Towards Effective Spatial Data Mining: Uncertainty, Condensity and Privacy

Bozhong Liu

Faculty of Engineering and Information Technology

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

April 2017

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: Bozhong Liu

Date: 04/21/2017

Acknowledgements

Firstly, I would like to express my sincere gratitude to my principal supervisor, Dr. Ling Chen, for her continuous support of my Ph.D study and related research, for her patience, kindly support and inspiring motivation. She is always supportive when I feel frustrated or despairing. Her guidance helps me in all the time of my research life. Also, I am very thankful to Prof. Zhu, for his immense knowledge and wonderful advices. Without their guidance, my research life would be much more difficult. I am also grateful to my supervisor in Shanghai Jiao Tong University, Prof. Qiu, for his grate support and encouragement. Without his effort, I might have lost this precious opportunity of studying in UTS.

Secondly, I would like to thank my fellow students for their suggestion, discussion, cooperation and of course friendship. Their patience and support help me in overcoming numerous obstacles I have been facing through my research. Especially, I am grateful to Chunyang Liu, Meng Fang, Zhe Xu, Zhibin Hong and Shirui Pan for their kindness and sincerity. They not only help me a lot when I came to Sydney, but also provide many nice advices for my research work. They make my Ph.D study more colorful and wonderful.

Last but not the least, I would like to thank my parents for supporting me spiritually throughout writing this thesis and my life in general.

Abstract

Spatial data mining (SDM) is a process of knowledge discovery that the observing data is related to geographical information. It has become an important data mining task due to the explosive growth and pervasive use of spatial data. It is more difficult to extract interesting and useful patterns from spatial datasets due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. Although existing methods can handle the spatial mining task properly, as the arrival of the big data era, new challenges for SDM are arising.

Firstly, traditional SDM methods usually focus on deterministic datasets, where spatial events occur affirmatively at precise locations. However, the inherent uncertainty of spatial data makes the mining process more difficult. Classical spatial data mining algorithms are no longer applicable or need delicate modification. Secondly, traditional SDM frameworks produce an exponential number of patterns, which makes it hard for users to understand or apply. To solve the condensity issue, novel techniques such as summarization or representation must be carefully investigated. Thirdly, spatial data usually involves an individual's location information, which incurs location privacy problem. It would be a challenge to protect location privacy with enhanced data security and improved resulting accuracy.

To address the uncertainty issue, we study the problem of discovering co-location patterns in the context of continuously distributed uncertain data, namely Probabilistic Co-location Patterns Mining (PCPM). We develop an effective probabilistic co-location mining framework integrated with optimization strategies to address the challenges.

To address the condensity issue, we investigate the problem of Representative Co-location Patterns Mining (RCPM). We define a new measure to quantify the distance between co-location patterns, and develop two efficient algorithms for summarization.

To address the privacy issue, we solve the problem of protecting Location Privacy in Spatial Crowdsourcing (LPSC). We propose a secure spatial crowdsourcing framework based on encryption, and devise a novel secure indexing technique for efficient querying.

The experimental results demonstrate the effectiveness and efficiency of our proposed solutions. The methods and techniques used in solving concrete SDM tasks can also be applied or extended to other SDM scenarios.

Contents

List of Tables

List of Figures