Identifying Optimal Spatial Groups for Maximum Coverage by Using Clustering Algorithms

By

Liyang Liu

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of E-Commerce

Faculty of Science and Technology
University of Macau

2012

Approved by _____
Supervisor

_____

_____

_____

Date _____

In presenting this thesis in partial fulfillment of the requirements for a Master's degree at the University of Macau, I agree that the Library and the Faculty of Science and Technology shall make its copies freely available for inspection. However, reproduction of this thesis for any purposes or by any means shall not be allowed without my written permission. Authorization is sought by contacting the author at

Address:

Telephone:
Fax:
E-mail:

Signature _____

Date _____

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

University of Macau

Abstract

Identifying Optimal Spatial Groups for Maximum Coverage by Using Clustering Algorithms

By Liyang Liu

Thesis Supervisor: Professor Simon Fong
Master of E-Commerce Technology

This Research investigates the effectiveness of spatial grouping technology in geographical situation through spatial clustering algorithms which help to obtain a spatial group with certain features. Due to constraints of traditional spatial group algorithms in the real world, and unique features of every different dataset, it is not feasible to obtain exact spatial groups which are independent of each other. As such, some group may contain dissimilar data. Spatial grouping can be considered as one of the most important unsupervised learning techniques so as every other problem of this kind; it deals with discovering a structure in a collection of unlabeled data. The grouping process is used extensively not only to organize and categorize data, but is also useful for data compression. And it is also useful for analytic model construction and for discovering relevant knowledge in data.

This research provides new method to achieve spatial groups for maximum coverage and to avoid overlap. The model is designed with flexibility such that it can support evaluation of each method in the model, in terms of the quality of the resultant groups that is expressed in multi-attributes.

TABLE of CONTENTS

LIST of FIGURES ..... **vi**

LIST of TABLES ..... **viii**

LIST of Abbreviations ..... **ix**

CHAPTER 1: Introduction **1**

1.1Research Background 1

1.2Motivation 2

1.3 Organization of the Thesis 4

CHAPTER 2: Literature Review **5**

2.1 Introduction 5

2.2 Clustering Technology Overview 5

2.2.1 Partition method 5

2.2.2 Hierarchical method 6

2.2.3 Density-based method 6

2.2.4 Grid-based method 6

2.3 Application of Clustering Technology 7

CHAPTER 3: Spatial Data Representation **9**

3.1 Introduction 9

3.2 Overview of spatial data representation 9

3.3 Spatial data Model 11

3.3.1 Vector Data Model 12

3.3.2 Raster Data Model 13

3.3.3 Imager Data Model 15

3.3.4 Comparison between Vector and Raster Data Model 15

3.4 Attributes used in data Model 16

CHAPTER 4: Proposed Methodology **17**

4.1 Introduction 17

4.2 Preprocessing of Image 18

4.3 Data Transformation 19

4.4 Grouping by clustering algorithms 20

4.5 Grouping by LP algorithm 23
CHAPTER 5: Experimental Results and Analysis 25
5.1 Introduction 25
5.2 Data Resource and Image Preprocessing 25
5.3 Experimental Results 30
5.3.1 Comparison Result of HC and KM Clustering 30
5.3.2 Result of Grouping 35
5.4 Experimental Analysis 52
5.4.1 Experimental Evaluation Method 52
5.4.2 Comparison Experimental Result 52
5.4.3 Discussion of G_{net} 63
5.5 LP case study 64
5.5.1 For construction 64
5.5.2 For destruction 68
CHAPTER 6: Conclusion 73
6.1 Strengths and Limitations 73
6.2 Future Research Directions 74
Bibliography 75
Publications 80
Appendix A Pseudo Codes 81
Appendix B Data Resources 87
VITA 91

LIST of FIGURES

Number *Page*

Figure 1 Thematic layer approach allows us to organize the complexity of the real world (*Courtesy for the use of image by http://planet.botany.uwc.ac.za/nisl/GIS/GIS_primer/page_13.htm*) 10

Figure 2 Representation of the real world and showing differences in how a vector and a raster GIS will represent this real world. (*Courtesy for the use of image by http://planet.botany.uwc.ac.za/nisl/GIS/GIS_primer/page_15.htm*) 12

Figure 3 Vector format.....13

Figure 4 Raster format in order list.....14

Figure 5 Raster data with center point 14

Figure 6 Raster format with 2^2 and 3^2 grids.....14

Figure 7 Comparison of raster and vector data (Courtesy for the use of image by http://planet.botany.uwc.ac.za/nisl/GIS/GIS_primer/page_16.htm).....16

Figure 8 Workflow of proposed methodology.....17

Figure 9 8-immediate neighbors 19

Figure 10 Pseudo codes of grid image and nominalization 20

Figure 11 Pseudo code using LP.....24

Figure 12 Representative traffic volume map of (a) dataset 1 (b) dataset 2 26

Figure 13 Result of skeleton extraction in dataset 1 using (a) "*bwmorph*" function (b) thinning algorithm 27

Figure 14 Result of skeleton extraction in dataset 2 using (a) "*bwmorph*" function (b) thinning algorithm 28

Figure 15 Clustering results for first dataset: top half uses KM clustering method, bottom half uses HC method. (a) Result of case (2) (b) Result of case (3) (c) Result of case (7) (d) Result of case (8) 33

Figure 16 Clustering results of case (2) in dataset 2: (a) KM method, (b) HC method....34
Figure 17 Comparison result of (i) Spatial clustering (ii) Spatial group by using (a) EM (b) KM (c) HC (d) XM (e) DBScan....40
Figure 18 Comparison result of (i) Spatial clustering and (ii) Spatial group by using (a) EM (b) KM (c) HC (d) XM (e) DBScan and Spatial group in LP method....47
Figure 19 Time trend using second dataset using: (a) first dataset (b) second dataset....55
Figure 20 Coverage of each cluster for six methods using: (a) dataset 1 (b) dataset 2....57
Figure 21 Density of each cluster for six methods using: (a) first dataset (b) second dataset....59
Figure 22 Proportions of Cluster Sizes (Balance) of first dataset using (a) KM (b) EM (c) DBScan (d) XM (e) HC (f) LP....61
Figure 23 Proportions of Cluster Sizes (Balance) of second dataset using (a) KM (b) XM (c) EM (d) DBScan (d) HC (f) LP....62
Figure 24 Population in Tibet (Courtesy for the use of image by *http://sikkim.nic.in/sws/gis_construc_table.html*)....65
Figure 25 (a) Spatial clustering in KM (b) spatial group in KM....66
Figure 26 Spatial groups in LP....67
Figure 27 Gettysburg battle action map (Courtesy for the use of image by *http://www.nps.gov/gis/gisday/gis_gallery.htm*)....69
Figure 28 (a) Spatial clustering in LP (b) spatial group in LP....70
Figure 29 Spatial groups in LP....71

LIST of TABLES

Number *Page*

Table 1 Comparison between *Bwmorph* function and thinning algorithm 29

Table 2 Time complexity (running times) 30

Table 3 Important statistic from figure 18 48

Table 4 Comparison for running time of first dataset 53

Table 5 Comparison for log-likelihood of first dataset 54

Table 6 Comparison for running time of second dataset 54

Table 7 Comparison for log-likelihood of second dataset 55

Table 8 Time trend of four datasets 55

Table 9 Coverage of each cluster for six methods of first dataset 58

Table 10 Coverage of each cluster for six methods of second dataset 58

Table 11 Density of each cluster for six methods of first dataset 60

Table 12 Density of each cluster for six methods of second dataset 60

Table 13 Statistic information of six methods of second dataset 63

Table 14 Best G_{net} for each method 64

Table 15 Statistic information of KM 68

Table 16 Statistic information of LP 68

Table 17 Statistic information of KM 72

Table 18 Statistic information of LP 72

LIST of Abbreviations

EM – E-means

DBMS – Database Management Software

HC – Hierarchical Clustering

KM – K-means

LP – Linear Programming

XM – X-means

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

Acknowledgments

I would like to express my deepest appreciation to my supervisor Prof. Simon Fong for his professional guidance, great patience and valuable advices throughout the whole thesis writing process Prof. Fong always encourages me when the research could not move forward. He provided many helpful ideas to me, and also taught me much related knowledge to improve my ability. Prof. Fong was also an important mentor and friend. He provided a critical but supportive perspective on my research. He would gently remind me of the importance of conciseness, clarity, readability and proper use of technical terms in writing both papers and thesis. I also thank my co-supervisor Prof. Andy Ip. His guidance leaded me successful to explore my research direction; his valuable suggestion assisted me in improving my research skills; his patience and kindness carried me on through the difficult time.

Specially, I would like to express my sincere appreciations to other professors in University of Macau, Prof. Gong Zhiguo, Prof. Li Xiaoshan, Prof. Guo Jingzhi, Prof. Robert, and Prof. Yiu Kwok Tham. Their impressed teaching inspirited me to make further research in the university. I also want to thank Mr. William Sio, who helped in setting up the experimental environment in Lab. I particularly thank University Research Committee, who offered me a student helper job during my research in Macau.

Finally, and the most important of all, I thank my parents, who have always been pound of their daughter. Whenever in troubles, their confidences in me have made me able to find a way out. Their encouragement always makes me go forwards.