Summarizing Data with Representative Patterns

Chunyang Liu

Faculty of Engineering and Information Technology

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

March 2016

To my parents and my wife.

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Chunyang Liu

Date: 05/03/2016

Acknowledgements

I benefited and learned a lot from my advisors, my friends, my colleagues and my family during PhD study in University of Technology, Sydney. I would like to take this good opportunity to appreciate their significant helps to me.

First of all, I would like to express my appreciation and gratitude to my academic supervisors Dr. Ling Chen and Prof. Chengqi Zhang. I benefited significantly from various discussions and communications with them. They always give me advices, encouragement, and sufficient freedom to think and explore. I also want to thank them for kindly offering me invaluable suggestions and experienced instructions on my research career and life. I cannot image how this thesis can be accomplished without their high scientific criterion, endless patience, generous support, and constant guidance.

Next, I wish to thank other researchers who have gave me helpful guidance and encouragement during my PhD study and on conferences: Prof. Jian Pei, A/Prof. Ivor W. Tsang, Dr. Lu Qin, and Dr. Yi Yang. From discussions with them I learned not only much knowledge but also their attitude for doing high quality research.

I have been fortunate to work in a center gathering the most brilliant researchers and best friends in the past four years: Guodong Long, Jing Jiang, Wei Bian, Tianyi Zhou, Meng Fang, Bozhong Liu, Zhe Xu, Mingming Gong, Ruxin Wang, Zhibin Hong, Sujuan Hou, Wei Wu, Haishuai Wang, Xueping Peng, Alan Wang, Zhenxing Qin, Shirui Pan, Jia Wu, Lianhua Chi, Maoying Qiao, Tongliang Liu, Weiwei Liu, Anjin Liu, Junfu Yin, Jinjiu Li, Changxing Ding, Nannan Wang, Naiyang Guan, Li Wan, Shengzheng

Wang, Xianhua Zeng, Zhijing Xu, Bo Du, Hongshu Chen, Ting Guo, Chao Ma, Shaoli Huang, Qiang Li, Dianshuang Wu, Zhiguo Long, Weilong Hou, Yi Ji, Ming Xie and many others. I enjoyed the invaluable friendships with them, their kindly support and accompany are always my source of strength and courage in both research and daily life. I am also grateful to my dearest friends Jiawang Liu, Weipeng Zhang, Mingjian Gao, Junfeng Ye, and Jia Chen since my college, and Sen Li and Song Li since my high school. They are the ones who have given me support during both joyful and stressful times, to whom I will always be thankful.

Finally, it is my greatest honor to thank my family: my parents and my wife. They always believe in me, encourage me, give me invaluable suggestions, and fully support all my decisions. No words could possibly express my deepest gratitude for their endless love, self-sacrifice and unwavering help. To them I dedicate this dissertation.

Abstract

The advance of technology makes data acquisition and storage become unprecedentedly convenient. It contributes to the rapid growth of not only the volume but also the veracity and variety of data in recent years, which poses new challenges to the data mining area. For example, uncertain data mining emerges due to its capability to model the inherent veracity of data; spatial data mining attracts much research attention as the widespread of location-based services and wearable devices. As a fundamental topic of data mining, how to effectively and efficiently summarize data in this situation still remains to be explored.

This thesis studied the problem of summarizing data with representative patterns. The objective is to find a set of patterns, which is much more concise but still contains rich information of the original data, and may provide valuable insights for further analysis of data. In the light of this idea, we formally formulate the problem and provide effective and efficient solutions in various scenarios.

We study the problem of summarizing probabilistic frequent patterns over uncertain data. Probabilistic frequent pattern mining over uncertain data has received much research attention due to the wide applicabilities of uncertain data. It suffers from the problem of generating an exponential number of result patterns, which hinders the analysis of patterns and calls for the need to find a small number of representative patterns to approximate all other patterns. We formally formulate the problem of *probabilistic representative frequent pattern (P-RFP)* mining, which aims to find the minimal set of patterns with sufficiently high probability to represent all other patterns. The bottleneck turns out to be checking whether a pattern can probabilistically represent another, which involves the computation of a joint probability of the supports of two patterns. We propose a novel dynamic programming-based approach to address the problem and devise effective optimization strategies to improve the computation efficiency.

To enhance the practicability of P-RFP mining, we introduce a novel approximation of the joint probability with both theoretical and empirical proofs. Based on the approximation, we propose an *Approximate P-RFP Mining* (APM) algorithm, which effectively and efficiently compresses the probabilistic frequent pattern set. The error rate of APM is guaranteed to be very small when the database contains hundreds of transactions, which further affirms that APM is a practical solution for summarizing probabilistic frequent patterns.

We address the problem of directly summarizing uncertain transaction database by formulating the problem as *Minimal Probabilistic Tile Cover* Mining, which aims to find a high-quality probabilistic tile set covering an uncertain database with minimal cost. We define the concept of *Probabilistic Price* and *Probabilistic Price Order* to evaluate and compare the quality of tiles, and propose a framework to discover the minimal probabilistic tile cover. The bottleneck is to check whether a tile is better than another according to the *Probabilistic Price Order*, which involves the computation of a joint probability. We prove that it can be decomposed into independent terms and calculated efficiently. Several optimization techniques are devised to further improve the performance.

We analyze the problem of summarizing co-locations mined from spatial databases. Co-location pattern mining finds patterns of spatial features whose instances tend to locate together in geographic space. However, the traditional framework of co-location pattern mining produces an exponential number of patterns because of the downward closure property, which makes it difficult for users to understand, assess or apply the huge number of resulted patterns. To address this issue, we study the problem of mining *representative co-location patterns* (RCP). We first define a covering relationship between two co-location patterns then formally formulate the problem of *Representative Co-location Pattern mining*. To solve the problem of RCP mining, we propose the *RCPFast* algorithm adopting the post-mining framework and the *RCPMS* algorithm pushing pattern summarization into the co-location mining process.

Contents

List of Tables

List of Figures