning properties from the constraints which improve the efficiency of the corresponding preference-based mining.

It is essential to study the theories and algorithms of preference-based co-location pattern mining in order to solve these challenges and issues. Preference-based co-location pattern mining refers to mining *constrained* or *condensed* co-location patterns instead of mining *all* prevalent co-location patterns. Specifically, this book includes problems such as *maximal* co-location pattern mining, *closed* co-location pattern mining, *top-k* co-location pattern mining, *non-redundant* co-location pattern mining, *dominant* co-location pattern mining, *high utility* co-location pattern mining, and *user-preferred* co-location pattern mining.

For the above problem areas, this book details the relevant research, the basic concepts, the resulting algorithms and their analysis, with experimental evaluation of each algorithm. These techniques come from the latest results of our research in recent years. For the convenience of readers, the chapters of this book are as integrated as possible, so the reading order of each chapter is quite flexible. You can find the corresponding chapter reading according to your interest. Of course, if you read the book from beginning to end, you will find a small amount of repetitive information intended to guarantee the relative independence of each chapter, but absolutely not information redundancy, as we describe the same content in different forms if possible. Indeed, readers are encouraged to read and study this book in order.

This book can be used both as a textbook for learners and as a good reference for professionals.

Although many people have contributed to this book, we first express our gratitude to our families. Without their encouragement and support, it would have been impossible to finish this book, and so this book is dedicated to them.

Secondly, we should sincerely thank Roger England, a colleague who helped correct the English of the book and gave a lot of valuable comments. We would also particularly like to thank Professor Xiaofeng Meng of Renmin University of China for his guidance and help. After reading the first draft of the book, he not only gave specific comments, but also gladly provided a foreword for the book; we would also like to thank Ms. Xin Li of Science Press of China and Mr. Wei Zhu of Springer. Their efforts have facilitated the smooth publication of the book.

Thanks must also go to the National Natural Science Foundation Committee and the Yunnan Provincial Department of Science and Technology for their long-term project funding (Nos.: 61966036, 61472346, 61272126, 61662086, 61762090, 2018HC019). Without their funding for specific research objectives, it would have been difficult to construct the systematic and forward-looking research results which this book conveys.

This book involves an academic discipline frontier, so numerous references have been given in the various chapters, but here we would like to particularly thank S. Shekhar, Y. Huang, and J.S. Yoo for their pioneering work in the relevant fields.

In the research work, although the authors invested a lot of effort, and in the writing of the book each chapter and every sentence has been carefully checked, although limited to our research depth and knowledge level, errors in the book are probably inevitable and we welcome the reader's criticisms and corrections.

Kunming, China
July 2021

Lizhen Wang
Yuan Fang
Lihua Zhou

Contents

1	Introduction	1
1.1	The Background and Applications	1
1.2	The Evolution and Development	5
1.3	The Challenges and Issues	7
1.4	Content and Organization of the Book	8
2	Maximal Prevalent Co-location Patterns	11
2.1	Introduction	11
2.2	Why the MCHT Method Is Proposed for Mining MPCPs	12
2.3	Formal Problem Statement and Appropriate Mining Framework	17
2.3.1	Co-Location Patterns	17
2.3.2	Related Work	19
2.3.3	Contributions and Novelties	21
2.4	The Novel Mining Solution	22
2.4.1	The Overall Mining Framework	22
2.4.2	Bit-String-Based Maximal Clique Enumeration	23
2.4.3	Constructing the Participating Instance Hash Table	28
2.4.4	Calculating Participation Indexes and Filtering MPCPs	30
2.4.5	The Analysis of Time and Space Complexities	32
2.5	Experiments	33
2.5.1	Data Sets	33
2.5.2	Experimental Objectives	34
2.5.3	Experimental Results and Analysis	34
2.6	Chapter Summary	47
3	Maximal Sub-prevalent Co-location Patterns	49
3.1	Introduction	49
3.2	Basic Concepts and Properties	51
3.3	A Prefix-Tree-Based Algorithm (PTBA)	54

3.3.1	Basic Idea	54
3.3.2	Algorithm	56
3.3.3	Analysis and Pruning	57
3.4	A Partition-Based Algorithm (PBA)	58
3.4.1	Basic Idea	58
3.4.2	Algorithm	62
3.4.3	Analysis of Computational Complexity	64
3.5	Comparison of PBA and PTBA	64
3.6	Experimental Evaluation	66
3.6.1	Synthetic Data Generation	67
3.6.2	Comparison of Computational Complexity Factors	67
3.6.3	Comparison of Expected Costs Involved in Identifying Candidates	69
3.6.4	Comparison of Candidate Pruning Ratio	69
3.6.5	Effects of the Parameter Clumpy	70
3.6.6	Scalability Tests	70
3.6.7	Evaluation with Real Data Sets	72
3.7	Related Work	75
3.8	Chapter Summary	77
4	SPI-Closed Prevalent Co-location Patterns	79
4.1	Introduction	79
4.2	Why SPI-Closed Prevalent Co-locations Improve Mining	81
4.3	The Concept of SPI-Closed and Its Properties	83
4.3.1	Classic Co-location Pattern Mining	83
4.3.2	The Concept of SPI-Closed	85
4.3.3	The Properties of SPI-Closed	86
4.4	SPI-Closed Miner	89
4.4.1	Preprocessing and Candidate Generation	89
4.4.2	Computing Co-location Instances and Their PI Values	93
4.4.3	The SPI-Closed Miner	93
4.5	Qualitative Analysis of the SPI-Closed Miner	95
4.5.1	Discovering the Correct SPI-Closed Co-location Set Ω	96
4.5.2	The Running Time of SPI-Closed Miner	96
4.6	Experimental Evaluation	96
4.6.1	Experiments on Real-life Data Sets	97
4.6.2	Experiments with Synthetic Data Sets	100
4.7	Related Work	104
4.8	Chapter Summary	105
5	Top-k Probabilistically Prevalent Co-location Patterns	107
5.1	Introduction	107
5.2	Why Mining Top- k Probabilistically Prevalent Co-location Patterns (Top- k PPCPs)	108

5.3	Definitions	110
5.3.1	Spatially Uncertain Data	110
5.3.2	Prevalent Co-locations	112
5.3.3	Prevalence Probability	113
5.3.4	<i>Min_PI</i> -Prevalence Probabilities	114
5.3.5	Top- <i>k</i> PPCPs	115
5.4	A Framework of Mining Top- <i>k</i> PPCPs	115
5.4.1	Basic Algorithm	115
5.4.2	Analysis and Pruning of Algorithm 5.1	116
5.5	Improved Computation of $P(c, \geq \min_PI)$	117
5.5.1	0-1-Optimization	117
5.5.2	The Matrix Method	118
5.5.3	Polynomial Matrices	122
5.6	Approximate Computation of $P(c, \geq \min_PI)$	125
5.7	Experimental Evaluations	128
5.7.1	Evaluation on Synthetic Data Sets	128
5.7.2	Evaluation on Real Data Sets	134
5.8	Chapter Summary	136
6	Non-redundant Prevalent Co-location Patterns	137
6.1	Introduction	137
6.2	Why We Need to Explore Non-redundant Prevalent Co-locations	139
6.3	Problem Definition	141
6.3.1	Semantic Distance	141
6.3.2	δ -Covered	143
6.3.3	The Problem Definition and Analysis	145
6.4	The RRclosed Method	148
6.5	The RRnull Method	150
6.5.1	The Method	150
6.5.2	The Algorithm	153
6.5.3	The Correctness Analysis	155
6.5.4	The Time Complexity Analysis	156
6.5.5	Comparative Analysis	157
6.6	Experimental Results	158
6.6.1	On the Three Real Data Sets	158
6.6.2	On the Synthetic Data Sets	161
6.7	Related Work	165
6.8	Chapter Summary	166
7	Dominant Spatial Co-location Patterns	167
7.1	Introduction	167
7.2	Why Dominant SCPs Are Useful to Mine	168
7.3	Related Work	171
7.4	Preliminaries and Problem Formulation	172

7.4.1	Preliminaries	173
7.4.2	Definitions	174
7.4.3	Formal Problem Formulation	179
7.4.4	Discussion of Progress	179
7.5	Proposed Algorithm for Mining Dominant SCPs	180
7.5.1	Basic Algorithm for Mining Dominant SCPs	180
7.5.2	Pruning Strategies	182
7.5.3	An Improved Algorithm	186
7.5.4	Comparison of Complexity	187
7.6	Experimental Study	188
7.6.1	Data Sets	188
7.6.2	Efficiency	189
7.6.3	Effectiveness	193
7.6.4	Real Applications	196
7.7	Chapter Summary	198
8	High Utility Co-location Patterns	201
8.1	Introduction	201
8.2	Why We Need High Utility Co-location Pattern Mining	202
8.3	Related Work	204
8.3.1	Spatial Co-location Pattern Mining	204
8.3.2	Utility Itemset Mining	205
8.4	Problem Definition	206
8.5	A Basic Mining Approach	208
8.6	Extended Pruning Approach	208
8.6.1	Related Definitions	209
8.6.2	Extended Pruning Algorithm (EPA)	210
8.7	Partial Pruning Approach	212
8.7.1	Related Definitions	212
8.7.2	Partial Pruning Algorithm (PPA)	217
8.8	Experiments	218
8.8.1	Differences Between Mining Prevalent SCPs and High Utility SCPs	218
8.8.2	Effect of the Number of Total Instances n	219
8.8.3	Effect of the Distance Threshold d	219
8.8.4	Effect of the Pattern Utility Ratio Threshold ξ	219
8.8.5	Effect of s in vss	219
8.8.6	Comparing PPA and EPA with a Different Utility Ratio Threshold ξ	220
8.9	Chapter Summary	221
9	High Utility Co-location Patterns with Instance Utility	223
9.1	Introduction	223
9.2	Why We Need Instance Utility with Spatial Data	224
9.3	Related Work	226

9.4	Related Concepts	228
9.5	A Basic Algorithm	231
9.6	Pruning Strategies	232
9.7	Experimental Analysis	236
9.7.1	Data Sets	236
9.7.2	The Quality of Mining Results	236
9.7.3	Evaluation of Pruning Strategies	237
9.8	Chapter Summary	240
10	Interactively Post-mining User-Preferred Co-location Patterns with a Probabilistic Model	241
10.1	Introduction	241
10.2	Why We Need Interactive Probabilistic Post-mining	242
10.3	Related Work	245
10.4	Problem Statement	246
10.4.1	Basic Concept	246
10.4.2	Subjective Preference Measure	247
10.4.3	Formal Problem Statement	247
10.5	Probabilistic Model	248
10.5.1	Basic Assumptions	248
10.5.2	Probabilistic Model	248
10.5.3	Discussion	251
10.6	The Complete Algorithm	252
10.6.1	The Algorithm	252
10.6.2	Two Optimization Strategies	253
10.6.3	The Time Complexity Analysis	254
10.7	Experimental Results	255
10.7.1	Experimental Setting	255
10.7.2	The Simulator	255
10.7.3	Accuracy Evaluation on Real Data Sets	257
10.7.4	Accuracy Evaluation on Synthetic Data Sets	262
10.7.5	Sample Co-location Selection	263
10.8	Chapter Summary	264
11	Vector-Degree: A General Similarity Measure for Co-location Patterns	265
11.1	Introduction	265
11.2	Why We Measure the Similarity Between SCPs	266
11.3	Preliminaries	268
11.3.1	Spatial Co-location Pattern (SCP)	268
11.3.2	A Toy Example	269
11.3.3	Problem Statement	270
11.4	The Method	270
11.4.1	Maximal Cliques Enumeration Algorithm	270
11.4.2	A Representation Model of SCPs	274

11.4.3	Vector-Degree: the Similarity Measure of SCPs	278
11.4.4	Grouping SCPs Based on Vector-Degree	279
11.5	Experimental Evaluations	279
11.5.1	Data Sets	280
11.5.2	Results	280
11.6	Chapter Summary	284
References	285