Preference Mining Techniques for Customer Behavior Analysis

By
Huy Quan Vu
BIT (Hons)

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Deakin University
September, 2014
I am the author of the thesis entitled

**Preference Mining Techniques for Customer Behavior Analysis**

submitted for the degree of **Doctor of Philosophy**

This thesis may be made available for consultation, loan and limited copying in accordance with the Copyright Act 1968.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: .................................................................
(Please Print)

Signed: .................................................................

Date: .................................................................

10 December 2014
DEAKIN UNIVERSITY
CANDIDATE DECLARATION

I certify the following about the thesis entitled (10 word maximum)

Preference Mining Techniques for Customer Behavior Analysis

submitted for the degree of Doctor of Philosophy

a. I am the creator of all or part of the whole work(s) (including content and layout) and that where reference is made to the work of others, due acknowledgment is given.

b. The work(s) are not in any way a violation or infringement of any copyright, trademark, patent, or other rights whatsoever of any person.

c. That if the work(s) have been commissioned, sponsored or supported by any organisation, I have fulfilled all of the obligations required by such contract or agreement.

I also certify that any material in the thesis which has been accepted for a degree or diploma by any university or institution is identified in the text.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: HUY QUAN VU
(Please Print)

Signed: 

Signature Redacted by Library

Date: 10 December 2014

Deakin University CRICOS Provider Code 00113B
To My Parents
# Table of Contents

Table of Contents

List of Tables

List of Figures

Acknowledgements

Publication List

Abstract

1 Introduction 1
  1.1 Motivation ................................. 3
    1.1.1 Decision Making Modeling .................... 3
    1.1.2 Travel Behavior Discovery .................... 5
    1.1.3 Emerging Preference Identification ............... 6
    1.1.4 Summary ................................ 8
  1.2 Research Objectives ............................ 9
  1.3 Research Problems ............................ 10
  1.4 Approaches ................................ 12
  1.5 Thesis Outline ................................ 15

2 Data Mining Techniques for Customer Behavior Analysis 17
  2.1 Data Mining in Tourism and Hospitality ............ 18
    2.1.1 Literature Review Methodology .................. 19
    2.1.2 Categorization Framework .................... 19
    2.1.3 Customer Behavior Analysis Categories ........... 21
    2.1.4 Data Mining Technique Categories ............ 25
# List of Tables

1.1 Hotel features used in existing studies. ........................................... 7

2.1 Journals included in this review. .................................................... 20

2.2 Distribution of articles according to the proposed categorization. .... 28

3.1 Structure of hotel rating data collections ........................................ 59

3.2 Shapley values .................................................................................. 60

3.3 Interaction Index values ...................................................................... 61

3.4 Structure of hotel rating data collections ........................................ 64

3.5 MAE values of the evaluated algorithms. ........................................... 66

3.6 Shapley values of hotel criteria. ........................................................ 68

3.7 Interaction indices for business tourists. ........................................... 70

3.8 Interaction indices for couple tourists. ............................................. 71

3.9 Interaction indices for family tourists. ............................................. 72

4.1 Photo search parameters. ................................................................. 84

4.2 Data collection .................................................................................. 85

4.3 Popularity of Areas of Interest. ....................................................... 88

5.1 Experiment results. ............................................................................ 119

7.1 A sample data set. ............................................................................. 173

7.2 A list of emerging hotel features. ..................................................... 173

7.3 Description of the Collected Data set. .............................................. 176
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.4</td>
<td>Emerging hotel features</td>
<td>179</td>
</tr>
<tr>
<td>7.5</td>
<td>Emerging trends of tourists’ concerns</td>
<td>180</td>
</tr>
</tbody>
</table>
List of Figures

4.1 Photo locations of Hong Kong inbound tourists. ..................... 86
4.2 Area of Interest for inbound tourists. ................................. 87
4.3 Tourist traffic flow in the Hong Kong metropolitan area. .......... 90
4.4 Travel route between Center Mong Kok and Times Square Towers. 92
4.5 Tourist activities in Areas of Interest. .................................. 94

5.1 Support Vector Classifier on representative data instances .......... 110
5.2 Performance of RIC with different values of \((\delta_{pos}, \delta_{neg})\) .......... 120
5.3 Representative photo examples. ............................................ 122
5.4 Photo collections and travel paths. ....................................... 127
5.5 Algorithm performance with a different numbers of photos in trunks. 128

6.1 A scheme for image reduction within \(3 \times 3\) blocks. .......... 134
6.2 Example of different clusters. ................................................. 138
6.3 Estimated fuzzy measure using the Minimum Spanning Tree approach. 142
6.4 Estimated fuzzy measure using the sub-additive Sugeno approach. ... 145
6.5 Estimated fuzzy measure using the supper-additive Sugeno approach. 147
6.6 Examples of basic components \(BC_1 \rightarrow BC_4\). ...................... 148
6.7 Estimated fuzzy measure using the decomposition approach. .......... 151
6.8 Fuzzy measure of different cluster shapes of the same cardinality. ... 154
6.9 Fuzzy measure values for cluster compactness. ........................ 155
6.10 Concentric circle pattern used to test image reduction operators. ... 156
6.11 Reduced circle image using cluster-based mode-like averaging functions. 157
6.12 A satellite image of a sea port. .............................. 159
6.13 Reduction of the Port of LA image using monotonic averages. . . . 161
6.14 Reduction of the Port of LA image using non-monotonic averages. . 162

7.1 Hotel Features’ Identification Process. ............................. 168
7.2 Output Hotel Feature Candidates. .................................. 177
7.3 Identified popular hotel features. .................................... 178
Acknowledgements

I would like to thank Dr. Gang Li, who was my Honors supervisor and also my PhD supervisor. His many suggestions, constant understanding and his encouragement have been the greatest gift I have received. He is the one who opened up the scientific research world to me, who made me believe that I could achieve more than I thought I could. He is the one who has witnessed my every moment of happiness, sadness, desperation and excitement in the past six years as a candidate.

I am also thankful to Prof. Gleb Beliakov and Dr. Vicky Mak, my PhD co-supervisors, for their useful suggestions and mentoring in studying mathematic and optimization techniques. I am also grateful to Prof. Rob Law at Hong Kong Polytechnic University for his support, in finding interesting applications for my proposed data mining techniques in Tourism and Hospitality Management.

I would like to thank my research mates and friends in our TULIP group (Team for Universal Learning and Intelligent Processing): Jia Rong, Yongli Ren, Shaowu Liu, and Rosanna Leung. You are my best friends forever.

Of course, I am very grateful to my parents for their patience and love. Without them this work would never have come into existence.

Melbourne, Australia                                      Huy Quan Vu
September, 2014                                      xii
Publication List

To date, I have a total of twenty papers published, which include thirteen journal articles, six conference papers and one book chapter. These are listed below:

Refereed Journal Articles


Refereed Conference Papers


**Book Chapter**

Abstract

Customer preference refers to how customers select goods and services in relation to factors such as taste, preference and individual choices. It governs the customers’ behavior of buying or not buying products. Due to its excellent ability in extracting and identifying useful information from large customer databases, Data mining has been a popular supporting tool for businesses to gain insight into customer behavior. However, it is not feasible that all tools and techniques to meet all situations and their requirements be developed even though rapid and accurate identification of consumer demands is essential to business management. The focus of this thesis is on novel data mining techniques for studying customer behavior and preference, and the tourism industry is used as a testbed to demonstrate our techniques.

The first topic is decision making modeling. The understanding of customer’s decision making process is important for improving products and designing focused marketing strategies. However, decision making is a complicated process, which usually involves the comparison of two or more criteria and evaluations against multiple factors according to their priorities. Traditional techniques often fail to capture such process closely to the real situation due to their natural assumption that the input criteria are independent of each other. In reality, the independence of criteria cannot be assumed, as some interactions among different criteria, including complement, and correlation, do exist. To tackle this problem, one of the contributions of this thesis is the fuzzy measure based preference mining technique. A new R package, the Rfmtool, is specifically developed for customer preferences analysis. It has the capability of considering multiple criteria simultaneously and accounts for all interactions among
different criteria, which outperforms existing modeling techniques. A case study in modeling the hotel selection process of tourists demonstrates the effectiveness of this technique.

The second topic is on travel behavior discovery. Behavior of tourists during their trips is important to decision makers in terms of destination development, transportation planning, and impact management at a tourism destination. Due to inefficiency of traditional data collection methods, such as survey and opinion poll, attention has been shifted to geo-tagged multimedia data, such as geo-tagged photos that are freely available on photo sharing sites. However, the geo-tagged photo data can be noisy or misleading, especially when many photos have been taken in transit rather than at the attractions themselves, and this makes it difficult to extract travel patterns.

There are an incredible number of travel photos available on the Internet, with only a small number of them tagged with geographical data. Another challenge is how to make use of such data resources for travel analysis as more and more applications for tourism services and recommendations are developed based on geo-tagged photos and satellite imaging. Terminal user devices are often insufficient in processing and displaying satellite images due to their high resolution. A solution is needed for satellite image processing to facilitate the development of tourism applications using GPS data. This thesis also aims to address such a challenge by the following contributions: 1) a new travel behavior mining framework that can effectively process the geo-tagged photo data to reveal preferred tourism attractions, and the travel route of tourists; 2) two travel trajectory construction techniques, named Representative Instance Classification and Bayesian Latent Dirichlet Allocation, for incorporating travel photos without GPS data into travel analysis; 3) three alternatives to satellite image reduction, which are Minimum-Spanning Tree approach, Sugeno-type Fuzzy Measure approach and Decomposing Fuzzy Measure approach. These three approaches can satisfy special characteristics and requirements for satellite image applications. The effectiveness of a geo-tagged photo analysis framework is presented in a travel behavior analysis case study at a tourism destination. A series of experiments were
conducted to evaluate the performance of trajectory construction and satellite image processing techniques.

The third topic is on emerging preference identification. Tourism managers are interested in understanding tourist preferences to improve their strategic planning, marketing, and product development. However, tourist preference is unstable and dynamic, which influences the performance of tourism businesses. The challenge of emerging preference identification is that analysts have no prior knowledge on what features should be included in their study. Large data samples are also required to identify emerging changes in customer preference patterns. These make traditional research methods inadequate. Aiming to address this challenge, this thesis proposes a text processing framework for identifying tourists’ concerns and the Positive and Negative Emerging Pattern Mining (PN-EPM) technique for emerging preference identification. The text processing framework helps identify tourists’ concerns automatically from online reviews, without requiring any prior knowledge of what should be considered. PN-EPM is efficient in capturing increasing changes and decreasing changes in customer preference. The performance of these techniques is demonstrated in the application of emerging hotel preference identification.

The developed methods in this thesis are valuable for researchers and business managers in gaining a deeper insight into customer behavior and for more appropriate business planning and decision making. This thesis also provides a basis for the work of data mining applications in customer behavior studies, especially in the tourism industry, which is currently the fastest growing industry.

Keywords: Machine Learning, Data Mining, Fuzzy Measure, Geo-tagged Photos, Tourism Management, Preference Mining, Behavior Analysis, Multi Instance Learning.
Chapter 1

Introduction

Customer preference refers to how customers select goods and services in relation to factors such as taste and individual choices [40]. In business, preference is an important factor that governs the customers’ intention of buying or not buying products. The understanding of customers’ preference and behavior is the key to effective marketing planning, which is crucial for the success of retail businesses [254].

Researchers have long sought to support the strategic plan of businesses through a good understanding of customer preferences. For instances, Heyd et al. [116] and Ominaga et al. [228] studied the customer preference of coffees to assist in the coffee product development task. Viney et al. studied the preferences for health and healthcare to predict the demand for health care [288]. Zhu and Ierland modeled the customer preferences for Novel Protein Food and Environmental Quality [322]. Le and Mustapha studied the customer preferences in telecommunication products to help network operators design the most profitable bundles together with their associated prices, and also to be competitive in the market [167]. Yang et al. predicted customer preference in the retrofit design of food and drink products [309]. Jolly et al. modeled the customer preference to evaluate factors affecting the consumption of roasted
peanut products [136].

Data mining has been considered as one of the best supporting tools for studying customer preference and behavior, due to its excellent ability in extracting and identifying useful information from large customer databases [35]. The fast development of computing technology has made a large number of data mining algorithms available. However, why are we still interested in developing preference mining techniques for customer behavior analysis? The concept of “customer” can be understood widely as “clients”, “buyers” or “purchasers” who are the recipients of goods, services, products, or even ideas, obtained from a seller, vendor, or supplier for a monetary or other valuable consideration [141, 268]. Due to the variety of the domains and contexts, it is infeasible to develop a number of tools and techniques that can meet all situations and their requirements. Rapid and accurate identification of customer demands is essential to production development and marketing, which are crucial to business management [258, 283].

The focus of this thesis is on preference mining techniques for studying customer behavior, and the tourism industry is used as a testbed to demonstrate our techniques. With hundreds of millions of people traveling around the world every year, the tourism industry is one of the most rapidly growing industries [85], and plays a significant role in the global economy. Over the past decade, the contribution of tourism has risen dramatically. The World Travel and Tourism Council commented that, in 2012, travel and tourism were directly and indirectly responsible for 9% (US $6.6 trillion) of global GDP and employed over 260 million employees worldwide [284]. Tourism appears to be one of the most concerned domains in customer behavior analysis [230, 258, 289], which presents great challenges for researchers in gaining
insightful knowledge to accommodate the growing demand of tourism business. In the context of the tourism domain, “customer” is usually refereed to as “tourist”. In this thesis, the term tourist is used.

1.1 Motivation

The motivation of this thesis is presented using the following scenarios in analyzing tourist behavior.

1.1.1 Decision Making Modeling

The understanding of customers’ decision making process is an effective strategy for improving products, and designing focused marketing strategies. For instance, a customer may give a low rating on a product for the quality and service criteria but still select it based on its cheap price. Other customers may select a product only if it can satisfy both the quality and service criteria. An insight into how criteria interact with each other in guiding customers’ intention can provide business managers with insightful knowledge into the preferences. However, such decision making process usually involves the comparison of two or more alternatives, that are evaluated against multiple factors according to their priorities. It is a challenging task model customer decision making process closely to the real situation.

Let take a simple example to illustrate this issue. A tourist wants to choose a hotel that is cheap, clean, and has good service. He/she makes the decision based on several hotel features such as Price, Cleanliness, and Service. The options have been narrowed down to four choices with the following utility values:
<table>
<thead>
<tr>
<th>Hotel</th>
<th>Price</th>
<th>Cleanliness</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotel A</td>
<td>0.7</td>
<td>0.6</td>
<td>0.9</td>
</tr>
<tr>
<td>Hotel B</td>
<td>0.6</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Hotel C</td>
<td>0.6</td>
<td>0.7</td>
<td>0.4</td>
</tr>
<tr>
<td>Hotel D</td>
<td>0.7</td>
<td>0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>

The tourist starts the decision-making process by ranking these hotels. The tourist can quickly suggest that hotel A is ranked higher than hotel D and hotel B is ranked higher than hotel C. This is because the service utility values for A and B are higher than D and C respectively, while the utility values for price and cleanliness are similar. Here, these rankings reflect some preferences over the criteria, and they follow the monotonicity of the preference relation. Besides, the tourist realizes that some unobvious comparison exists between hotel pairs A vs. B and C vs. D. For instance, the utility value for the price of hotel A is higher than hotel B, whereas the utility value of cleanliness of hotel B is higher than hotel A. The tourist might adopt the following reasoning: *If a hotel has good service, it is considered more important if it also has a good price, so hotel A is ranked higher than hotel B; if a hotel has poor service, then it is more important that it is clean, so hotel C is ranked higher than hotel D.* This leads to the ranking order A > B > C > D.

The reasoning made by the tourist is an example of a complex decision-making process in a real situation. The problem originates from the fact that criteria are not mutually independent; and that some interaction does occur between them. In addition, the weight of criteria should be considered in the group of criteria. If such a complex decision making process can be modeled, the preference of consumers can be revealed and analyzed effectively.
1.1.2 Travel Behavior Discovery

Behavior of tourists during their trip is important to decision makers in destination development, transportation planning, and impact management at a tourism destination [176]. The travel pattern discovery process can be supported by different types of data, which are traditionally collected by survey and opinion poll methods [16, 207]. Thus, collected data are limited in the number of responses or in the scale of geographical areas [319].

Recently, advances in multimedia and mobile technology allow massive amounts of user generated data to be created, such as travel photos. Many photo capturing devices, such as smart phones and tablets, have built in GPS technology, which enables geographical information (latitude and longitude coordinates) to be stored in the meta data area of each photo taken. Those geo-tagged photos, embedded with time and geographic information, implicitly carry the spatial-temporal movement trajectories of the photo shooters. With millions of geo-tagged photos available, online databases such as Flickr have been a rich data resource for mining tourist travel patterns [170,320]. Since the use of geo-tagged multimedia data in travel research is a relatively new approach, a natural question arises as to what data mining techniques can be used effectively to discover the travel preference of tourists? and How can the mined knowledge be used to support the business of tourism in a real situation?

Although a massive amount of travel photos are available on the Internet, only a small number of them are tagged with geographical data. This is due to the fact that photo capturing devices with the GPS tagging function have only been available in recent years. Previous to this, many tourists did not own such cameras or enable
its GPS function, which left many travel photos taken without GPS tagging information. Another emerging challenge is that if such a massive number of available travel photos without GPS information can be utilized for travel behavior analysis, tourism researchers and business managers can maximize the potential of such data resources to fully capture the travel behaviors of tourists.

Some attempts have already been made to develop tourism applications using geotagged photos, such as automatic intelligence services to make recommendations on “where to go”, “what to see” [200], or applications that equips tourists with knowledge mined from travelogues [109]. Those applications often involve the analysis and display of satellite images on end user devices such as smart-phones, tablets or laptop computers. These screens are being used to display images captured at higher resolutions than the screen is capable of displaying, which makes it insufficient for many end user devices to process and display them effectively. There is a need for a solution that processes the satellite image which overcomes the limitation of end users to facilitate the development of tourism applications using GPS data.

1.1.3 Emerging Preference Identification

Tourism managers continue to find ways to understand tourist preferences, with the aim of improving their strategic planning, marketing, and product development. However, tourist preference is unpredictable and dynamic. For example, in regards to hotel feature preference, tourists once preferred having a telephone in their room. During that time, charging for telephone usage used to be a significant source of revenue, however usage has declined to a point wherein investing in this facility resulted
in losses for many hotels that offered this facility [127]. Today, hotels gain significant customer satisfaction by offering free wireless Internet [41]. These changes in tourists' concerns can affect the performance of tourism businesses. Several studies have analyzed hotel features to acquire knowledge on tourist preference, as shown in Table 1.1.

### Table 1.1: Hotel features used in existing studies.

<table>
<thead>
<tr>
<th>Hotel Features</th>
<th>Related Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleanliness, Location, Room, Service, Sleep Quality, Value</td>
<td>[189]</td>
</tr>
<tr>
<td>Shabby Bed, Clean Rats, Friendly Staff, Limited Parking, Good Room</td>
<td>[36]</td>
</tr>
<tr>
<td>Personalization, Warm Welcoming, Special Relationship, Straight from the Heart, Comfort</td>
<td>[15]</td>
</tr>
<tr>
<td>Promenade and Comfort, Security and Protection, Network Services, Pleasure, Hotel Staff and their Services, News and Recreational Information, Cleanliness and Room Comfort, Expenditure, Room Facilities, Car Parking</td>
<td>[266]</td>
</tr>
<tr>
<td>Location, Size and Diversity, Characteristics of the Lobby, Characteristics of the Rooms, Parking</td>
<td>[214]</td>
</tr>
<tr>
<td>Type of Building, Location, Number of Bedrooms, Price per Room, Hiring Horses, Play Area, Meal Service, Swimming-pool, Sports Facilities, Mini-farm, Bathroom, Type of Rent, ‘Q’ Quality Award, Booking</td>
<td>[6]</td>
</tr>
<tr>
<td>Problem Solving Abilities by Service Personal, Price Level, Sanitary Hot Spring Environment, Convenience Traffic Route/Shuttle, Special Promotions, Convenience of Reservation Procedure, Food and Beverages Service</td>
<td>[121]</td>
</tr>
<tr>
<td>Staff Service Quality, Room Quality, General Amenities, Business Services, Value, Security, IDD Facilities</td>
<td>[61]</td>
</tr>
<tr>
<td>Location, Price, Facilities, Cleanliness</td>
<td>[191]</td>
</tr>
<tr>
<td>Cleanliness, Location, Room Rate, Security, Service Quality, Reputation of Hotel</td>
<td>[11]</td>
</tr>
</tbody>
</table>

The majority of existing studies have focused on identifying and analyzing the most valuable hotel features that significantly affect the hotel selection process for tourists. These studies examine several commonly mentioned hotel features, such as price, location, room quality, staff, and service [51,189,270]. However, the natural assumption that popular or frequently mentioned hotel features are worth studying is one limitation of these studies. Infrequently used hotel features may also be interesting despite being mentioned less because these features are new and have
already gained increasing interest from tourists. The wireless Internet facility is one such feature \cite{41}. Wireless technology has only become widely used in recent years, with many wireless Internet-enabled devices available to users. The demand for wireless Internet is growing significantly, even though the traditional hotel features remain more popular. As such, managers must effectively identify features that are becoming important to tourists. If such emerging preferences are effectively captured, insights into tourists’ interests can be discovered, which can enable tourism managers to gain a better understanding of the rapid changes in tourist preferences.

1.1.4 Summary

The above examples have illustrated several important issues in customer behavior analysis in a tourism context. Tourism managers and researchers have a desire for insightful knowledge into tourists’ decision making process, travel behavior and emerging preference for effective business management. Unfortunately, such desires have not been fully satisfied, because existing tools and techniques are incapable of performing the analysis tasks effectively. For decision making process analysis, there is lack of techniques to capture the complex decision making process of the customer to the real situation. For travel behavior analysis, no framework was reported to effectively discover the travel preference of tourists from the fruitful data resources; the geo-tagged photos. Very limited tools and techniques support the development of tourism applications using GPS technology. For emerging preference identification, no method was reported that can capture and effectively identify the rapid changes in the preference of tourists. These research gaps, if fulfilled, can better support the
customer behavior analysis tasks and a deeper understanding into customer preferences can be achieved. Thus businesses can improve their performance by efficient strategic planning and decision making.

1.2 Research Objectives

The goal of this thesis is to improve the effectiveness of customer behavior analysis tasks by developing new tools and techniques for better modeling the customers’ decision making process, to efficiently analyse their travel behavior, and for more effective identification of their emerging preference. The aims of this thesis is to address the challenges in analyzing and understanding customer behavior, from decision making, travel planning, travel activities, to changes in preference with tourism as a testbed. To achieve this goal, the following research objectives are the targeted of this thesis:

**Multiple Criteria Decision Making Modeling:** A better approach to modeling the complex decision making process of customers so that a deeper insight into customer behavior and their preference can be effectively discovered.

**Geo-tagged Photos’ Analysis:** New data mining techniques that can effectively extract hidden travel preferences of tourists from Geo-tagged Photo data resources are introduced. New methods to incorporate ungeotagged travel photos for the travel behavior analysis are presented. The challenge in tourism application development using GPS technology are addressed with new approaches for satellite image processing.
Emerging Preference Identification: A new approach to automatically capture and identify emerging changes in tourist preference are proposed so that better business strategies can be developed.

Applications in Tourism Management: The effectiveness of the proposed methods are presented in several real-world case studies of analyzing customer behaviors in the context of the tourism industry.

1.3 Research Problems

Customer decision making typically involves the comparison of two or more alternatives, which are evaluated against various factors according to their priorities. There are different types of interactions between criteria in the decision making process, such as independence, complement, and correlation [102]. Effective modeling of Multi-criteria Decision Making (MCDM) should consider multiple criteria simultaneously and account for such interaction. Unfortunately, traditional approaches to MCDM in the studying customer behavior are assumed to be too strong on the assumption that the criteria are independent. Our research problems in MCDM process modeling are:

- How to consider the interaction between multiple criteria into the modeling customer MCDM processes?
- How to effectively extract and present the preference of the customer to support business management tasks?

Geo-referenced multimedia data such as travel photos have emerged as a potential gold mine for studying travel behavior analysis. The massive volume of shared photos
available is a comprehensive resource for studying travel behaviors. Nevertheless, the data can be noisy or misleading, especially when many photos have been taken in transit rather than at the attractions themselves. The sequence of photos, in some situations, can be used to infer personal activities, intuition and goals. However, photos are static media, while travel behaviors are dynamic. These factors make it hard for traditional data mining approaches that analyze and extract useful knowledge for geo-tagged data resources. In addition, there are an incredible number of travel photos available on the Internet but only a small number of them are tagged with geographical data. This prevents tourism researchers and business managers from fully capturing and the understanding travel behavior of tourists. In view of application development, tourism applications based on geo-tagged photos usually involve the processing of a large amount of satellite images. The resolutions of satellite images are captured at high resolution, which makes them insufficient for many end user devices to process and display effectively. The ubiquitous nature of small display screens in such devices has reinforced the need for robust image processing techniques. However, it is important that the preprocessing of the image does not cause the loss of fine image details that may convey important information relevant to the analysis of the image content. Our research problems in using geo-tagged photos for travel behavior analysis and applications are presented as follows:

- **What techniques can be used to process and analyze the noisy geo-tagged photo data for extracting useful knowledge about the travel preference of tourists?**

- **How can the large amount of ungeo-tagged photos be transformed and incorporated into travel preference analysis?**
How can the satellite images be represented and processed effectively on end user devices?

The problem of emerging preference identification is different from traditional approaches where analysts usually have no prior knowledge on what should be included in the analysis. Traditional methods for data collection in hotel preference analysis is to use survey, with hotel features represented by short questions or a set of keywords. However, in the case of emerging hotel preference identification, such interested features are unknown. It is difficult, if not impossible, to define questions or keywords for analysis without any prior knowledge about what should be included. The analysis of emerging preferences also requires large data sample, so any emerging changes in customer preference patterns can be captured. This makes traditional approaches using statistical techniques insufficient. Our problems in emerging preference identification are:

- How to capture customer preference without prior knowledge of what should be included for analysis?
- How to effectively identify emerging changes in customer preference?

### 1.4 Approaches

In order to address the above problems, the following approaches are presented in this thesis:

**Preference Mining using Fuzzy Measure:** This thesis addresses the challenges of MCDM modeling by using fuzzy set theory and deploying a new aggregation
function, the **Choquet Integral**. Two metrics computed from the fuzzy measure, the **Shapley value** and the **Interaction Index**, can provide insight into the preferences of customers and explore the interaction among criteria. A new \( R \) software package, named **Rfmtool**, is specifically developed for customer preferences analysis. The proposed approach can consider multiple criteria simultaneously and account for all interactions among different criteria, which outperforms existing MCDM modeling techniques.

**Geo-tagged Photos Analysis Techniques**: For the challenges in geo-tagged photos analysis, several objectives are defined as follows:

- **Travel Behavior Mining Framework**: A general framework for mining preferences from geo-tagged travel photos is presented. Two data mining techniques, **Density Clustering** and **Markov Chain**, are incorporated in this framework for identifying preferred tourism attractions, as well as revealing the travel flow of tourists. The introduced framework has the potential to benefit tourism researchers by providing useful practical applications for destination development, transportation planning, and impact management.

- **Travel Diary Construction**: The challenges of incorporating the massive amount of ungeotagged photos for travel analysis are addressed from two perspectives: **visited location verification** attempts to verify if a tourist has actually visited a hypothesis location based on the content of his/her travel photos; **travel diary construction** attempts to estimate the most likely travel routes or paths the tourists took based on the content and temporal information in their travel photo collections. Two novel machine
learning techniques are proposed to accomplish the task; the Representative Instance Classification (RIC) and the Bayesian Latent Dirichlet Allocation (BA-LDA).

- **Satellite Image Reduction**: The challenge of end user tourism application development is addressed by using image reduction, and propose three alternatives to image reduction techniques using fuzzy set theory that can satisfy special characteristics and requirements for satellite image applications. These include the Minimum-Spanning Tree approach (MST), Sugeno-type Fuzzy Measure approach (SFM) and the Decomposing Fuzzy Measure Approach (DFM).

**Positive and Negative Emerging Pattern Mining**: This thesis addresses the problem of emerging preference identification by proposing a method named Positive and Negative Emerging Pattern Mining (PN-EPM). This technique is specifically proposed for detecting any increasing or decreasing changes in customer preference. New interesting knowledge can thus be discovered in a rapid and comprehensive manner.

**Applications in Tourism Management**: The use of developed techniques is demonstrated in practical tourism applications. The challenges in tourism management of a representative tourism industry, such as Hong Kong, is addressed with the following aims: to explore pre-trip behavior by discovering tourist accommodation selection preferences when planning for their trips, to explore the in-trip behavior by studying the tourist travel behavior at tourism destinations, to explore changes in tourist behavior by identifying emerging changes in tourist demand for tourism products. The analysis outcome will
contribute toward understanding the Hong Kong tourism industry, and help tourism and hospitality managers create better tourism products, attract more customers and generate more benefits.

\subsection{Thesis Outline}

An overview of the organization of this thesis is presented as below:

\textbf{Chapter 2} A \textit{Literature Review} provides an overview about the applications of \textit{Data Mining} techniques in \textit{Customer Behavior Analysis}. Emphasis is put on current and emerging issues in in the tourism industry, which defines the theme of technological development and applications in later chapters of this thesis.

The main contributions of this thesis are reported in Chapters 3, 4, 5, 6 and 7. Based on the defined objectives, these five chapters can be further divided into three parts. The first part (Chapter 3) introduces a new MDCM modeling technique based on fuzzy measure theory. The second part (Chapters 4, 5 and 6), which is also the central part of this thesis, aims to present tools and techniques for mining travel preferences of tourists from geo-tagged photos. The third part (Chapter 7) presents a new data mining technique for identifying emerging change in customer’ preferences.

\textbf{Chapter 3} presents a \textit{fuzzy measure based preference mining techniques} and the development of a \texttt{R} package, the (\texttt{Rfmtool}), for effective modeling of the multi-criteria decision making process. The advantage of this approach is demonstrated and compared with other approaches in an application for discovering tourists’ hotel selection preferences.
Chapter 4 introduces a general framework for travel behavior mining from geotagged photos, using density clustering and markov chain techniques. An application for analyzing inbound tourists’ travel behavior demonstrates its efficiency in providing comprehensive and insightful knowledge for supporting tourism management tasks.

Chapter 5 proposes novel machine learning methods for travel diary construction from travel photos, which leverages the challenges in using ungeotagged photos for travel analysis. Two techniques are presented, including the Representative Instance Classification for visited location verification from travel photos and the Baysian Latent Dirichlet Allocation for travel diary construction from travel photo sequences.

Chapter 6 proposes three alternatives to the cluster compactness measure. These are the Minimum-Spanning Tree approach, the Sugeno-type Fuzzy Measure approach and the Decomposing Fuzzy Measure approach. All three alternatives are designed to accommodate the special requirements of Satellite Image Reduction.

Chapter 7 presents the Positive and Negative Emerging Pattern Mining technique. The effectiveness of this approach is demonstrated in an application of emerging hotel preference identification.

Chapter 8 concludes the thesis by summarizing both theoretical and practical contributions and indicating some possible avenues for future research.
Chapter 2

Data Mining Techniques for Customer Behavior Analysis

The aim of customer behavior analysis (CBA) is to provide business managers with insight into the behavior of customers, to support organizations to discriminate better and more effectively allocate resources to the most profitable group of customers [35, 275]. The analysis of customer behavior is built upon the foundation of customer data and information technology tools, especially the Data Mining tools [223]. Many organizations have collected and stored a massive amount of data on their customers, suppliers, and business partners. Unfortunately, sometimes they are unable to discover valuable information and transform it into useful knowledge [35]. To address the need to discover the hidden knowledge from the enormous amount of data, data mining research started as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases” [88]. Later on, Turban viewed data mining as “the process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from large databases” [283]. A similar definition referring to data mining as the process of extracting or detecting
hidden patterns or information from large databases is also provided in [5,34]. Data mining tools are excellent in extracting and identifying useful information and knowledge from large customer databases, and are the most common supporting tool for business decision making [35]. The use of data mining techniques in CBA is an emerging trend in the global economy, in order to acquire and retain potential customers and maximize customer value [223]. As such, the research and development of data mining techniques to support CBA is a research direction worth pursuing.

In order to establish the background of this work, this chapter is devoted to a comprehensive review of literature related to data mining techniques in customer behavior analysis. Since, the application case studies are in the tourism domain, the review will be mainly on publications in academic journal on tourism and hospitality management. Section 2.1 reviews existing data mining techniques and their applications in tourism and hospitality management. Section 2.2 investigates current issues in tourism management and discuss limitations of traditional approaches. Section 2.3 summarizes this chapter and provides an indication and motivation for the later chapters of this thesis.

### 2.1 Data Mining in Tourism and Hospitality

This section firstly describes our approach for conducting the literature review, and then presents a categorization framework to present the work in data mining applications for CBA.
2.1.1 Literature Review Methodology

This literature review examines articles that were published by major tourism and hospitality journals, from 2007 to 2014. Since there is no standard list of research journals in this field, the focus of the review is on publications in top tier journals according to certain lists of popular journal rankings [71, 209, 256]. Table 2.1 presents the journal names included in this review. Most of these journals are ranked A/A* according to Excellence Research Australia ranking framework 2010 (ERA ranking 2010), which is proposed by the Australian Research Council [71]. In addition, three tier B journals are also included, as they are highly regarded by tourism and hospitality researchers [129, 209]. These journals are “Current Issues in Tourism”, “International Journal of Contemporary Hospitality Management” and “Journal of Vacation Marketing”. This review only account for refereed research articles, whereas editor prefaces, case studies, research notes and conference or book reviews are excluded. The content of these research articles are studied to construct an overview about the research topics in CBA and the related data mining techniques, which are presented in the subsequent section.

2.1.2 Categorization Framework

Although there exists several categorization frameworks in customer relationship management [158, 235, 272], no existing survey focuses on customer behavior analysis. Through an information search and a content analysis of the articles, the CBA are classified into the following five dimensions according to their targeted knowledge about customers:
Table 2.1: Journals included in this review.

<table>
<thead>
<tr>
<th>Journal Name</th>
<th>ERA Ranking 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Annals of Tourism Research</td>
<td>A*</td>
</tr>
<tr>
<td>2. Cornell Hospitality Quarterly</td>
<td>A</td>
</tr>
<tr>
<td>3. Current Issues in Tourism</td>
<td>B</td>
</tr>
<tr>
<td>4. International Journal of Hospitality Management</td>
<td>A</td>
</tr>
<tr>
<td>5. International Journal of Contemporary Hospitality Management</td>
<td>B</td>
</tr>
<tr>
<td>6. Journal of Business Research</td>
<td>A*</td>
</tr>
<tr>
<td>7. Journal of Leisure Research</td>
<td>A</td>
</tr>
<tr>
<td>8. Journal of Hospitality and Tourism Research</td>
<td>A</td>
</tr>
<tr>
<td>9. Journal of Service Research</td>
<td>A</td>
</tr>
<tr>
<td>10. Journal of Sustainable Tourism</td>
<td>A</td>
</tr>
<tr>
<td>11. Journal of Tourism Studies</td>
<td>A</td>
</tr>
<tr>
<td>12. Journal of Travel and Tourism Marketing</td>
<td>A</td>
</tr>
<tr>
<td>13. Journal of Travel Research</td>
<td>A*</td>
</tr>
<tr>
<td>15. Tourism Analysis</td>
<td>A</td>
</tr>
<tr>
<td>16. Tourism Management</td>
<td>A*</td>
</tr>
</tbody>
</table>

- **Tourist Profile**

- **Tourist Behavior**

- **Tourist Preference**

- **Tourist Perception**

- **Tourist Demand**

These dimensions aim to explore different aspects of customer behavior in a tourism context, and share the common goal of creating a deeper understanding of customers to help improve business performance. These tasks are supported by a wide range of data mining techniques by discovering hidden customer characteristics and behavior from data. Data mining is in fact a wide area, which contains a large number of tools and techniques. Different researchers categorize data mining techniques differently and even the categories have different names in different contexts [95, 108, 273, 283]. Data mining techniques are categorized based on a comprehensive analysis of the topics and methodologies in the journals listed in Table 2.1.
We noticed that not all available data mining techniques are applied in tourism and hospitality areas, while, some techniques may be used differently in different applications. The following data mining technique categories are proposed to make it convenient for presenting data mining applications in the CBA context:

- **Association**
- **Clustering**
- **Prediction Models**
- **Forecasting**
- **Visualization**
- **Computational Intelligence**
- **Statistical Learning**

A brief description about each of the proposed *Customer Behavior Analysis Categories* and *Data Mining Techniques Categories* are provided in the subsequent sections.

### 2.1.3 Customer Behavior Analysis Categories

In the context of this study, CBA is referred to as the study and exploration of customers in relation to tourism activities. Therefore, Only works that bring more knowledge about customer behavior to improve strategic planning and decision making for tourism businesses are considered. The details of five CBA categories are presented as follows:
Tourist Profile  The focus of this category is on the characteristics of tourists. This knowledge can help tourism managers develop tourism productions or services that are suitable for different groups of tourists. The main task of this category is tourist profile construction, which helps identify different groups of tourists, and their characteristics. For instance, Li et al. constructed and analyzed international tourists’ profile to Hong Kong [179], while, Lo et al. focus on leisure tourist profiles to generate insights into the nature of consumer markets at this location [190]. Tourist motivation profiles were constructed to identify the drive for them to travel to rural areas in Korea [183], or to rural community-based festivals in Hong Kong [232]. Lee and Lee compared demographic profiles between Korean and Japanese leisure tourists [169], and Xu et al. compared profiles between British and Chinese international student tourists [307]. Rong et al. profiled purchasers and browsers for online hotel web services [253]. Park and Reisinger constructed a profile for Western, Asian, and Hispanic tourists to reveal their shopping characteristics [233].

Tourist Behavior  This category aims to study tourists’ behavior when participating in tourism activities. It contains several sub-topics, including decision making, travel behavior, and expenditure behavior. Among them, Decision making analysis focuses on discovering how tourists make decisions and what factors influencing such decisions. Decision making is usually about selecting an option over several alternatives. For instances, Lyons et al. studied the selection process of holiday destinations for Irish tourists [195]. Crouch et al. focused on modeling consumer choice behavior in space tourism [74]. Jin et al. investigate how option framing affects tourists’ choice of package-tour services [135].
Yoon et al. assessed consumers’ brand selection process in a service context [313]. *Travel behavior* analysis focuses on the actual activity of tourists during their trips such as the movement patterns of tourists [48, 76, 305], visitor activity along roads and hiking trails [300], or the actual length of stay of tourists [19, 240]. *Expenditure behavior* analysis focuses on the consumption of tourism products and services such as accommodation, food, transport, shopping and entertainment [80]. For instance, Bojanic examined the shopping expenditures of Mexican visitors [38] and Coussement et al. predicted the customer churn in online gambling [72].

**Tourist Preference** This category studies the interest and expectations of tourists towards tourism destinations, products or services. For example, Hsu et al. analyzed the preference for tourists’ choice of destination [123]. Konu et al. focused on the preferred criteria of tourists when selecting a ski destination [157]. Albaladejo-Pina and Díaz-Delfa studied tourist preferences for rural house stays [6]. Gautam investigated the preference of consumers regarding tourism services in an Indian context [93]. Okazaki and Morikazu examined the media choice for travel information searches in Japan [226], while Kim et al. examined Japanese tourists’ shopping preferences [151], and Corsi et al. explored the expectation of tourists towards wine menus in restaurants [68].

**Tourist Perception** The focus of this category is on *tourist opinion* and *tourist experience*. *Tourist opinion* analysis studies tourists’ thinking, attitudes and intentions. For instance, Lee et al. analyzed tourists’ attitudes towards textiles and apparel-related cultural products [172]. Huang et al. examined tourist
intentions of buying or purchasing travel products [125]. Jang et al. studied American customers’ perceptions about the attributes of Asian foods [132]. Ahmad and Richard aimed to understand consumer’ brand categorization [4]. Chiu et al. studied the opinions of hotel customers [59]. Tourist experience analysis focuses on tourist feelings, emotions, and their satisfaction when participating in tourism activities. Zabkar et al. studied visitor satisfaction about Slovenia as a tourism destination [314]. Hasegawa studied tourist satisfaction and how they felt about the trip [110]. Chen and Chen studied the quality of the experiences, perceived value, satisfaction and behavioral intentions for heritage tourists [53]. Wu et al. studied impression of tourists when shopping at Beijing Silk market [304].

**Tourist Demand** The main task of this category is to identify and forecast of tourism volume in the form of inbound travel to a tourist destination, or outbound travel to other locations. Such knowledge is helpful for tourism managers in allocating resources to accommodate the need of tourists in terms of traffic, transportation, accommodation, and tourism packages. Some works focus on the overall tourism arrivals to analyze the demand of tourists [65,105]. Others directly investigate tourist demand on particular tourism products or services, such as inbound expenditure by different purpose of visit [45], or a study of the factors that influence tourism demand, such as seasonal variation [160], or the impact of the climate on tourism demand [96].
2.1.4 Data Mining Technique Categories

The analysis of customer behavior, as presented Section 2.1.3, are accomplished with the help of data mining techniques. Each task may be achieved by one or more techniques to discover hidden customer characteristics and behavior from databases. Depending on the analysis purpose, the data mining techniques are classified into the following categories:

**Association** Association analysis aims to reveal relationships between items, that exist together in a given record [5]. It is usually used in market basket analysis and cross selling programs, where association modeling is required [134]. In a tourism and hospitality context, the most frequently used algorithm is *Apriori* for association rule mining [166, 189, 287].

**Clustering** is the task of segmenting a heterogeneous population into a number of more homogeneous clusters [34, 95]. Clustering is an unsupervised process where the clusters are unknown at the starting time of the algorithms. The most popular tool for cluster analysis is *K-means* clustering and hierarchical cluster analysis [173, 183, 184, 242, 276].

**Predictive Modeling** Techniques in predictive modeling include two main types: classification and regression. *Classification* aims at constructing a model for predicting future customer behaviors through classifying records into classes based on certain features [57, 217]. Decision tree is the most common tool used for classification in tourism [73, 155]. *Regression* is a kind of statistical estimation technique that maps each data record to a real value [95]. Regression
is used for curve fitting, prediction, modeling of causal relationships, and testing a hypothesis about a variable relationship. *Linear regression* and *logistic regression* are the most commonly used techniques [94, 170].

**Forecasting** estimates the future based on patterns of records, that deal with continuously valued outcomes [5, 34]. It is relating to modeling and the logical relationship of the model at a future time. Demand forecast is a typical example of a forecasting application. The focus of forecasting in tourism and hospitality is on time series data analysis, with popular techniques including the *autoregressive moving average*, and *autoregressive conditional heteroskedasticity* [69, 267].

**Visualization** refers to the presentation of data so that complex patterns can be viewed by users. It is often used in conjunction with other data mining models to provide a deeper understanding of the relationships [283]. Tools for visualization are various depending on the analysis, with *multidimensional scaling* (MDS) [185, 247] and *principal component analysis* (PCA) [146] are some of the tools used in tourism.

**Computational Intelligence** is a set of computational approaches to complex real-world problems, which primarily includes *evolutionary/genetic computing, fuzzy computing* and *neural computing* [86]. Among these, *fuzzy computing* appears to be used most often. Some frequently used fuzzy computing techniques are *genetic fuzzy systems* [106], the *fuzzy analytic hierarchy process* [293] and the *fuzzy set* [266].
Statistical Learning refers to a large collection of statistical analysis techniques to form an analysis of variables and hypothesis testing. Techniques widely used in tourism and hospitality literature are analysis of variance (ANOVA) [78, 145, 152], multivariate analysis of variance (MANOVA) [107, 211], factor analysis [113, 271, 299, 301], structural equation modeling (SEM) [107, 138, 140], correlation analysis [50, 172], and some form of regression analysis [42, 94, 113].

As a matter of fact, different data mining techniques often need to be integrated to support, forecast or validate the effect of a business strategy. The categorization of data mining models can also be based on the major issues being considered. For instance, principal components analysis can be used as a statistical learning method as in [233] or a visualization method as in [300]. Logistic regression can be used as a classifier for classifying a tourist profile [81] or as a statistical analysis tool for analyzing the relationship between variables [234]. Similarly for multiple linear regression, it can be used as a predictor [170], or as a statistical learning technique [50].

Moreover, it is noticed that several data mining techniques are frequently used together to perform a specific customer behavior analysis task. For instances, in the case of tourist profile analysis, clustering can be applied to the segment customer, followed by a statistical learning model on each cluster to study the characteristics of tourists [12, 311]. Correspondence analysis is usually used together with multidimensional scaling analysis in analyzing tourist behavior [185]. Latent class analysis can be used with structural equation modeling for determining the length of tourist stay [9].
2.1.5 Overview of Data Mining Techniques in CBA

This section presents an overview about data mining (DM) techniques and their use in CBA. It should be noted that Statistical Learning methods such as ANOVA, MANOVA, SEM, Factor Analysis, SEM, and Correlation Analysis, etc. have been used widely and are well-known by tourism and hospitably researchers. For regression analysis, its techniques have been used widely for modeling and analyzing the relationship between the variables and hypothesis testing. The focus is on the works that use regression techniques for prediction purposes and classify them into the predictive modeling category. In addition, techniques such as choice modeling [144,212], and Importance-Performance analysis [132,169] are frequently used in analyzing customer preference. However, their origins are not from the data mining area, thus, they are not included in this section. Table 2.2 shows the works, that used data mining techniques for addressing CBA in Tourism and Hospitality.

Table 2.2: Distribution of articles according to the proposed categorization.

<table>
<thead>
<tr>
<th>CBA categories</th>
<th>DM categories</th>
<th>DM techniques</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tourist Profile</td>
<td>Association</td>
<td>Association rule mining</td>
<td>[166, 177, 254]</td>
</tr>
<tr>
<td></td>
<td>Clustering</td>
<td>K-means</td>
<td>[7, 24, 30, 82, 124, 146, 149, 163, 173, 183, 184, 190, 196, 204, 207, 211, 218, 220, 232, 239, 242, 307]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hierarchical clustering</td>
<td>[92, 124, 173, 183, 184, 190, 196, 204, 218, 220, 232, 307]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Biclustering</td>
<td>[83]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Expectation Maximization</td>
<td>[44]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Self-Organizing Map</td>
<td>[179, 253]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Latent class clustering</td>
<td>[10]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic regression</td>
<td>[73]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chi-square Automatic Interaction Detection</td>
<td>[73]</td>
</tr>
<tr>
<td>Tourist Behavior</td>
<td>Association</td>
<td>Association Rule Mining</td>
<td>[278, 287]</td>
</tr>
<tr>
<td></td>
<td>Clustering</td>
<td>Semi-Markov processes</td>
<td>[48]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K-means</td>
<td>[38, 185, 225, 248, 276, 290]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hierarchical clustering</td>
<td>[248]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Expectation Maximisation</td>
<td>[305]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic regression</td>
<td>[205]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multinomial logit model</td>
<td>[103]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Latent class analysis</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Classification And Regression Tree</td>
<td>[72]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fuzzy set</td>
<td>[62, 122, 302]</td>
</tr>
<tr>
<td>Field</td>
<td>Technique</td>
<td>References</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>Fuzzy analytic hierarchy process</td>
<td>[293]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fuzzy decision making trial and evaluation laboratory</td>
<td>[54]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forcasting</td>
<td>Survival analysis</td>
<td>[19, 97, 213, 240, 277]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spatial Origin-Destination travel flow model</td>
<td>[76]</td>
<td></td>
</tr>
<tr>
<td>Visualization</td>
<td>Network graphs</td>
<td>[117]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Principle component analysis</td>
<td>[300]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multidimensional scaling</td>
<td>[185]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graphical log-linear model</td>
<td>[3]</td>
<td></td>
</tr>
<tr>
<td>Tourist Preference</td>
<td>Association rule mining</td>
<td>[189]</td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>K-means</td>
<td>[114, 150]</td>
<td></td>
</tr>
<tr>
<td>Predictive Modeling</td>
<td>Logistic regression</td>
<td>[93, 104]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chi-square Automatic Interaction Detection</td>
<td>[151]</td>
<td></td>
</tr>
<tr>
<td>Computational Intelligence</td>
<td>Fuzzy set</td>
<td>[123, 266]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fuzzy membership function</td>
<td>[266]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fuzzy measure</td>
<td>[178]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technique for Order of Preference by Similarity to Ideal Solution</td>
<td>[123]</td>
<td></td>
</tr>
<tr>
<td>Visualization</td>
<td>Correspondence analysis</td>
<td>[159]</td>
<td></td>
</tr>
<tr>
<td>Tourist Perception</td>
<td>K-means</td>
<td>[8, 12, 119, 219, 240, 252, 311]</td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>Hierarchical clustering</td>
<td>[8, 247, 249, 280]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biclustering</td>
<td>[221]</td>
<td></td>
</tr>
<tr>
<td>Predictive Modeling</td>
<td>Logistic Regression</td>
<td>[137, 155]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>[4, 59]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ifthen-else rules</td>
<td>[4]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fuzzy rule based classification</td>
<td>[4]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chi-square Automatic Interaction Detection</td>
<td>[286]</td>
<td></td>
</tr>
<tr>
<td>Computational Intelligence</td>
<td>Fuzzy set</td>
<td>[187]</td>
<td></td>
</tr>
<tr>
<td>Visualization</td>
<td>Artificial neural network</td>
<td>[216]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multidimensional scaling</td>
<td>[13, 247]</td>
<td></td>
</tr>
<tr>
<td>Leximancer text analytics</td>
<td>[304]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tourist Demand</td>
<td>Forecasting</td>
<td>Autoregressive moving average</td>
<td>[45, 63–65, 69, 70, 98, 118, 156, 161, 248, 260, 267]</td>
</tr>
<tr>
<td></td>
<td>Autoregressive conditional heteroskedasticity</td>
<td>[91, 120, 128, 259]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vector error correction</td>
<td>[96, 128, 260]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Latent cycle component</td>
<td>[105]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explanatory variables model</td>
<td>[105]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exponential smoothing</td>
<td>[18, 45, 64, 70]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nave 2 model</td>
<td>[70]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discrete-time Markov chains</td>
<td>[20]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Meta-regression analysis</td>
<td>[238]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Basic Structural Model</td>
<td>[160]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gini index decomposition</td>
<td>[90]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Survival analysis</td>
<td>[159]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Piecewise linear models</td>
<td>[66]</td>
<td></td>
</tr>
<tr>
<td>Predictive Modeling</td>
<td>Linear regression</td>
<td>[17, 171]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Support vector regression</td>
<td>[55]</td>
<td></td>
</tr>
<tr>
<td>Computational intelligence</td>
<td>Genetic fuzzy systems</td>
<td>[106]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Genetic algorithm</td>
<td>[115]</td>
<td></td>
</tr>
</tbody>
</table>

Remark: Each article may have used more than one data mining techniques

Application of data mining techniques in CBA is an emerging topic, which has received increasing attention of practitioners and academics. The identified articles are related to the applications of data mining techniques in CBA (between 2007 and
mid 2014), and proposed a classification framework to summarize current progress in the field. Many attempts have been made in different aspects of CBA but mostly on the tourist profile, tourist behavior and tourism demand. Tourist preference is an interesting aspect, because preference is an important factor that governs the customer's behavior and intention, however it has the least number of publications. Among the journals, the scope of *Tourism Management* appears to fit well with our research focus as it has the most papers on the topic of data mining applications in customer behavior analysis. Several CBA issues are closely investigated in the tourism context, to identify the challenges and research gaps in Section 2.2.

### 2.2 Current CBA Issues in Tourism Management

The previous section has provided an overview about the applications of Data Mining techniques in a tourism and hospitality context. This section is devoted to investigating into several CBA issues to identify the current challenges presenting to tourism managers as well as the limitations of existing techniques. Three important issues in tourism management are discussed, including *Accommodation Preference*, *Travel Behavior* and *Online Behavior Analysis*. Since our case studies are on a major tourist destination such as *Hong Kong*, some background are provided on the current situation of the Hong Kong tourism industry.
2.2.1 Accommodation Preference

2.2.1.1 Issue in Accommodation Preference Analysis

The choice of accommodation is a high priority for most overseas tourists. It is also an example of a complicated decision-making process [266]. Hotel managers have long sought to identify the factors influencing the selection of tourists [191, 206]. Many studies have been conducted to study the selection criteria that affect the choice intentions of customers. For instance, Lockyer identified factors such as location, price, facilities and cleanliness as having a strong influence on tourists’ hotel selections [192]. Other criteria of interest are location, size of guest rooms, staff, facilities, and breakfast [270]. Merlo and de Souza Joao identified three attributes of the low-priced hotel segment that are more valuable in terms of improving consumer satisfaction: cleanliness, silence, and air conditioning [214]. Ariffin and Maghzi demonstrated that expectations of hotels are influenced by personal factors such as gender, purpose of stay, nationality, culture, and the private domain of hospitality [14].

Attention has been paid to identifying the relative importance of each factor in determining tourists’ overall satisfaction levels. Once these criteria have been identified and evaluated, managers can develop their practices and focus on what is important for customers so that service quality as well as customer satisfaction can be improved [131]. For instance, Choi and Chu showed that service quality, room quality, and value were the most influential factors in determining tourists’ overall impression of the selected hotels [61]. Aiming to identify the factors most likely to influence vacationers’ perceptions, Shergill and Sun examined overall facilities,
room facilities, and services in New Zealand hotels [261]. More recently, Tsai et al. compared the importance of ratings assigned to various hotel selection criteria by Mainland Chinese and foreign tourists to Hong Kong [282].

Despite the considerable research efforts in this area, hotel managers who want to understand tourists’ behavior and decision making processes for informed effective planning still face the major barriers of modeling tourists’ multi criteria decision making (MCDM) process. For instance, one tourist may give a hotel a low rating for the room quality and service criteria but still selects it on the basis that it is clean. Other tourists may select a hotel only if it satisfies both the room quality and service criteria. An understanding of how such criteria interact in guiding customers’ decision making intentions can provide managers an insight into their preferences. Modeling MCDM requires the simultaneous consideration of multiple criteria, whereas most studies to date have focused mainly on evaluating these independently using various statistical techniques. Therefore, there is still a strong demand for a technique that will enable the exploration of tourists’ MCDM process.

2.2.1.2 MCDM Modeling Techniques

The field of tourism research has witnessed a number of attempts to apply fuzzy techniques to the modeling of the MCDM process of tourists. In an early study, a fuzzy expert system using fuzzy logic was developed to assist tourists in making their hotel selection [222]. Later, the fuzzy analytic hierarchy process and triangular fuzzy numbers were used to consolidate decision makers’ assessments of criteria weighting in modeling hotels location selection of international tourists in Taiwan [62]. Similarly, a fuzzy method called DEMATEL was applied to linguistic information for group decision
making and a cause and effect model was developed for expectations of service quality at hot springs [52]. Recently, fuzzy set theory was combined with the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) to identify the factors influencing tourists’ choice of destination and to evaluate their preferences [123]. Using a fuzzy logic approach, a prototype system called the Tourism Advisory System was proposed to assist tourists in planning their travel [224]. A fuzzy membership function based on the intrinsic vagueness of the selection process was proposed for analyzing hotel selection factors [266]. The models of the fuzzy membership function can be seen as logical models that use “if-then” rules to establish qualitative and quantitative relationships among variables. Due to the nature of rule-based models, the use of information can be expressed in natural language, thus making the model transparent for interpretation.

In general, techniques such as fuzzy logic, fuzzy rule, and fuzzy number have been incorporated into the analysis of tourists’ preferences. These techniques allow the assessment of preference through the weighting and aggregation of criteria, which helps to identify the important criteria in the decision making process of tourists. Nevertheless, these techniques have a limitation that derives mainly from their natural assumption that the input criteria are independent of each other. As shown in [62, 123], a weight is assigned for each input criterion in the modeling process and each criterion is interpreted independently. Such an assumption is not always true in reality, where the independence of criteria cannot be assumed, and some interactions among different criteria, including the complement and the correlation [102]. For instance, two correlated criteria, such as room quality and cleanliness, may refer to the same concept. It is impossible to discover such interactions through an
interpretation of their importance via the weight assigned to each criterion. In order to take all interactions among attributes into account, the use of a fuzzy measure in the calculation of the aggregation function for MCMD modeling has been proposed [102]. Choquet Integral is one such technique that enables the importance not only of each criterion but also of each group of criteria to be considered [30]. In addition, its properties, the Shapley value and the Interaction index, offer good potential in terms of providing representations of the overall importance of each criterion and the interaction between the criteria.

2.2.2 Travel Behavior

2.2.2.1 Issues in Travel Behavior Analysis

Tourism managers have a high demand for insight into tourist travel behavior for destination management, product development and attraction marketing [186]. For instances, in transportation planning, providers can meet the needs of tourists and better coordinate their travel with local transportation flows, if they know about tourist preferences, daily itineraries and factors that influence those itineraries. Movement information can also be used to identify bottlenecks and unnecessary barriers in the flow between accommodation locations and places of attractions, or any other destinations [243]. In tourism location development, knowledge about tourists’ location preferences can be used to better define existing attractions, plan new ones and market them more effectively [176]. The identification of actual routes taken by tourists during their trips can help define the boundaries of districts and nodes, as well as their most appropriate gateways. This information can be used to develop new attractions and products along common routes and in the districts and destination
nodes [49]. In impact management, it is important for tourism managers to identify the time and space of the routes and the destinations which tourists visit most frequently, so that appropriate plans can be developed to prevent the overload of capacity, which has the potential to cause negative social, environmental, or cultural impacts [176].

Tourism researchers and business managers still find it difficult to fully capture and understand travel behavior of international tourists. The first barrier is capturing tourist travel information. Survey and opinion poll are popular methods to collect travel data from tourists [16,208]. These approaches are usually time consuming, and limited in the number of responses as well as the scale of captured information. The collected data is unable to reflect the travel patterns closely to real situations [319]. There is high demand for an efficient method to capture the comprehensive travel behavior of tourists. The second barrier is the lack of understanding about travel preferences. It is a natural assumption that travel behavior varies between different groups of tourists [21]. For instance, tourists coming from different countries may have different preferences on their length of stay and attractions to visit, whereas the length of stay may affect the travel activities of tourists [175]. There have been limited attempts to analyze their travel behaviors, which take preferences between different groups of tourists into consideration.

2.2.2.2 Travel Behavior Analysis Techniques

In the work on analysis of travel behavior, , as identified in Section 2.1, most focus on predicting the length of stay of tourists, such as logistic regression [205], multinomial logit model [103], latent class analysis [9], and survival analysis [19,97,213,240,277].
Others study the travel behavior of tourists based on their profiles and by using techniques such as k-means and hierarchical clustering [248], network graphs [117], and principle component analysis [300]. Only a few attempts has been made in studying the actual travel patterns or movement of tourists at a destination. For example, Xia et al. introduced a clustering method using expectation maximisation for tourist market segmentation based on the dominant movement patterns of tourists [305]. Deng and Athanasopoulos utilized the anisotropic dynamic spatial lag panel Origin-Destination travel flow model to understand Australian domestic and international inbound travel patterns [76]. Cecillia et al. used semi-Markov processes to model the spatial and temporal movements of tourists [48]. Versichele et al. mined tourist attraction visits through association rule mining on Bluetooth tracking data [287].

Tourism researchers and business managers still find it difficult to fully understand the travel behavior of international tourists. This is due to the inefficiency in capturing tourist travel information. The collected data is unable to reflect the travel patterns close to a real situation [319]. Even with the support of GPS technology, participants are usually required to carry mobile devices for recording their movements while traveling. The data collected is therefore limited in terms of the number of responses or the scale of the geographical areas included. There is a high demand for an efficient method to capture the comprehensive travel behavior of tourists.

When the data collected is not reliable or adequate, it is necessary to resort to online data. The development of computer science research on geo-tagged photos presents a new direction for capturing tourist movement data in an efficient and timely manner. With millions of geo-tagged and time-stamped photos available on a global scale, photo-sharing sites are potential gold mines for researchers seeking to
extract data and study travel behavior. Some examples of potential questions asked by managers are “What are the preferred locations for each group of tourists when visiting a tourist destination?”, “When do they prefer to visit such locations?” and “What travel routes are they likely to take when visiting different locations?”. Unfortunately, none of the existing works in the tourism area attempts to discover such insightful knowledge. As such, it is still a challenging task for tourism managers to develop effective destination management and transportation plans to accommodate the increasing demand of international tourists.

While the massive volume of shared photos available online is a comprehensive resource for studying travel behaviors, the data can be noisy or misleading [320], especially when many photos have been taken in transit rather than at the attractions themselves. In addition, when tourists move from one attraction to another, the sequence of photos implies mobility information, which is key to inferring personal activities, intuition, and goals. However, photos are static media. They should be transformed into a suitable representation for the travel analysis task. As such, there is a demand for new data mining techniques to effectively analyze the geo-tagged photos for discovering travel behavior.

2.2.3 Hotel Feature Preference Analysis

2.2.3.1 Issues in Hotel Feature Preference Analysis

Tourists have different expectations and/or preferences, depending on their destination, purpose and mode of travel, as well as previous accommodation experiences [188,189]. A comprehensive understanding of customer requirements can help hotel managers gain a lead in the market in terms of strategic planning, marketing,
and product development [298].

Among these criteria, the most important has to do with hotel features (i.e., attributes or factors) that most tourists seriously consider. The most valuable hotel features that significantly affect a tourists’ selections include location, price, facilities, and cleanliness [192]. Other features, such as the size and type of building, quality of service and a quiet environment, are also important to some people [6, 214]. Merlo and de Souza Joao examine specific hotel features, such as air conditioning in bedrooms [214]. Sohrabi et al. present another list of important hotel features, including promenade, comfort, security, network, pleasure, news, recreational information, expenditure, room facilities, and car parking [266].

Advances in Internet technology enable tourists to share their travel-related experiences, opinions, and concerns on many online platforms [201]. Thus, researchers are now shifting their attention to this data source as a way of mining tourist preference in a cheap, efficient, and non-intrusive manner. For instance, Stringam and Gerdes use a corpus-based approach to analyze guest comments on online hotel distribution sites as well as to identify frequently used words, patterns of word usage, and their relationship to the hotel features rating [270]. Furthermore, descriptive statistical data are used to assess the importance and effect on ratings of several features, including location, size of guest rooms, staff, facilities, and breakfast offerings [269]. Chaves et al. show that room, staff, location, cleanliness, friendliness, and helpfulness are the most frequently used words in online reviews of small and medium hotels in Portugal [51]. Others analyze comments collected from TripAdvisor.com and changes in hotel customers’ expectations according to travel mode by
using the association rule mining technique [188, 189].

Even though researchers have not yet been able to completely satisfy the increasing demand of hotel managers for knowledge about tourists’ hotel preferences, it is a natural assumption that these preferences are not stable, and subjected to change. These changes in the concerns of tourists can have a huge influence on the performance of hotel businesses. It is necessary for managers to be able to quickly and effectively identify features that are becoming important to tourists. Limited effort has been made to address this issue.

2.2.3.2 Hotel Feature Preference Analysis Techniques

A popular approach in hotel feature analysis is using survey, wherein hotel features are represented by short-answer questions or a set of keywords. Due to the increasing interaction among tourists, studies are increasingly utilizing observation data collected from online resources (e.g., blogs, travel websites, and social media) through online reviews. Traditional statistics-based data analysis models are ineffective in extracting information from text-based reviews and comments. Thus, new techniques and approaches have been proposed. For instance, manual content analysis is used to study tourist characteristics and communications about Australia as a tourism destination [47, 296]. Another study employed the narrative structure analysis to identify key marketing elements, including characterization, space categorization, and evaluation of the product experience [285]. Manual methods are time consuming and incapable of obtaining the overall differences in tourist preferences. Although automated approaches, such as corpus-based semantic analysis [244] and stance-shift analysis [75], are also employed, such methods require users to have a background in
linguistics and access to expensive software [46].

The identification of emerging features is different from traditional approaches to hotel feature analysis because analysts have no prior knowledge on what features should be included in the study. Large data samples are also required to identify emerging changes in customer preference patterns. Traditional research methods, such as surveys, opinion polls or focus groups, are inadequate. Therefore, resorting to available online data, such as online reviews generally expressed as textual comments, is necessary. These reviews contain abundant information on user opinions, experiences, or concerns, and are considered potential gold mines from which tourism researchers can gain insights into the behavior of tourists [231]. Unfortunately, none of the existing techniques, as shown in Table 2.2, has the capability to detect the emerging preference of tourists.

In hotel preference analysis, each feature of hotel can be treated as an item, and a set of hotel features associated with a tourist is an item set. Identifying emerging changes in a tourist response to such features can be typically formulated as a problem of Emerging Pattern Mining (EPM). Originally proposed by Dong and Li [84], EPM can capture emerging trends in time-stamped databases or sharp contrasts between data sets or groups. EPM is mainly applied in bio-informatics. For instance, Li and Wong attempt to find groups of genes using EPM and apply these on a colon tumor dataset [181]. Li et al. develop an interpretable classifier on an acute lymphoblastic leukemia micro array dataset [180]. Wang et al. adopted the EPM technique to mine local conserved clusters from gene expression data [291]. Sherhod et al. utilized the jumping EPM to develop a method for an automatic toxicity alert [262], while Huang et al. used this technique in mining the changes of medical behavior for clinical path
ways of bronchial lung cancer [126]. It is promising to adopt EPM concepts into emerging hotel preference identification.

2.2.4 Hong Kong Tourism Industry

Our case studies focus on the typical tourist destination of Hong Kong, which is a major Asian tourist destination. According to the annual report from the Hong Kong Tourism Board [2], in 2012, Hong Kong attracted more than 48.6 million visitors, 16% more than in 2011. Total tourism expenditure associated to inbound tourism reached HK$296 billions. Over the same period, the average hotel occupancy rate in Hong Kong increased significantly and reached 89%, with the average daily room rates for all hotels also experiencing remarkable growth (up to HK$1,489). Such expansion has brought many new opportunities and challenges for researchers seeking to gain an insight into customers’ behavior in order to support hotel managers in their business planning and decision making.

Being the center of East Asia, along with the development of Mainland China as its closest neighbor, Hong Kong Special Administrative Region has been an international business, trade and financial hub. Hong Kong has developed into a modern, exciting and multinational services economy, supporting the role of the city as a global business platform. According to the Hong Kong Census and Statistics Department [1], the number of inbound tourists exceeds 29.59 millions in 2009, and brought Hong Kong $162.8 billion in tourism expenditure. The steady growing trend of Hong Kong outbound travel has attracted attention from local as well as international companies in travel and tourism marketing. Here, the term of inbound travel refers to any leisure trips from outside to Hong Kong, although since 1997, Hong Kong became a part of
China. For tourism and hospitality management, customer satisfaction is the key issue when designing any kind of tourist product or service. There has been a large amount of research in understanding the behavior of consumers on the aspect of intentions, preferences, decision making, satisfaction and willingness to revisit. Among them, attracting the attention of visitor was considered a crucial role in the success of a tourism destination in terms of motivation to revisit and resources for local communities [168]. Indeed, travel motivation has been an emerging issue in tourism planning and marketing for many years. It can be very advantageous to understand the reason people travel and what factors lead to their behavior of choosing Hong Kong as a travel destination.

In the Hong Kong tourism industry, although many marketing strategies have targeted the promotion of delicious Hong Kong food, shopping and transportation have been considered as significant and important factors in tourist satisfaction for many years [310]. Researchers have been trying to understand this special relationship between tourism and the retail industry in Hong Kong. Tsan and Asli pointed out that customers from Mainland China tend not to follow the suggestions of sales people when making their purchasing decisions. Instead, they look for brands that are new to them in Hong Kong or make them look good and feel positive about themselves [60, 274]. Rob [251] and Terry Lam [164] have both found that word of mouth is more important than advertising in attracting customers from Taiwan and Mainland China, which means that tourists with adequately good experiences are more likely to re-patronize and help increase the returning potential of tourists to Hong Kong. They suggested the Hong Kong government should pay more attention to the hospitality and tourism industry and proper training should be provided for
employees to enhance the quality of service.

2.3 Summary

Knowledge about customer behavior is the key to effective planning and decision making for successful business management. The tasks of *customer behavior analysis* requires advanced techniques, such as data mining. Such techniques can handle a massive amount of data so that a wider and deeper knowledge can be effectively achieved. This chapter has reviewed the literature about data mining applications in tourism and hospitality management and identified several potential research issues on customer behavior in the tourism domain.

MCDM modeling is an important approach to studying tourist preferences. Traditional techniques often fail to closely capture the real MCDM of customers due to their natural assumption that the input criteria are independent of each other. There is a demand for a better method to model the MCDM process close to the real situation. Besides, the travel behavior of tourists is important to decision makers in destination development, transportation planning, and impact management at a tourism destination. Geo-tagged photo data has appeared as a new approach to capture travel behavior in a comprehensive and efficient manner. However, the geo-tagged photo data is noisy and misleading, which requires novel techniques to reveal hidden travel behaviors. Furthermore, it is important for business managers to identify the emerging preference of tourists, for more effective planning and decision making. Traditional techniques are unable to fully capture the emerging pattern hidden in customer data. Alternative approaches are needed to improve their capability to address the new challenges of emerging preference analysis.
This study aims to develop novel data mining techniques for addressing the challenges of *customer behavior analysis* in the context of the tourism industry. The challenge in travel preference analysis is tackled by proposing a new MCDM modeling approach based on the fuzzy measure in Chapter 3. It has the capability of considering multiple criteria simultaneously and accounting for all interactions among different criteria, which outperforms existing MCDM modeling techniques. The problem of discovering travel preferences are addressed using a new data mining framework for geo-tagged photos in Chapters 4. This framework is able to effectively process the geo-tagged photo data to reveal preferred tourism attractions, and travel routes of tourists. New supporting tools are proposed to support the travel analysis task using geo-tagged photo data (Chapter 5), and facilitate the development of tourism applications (Chapter 6). A new approach based on emerging pattern mining is proposed to address the problem of emerging hotel preference identification in Chapter 7. The developed techniques contribute towards the advanced development of data mining technology, while the discovered knowledge extends the understanding of tourism managers and practitioners in regards to the behavior of tourists.
Chapter 3

Fuzzy Measure Based Preference Mining

Customer preferences have always been of interest to researchers who support the strategic planning and decision making of business managers. Over the past decade, fuzzy decision support has emerged as a means of providing effective tools and techniques for solving MCDM problems. For instance, Grabisch and Roubens used the aggregation function, which is a fuzzy decision-support technique, to support the MCDM process in a game theory context [102]. Lu and colleagues proposed a group MCDM method to evaluate non-woven cosmetic product prototypes [194]. Furthermore, a fuzzy system called Decider was implemented to increase overall satisfaction with a final decision across groups of respondents and it also dealt with the uncertainty in solving an MCDM problem [197]. Recently, more applications of fuzzy decision support for MCDM have been created [23,77].

Although a number of works have incorporated fuzzy decision support techniques such as fuzzy logic, fuzzy rules, and fuzzy number into the analysis of tourists’ preferences, they have a common limitation from their assumption that the input criteria are independent of each other. However in real situations, interactions may exist among
different criteria, including independence, complement, and correlation [101]. For instance, two criteria which have a correlating relationship, such as room quality and cleanliness, may refer to the same concept. It is not possible to discover such interactions through an interpretation of their importance via the weight assigned to each criterion. In order to account for the interactions among criteria, fuzzy measure theory has been proposed to aggregate interacting criteria [102]. Fuzzy measure is usually used in the computation of aggregation function for modeling the MCDM process [30]. Choquet Integral is one such technique that enables the importance not only of each criterion but also of each group of criteria to be considered. This chapter investigates the use of fuzzy measure to address the challenges in the MCDM of tourists, and introduces a new method for travel preference analysis.

This chapter starts with an introduction to aggregation function and fuzzy measure (Section 3.1). It then presents the approach of mining customer preferences based on fuzzy measure, and a R software package, named Rfmtool, which has been specifically designed for customer preference analysis (Section 3.2). The proposed approach is demonstrated in an application of modeling the hotel selection preference of tourists (Section 3.3). Finally, the conclusion and contribution are presented (Section 3.4).

### 3.1 Background

This section provides a formal definition of the aggregation function for MCDM modeling, followed by fuzzy measure concepts. An aggregation function named Choquet Integral, which uses fuzzy measure in its computations, is then introduced.
3.1.1 Aggregation Functions

Aggregation is the process of combining multiple inputs into a single output which in some sense represents all the inputs. The numerical function performing this process is called the aggregation function, which is defined as a function of \( n > 1 \) arguments \( f : [0, 1]^n \rightarrow [0, 1] \), with the following properties:

\[
 f(x_1, x_2, \ldots, x_n) = \begin{cases} 0, & \text{if } x_i = 0, \forall i \\ 1, & \text{if } x_i = 1, \forall i \end{cases} \quad (3.1.1)
\]

\[
 x \leq y \text{ implies } f(\vec{x}) \leq f(\vec{y}), \forall \vec{x}, \vec{y} \in [0, 1]^n \quad (3.1.2)
\]

where \( \vec{x} = \{x_1, x_2, \ldots, x_n\} \) are the inputs. The first constraint says that if all the inputs are 0 (1) then the aggregated output has to be 0 (1) as well. The second constraint says that if every value in \( \vec{x} \) is no greater than the corresponding value in \( \vec{y} \), then the aggregated output of \( \vec{x} \) must also be no greater than that of \( \vec{y} \).

When considering numerical values, an aggregation function \( f \) takes \( n > 1 \) input values, usually interpreted as degree of membership, level of preferences, strength of evidence etc., and then combines them into a single output value. In addition to the above two constraints that all aggregation functions must satisfy, other constraints exist that further divide aggregation functions into four main classes: averaging, conjunctive, disjunctive, and mixed. Among them, averaging aggregation is one widely used class of functions, which is bounded by its value by \( \min(x) \leq f(x) \leq \max(x) \). The term “average” is commonly employed in everyday language when referring to the arithmetic mean (AM). In case some criteria are considered as more important than others, it is common to consider the aggregation function to be additive and to
take the form of a weighted arithmetic mean (WAM) [279], or an ordered weighted averaging (OWA) [308].

AM is the function of $n$ values:

$$AM(x) = \frac{1}{n} (x_1 + x_2 + \cdots + x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \quad (3.1.3)$$

WAM is a linear function with respect to a positive valued weighting vector $w$ with $\sum_{i=1}^{n} w_i = 1$:

$$WAM_w(x) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \sum_{i=1}^{n} w_i x_i \quad (3.1.4)$$

or a given weighting vector $w$ with $\sum_{i=1}^{n} w_i = 1$, $w_i \geq 0$, OWA is defined by

$$OWA_w(x) = \sum_{i=1}^{n} w_i x(i) \quad (3.1.5)$$

Although WAM has been widely used, it can only be used in the presence of independent criteria, which is not appropriate for the aggregation of interacting criteria in the MCDM process [202]. The OWA can, to some extent, account for the interaction among criteria, but it only considers the relative weighting of individuals, whereas the evaluation of the MCDM process is usually performed under conditions in which there are relative influences within groups of criteria. Hence, it is necessary to consider alternatives, such as the fuzzy measure as introduced below.

### 3.1.2 Fuzzy Measure

Fuzzy measure theory, derived from classical measure theory, is capable of modeling informational uncertainty [295]. Fuzzy integrals, such as the Choquet integral [101], utilizes fuzzy measures to aggregate multiple input values into a single representative
output value. *Fuzzy measures* and *fuzzy integrals* have been applied in various applications from decision-making to classification [101]. Though two concepts are highly correlated, a *fuzzy integral* is indeed independent from the *fuzzy measure* used. *Fuzzy measure* models uncertainty and measures the criteria in a better way. *Fuzzy integral* is an operation applied on a constructed *fuzzy measure*. Different *fuzzy integrals* and *fuzzy measures* possess their own special properties. The remainder of this section is devoted to the concepts of *fuzzy measures*, follows by an illustration of how *fuzzy integrals* can be used to perform aggregation tasks by using *fuzzy measures*.

**Definition 1 (Fuzzy Measure).** Let \( X = \{x_1, x_2, ..., x_n\} \) be the criteria. A *fuzzy measure* is a set function \( v : 2^X \rightarrow [0, 1] \) on all possible combinations of \( n \) criteria, which satisfies: (i) \( v(\emptyset) = 0 \) and \( v(X) = 1 \). and (ii) Monotonic, \( v(A) \leq v(B) \) whenever \( A \subseteq B \subseteq X \).

The *fuzzy measure* can also be represented as a Hasse diagram [30], for example, in the case \( X = \{x_1, x_2, x_3\} \):

\[
\begin{array}{cccc}
v(\{x_1,x_2,x_3\}) & v(\{x_1,x_2\}) & v(\{x_1,x_3\}) & v(\{x_2,x_3\}) \\
v(\{x_1\}) & v(\{x_2\}) & v(\{x_3\}) & v(\emptyset) \\
\end{array}
\]

There exist different types of *fuzzy measures* including but not limited to *Additive Fuzzy Measure*, *Sugeno Fuzzy Measure*, and *Decomposable Fuzzy Measure* [30].

**Additive Fuzzy Measure:** A *fuzzy measure* \( v \) is called *additive* if for any \( A, B \subseteq X \), \( A \cap B = \emptyset \) satisfies

\[
v(A \cup B) = v(A) + v(B)
\]
Additive fuzzy measure is the same as probability measures used in probability theory.

**Sugeno Fuzzy Measure:** Given a parameter $\lambda \in [-1, \infty)$, a Sugeno fuzzy measure $v$ is that for all $A, B \subseteq X$, $A \cap B = \emptyset$ satisfies

$$v(A \cup B) = v(A) + v(B) + \lambda v(A)v(B)$$

**Decomposable Fuzzy Measure:** A decomposable fuzzy measure $v$ is a fuzzy measure which for all $A, B \subseteq X$, $A \cap B = \emptyset$ satisfies

$$v(A \cup B) = f(v(A), v(B))$$

These fuzzy measures possess their own special properties and can be applied depending on the application [30]. However, all fuzzy measures possess the monotonicity property

$$v(A \cup \{i\}) - v(A) \geq 0, \forall A|i \notin A, i = 1, \ldots, n$$

Fuzzy measures can also be represented alternatively by applying some invertible linear transformations [30]. These alternative representations are also capable of measuring the interactions and importance of criteria, however, computations on these representations can be easier.

Here the Mobius representation is described using an example. Let $v$ be a fuzzy measure. The Mobius transformation of $v$ is a set function defined as:

$$M(A) = \sum_{B \subseteq A} (-1)^{|A\setminus B|} v(B), \forall A \subseteq X \tag{3.1.6}$$

Mobius transformation is invertible. The original fuzzy measure can be obtained using the Zeta transformation [30].

$$v(A) = \sum_{B \subseteq A} M(B), \forall A \subseteq X \tag{3.1.7}$$
To illustrate with an example, let the following Hasse diagram represents a general fuzzy measure.

\[
\begin{array}{ccc}
1 & 0.5 & 0.6 \\
0.5 & 0.4 & 0.1 \\
0 & & \\
\end{array}
\]

After applying Mobius transformation the fuzzy measure becomes:

\[
\begin{array}{ccc}
0 & & \\
-0.4 & 0 & 0.4 \\
0.5 & 0.4 & 0.1 \\
0 & & \\
\end{array}
\]

Each entry of general representation is the sum of all subset entries of Mobius representation. For example, \( v(\{x_1, x_2\}) = M(\{x_1, x_2\}) + M(\{x_1\}) + M(\{x_2\}) \). Therefore the entry in general representation measures not only the set itself but also its subsets. However, in Mobius representation, each entry is a measurement of the set itself and the sum of all entries are equal to 1.

Mobius transformation is helpful in expressing various quantities, like the interaction, in a more compact form. It also serves as an alternative representation of a fuzzy measure, and makes computation easier in some cases.
3.1.3 Choquet Integral

The fuzzy measure is usually adopted in the computation of an aggregation function for modeling MCDM. The Choquet Integral is one such function, which can be represented by:

\[ C_v(x) = \sum_{i=1}^{n} x_i \ast [v(\{j \mid x_j \geq x_i\}) - v(\{j \mid x_j \geq x_{(i+1)}\})] \] (3.1.8)

where \(x_1, x_2, \ldots, x_n\) is a non-decreasing permutation of the input \(x\), and \(x_{(n+1)} = \infty\) by convention. Both the inputs and the outputs are usually defined on the unit interval \([0, 1]\), however, other choices are also possible [30].

Alternatively, Equ. 3.1.8 can also be written, by rearranging the terms of the sum, as:

\[ C_v(x) = \sum_{i=1}^{n} [x_i - x_{i-1}] \ast v(H_i). \] (3.1.9)

Let \(M\) be the Mobius representation of the fuzzy measure, with Choquet Integral computed as:

\[ C_{M_v}(x) = \sum_{A \subseteq N} M(A) \ast \min_{i \in A} x_i \] (3.1.10)

Example 1. Given an input \(x = (0.6, 0.3, 0.8)\) and the fuzzy measure values, the result obtained from the inputs \(x\) using a general fuzzy measure following Equ. 3.1.9 is:

\[ C_v(x) = 0.3 \ast v(\{1, 2, 3\}) + [0.6 - 0.3] \ast v(\{1, 3\}) + [0.8 - 0.6] \ast v(\{3\}) \]
\[ = 0.3 \ast (1) + 0.3 \ast (0.8) + 0.2 \ast (0.4) \]
\[ = 0.62 \]
The result obtained from the inputs $x$ using the *Mobius* fuzzy measure, following Equ. 3.1.10 is:

$$C_{M_v}(x) = M(1) \ast 0.6 + M(2) \ast 0.3 + M(3) \ast 0.8$$

$$+ M(12) \ast \min(0.6, 0.3) + M(13) \ast \min(0.6, 0.8)$$

$$+ M(23) \ast \min(0.3, 0.8) + M(123) \ast \min(0.6, 0.3, 0.8)$$

$$=(0.3) \ast 0.6 + (0.5) \ast 0.3 + (0.4) \ast 0.8 + (-0.2) \ast 0.3$$

$$+ (0.1) \ast 0.6 + (-0.2) \ast 0.3 + (0.1) \ast 0.3$$

$$=0.62$$

There is an issue with the fuzzy measure that the complexity of the model grows exponentially ($2^n$) with the input criteria $n$. Grabisch developed the concept of a $k$-additive fuzzy measure to deal with this problem by reducing the number of variables to define the fuzzy measure [100]. The interactions between the criteria are only considered for the subsets of $k$ elements or less, which allows for a trade-off between modeling ability and complexity. Users can decide how complex a fuzzy measure is to be considered by choosing a $k$-additive value ($1 \leq k \leq n$). It should be noted that the traditional aggregation function like the *Weighted Average Mean* (WAM) is equivalent to the *Choquet Integral* with 1-additive fuzzy measure. When $k = n$, the fuzzy measure is said to be *unrestricted*. Thanks to the use of *fuzzy measure* in its computation, the *Choquet Integral* is able to account for all possible interactions between criteria to model MCDM process closely to the real situation. The next section presents our preference mining approach based on the fuzzy measure and a software package specifically designed for customer preference analysis.
3.2 Methodology

This section firstly presents our approach for mining preferences using the fuzzy measure, followed by a description of the \texttt{Rfmtool}, which is a \texttt{R} software package for preference analysis. The use of this technique and the package is then illustrated in a sample case study.

3.2.1 Fuzzy Measure Preference Mining

\textit{Choquet Integral} has been widely used as an aggregation function for modeling the multi-criteria decision making process. However, there is little understanding on how to mine useful information from the fuzzy measures for analyzing customer preferences and their behavior. This study explores the use of the fuzzy measure from the perspective of preference mining by utilizing its two properties: \textit{Shapley} value and \textit{Interaction Index}.

The \textit{Shapley} value measures the overall importance of each criteria in terms of its contribution to the score of every coalition. Let $v$ be a fuzzy measure. The \textit{Shapley} value for every input $i \in N$ is

$$
\phi_i = \sum_{A \subseteq N \setminus \{i\}} \frac{(n - |A| - 1)! |A|!}{n!} [v(A \cup \{i\}) - v(A)]
$$

The \textit{Shapley} value is the vector $\Phi(v) = (\phi_1, \ldots, \phi_n)$. The index $\phi_i$ can be interpreted as an average of the contributions of criteria $i$ in all groups.

The \textit{Interaction Indexes} reflect the behaviors of criteria in groups, and they can be
considered as a measurement of the interaction between the criteria in the decision-making process. For every set \( A \subseteq N \):

\[
I(A) = \sum_{B \subseteq N \backslash A} \frac{(n - |B| - |A|)! |B|!}{(n - |A| + 1)!} \sum_{C \subseteq A} (-1)^{|A|} v(B \cup C)
\]

(3.2.2)

It should be noted that \( I(A) \in [-1, 1] \) and the computation of the Interaction Index can include all combination of criteria. However, the Interaction Index \( I_{ij} \) for each pair \( A = \{i, j\} \) of criteria is used most often, due to its convenience in interpretation. If there exists a correlation between a pair of criteria \( x_i \) and \( x_j \), then we have \( I_{ij} < 0 \). If they have a complementary relationship, we have \( I_{ij} > 0 \). When \( x_i \) and \( x_j \) are independent or have little interaction, we have \( I_{ij} \approx 0 \).

Here the key problem is the identification of the fuzzy measure by fitting the Choquet Integral to data. Given a data set containing \( M \) records in pairs \((x_1, y_1), \ldots, (x_M, y_M)\), where \( x_i \) is composed of \( n \) criteria \( x_{k(1)}, x_{k(2)}, \ldots, x_{k(n)} \), and \( y_k \) is the observed aggregated value. A Mobius fuzzy measure \( M_v \) can be estimated from this data set to minimize the least absolute deviation (LAD):

\[
\text{minimize} \sum_{k=1}^{K} |C_{M_v}(x_k) - y_k|
\]

(3.2.3)

According to [28], the LAD problem can be converted into a Linear Programming (LP) problem by using the auxiliary variables \( r_k^+, r_k^- \geq 0 : r_k^+ - r_k^- = C_{M_v}(x_k) - y_k \), in which case \( r_k^+ + r_k^- = |C_{M_v}(x_k) - y_k| \). If we denote \( C_{M_v}(x_k) = f(x_k; w) \), where \( w \) is the Mobius fuzzy measure, to be estimated from the data, then the complete LP
problem for estimating \( w \) satisfying constraints on the Mobius fuzzy measure is:

\[
\begin{align*}
\text{minimize} & \quad \sum_{k=1}^{K} r^+_k + r^-_k \\
\text{subject to} & \quad r^+_k - r^-_k - f(x_k; w) = -y_k, \ k = 1, \ldots, K, \\
& \quad \sum_{i \in B \subseteq A} w_B \geq 0, \ \forall A \subseteq N, \ \forall i \in A, \\
& \quad \sum_{A \subseteq N} w_A = 1, \\
& \quad r^+_k, r^-_k \geq 0, \\
& \quad w_{\{i\}} \geq 0, \ i = 1, \ldots, n, \text{and other } w_A \text{ unrestricted.}
\end{align*}
\]

(3.2.4)

Once estimated, the Mobius fuzzy measure can be used to compute the Shapley value and Interaction Index for preference analysis.

3.2.2 Rfmtool

This section provides a detailed description of our R software package, the Rfmtool. This tool was developed to perform operations on the Choquet Integral and fuzzy measures for preference mining applications. The Rfmtool package is distributed as a standard R package containing source code files, a data sample, and examples. The package can be downloaded from http://www.tulip.org.au/resources/rfmtool. There are two distribution files: the file Rfmtool.zip is for installation and running on the Windows platform, and the file Rfmtool.tar.gz is for the Linux platform.

It should be noted that the routine code for the fuzzy measure operation is written in C/C++ by Beliakov [26]. The source files of the Rfmtool also include the source code of the Lp.solve library [33] because the operation of the Rfmtool depends on this library for solving the related linear programming problem. A “wrapper” function
is designed that allows the operations and data input/output to be performed in the $R$ environment.

The following core functions are provided in the `Rfmtool` package:

**`fm.fitting(data, kadd)`** This estimates the fuzzy measure based on the empirical data following Equ. 3.2.4. The first input argument (`data`) is the empirical data set in pairs $(x_i, y_i)$, where $x_i \in [0, 1]^n$ is a vector containing the utility values of $n$ input criteria $\{x_{i1}, x_{i2}, \ldots, x_{in}\}$, $y_i \in [0, 1]$ is a single aggregated value given by decision makers. The second argument (`kadd`) is the value of $k-\text{additivity}$ for reducing the complexity of fuzzy measures [100]. `kadd` is defined as an optional argument, and its default value is $kadd = n$. The estimated fuzzy measure $v$ is in the Mobius representation and stored in an array of size $2^n$ following binary ordering, to make it convenient for computational purposes.

**`fm.Choquet(x, v)`** This calculates the Choquet Integral based on the general fuzzy measure, following Equ. 3.1.9. The argument `x` is a vector containing the input criteria, and `v` is a vector of the general fuzzy measure.

**`fm.ChoquetMob(x, Mob)`** This calculates the Choquet Integral based on the Mobius fuzzy measure, following Equ. 3.1.10. The argument `x` is a vector containing the input criteria, and `Mob` is a vector of the Mobius fuzzy measure.

**`fm.Mobius(v)`** This transforms the general fuzzy measure `v` into the Mobius representation, following Equ. 3.1.6.

**`fm.Zeta(Mob)`** This transforms the Mobius fuzzy measure `Mob` to the general representation, following Equ. 3.1.7.
**fm.Shapley(v)** This calculates the *Shapley* value from the general fuzzy measure *v*, following Equ. 3.2.1.

**fm.Interaction(Mob)** This calculates the *Interaction Index* from the *Mobius* fuzzy measure *Mob*, following Equ. 3.2.2.

During the operation, users can view a list of all functions included in this package for reference at any time by typing:

```
> fm()
```

Users can view the manual by using the help functions.

```
> help(Rfmtool)
```

### 3.2.3 An Illustrative Example

We explore the operations of the *Rfmtool* package via a simple example. Suppose a company wants to know which preference of customers to support during their product development task. A data set was collected with customer ratings on three product selection criteria (*price*, *quality*, and *service*). The rating takes a value between the unit interval [0, 1]. For instances, the customer give a value of 1 for the *price* criterion if he/she think that the product has a good *price*. A value of 0 is given for the *quality* criterion if the customer think that the product is of bad *quality*. In addition, an overall rating which represents the final decision of the customer. A high value such as 1 indicates that the customer prefers or likes the product, and a low value such as 0 indicates otherwise. The sample data set is shown in Table 3.4.
Table 3.1: Structure of hotel rating data collections

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Price</th>
<th>Quality</th>
<th>Service</th>
<th>Overall Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>0.6</td>
<td>0.9</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>R2</td>
<td>0.8</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>R3</td>
<td>0.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>R4</td>
<td>0.7</td>
<td>1.0</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>R5</td>
<td>0.3</td>
<td>0.6</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>R6</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>R7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>R8</td>
<td>0.3</td>
<td>0.8</td>
<td>0.9</td>
<td>0.6</td>
</tr>
<tr>
<td>R9</td>
<td>0.6</td>
<td>0.8</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>R10</td>
<td>0.3</td>
<td>0.6</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

The data set is saved in a file `data.txt` containing only the rating values. We can apply the `Rfmtool` to this data set as follows:

```r
> # Load the Rfmtool package
> library("Rfmtool")

> # Load data from files.
> data <- as.matrix(read.table("data.txt"));
> # Note that the data.txt file contain only the rating values.
> # The record ID and labels for product features are not included.

> # Estimate fuzzy measure (in Mobius representation) from data sets.
> # Here, the kadd value is not specified, thus kadd is assigned with
> # the default value as kadd = n = 3.
> Mobfuzzy <- fm.fitting(data)
[1] 0.00 0.00 0.50 0.25 0.50 0.00 -0.50 0.25
> # Note that estimated fuzzy measure is in Mobius representation,
> # and it is stored in an array containing 2^3 = 8 values.

> # Transform the estimated Mobius fuzzy measure into general
> # fuzzy measure by calling Zeta transform function.
> Genfuzzy <- fm.Zeta(Mobfuzzy)
[1] 0.00 0.00 0.50 0.75 0.50 0.50 0.50 1.00

> # User can try converting the general fuzzy measure back to
> # Mobius fuzzy measure by calling Mobius transform function.
> fm.Mobius(Genfuzzy)
[1] 0.00 0.00 0.50 0.25 0.50 0.00 -0.50 0.25
> # The result is expected to be exactly the same as in the Mobfuzzy array.

> # User can try computing the Choquet Integral from general
> # fuzzy measure for a new input x with rating values as (0.8, 0.4, 0.6).
> x <- c(0.8, 0.4, 0.6)
> fm.Choquet(x,Genfuzzy)
```
We can discover the behavior of customers by interpreting the output values of the Shapley and Interaction Index. It is convenient to associate the computed Shapley values with the labels of the input criteria as in Table 3.2.

The preference of the customer is interpreted as follows. Product quality is the most important criterion for customers as shown by the highest value of 0.4583. The price is considered least as indicated by the lowest value of 0.2083.

Table 3.2: Shapley values

<table>
<thead>
<tr>
<th></th>
<th>price</th>
<th>quality</th>
<th>service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapley</td>
<td>0.2083</td>
<td>0.4583</td>
<td>0.3333</td>
</tr>
</tbody>
</table>

We can discover the behavior of customers by interpreting the output values of the Shapley and Interaction Index. It is convenient to associate the computed Shapley values with the labels of the input criteria as in Table 3.2.

The preference of the customer is interpreted as follows. Product quality is the most important criterion for customers as shown by the highest value of 0.4583. The price is considered least as indicated by the lowest value of 0.2083.

Table 3.2: Shapley values

<table>
<thead>
<tr>
<th></th>
<th>price</th>
<th>quality</th>
<th>service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapley</td>
<td>0.2083</td>
<td>0.4583</td>
<td>0.3333</td>
</tr>
</tbody>
</table>
Insight into customer behavior can also be understood by interpreting the Interaction Index. As shown in the above code, the Interaction Index values are computed for every subset of input criteria. However, it is suggested the values of pair-wise interactions be used for preference analysis as they are easy to interpret and understand. The extracted pairwise Interaction Index values for analysis are \{1, 2\} \((0.375)\), \{1, 3\} \((0.125)\), and \{2, 3\} \((-0.3750)\). For the purpose of interpretation, they are presented as an interaction matrix in Table 3.3.

The behavior of customers can be analyzed as follows. There is a significant complementary relationship between the \texttt{price} and \texttt{quality} criteria as indicated by a positive Interaction Index value \(0.375\). This means the customers will be more interested in the product that is cheap but good quality. On the contrary, the pair \texttt{quality} and \texttt{service} appears to have a correlating relationship as indicated by a negative Interaction Index value \((-0.375)\). Apparently, the preference of the customers does not increase even if such product has good \texttt{quality} and good \texttt{service}. In addition, there is a slight complement of \(0.125\) between the pair \texttt{price} and \texttt{service}.

It should be noted that the value of Interaction Index reflect the degree of interaction. The consideration of which value to report is depending on the actually application and the desire of analyst.

<table>
<thead>
<tr>
<th></th>
<th>price</th>
<th>quality</th>
<th>service</th>
</tr>
</thead>
<tbody>
<tr>
<td>price</td>
<td>—</td>
<td>0.375</td>
<td>0.125</td>
</tr>
<tr>
<td>quality</td>
<td>—</td>
<td>—</td>
<td>-0.375</td>
</tr>
<tr>
<td>service</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

In summary, \texttt{Rfmtool} provides a tool for performing operations on the fuzzy measures and for computing its related metrics. When it comes to real-life applications
for preference discovery, the role of users is important when interpreting the computed Shapley values and the Interaction Index values into meaningful information. Section 3.3 provides a real application, which can assist users to better understand the proposed approach and the use of this package.

### 3.3 Application: Discovering Tourists’ Hotel Selection Preferences

As discussed in Section 2.2.1, a technique for the effective modeling of tourists’ MCDM process and constructing tourists’ preference profile is important for tourism managers to identify their target customers for customized products and marketing strategies. This section demonstrates the use of the fuzzy measure based preference mining technique in a real life application. The case focuses on the modeling of the behavior of international tourists in the hotel selection process. This section firstly describes the process of online hotel review collections, which is followed by the experiment design description. Results from model evaluations and preference mining are presented and analyzed. The last subsection contains a summary of this case study with managerial implications on how to improve the satisfaction level of tourists.
3.3.1 Data collection

The data set used in this study was collected from TripAdvisor\(^1\), a well-known travel review web site for opinion analysis of tourists [37, 292]. Each review contains ratings for popular hotel criteria including value for money (Value), hotel location (Location), quality of sleep (Sleep), quality of room (Room), room cleanliness (Cleanliness), and additional service (Service), as well as an overall rating. These ratings are in scale from 1 (very unsatisfied) to 5 (very satisfied). A professional data extraction software, named Visual Web Ripper\(^2\), are used to extract such user ratings of hotel criteria together with the reviewer’s demographic data such as travel types (business, family, couple) and countries of origin. The software navigated through all listed hotels in Hong Kong and extracted all review ratings. Approximately 12,000 tourists rating records were collected.

3.3.2 Experimental Design

In the tourism research, it is suggested that the behavior of people and their decisions are guided by the profound effects of national culture [245], and therefore there is a link between cultural values and hotel ratings [174]. Furthermore, people from different continents come from different backgrounds [142]. Because of this, the regions are grouped according to reviewers’ continent of origin as a preparatory step in the profile-construction task. It was interesting to note that the majority of people who traveled to Hong Kong and posted review comments came from North America, Europe, Asia, and Oceania, with only a few people from South America and Africa. For this reason,

\(^{1}\text{www.tripadvisor.com}\)
\(^{2}\text{www.visualwebripper.com}\)
only the first four continents are considered in this research.

Most of the reviews on TripAdvisor for Hong Kong hotels were created in 2010 and 2011 and some reviewers did not provide ratings for all six criteria, which resulted in missing values. Only the reviews from 2011 are taken into consideration and removed the records where data was missing. The data for 2011 was used because this was the latest data set, thus the findings were more up-to-date. This left us with 5443 instances. Table 3.4 shows the structure of the collected data set with respect to travel types and regions.

Table 3.4: Structure of hotel rating data collections

<table>
<thead>
<tr>
<th>Travel Type</th>
<th>Region</th>
<th>Number of Instances</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>Asia</td>
<td>344 instances</td>
<td>$D_1$</td>
</tr>
<tr>
<td>Europe</td>
<td>349 instances</td>
<td>$D_2$</td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>176 instances</td>
<td>$D_3$</td>
<td></td>
</tr>
<tr>
<td>Oceania</td>
<td>86 instances</td>
<td>$D_4$</td>
<td></td>
</tr>
<tr>
<td>Couple</td>
<td>Asia</td>
<td>828 instances</td>
<td>$D_5$</td>
</tr>
<tr>
<td>Europe</td>
<td>986 instances</td>
<td>$D_6$</td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>433 instances</td>
<td>$D_7$</td>
<td></td>
</tr>
<tr>
<td>Oceania</td>
<td>500 instances</td>
<td>$D_8$</td>
<td></td>
</tr>
<tr>
<td>Family</td>
<td>Asia</td>
<td>995 instances</td>
<td>$D_9$</td>
</tr>
<tr>
<td>Europe</td>
<td>246 instances</td>
<td>$D_{10}$</td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>214 instances</td>
<td>$D_{11}$</td>
<td></td>
</tr>
<tr>
<td>Oceania</td>
<td>286 instances</td>
<td>$D_{12}$</td>
<td></td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td><strong>5443 instances</strong></td>
<td></td>
</tr>
</tbody>
</table>

Since the Choquet Integral (CI) claims to model the decision making process of tourists better than previous typical methods, such as the Arithmetic Mean (AM), Weighted Arithmetic Mean (WAM) [279], and Ordered Weighted Averaging (OWA) [308], its performance is evaluated on these data sets. Note that the input data values for the CI should be in the range $[0, 1]$. Therefore, all hotel rating scores are normalized into this range before fitting them into the `fm.fitting()` function. The mean
absolute error (MAE) was adapted to measure the prediction error:

\[ MAE = \frac{1}{M} \sum_{i=1}^{M} |y_i - f(x_i)| \]  

(3.3.1)

where \( M \) is the total number of instances to be evaluated, \( y_i \) is the value of the \( i \)-th input instance to be predicted and \( f(x_i) \) is the predicted value of \( y_i \).

To mine the hotel criteria preferences of tourists to Hong Kong, the following cases are analyzed:

**Preference profile construction:** A detailed profile of hotel preferences is constructed with respect to both travel type and region of origin of tourists to Hong Kong by analyzing Shapley values.

**Interaction analysis of selection criteria:** Since a major advantage of the CI is its ability in assessing the interaction between criteria, its usage is demonstrated by analyzing the interaction index for each travel group as defined in the previous case.

### 3.3.3 Result and Analysis

#### 3.3.3.1 Model evaluation

This section evaluates the performance of the CI against other algorithms (AM, WAM, and OWA) in modeling tourists’ decision making. Particularly for CI, we also consider the impact of \( k \)-additive by performing the experiment with different values of \( k \). Each sub-dataset in Table 3.4 was inputted into these algorithms using the 10-fold cross-validation strategy. The algorithms were fitted to the dataset to minimize the absolute difference between predicted and observed values. MAE values reflecting the prediction error are presented in Table 3.5.
Table 3.5: MAE values of the evaluated algorithms.

<table>
<thead>
<tr>
<th>Sub-dataset</th>
<th>AM</th>
<th>WAM</th>
<th>OWA</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>k = 1</td>
</tr>
<tr>
<td>D1</td>
<td>0.0729</td>
<td>0.0739</td>
<td>0.0733</td>
<td>0.0740</td>
</tr>
<tr>
<td>D2</td>
<td>0.0645</td>
<td>0.0638</td>
<td>0.0639</td>
<td>0.0639</td>
</tr>
<tr>
<td>D3</td>
<td>0.0600</td>
<td>0.0614</td>
<td>0.0592</td>
<td>0.0607</td>
</tr>
<tr>
<td>D4</td>
<td>0.0643</td>
<td>0.0683</td>
<td>0.0632</td>
<td>0.0675</td>
</tr>
<tr>
<td>D5</td>
<td>0.0707</td>
<td>0.0702</td>
<td>0.0666</td>
<td>0.0703</td>
</tr>
<tr>
<td>D6</td>
<td>0.0588</td>
<td>0.0577</td>
<td>0.0573</td>
<td>0.0576</td>
</tr>
<tr>
<td>D7</td>
<td>0.0596</td>
<td>0.0589</td>
<td>0.0607</td>
<td>0.0588</td>
</tr>
<tr>
<td>D8</td>
<td>0.0593</td>
<td>0.0570</td>
<td>0.0595</td>
<td>0.0573</td>
</tr>
<tr>
<td>D9</td>
<td>0.0714</td>
<td>0.0707</td>
<td>0.0701</td>
<td>0.0707</td>
</tr>
<tr>
<td>D10</td>
<td>0.0638</td>
<td>0.0596</td>
<td>0.0627</td>
<td>0.0598</td>
</tr>
<tr>
<td>D11</td>
<td>0.0721</td>
<td>0.0701</td>
<td>0.0746</td>
<td>0.0701</td>
</tr>
<tr>
<td>D12</td>
<td>0.0589</td>
<td>0.0565</td>
<td>0.0578</td>
<td>0.0566</td>
</tr>
<tr>
<td>Overall</td>
<td>0.0647</td>
<td>0.0640</td>
<td>0.0641</td>
<td>0.0639</td>
</tr>
</tbody>
</table>

As shown in Table 3.5, the overall MAE values indicate that the performance of CI (when \( k > 1 \)) appears to outperform other algorithms. Its performance is similar to WAM only when \( k = 1 \), because the fuzzy measure was restricted to individual criterion, and for this reason, the CI was reduced to the WAM. Here, the overall values are the averages of the MAE values for each algorithm on the sub-datasets. The overall performance of CI increased as \( k \) increased, and it achieved the best performance when \( k = 6 \). This was the result of the fuzzy measure being less restricted, where more possible combinations of input criteria were considered. This evidence supports the claim that the use of the CI with less restricted fuzzy measures can model the decision making process in a way that is closer to reality. It should be noted that the maximum value of \( k \) is the number of criteria under consideration, which is 6 in our case study. In different applications when more than 6 criteria are considered, a value greater than 6 can be considered for evaluating the performance of CI.

For each individual dataset, the best performance is indicated by the lowest MAE.
value, which is highlighted by an underline. CI achieved the lowest prediction error on every sub-dataset. We also noticed that the performance of CI on each individual dataset was not necessarily increasing when $k$ increased; its best performance was not always achieved with $k = 6$. This reflects the fact that the model was getting more complicated as $k$ increased, while the available data for some sub-datasets was limited and may not have fully covered the domain of the model. The use of $k < 6$ could be considered in modeling the decision making of tourists in different groups to reduce the complexity of the CI model. This is despite the fact the overall performances of CI were still significantly better than the other algorithms when $k > 1$.

3.3.3.2 Preference profile construction

For business managers, the study of customer profiles is crucial in designing effective marketing strategies. In this application, a hotel preference profile of tourists was constructed from different regions and with different traveling purposes. The collected dataset was divided into subsets according to travel type and region, as shown in Table 3.4, and inputted into the `fm.fitting()` function. The fuzzy measure was set to be unrestricted ($k = 6$) because, in general, this allows the CI to model the decision making process of tourists. The Shapley values, indicating the importance of hotel criteria for different travel groups, are computed using the `fm.Shapley()` function, as shown in Table 3.6.

The hotel preference profile of visitors traveling to Hong Kong can be constructed as follows:

**Business Group:** In general, there are significant variations among the Shapley values of hotel criteria, and the preferences of people from different regions are
Table 3.6: Shapley values of hotel criteria.

<table>
<thead>
<tr>
<th>Travel Type</th>
<th>Region</th>
<th>Hotel Criteria</th>
<th>Value</th>
<th>Location</th>
<th>Sleep</th>
<th>Room</th>
<th>Cleanliness</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>Asia</td>
<td>0.192</td>
<td>0.125</td>
<td>0.108</td>
<td>0.075</td>
<td>0.142</td>
<td>0.358</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Europe</td>
<td>0.117</td>
<td>0.100</td>
<td>0.083</td>
<td>0.308</td>
<td>0.075</td>
<td>0.317</td>
<td></td>
</tr>
<tr>
<td></td>
<td>North America</td>
<td>0.217</td>
<td>0.217</td>
<td>0.017</td>
<td>0.183</td>
<td>0.100</td>
<td>0.267</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>0.317</td>
<td>0.233</td>
<td>0.200</td>
<td>0.033</td>
<td>0.067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Couple</td>
<td>Asia</td>
<td>0.250</td>
<td>0.133</td>
<td>0.150</td>
<td>0.150</td>
<td>0.183</td>
<td>0.133</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Europe</td>
<td>0.208</td>
<td>0.125</td>
<td>0.192</td>
<td>0.200</td>
<td>0.133</td>
<td>0.142</td>
<td></td>
</tr>
<tr>
<td></td>
<td>North America</td>
<td>0.208</td>
<td>0.092</td>
<td>0.117</td>
<td>0.258</td>
<td>0.133</td>
<td>0.192</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>0.133</td>
<td>0.108</td>
<td>0.092</td>
<td>0.283</td>
<td>0.158</td>
<td>0.225</td>
<td></td>
</tr>
<tr>
<td>Family</td>
<td>Asia</td>
<td>0.208</td>
<td>0.125</td>
<td>0.083</td>
<td>0.242</td>
<td>0.150</td>
<td>0.192</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Europe</td>
<td>0.139</td>
<td>0.089</td>
<td>0.103</td>
<td>0.231</td>
<td>0.228</td>
<td>0.211</td>
<td></td>
</tr>
<tr>
<td></td>
<td>North America</td>
<td>0.125</td>
<td>0.192</td>
<td>0.100</td>
<td>0.150</td>
<td>0.208</td>
<td>0.225</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>0.100</td>
<td>0.100</td>
<td>0.217</td>
<td>0.250</td>
<td>0.067</td>
<td>0.267</td>
<td></td>
</tr>
</tbody>
</table>

also different. More specifically, Asian tourists care most about the service criterion while paying very little attention to room quality. On the other hand, European tourists value both room quality and service but do not care much about sleep quality or cleanliness. The sleep quality criterion is also considered as unimportant by North American tourists. For tourists from Oceania, the preferred criterion is value for money, whereas cleanliness and service are ranked as being of lowest importance. It is also interesting to see that while the service criterion is highly ranked by tourists from most regions (as shown by its Shapley values of 0.25 or higher), tourists from Oceania do not care about this criterion at all (value of less than 0.1).

**Couple Group:** Since most of the Shapley values fall between 0.1 and 0.25, the preferences of couples are quite well-distributed across the criteria. However, it can be seen that Asian couples pay close attention to value for money when they are traveling with their partners and that cleanliness is relatively important.
to couples from North America and Oceania. North American couples pay the least attention to location criteria, and sleep quality is the least important criterion for couples from Oceania.

**Family Group:** Among tourists accompanied by their families, it appears that there is no significant preference as most of the Shapley values are below 0.25. Among Asian families, the sleep quality criterion is the lowest ranked, while the location criterion is the least important to European families. For all families except those from Oceania, hotel service is considered important but cleanliness is not.

### 3.3.3.3 Interaction analysis of selection criteria

Another advantage of the fuzzy measure is its ability to provide insights into the interaction between criteria through the interpretation of the interaction index. The interpretation of the interaction index for more than 2 criteria is complicated, thus 2-additive fuzzy measures are suggested to be sufficient for a semantic analysis [99]. The sub-datasets were inputted into the `fm.fitting()` function with \( k = 2 \), then the interaction indices for every pair of criteria were computed using the `fm.Interaction()` function. It should be noted that the overall modeling capability of CI in this case was still better than other algorithms as demonstrated in Section 3.3.3.1. The pairwise interaction indices corresponding to different travel groups (business, couple, and family) are presented in Tables 3.7, 3.8 and 3.9.

From the pair-wise interactions of the hotel criteria in the tables, we can see there are some significant interactions between different criteria in the selection processes of tourists. The interactions were also quite different across groups. This indicates that
people from different regions and with different travel purposes engage in different decision-making processes. Accordingly, we make the following observations:

**Business Tourists:** Hotel selection criteria appears to have some interactions for business tourists from Asia but no considerable interaction was found. For business tourists from Europe, the service criterion appears to be correlated with value for money. Tourists’ preferences for a hotel do not increase even if it satisfies both of these criteria. On the contrary, the service criterion has a significantly positive interaction with room in the decision-making processes of North American business tourists. This positive index indicates that the preference of these tourists for a hotel will increase significantly only if it can

### Table 3.7: Interaction indices for business tourists.

<table>
<thead>
<tr>
<th>Region</th>
<th>Interaction Index</th>
<th>Location</th>
<th>Sleep</th>
<th>Room</th>
<th>Cleanliness</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td>Value</td>
<td>0.000</td>
<td>-0.212</td>
<td>-0.182</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.060</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sleep</td>
<td>0.000</td>
<td>0.000</td>
<td>0.182</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room</td>
<td>-0.091</td>
<td>-0.121</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleanliness</td>
<td></td>
<td></td>
<td>0.152</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>Value</td>
<td>0.000</td>
<td>0.053</td>
<td>-0.015</td>
<td>0.068</td>
<td>-0.429</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sleep</td>
<td>0.000</td>
<td>0.000</td>
<td>0.145</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleanliness</td>
<td></td>
<td></td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>Value</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.300</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>0.000</td>
<td>0.000</td>
<td>0.100</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sleep</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room</td>
<td>0.000</td>
<td>0.000</td>
<td>0.500</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleanliness</td>
<td></td>
<td></td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oceania</td>
<td>Value</td>
<td>0.000</td>
<td>0.000</td>
<td>-1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sleep</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleanliness</td>
<td></td>
<td></td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.8: Interaction indices for couple tourists.

<table>
<thead>
<tr>
<th>Region</th>
<th>Interaction Index</th>
<th>Location</th>
<th>Sleep</th>
<th>Room</th>
<th>Cleanliness</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td>Value</td>
<td>0.000</td>
<td>-0.014</td>
<td>-0.321</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>0.014</td>
<td>0.017</td>
<td>0.038</td>
<td>-0.027</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sleep</td>
<td>-0.014</td>
<td>0.171</td>
<td>0.014</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room</td>
<td>-0.055</td>
<td>-0.028</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleanliness</td>
<td></td>
<td></td>
<td></td>
<td>0.291</td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>Location</td>
<td>0.000</td>
<td>-0.028</td>
<td>-0.285</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Sleep</td>
<td>-0.115</td>
<td>0.093</td>
<td>0.000</td>
<td>-0.093</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room</td>
<td>0.000</td>
<td>0.018</td>
<td>0.161</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleanliness</td>
<td>-0.071</td>
<td>0.162</td>
<td></td>
<td>0.042</td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>Location</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.375</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Sleep</td>
<td>0.000</td>
<td>0.000</td>
<td>0.125</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleanliness</td>
<td></td>
<td></td>
<td></td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>Oceania</td>
<td>Location</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.333</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Sleep</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room</td>
<td>0.000</td>
<td>0.000</td>
<td>0.167</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleanliness</td>
<td></td>
<td></td>
<td></td>
<td>0.278</td>
<td></td>
</tr>
</tbody>
</table>

satisfy all of those criteria. For business tourists from Oceania, there appears to be no interaction between most pairs of the criteria. However, the exception was for the pair value vs. room quality, where strong negative interactions were found, as shown by the strong negative interaction index.

**Couple Tourists:** It is interesting to see that, value for money showed strong redundancy with the room quality criterion for couples from all four evaluating continents. These were indicated by strong negative interaction indexes. Apparently, the preference of these tourists does not increase if such a hotel also offers good value for money and high quality rooms. In contrast, slight positive interactions were found between cleanliness and service for tourists.
Table 3.9: Interaction indices for family tourists.

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>Sleep</th>
<th>Room</th>
<th>Cleanliness</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>-0.006</td>
<td>0.000</td>
<td>-0.178</td>
<td>0.045</td>
<td>-0.149</td>
</tr>
<tr>
<td>Location</td>
<td>0.006</td>
<td>-0.034</td>
<td>0.006</td>
<td>0.066</td>
<td></td>
</tr>
<tr>
<td>Sleep</td>
<td>-0.164</td>
<td>0.000</td>
<td>0.068</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Room</td>
<td>0.028</td>
<td>0.058</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cleanliness</td>
<td></td>
<td></td>
<td></td>
<td>0.124</td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.333</td>
</tr>
<tr>
<td>Location</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Sleep</td>
<td>0.000</td>
<td>0.125</td>
<td>0.062</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Room</td>
<td>0.208</td>
<td></td>
<td>-0.042</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cleanliness</td>
<td>0.104</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.250</td>
<td>-0.125</td>
<td>0.000</td>
</tr>
<tr>
<td>Location</td>
<td>0.042</td>
<td>0.000</td>
<td>0.042</td>
<td>0.097</td>
<td></td>
</tr>
<tr>
<td>Sleep</td>
<td>0.000</td>
<td>0.083</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Room</td>
<td></td>
<td></td>
<td></td>
<td>-0.042</td>
<td></td>
</tr>
<tr>
<td>Cleanliness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.167</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Location</td>
<td>0.000</td>
<td>0.167</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Sleep</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Room</td>
<td></td>
<td>0.000</td>
<td>-0.167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cleanliness</td>
<td></td>
<td></td>
<td></td>
<td>0.333</td>
<td></td>
</tr>
</tbody>
</table>

from Asia and Oceania.

**Family Tourists:** Most of the hotel criteria pairs for this travel group show little interaction, as indicated by the low interaction indices. Only a few significant interactions were found. For instance, in the case of North American families, value for money had a slight negative interaction with the room quality criterion. For families from Oceania, cleanliness had a complementary effect on the service criterion.
3.3.4 Implication

The detailed analysis of the preference profiles in Section 3.3.3.2 highlights the important criteria for tourists from different regions and groups. To be more specific, we found that room quality is significant for business tourists from Europe and couples from North America and Oceania. Additionally, service is the focus of business people from Asia, Europe, and North America and families from Oceania. Thus, hotel managers should give a high priority to improving these aspects of their offering. In addition, their marketing strategies should be carefully designed to draw tourists' attention. The value for money criterion was important to Asian couples and business people from Oceania, thus, cheaper packages with extra benefits could be designed to attract more tourists from these groups.

On the other hand, as presented in Section 3.3.3.3, the advantage of the fuzzy decision-support technique using the fuzzy measure is its power to assess the interaction between criteria. For instance, in Table 3.7, the positive interaction between the room and service criteria for business tourists from North America suggests that hotels must improve both of these criteria at the same time in order to satisfy the expectations of this group. In addition, value for money and room quality criteria does not need to be improved as there appears to be strong negative interactions between them for most couples as shown in Table 3.8. Such information is useful in enabling hotel managers to decide what to focus on in order to achieve the best outcome with minimal effort.
3.4 Summary

The effective modeling of tourists’ MCDM process has always been of interest to researchers working to support managers’ strategic planning and decision making. Efforts have been made to study the preferences by proposing techniques and modeling the MCDM of customers. However, traditional techniques are unable to effectively perform this task, because they fail to consider multiple criteria simultaneously. There remains a high demand for a technique that allows for the exploration of the MCDM process.

This chapter has addressed such a demand by introducing a new preference mining technique based on fuzzy measure, for effective modeling of the MCDM process. By using the Shapley and Interaction Index values, the insight into customer preferences as well as the interaction among criteria, has been explored. As a consequence, managers can allocate their limited resources to improve the aspects of their hotels which are significant to different groups of tourists. They can have more confidence in their decision making while reducing the investment risk. The Rfmtool package is also introduced for customer preference analysis. The operations of this package can be performed easily in the R environment, which is freely available for the research community. An application of modeling the international tourists behavior in the hotel selection process has also demonstrated the use of the proposed technique in practice. In the next chapter, the in-trip behavior of tourists are examined and a framework is proposed to study their travel preference at their destination.
Chapter 4

Travel Behavior Mining from Geo-tagged Photos

Travel behavior patterns during a trip are important to decision makers in destination development, transportation planning, and impact management at a tourism destination. The behavior pattern discovery process can be carried out on different types of data, which are traditionally collected by survey and opinion poll methods. Since the Geographic Information Systems (GIS) were the first to be introduced [165], many research works have been devoted to studying the movement patterns of tourists. For instance, Wu and Carson adopted the GIS application to identify the multiple destination travel behavior of travelers in South Australia [303]. Mckercher et al. used GIS to examine the movements of tourists within an urban destination in Hong Kong, and identified 78 discrete movement patterns [207]. Others used the Global Positioning System (GPS) to gain an understanding about tourists’ experiences and mobility [316], or to explore visitor movement patterns in natural recreational areas [229]. Mckercher et al. compared and contrasted travel behaviors between first-time visitors and repeat visitors to Hong Kong using GPS Tracking and GIS Analysis [210]. These approaches are usually time consuming and are unable to closely reflect the
travel behavior of tourists. Even with the support of technology such as GIS, participants are usually required to carry mobile devices for recording their movement while traveling. Therefore, the collected data are limited in the number of responses or in the scale of geographical areas. When the data collected are unreliable or inadequate, it is necessary to resort to data available online.

Recently, the advances in Multimedia and Mobile technology has allowed massive amounts of user generated data, such as travel photos, to be created. Many of the photo capturing devices, such as smart phones and tablets have built in GPS technology, which enables Geographical information (latitude and longitude coordinates) to be stored in the meta data area of each taken photo. These geo-tagged photos, embedded with time and geographic information, implicitly carry the spatial-temporal movement trajectories of the photo shooters. With millions of geo-tagged photos available, online databases such as Flickr have been a rich data resource for mining tourist travel patterns [170, 320]. It will, therefore, be advantageous for tourism researchers to adopt these advanced technological developments when studying travel behavior.

This chapter aims to address the shortcomings in the existing literature on tourist travel behavior by utilizing the geo-tagged photos that are available on social networking sites. Firstly, a method for constructing the data collection is presented, which captures travel information from geo-tagged photos on Flickr. Two relatively new data-mining techniques are then described for processing and analyzing this data to yield information about the travel behaviors of tourists. The effectiveness of the introduced technique is presented in an application focusing on inbound tourism in Hong Kong, which is a major tourism destination in the Asia Pacific region. The
study aims to discover the locations in which tourists are most interested and the routes they take when visiting Hong Kong. This method and the associated findings are of potential benefits to tourism researchers worldwide who are interested in travel behavior.

Having set the context for this chapter, the remaining parts are organized as follows. Section 4.1 provides our problem statement for the task of analyzing travel preferences based on geo-tagged photos. Section 4.2 presents our framework for extracting and analyzing geo-tagged photos for mining travel preferences. Section 4.3 demonstrates the effectiveness of this framework in an application of inbound tourists to Hong Kong. Finally, Section 4.4 outlines the research directions for later chapters.

4.1 Problem Statement

A number of specific challenges need to be addressed in order to analyze this kind of data. While the massive volume of shared photos available is a comprehensive resource for studying travel behaviors, the data can be noisy or misleading, especially when many photos have been taken in transit rather than at the attractions themselves. In addition, the sequence of photos, especially in certain situations where the tourist moves from one attraction to another, implies mobility information, which is a key to inferring personal activities, intuitions, and goals. However, photos are static media. The photos should be transformed into a suitable representation for the travel analysis task. Two relatively new data-mining techniques, based on Density Clustering and the Markov Chain, are adopted to tackle these issues. Our specific objectives can be defined as follows:
• to introduce a framework for effectively extracting geographical information from geo-tagged photos posted online and using this to analyze tourist travel behavior;

• to identify the attractions of interest to tourists with different profiles who are visiting a tourist destination such as Hong Kong; and

• to identify the travel behaviors of tourists, travel route and travel time, in order to support traffic management and the product development of tourism businesses.

4.2 Methodology

Our framework for travel behavior analysis is carried out in three stages:

1. **Geographic Data Extraction:** This stage extracts the geographic data from online geo-tagged photo data resources for a location of interest.

2. **Tourism Attraction Identification:** This stage identifies a tourism attraction at a location of interest by using the density clustering technique.

3. **Travel Flow Analysis:** This stage explores the travel patterns between tourism attractions by using the Markov Chain technique.

A detailed description of each of these stages is presented in the subsequent sections.
4.2.1 Geographic Data Extraction

Geo-tagged photos are available for public view through web applications such as Flickr, but are not directly downloadable. They must be accessed via Flickr’s Application Programming Interface (API), for which documentation is available at http://www.flickr.com/services/api. One of the challenges in extracting data from this source is that it is impossible to identify individual owners whose photos should be downloaded. Therefore, this task is addressed using a similar approach as in [320]. Geo-tagged photos taken in Hong Kong are searched, and the owners’ information are extracted in order to retrieve the photo information.

A bounding box for the region is defined from which we want to extract geo-tagged photos. Let $x_{\min}, y_{\min}, x_{\max}, y_{\max}$ be its geographical coordinates for the minimum longitude, the minimum latitude, the maximum longitude, and the maximum latitude, respectively. A set of $M$ seed photos are randomly extracted using Flickr’s photo search function with the bounding box. Those photos are returned from the search engine together with their owner’s identification number (ownerID). It is possible for multiple photos to be uploaded by the same owner. A list of owners $< o_1, o_2, \ldots, o_n >$, who have uploaded photos taken in the target area, are then constructed. Their ownerIDs are used to retrieve the user demographic information such as country of origin using the Flickr user search function. This helps us differentiate Hong Kong residents from international travelers. Based on the ownerIDs of the targeted tourist group and the bounding box coordinates, The entire collection of their shared photos are then extracted to ensure their travel activities are captured completely. It should be noted that in addition to the spatial information, each photo is usually stamped with the time and date on which it was taken. This temporal information is also
useful in constructing a movement trajectory for travel pattern analysis. The Flickr photo search function also allows the user to specify temporal information to limit the search. If values for the minimum \( t_{\text{min}} \) and maximum time \( t_{\text{max}} \) are provided, only photos taken between that period are returned.

4.2.2 Tourism Attraction Identification

Geographical location is one important memory cue for recalling previous trips. Many photos taken by tourists may relate to a single day’s activity, and some may also be taken on the journey there. According to [320], a tourism attraction is defined as a spatial extent within a geographical location through which considerable volumes of tourist movement trajectories pass, or where many tourists visit and take photographs. Previous work proposes a Density-based spatial clustering of applications with noise \( \text{DBSCAN} \) \[87\] to perform this task \[153, 170\]. However, a shortcoming of this technique is that generic photo points are treated as being of equivalent importance, whereas the owners of photos are the main factor determining the importance of a cluster in the geo-tagged data analysis context. Thus, an probabilistic version of \( \text{DBSCAN} \), named \( \text{P-DBSCAN} \) \[154\], is adopted to perform this task. \( \text{P-DBSCAN} \) was specifically developed for clustering geo-tagged photos by considering the information about the photo owners into this computation. The details of this technique are described below.

Suppose \( D \) is a collection of geo-tagged photos. Each photo is a point \( p \) referenced by a value pair \( < x_p, y_p > \) for longitude and latitude coordinates. The distance between two photo points \( p \) and \( q \) is denoted by \( \text{Dist}(p,q) \). The neighborhood of a
photo point \( p \), denoted by \( N_\delta(p) \), is defined by

\[
N_\delta(p) = \{ q \in D, Owner(q) \neq Owner(p) \mid Dis(p, q) \leq \delta \} \tag{4.2.1}
\]

where \( Owner(q) = (a_i \in O) \) is an ownership function. In another words, a photo \( q \) is in the neighborhood of another photos \( p \) if they belong to different users and the location of photo \( q \) is within a neighborhood radius \( d \) from photo \( p \). Let \( NeighborOwner(p) \) be the owner number of the neighbor photos \( N_\delta(p) \), and \( \lambda \) be the owner number threshold. A photo \( p \) is called core photo, if its neighbor photos belong to at least a minimum number of owners \((NeighborOwner(p) \geq \lambda)\).

The clustering process of \( P-DBSCAN \) starts with a set of unprocessed photos \( P = \{ p_1, p_2, \ldots \} \). For each photo \( p_i \), if it is not a core photo, it is marked as noise. Otherwise, it is assigned to a cluster \( c_j \), and all of its neighbors \( N_\delta(p_i) \) are put into a queue for further processing. Each photo \( q_{ij} \in N_\delta(p_i) \) is then processed and assigned to the current cluster \( c_j \) until the queue is empty. This process repeats for the rest of the unprocessed photos in \( P \). After clusters of photos are obtained, their geographical coordinates are examined to determine the name and the spatial extent of tourist attractions, which are defined as Areas of Interest (AOI). Here, the values of \( \delta \) and \( \lambda \) are determined based on the scale of specific applications. If the studied region is at macro level such as a country, then AOI can be defined as big as a city. Large values can be assigned to the \( \delta \) and \( \lambda \). If a region is at a micro level, such as a park, then AOI can be at a much smaller scale, with \( \delta \) and \( \lambda \) taking smaller values.

### 4.2.3 Travel Flow Analysis

Tourism managers are interested in not only where tourists travel but also how they get there. This section describes a method, based on the Markov Chain, for mining
such travel patterns and routes taken by tourists between the main attractions [306].

Let \( A = \{a_1, a_2, \ldots, a_m\} \) denote an AOI as identified by using \( P-DBSCAN \). The travel trajectory of a tourist from time \( t = 1 \) to \( t = k \) is defined as \( T = \{a_1^{t_1}, a_2^{t_2}, \ldots, a_i^{t_k}\} \). The probability of a tourist moving to a location \( a_i \) is computed by:

\[
P(a_i^{t_n} | a_i^{t_{n-1}}, a_i^{t_{n-2}}, \ldots, a_i^{t_0}) = P(a_i^{t_n} | a_i^{t_{n-1}}) \quad (4.2.2)
\]

Equation 4.2.2 implies the conditional independent assumption Markov Chain, which can be used to model how tourists flow from one location to another. The transition probability of a tourist moving from attraction \( a_i \) to attraction \( a_j \) (\( i \neq j \)), between time \( t = n \) to \( t = n + 1 \), is computed by:

\[
P(a_j(n+1)|a_i(n)) = \frac{P(a_j(n+1) \cap a_i(n))}{P(a_i(n))} \quad (4.2.3)
\]

where the numerator is the probability of a tourist visiting both locations \( a_i \) and \( a_j \), with \( a_i \) being visited first and then followed by \( a_j \). The event \( a_i(n) \) can be expressed as a combination of mutually exclusive events \( \cup_{j=1}^{k}(a_i(n) \cap a_j(n+1)) \), thus the denominator is computed by:

\[
P(a_i(n)) = \sum_{j=1}^{m} P(a_i(n) \cap a_j(n + 1)) \quad (4.2.4)
\]

The transition probability for all possible travel routes among different attractions can be computed and presented in a one-step transition probability matrix \( P \):

\[
P = \begin{pmatrix}
0 & P(a_2(n+1)|a_1(n)) & \cdots & P(a_m(n+1)|a_1(n)) \\
P(a_1(n+1)|a_2(n)) & 0 & \cdots & P(a_m(n+1)|a_2(n)) \\
\vdots & \vdots & \ddots & \vdots \\
P(a_1(n+1)|a_m(n)) & P(a_2(n+1)|a_m(n)) & \cdots & 0
\end{pmatrix} \quad (4.2.5)
\]
where the value of each entry $p_{ij} \in P$ reflects how likely tourists are to travel from attraction $a_i$ to attraction $a_j$.

In our framework, consideration of the photo owner number in the clustering process $P$-DBSCAN is able to eliminate the noise of photos taken while traveling. Attention can thus be focused on those Areas of Interest which are visited by many tourists. The use of the Markov Chain allows for the consideration of dynamic travel patterns between locations, and models the travel flow of tourists. This framework helps address the challenges in processing and analyzing geo-tagged photos. Thus, more detailed analysis can thus be performed to fully explore the travel behavior of tourists. Section 4.3 presents an application of the proposed framework in travel behavior analysis.

4.3 Application: Travel Behavior Analysis

This section addresses the challenges in the study of travel behavior (as outlined in Section 2.2.2) in an application for Hong Kong inbound tourists using the geo-tagged photos and the proposed framework. This section begins with a description of our data collection followed by an analysis and the results. The practical implications of the findings are then discussed.

4.3.1 Data Collection

The data set used in this work was collected from Flickr following the method described in Section 4.2.1. Since the focus is on Hong Kong inbound travelers, The seed photos are searched by providing a bounding box for the Hong Kong area in the
photo search function, as shown in Table 4.1.

Table 4.1: Photo search parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{\text{min}}$</td>
<td>113.887603</td>
<td>the minimum longitude of the bounding box</td>
</tr>
<tr>
<td>$y_{\text{min}}$</td>
<td>22.215377</td>
<td>the minimum latitude of the bounding box</td>
</tr>
<tr>
<td>$x_{\text{max}}$</td>
<td>114.360015</td>
<td>the maximum longitude of the bounding box</td>
</tr>
<tr>
<td>$y_{\text{max}}$</td>
<td>22.51446</td>
<td>the maximum latitude of the bounding box</td>
</tr>
<tr>
<td>$t_{\text{min}}$</td>
<td>1/1/2011</td>
<td>the earliest photo taken date</td>
</tr>
<tr>
<td>$t_{\text{max}}$</td>
<td>——</td>
<td>latest photo taken date</td>
</tr>
</tbody>
</table>

The coordinates of the bounding box are in decimal degree form, which can be determined manually by using Google Maps (http://maps.google.com). These values were selected to ensure that the bounding box covers the entire geographical area of Hong Kong. Only photos taken from 2011 to 2013 were retrieved. Accordingly, only the date parameter ($t_{\text{min}}$) for the earliest photo taken is provided.

From these seed photos, a list of owner identification numbers was extracted based on the approach in Section 4.2.1. Those indicating that the owner’s origin was not from Hong Kong were treated as labeling inbound travelers and hence were used to retrieve their entire photo collection. The meta data tags of each photo contained photo identification number photoID, owner identification number OwnerID, owner location of origin, date and time taken, and GPS location (longitude and latitude). The GPS location indicates where the photo was taken, which reflects the tourist’s travel footprint. If a user takes many photos at the same location, only one of those was kept, while the others were discarded.

Since people from different countries tend to have different travel preferences [175], tourists are grouped based on their locations of origin to examine the behavior of travelers with different profiles. We observed that the majority of tourists visiting Hong Kong are from countries in the Asia-Pacific, European, and North American regions.
The European and North American tourists are mainly from English-speaking countries such as the U.S., Canada, and the United Kingdom. They are assumed to be more similar to one another in terms of background than tourists from Asia. Therefore, we group tourists from Europe and North America together and refer to them from now on as Western tourists, and similarly group tourists from Asia together and label this group as Asian tourists. We then arrive at a data set containing 29,443 photos collected from 2,100 Hong Kong inbound tourists as shown in Table 7.3.

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Tourists</th>
<th>Number of Photos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western Tourist</td>
<td>1,036 Tourists</td>
<td>15,990 photos</td>
</tr>
<tr>
<td>Asian Tourist</td>
<td>1,064 Tourists</td>
<td>13,453 photos</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,100 Tourists</strong></td>
<td><strong>29,443 photos</strong></td>
</tr>
</tbody>
</table>

The following sections demonstrate how the spatial and temporal information in geo-tagged photos can help to address the challenges of travel analysis. More specifically, three major analyses were performed to examine the behavior of inbound tourists in Hong Kong, namely AOI Identification, Tourist Flow Analysis and Temporal Activity Analysis.

### 4.3.2 Findings and Analysis

#### 4.3.2.1 AOI Identification

This section presents our findings on the identification of the AOI for Hong Kong inbound travelers, to support managers in developing their tourism locations. Firstly, the data collection is inspected by visualizing the photo locations by incorporating their GPS information into Google Earth, as shown in Fig. 4.1. Each photo is indicated by a yellow dot on the satellite image.
The geo-tagged photos appear all over the regions of Hong Kong, which indicates that the raw geographical information is noisy and may be misleading if directly used for analysis. The density clustering algorithm are applied, as described in Section 4.2.2, to remove the noise and identify the locations in which tourists are most interested. An advantage of this density clustering approach is that users are not required to provide the number of clusters in the algorithm, with the clusters automatically determined based on the density of the points and the number of visitors. The neighborhood radius value is set to $\delta = 0.002$, which is approximately equivalent to 150 meters. The minimum owner ($\lambda$) is set to a value of 10% of the tourist number. A total of seven clusters were found, indicating there were seven areas of interest for Hong Kong inbound tourists, as shown in Fig. 4.2. Four of them were in the Hong Kong metropolitan area, including Center Mong Kok, the Tsim Sha Tsui area, Hong Kong Central, and Times Square Towers. Two areas of interest were found in the countryside, namely the Peak Tower and the Tian Tan Buddha Statue. Hong Kong
International Airport is also an AOI, as indicated by the many photos taken within its terminals. These findings are consistent with the fact that the metropolitan area of Hong Kong is well known as a major destination for visitors. P-DBSCAN is shown to be effective in removing noise from geo-tagged photos and identifying AOI to inbound tourists.

![Figure 4.2: Area of Interest for inbound tourists.](image)

In order to identify the location preferences of different groups of tourists, their profiles were taken into account. The clusters were examined with respect to the two groups, Asian and Western tourists, and the proportion of holiday makers visiting each area was used as a measure of its popularity, as shown in Table 4.3. The AOIs are presented in ranking order from the most to the least popular.

- **Hong Kong Central** and the **Tsim Sha Tsui** area are the two most popular destinations for tourists in both groups, as shown by their first and second ranking. **Hong Kong Central** is the central business district and the heart of Hong Kong, while the **Tsim Sha Tsui** area is a major tourist hub in the
metropolitan area with many shops, restaurants, and a good view of Victoria Harbor.

- Western tourists show more interest in the Peak Tower (third) and Center Mong Kok (fourth) than in Times Square Towers (fifth) and Hong Kong International Airport (sixth). Asian tourists show the opposite trend, as indicated by their ranking of Times Square Towers and Hong Kong International Airport as more popular than the Peak Tower and Center Mong Kok.

- It is interesting to see that the Tian Tan Buddha Statue is identified as an AOI for Western but not Asian tourists. A possible explanation is that most Asian countries are likely to contain similar cultural or religious sites such as temples, shrines, or Buddha statues. Asian visitors may, therefore, not be interested in seeing similar things when visiting Hong Kong. On the other hand, Western visitors are likely to be interested in exploring Asian culture, and the Tian Tan Buddha Statue is one such symbol.

Table 4.3: Popularity of Areas of Interest.

<table>
<thead>
<tr>
<th>Group</th>
<th>Area of Interest</th>
<th>Percentage (%)</th>
<th>Popularity Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian Tourist</td>
<td>Hong Kong Central</td>
<td>40.35</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Tsim Sha Tsui Area</td>
<td>38.80</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Times Square Towers</td>
<td>20.08</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Hong Kong International Airport</td>
<td>18.34</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>The Peak Tower</td>
<td>14.19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Center Mong Kok</td>
<td>10.52</td>
<td>6</td>
</tr>
<tr>
<td>Western Tourist</td>
<td>Hong Kong Central</td>
<td>47.92</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Tsim Sha Tsui Area</td>
<td>44.64</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>The Peak Tower</td>
<td>19.74</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Center Mong Kok</td>
<td>15.14</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Times Square Towers</td>
<td>12.22</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Hong Kong International Airport</td>
<td>10.99</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Tian Tan Buddha Statue</td>
<td>10.43</td>
<td>7</td>
</tr>
</tbody>
</table>
4.3.2.2 Travel Flow Analysis

Knowledge of tourist flow is important in transportation planning and traffic management, especially for metropolitan areas with a high density of traffic. An overview of how tourists move from one area to another and what routes they prefer to take will be beneficial for developing appropriate travel management plans. To address such challenges, this part of the analysis focuses on tourist movements between the AOIs in the metropolitan district. The GPS information from the photos is concatenated according to the time taken on a daily basis. Tourist trajectories are analyzed between AOIs by utilizing the Markov Chain technique. A one-step transition probability matrix is computed for each travel group, as described in Section 4.2.3. A high value for $P(a_j|a_i)$ suggests that tourists are likely to visit $a_j$ right after $a_i$. To make it easy for interpretation purposes, the transition probability and transition direction are displayed in cases of Asian Tourists and Western Tourists, as shown in Fig. 4.3. Only flows with a $P(a_j|a_i) \geq 0.3$ are shown in Fig. 4.3. Our finding about the travel flow of tourists can be summarized as follows:

- Fig. 6.2a shows that Asian tourists are likely to flow to Hong Kong Central from the surrounding areas, as shown by the red arrows and a high probability value of around 0.6. Tourists in Center Mong Kok tend to visit Tsim Sha Tsui next (red arrow with a probability of above 0.5), while some chose to travel on directly to Hong Kong Central. Considerable tourist flow is found from Hong Kong Central and the Times Square Towers to Tsim Sha Tsui with probabilities of 0.44 and 0.307, respectively.
Fig. 6.2b shows some differences between Western and Asian tourists. Namely, tourists from Asian countries tend to visit Hong Kong Central right after Center Mong Kok (red arrow with a probability of 0.678), while some travel in the opposite direction. Western tourists are more likely to visit Tsim Sha Tsui (probability of 0.579) than Hong Kong Central (probability of 0.316) after Times Square Towers. Some tourists visited Tsim Sha Tsui immediately after the Peak Tower.

It should be noted that Fig. 4.3 indicates the travel flow of tourists from one location to another, but not necessarily the actual routes taken. Such routes can be revealed by analyzing the photos taken while traveling between AOIs. To demonstrate this, we inspect the common travel routes taken by tourists between more widely
separated areas such as Center Mong Kok and Times Square Towers. Firstly, tourists who visited both locations on a single day are identified, then the photos taken while traveling from one place to another are retrieved and plotted in Fig. 4.4. The yellow dots denote photos which represent the footprint of tourists during their trip. It can be seen that most tourists traveled to and from these locations via Tsim Sha Tsui and Hong Kong Central. To be more specific, they traveled along Nathan Road between Centre Mong Kok and Tsim Sha Tsui. Direct ferry services were used to travel between Tsim Sha Tsui and Hong Kong Central. To go from Hong Kong Central and Times Square Towers, tourists used Hennessy Road. An interesting finding about tourist preferences is that a direct ferry service which departs from Wan Chai, which is near Times Square Towers and goes to Tsim Sha Tsui. But most tourists did not choose this option and preferred to traveling to Tsim Sha Tsui via the ferry line leaving from Hong Kong Central.

4.3.2.3 Temporal Activity Analysis

Tourism managers are not only interested in knowing where and how tourists travel, but also when they visit different places. This is crucial in transportation planning and destination management, to avoid problems of overloading when too many people visit the same place in a short time frame. To address this, the visiting patterns of tourists are examined based on the spatial and temporal information in their photos. More specifically, the probability of tourists appearing at a particular AOI over a 24-hour period, are computed, as shown in Fig. 4.5. The horizontal axis represents time, and the vertical axis probability. Here, only six AOIs common to both Asian and Western tourists were included, because these are the locations that receive many
visits. These findings are summarized as follows:

- Fig. 4.5a shows the presence of tourists at *Hong Kong International Airport* on a given day. Asian tourists appear to be arriving or departing during the daytime (10:00 to 17:00), whereas Westerners may travel at any time.

- Fig. 4.5b shows that tourists in both groups are likely to visit *Hong Kong Central* between 11:00 and 16:00 hours. In particular, the peak time for Asians is at noon, while the busiest time for Westerners is 15:00.

- Tourists are likely to visit the *Tsim Sha Tsui* area in the late afternoon and evening, as shown in Fig. 4.5c. Asian tourists are most likely to be there at 17:00 and 21:00 hours. Western tourists visit most often at around 20:00. Similarly, the busy time for *Times Square Towers* is in the afternoon, as shown in Fig. 4.5d.
Similar visiting patterns to Center Mong Kok are found for both Asian and Western visitors, as shown in Fig. 4.5e. Both demonstrate a peak visiting time at this location of 16:00. They also have relatively similar visiting patterns for the Peak Tower as in Fig. 4.5f, where the peak for Asian tourists is 19:00.

4.3.3 Implications

The analysis of AOI in Section 4.3.2.1 highlights some differences in travel preferences for different tourist groups. Different travel packages and tour routes can be developed accordingly to meet the needs of inbound tourists. For instances, Western, but not Asian, tourists show an interest in Tian Tan Buddha Statue. Tourism managers can thus develop marketing plans and transportation arrangements to promote this attraction to Western visitors. The popularity ranking given in Table 4.3 can also help tourism developers to design local tour recommendations for tourists. Tours to Times Square Towers can be recommended to Asian visitors before the Peak Tower and Center Mong Kok, whereas tours to the Peak Tower and Center Mong Kok can be recommended to Westerners in preference to Times Square Towers.

The analysis of tourist flow in Section 4.3.2.2 offers some practical implication for traffic management. For example, the traffic flow captured in Fig. 4.3 suggests that additional transportation to Hong Kong Central could be arranged for Asian visitors, as they are likely to visit this location from surrounding areas. On the other hand, more direct transport can be arranged for Western tourists to travel from Center Mong Kok to Hong Kong Central, or from Times Square Towers to the Tsim Sha Tsui area. In addition, Fig. 4.4 shows the current travel routes taken by tourists between Center Mong Kok and Times Square Towers. Most people tend to use the
Figure 4.5: Tourist activities in Areas of Interest.
ferry service from *Hong Kong Central* to *Tsim Sha Tsui*, even though there is a direct ferry service leaving from *Wan Chai*, which is near the *Times Square Towers*. Traffic managers can develop appropriate management plans to encourage tourists to use the direct service. This can reduce the traffic in *Causeway Bay* and the ferry line between *Hong Kong Central* and *Tsim Sha Tsui*, thus helping to avoid traffic congestion in the future as more and more visitors travel to Hong Kong.

The tourist activity analysis in Section 4.3.2.3 reflects the visiting patterns of tourists to particular AOIs. The knowledge of where tourists are likely to be at different times of the day can support destination promotion and management. For instance, support services and transportation can be arranged at *Hong Kong International Airport*, when many tourists are arriving and departing (Fig. 4.5b). Different travel itineraries can be developed to avoid overcrowding in *Hong Kong Central* around noon (Fig. 4.5b), or at *Center Mong Kok* in the afternoon (Fig. 4.5e).

### 4.4 Summary

Knowledge of travel behaviors is crucial to help tourist managers construct strategic plans and make decisions that will create a sustainable tourism industry. This chapter approached the task of mapping tourist travel behavior by exploiting the socially generated and user-contributed geo-tagged photos publicly available on the Internet. A method for constructing a travel dataset from geo-tagged photos are presented. A dataset containing thousands of photos with temporal and geographic information attached enables us to capture the movement trajectories of tourists on a larger scale. Two techniques are introduced, the *P-DBSCAN* and the *Markov Chain*, to mine travel behavior patterns from this dataset. The effectiveness of these approaches are
demonstrated in an application on Hong Kong inbound tourism and discovered the locations of interest to travelers, their travel patterns, and their daily activities. Practical implications are offered to support Hong Kong tourist managers in destination development, transportation planning, and impact management.

Although there are a massive volume of travel photos available on the Internet, only a small number of them have location data available. The next chapter presents two approaches that allow tourism managers to incorporate the ungeotagged photo data into travel behavior analysis, in order to achieve the full potential of this valuable data resources.
Chapter 5

Travel Diary Construction from Travel Photos

Advances in Multimedia and Mobile technology has allowed massive amounts of user generated data, such as travel photos, to be created. Many of these photos have GPS information available, which means they are fruitful data resources for researchers to study the travel behavior of tourists [170]. Although a massive amount of travel photos are available on the Internet, only a small number of these are tagged with geographical data. This prevents tourism researchers and business managers from fully capturing and understanding the travel behavior of tourists.

Several attempts have been made to estimate the geographical location of photos. Hays and Efro used a collection of 6 million geo-tagged photos as a reference set to find the location of ungeo-tagged photos [112]. Their algorithm was basically photo matching, where each photo was represented using several sets of visual descriptors. Kalogerakis et al. [139] improved the work of Hays and Efro [112] by incorporating Human Travel Prior to the estimation process. Given a sequence of ungeo-tagged photos, their locations were estimated based on the reference geo-tagged photos and the past travel trajectory of tourists. Their work was done on the global scale, and
the estimation location is bins on the grid of the world map. Another attempt is in [257], where photos were put on to a map based entirely on photo tag.

As demonstrated in Chapter 4, tourism managers are interested in knowing where tourists visit during their trip to a tourist destination so they can develop an effective business plan. It may not be necessary to estimate the exact GPS location of a particular photo but it is crucial for tourism managers to know where tourists visit during their trip. We are also interested in estimating the locations and travel paths visited by tourists, rather than predicating the GPS coordinate of the travel photos. Two scenarios are of interest to tourism managers: 1) visited location verification, and 2) travel diary construction. In the former scenario, tourism managers are interested in whether a tourist has visited a predefined location, such as a natural beach, or temple, or a particular sight seeing location. In the later scenario, tourist managers would like to reconstruct a travel diary describing locations and paths visited by tourists while traveling. In this chapter, two novel methods are proposed to address such scenarios, which can support the travel behavior analysis of tourism managers. The visited location verification task is tackled by a hybrid multi instance learning techniques, named Representative Instance Classification (RIC). The task of travel diary construction is addressed by a method based on the Bayesian approach and Latent Dirichlet Allocation (BA-LDA).

Having set out the background for undertaking this work, this chapter is structured as follows. Section 5.1 provides our problem statement for developing the RIC for identifying visited locations and BA-LDA for travel diary construction. Section 5.2 describes the details on the RIC and BA-LDA algorithms. Section 5.3 presents our
experiments to validate the performance of our methods, in visited location verification and travel diary construction. Section 5.4 concludes this chapter and envisions the directions for later work.

5.1 Problem Statement

5.1.1 RIC for Visited Location Verification

The motivation for this approach is based on the observation that different tourists may visit different places during their trip, but if they all visited at least one of the same locations and take photos, some photos in their collections are likely to have similar content. For instance, if they all visited a natural beach, one or more photos in their collection may contain common elements such as sea, sky, sand, or trees, etc. If we can build a model to capture such similarity in the photo collection taken by tourists at the location, it is possible to verify whether other tourists have also visited such a location based on their photo collection.

Nevertheless, this is a challenging task for the following reasons. Firstly, there is no clear indication about which scene is best at describing a particular location. This is entirely based on interest of the tourists to take the photos. Secondly, it may be possible to know if tourists have visited a location based on their tags or comments, however there is usually no indication on which photos were actually taken at the area of interest (AOI) or which photos best describe such an AOI. There exist many photos taken at different locations with different scenes within the same travel photo collection.
Recent advances in machine learning, especially the MIL, have provided an alternative framework to address these issues. In MIL, the training set comprises labeled bags that are composed of unlabeled instances, and the task is to predict the labels of unseen bags. The location verification problem for travel photos, can also be investigated from the perspective of multi-instance. More specifically, each photo collection of a tourist is regarded as a bag, while each photo is regarded as an instance. A collection labeled as visited a location means that at least one of its photos is taken at that location. This perspective makes the MIL appropriate. Section 5.2.1 provides the background on the MIL framework and details about the proposed RIC algorithm for visited location verification.

5.1.2 BA-LDA for Travel Diary Construction

This technique aims to estimate the visited locations to construct travel diaries for the travel behavior analysis task. A natural assumption is that the temporal information (time and date) is available, so that the photos can be arranged in time order as a photo sequence. Our task is to estimate the visited locations of tourists based on the photo sequences. Several challenges need to be addressed for this task. Firstly, location entity is a high level concepts. For instance, a city scene photo is usually indicated by the presence of concepts, such as road, building, car, etc. The estimation of photo locations should consider these concepts, which can not be done efficiently using low level photo features, such as color and textures. Secondly, the photos taken in the same location can have different concepts due to viewing direction. For instance, a photo taken at a river bank will have the concept of building, car, road, if the tourist faces the camera toward the bank. It may contain semantic concepts
such as the river, sky, boat, if he/she faces the camera toward the river. The task of travel diary construction should consider the mixture of scenes at a potential location. Thirdly, the estimation of locations for photo sequences should take the order of photo sequences and timing constraint into consideration.

Photos are assumed to be taken close to each other in temporal order are in the same AOI. Here, we define a photo chunk \((\text{Chunk}_i)\) as a set of photos each taken in the same AOI and relatively close in time. A photo sequence is a collection of photo chunk \(D = \{\text{Chunk}_1, \text{Chunk}_2, \ldots\}\) belonging to users, and arranged in timing order. The time interval between photos in different photo chunks is much longer than photos in the same photo chunk. Our task is to estimate the location of each chunk in a photo sequence to reveal the travel trajectory of tourists at a tourist destination. Details on BA-LDA are presented in Section 5.2.2

5.2 Methodology

5.2.1 RIC Technique

This section firstly provides some background on Multi-Instance Learning. Details about RIC are then described, followed by a justification on the appropriateness of RIC for location verification task.

5.2.1.1 Multi-Instance Learning

The problem of location identification can be formulated into a classification problem. In traditional supervised learning, each training sample is represented by an instance (or feature vector) and labeled with a class. Although this formalization is
prevailing and successful in a large number of applications, there exist many real-world applications, which do not fit this framework well. For example, in the task of location identification from a photo collection, a photo collection is labeled positive, if it contains at least one photo taken at an AOI, thus indicating that the tourist has been to that place. A photo collection is labeled negative, with none of the photos taken at the AOI, reflecting the fact that the tourist did not visit that place. Such a task is difficult because although the label of the collection is known, that of the individual photo is not available. It has been shown that learning algorithms ignoring the characteristics of multi-instance problems do not work well in this scenario [133].

The term multi-instance learning was coined by Dietterich et al. [79] for drug activity prediction. In multi-instance learning, the training set is composed of many bags. Each bag contains many instances. A bag is positively labeled, if it contains at least one positive instance, otherwise, it is negatively labeled. The task is to construct a model to correctly label unseen bags. As a new learning framework, multi-instance learning has attracted significant attention. A representative multi-instance learning algorithm is the Diverse Density algorithm [203], where the diverse density at a point in the feature space is defined to be a measure of how many different positive bags have instances near that point, and how far the negative instances are from that point. Thus, the task of multi-instance learning can be transformed into an optimization task to find the point with the maximum diverse density. More recently, many other algorithms have been proposed, such as Wang and Zucker’s extended KNN algorithm [294], Ruffo’s multi-instance decision tree Relic [255], Chevaleyre and Zuchker’s ID3-MI and RIPPER-MI algorithms [58], Zhou and Zhang’s multi-instance neural network BP-MIP [321], and Zhang and Goldman’s ED-DD algorithm [317],
5.2.1.2 RIC

Assume that the features and visual characteristic of the photo have been extracted. Each photo is presented as a vector of features, and called an instance, and a photo collection of a tourist is a bag of instances. Let $\mathcal{D} = \{X, Y\}$ denote the training data set in which $X$ is training data containing $n$ bags: $X = \{B_1, B_2, \ldots, B_n\}$, and $Y$ is a set of labels: $Y = \{L_1, L_2, \ldots, L_n\}$. For each bag $B_i$, it contains a set of $m$ instances $B_i = \{I_{i1}, I_{i2}, \ldots, I_{im}\}$, and its class label is $L_i$. Let $\text{net}$ be a neural network model for multi-instance learning, with $\delta_{pos}$ and $\delta_{neg}$ as the thresholds for positive bags and negative bags, respectively. The pseudo-code of the RIC algorithm is as shown in Alg. 1.

The RIC algorithm works in two steps. The first step is to identify representative instances from bags into a new data set. The second step is to construct a supervised learning model from the new data set.

**Identify Representative Instances from Bags:**

We first provide a brief introduction to a multi-instance learning classifier called $\text{BP-MIP}$ [321], which is adopted to identify representative instances from bags. The idea behind this method is to build a classification model using a neural network on a multi-instance training data set ($\mathcal{D}$) to minimize the global error:

$$Err = \frac{1}{2} \sum_{i=1}^{n} \left( \max_{1 \leq j \leq m} o_{ij} - \delta_i \right)^2$$  \hspace{1cm} (5.2.1)

where $o_{ij}$ is the actual neural network output upon the instance $I_{ij}$, $\delta_i$ is $\delta_{pos}$ or $\delta_{neg}$ according to the class label of bag $B_i$.  

etc.
Algorithm 1 Representative Instance-based Classification (Training)

Require: Training data $X$ and label $Y$, user specified learning rate $\alpha$ and $\beta$, threshold of positive bag $\delta_{pos}$, threshold of negative bag $\delta_{neg}$, number of iteration $\text{iter}$

Ensure: Trained RIC model $S$

1: initialize a neural network model $\text{net}$; 
2: $\text{netUpdate} = \text{true}$; 
3: while $\text{iter} > 0$ and $\text{netUpdate}$ do 
4: $\text{netUpdate} = \text{false}$; 
5: for each bag $B_i$ in $X$ do 
6: $O_i \leftarrow$ the vector of outputs for instances in $\text{net}(B_i)$; 
7: if $B_i$ is positive bag and $\max(O_i) < \delta_{pos}$ then 
8: update $\text{net}$ according to $\alpha$, $\beta$ and the bias $|\max(O_i) - \delta_{pos}|$; 
9: $\text{netUpdate} = \text{true}$; 
10: else if $B_i$ is negative bag and $\max(O_i) > \delta_{neg}$ then 
11: update $\text{net}$ according to $\alpha$, $\beta$ and the bias $|\max(O_i) - \delta_{neg}|$; 
12: $\text{netUpdate} = \text{true}$; 
13: end if 
14: end for 
15: $\text{iter} \leftarrow \text{iter} - 1$; 
16: end while 
17: $\mathcal{D}_{RI} = \phi$; 
18: for each bag $B_i$ in $X$ do 
19: $\hat{O}_i \leftarrow$ a set of output from $\text{hat}(B_i)$; 
20: $\text{index} = \arg\max_j \hat{o}_{ij}$; 
21: representative instance $x_i \leftarrow B_i[\text{index}]$; 
22: representative label $y_i \leftarrow L_i$; 
23: $\mathcal{D}_{RI} = \mathcal{D}_{RI} \cup \langle x_i, y_i \rangle$; 
24: end for 
25: $S \leftarrow$ train a SVM model on $\mathcal{D}_{RI}$; 
26: return $S$
The Equ. 5.2.1 means that if at least one instance of a positive training bag is predicted as positive or all instances of a negative bag are predicted as negative, then the error on the concerned bag is zero and the weights of the network will not be updated. Otherwise, the weights will be updated according to the error on the instance whose corresponding actual output is the maximal among all the instances in the bag. Here, the maximum output and the minimum values of the neural network model are 1 and 0 respectively. When a trained BP-MIP network is used in prediction, a bag is labeled as positive if the output of at least one instances within the bag is greater or equal to 0.5. This way, the model can evaluate the likelihood of a positive for each instance in a bag. Hence, it is possible to make use of this property for identifying the most positive instance from each bag as the representative instance.

The process of identifying representative instances from bags can be described in Alg. 1. In each training iteration, the training bags are fed to the neural network \( \text{net} \) one by one. For the instance \( I_{ij} \) in the bag \( B_i \), an output value \( o_{ij} \) is generated. The maximal output \( \max_{1 \leq j \leq m} o_{ij} \) is considered as the predicted output for the bag \( B_i \). If bag \( B_i \) has been predicted correctly then the network weight will not be changed, otherwise, the weights are updated according to the error on the instance whose corresponding actual output is the maximal in \( B_i \). The training process iterates until the error \( Err \) decreases to a threshold, or the number of training iterations reaches a threshold. Since, BP-MIL is used to find the best set of representative instances rather than predicting a new bag label, the threshold of \( Err \) is set to 0 so that the global error is minimized.

After the training process of the BI-MIP model is finished, the training bags are fed into the trained model which is denoted as \( \hat{\text{net}} \). For each bag \( B_i \), the training
model will evaluate $\hat{O}_i = \{\hat{o}_{i1}, \hat{o}_{i2}, \ldots, \hat{o}_{im}\}$ as assigned for each instance $I_{ij}$ in $B_i$. The instance with maximal corresponding output value $\hat{o}_{max}$ is then selected to represent the bag $B_i$. $N$ representative instances are extracted from $N$ bags and form a single training data set $D_{RI} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where each instance $x_i$ is the instance in $B_i$ with the maximal corresponding output value from the BI-MIP model, and $y_i$ is the label of each instance $x_i$ whose value is assigned from the label of bag $B_i$. This means the multi-instance learning model BI-MIP transforms the classification problem of bags back to a traditional classification problem.

**Supervised Learning:**

The goal of this step is to build a supervised classifier on representative instances identified from a training data set. Based on the data set $D_{RI}$ with representative instances, any classification method can be utilized. In this work, Support Vector Machine (SVM) model is used to find an optimal hyperplane which separates the positive and negative instances [148]. Once the RIC model is constructed, it is used to classify whether a new tourist visited an AOI or not based on his/her photo collection, and to identify the instance (photo) that best describes the location.

For a test bag $B'$, its instance $I'_i$ is fed into the trained RIC model $S$. A vector of predicted labels $L' = \{l_1, l_2, \ldots, l_m\}$ can be generated for bag $B'$. If any instance in the bag is identified as positive, then the bag is considered as having visited the location, otherwise, if no instance is identified as positive, the bag is labeled negative. The process of scoring new testing bags is presented in Algorithm 2.
Algorithm 2 Representative Instance-based Classification (Testing)

Require: Testing bag $B'$, trained RIC model $S$.
Ensure: Predicted label $L'$

1: for each instance $I'_i$ in $B'$ do
2: \quad $l_i \leftarrow S(I'_i)$;
3: end for
4: if $\exists l_i$ is positive then
5: \quad $L' \leftarrow$ positive;
6: else
7: \quad $L' \leftarrow$ negative;
8: end if
9: return $L'$

5.2.1.3 Justification

Suppose $X = \{B_1, B_2, \ldots, B_n\}$ is the input space where $B_i$ is a bag composed by a set of instances $I_{i1}, I_{i2}, \ldots, I_{im}$, and the set of class labels for bags is $Y = \{L_1, L_2, \ldots, L_n\}$.

In this work, our focus is on binary cases $Y = \{0, 1\}$ corresponding to the class labels of visited and non-visited locations, with labels attached to bags rather than each instance within the bags. These labels are interpreted as: if $L_i = 0$, then $l_{ij} = 0$ for all $1 \leq j \leq m$, and no instance in the bag is positive. In contrast, if $L_i = 1$, then at least one instance $I_{ij} \in B_i$ is positive ($l_{ij} = 1$). Therefore, the relationship between the bag label $L_i$ and instance labels $l_{ij}$ can be expressed as $L_i = \max_{1 \leq j \leq m} l_{ij}$.

Let $F: X \rightarrow Y$ be the discriminant function for a multi-instance data set. The prediction for a bag $B_i$ takes the form $L'_i = \text{sgn} \max_{1 \leq j \leq m} F(I_{ij})$. Notice that the hyperplane is defined by the representative instances, which are the “most positive” instance in each positive bag and the “least negative” instance in each negative bag. The problem of learning $F$ for multi-instance data can be thus transformed into traditional supervised learning by considering only these instances. Since, the supervised learning is a standard process, our major task is to identify such representative instances from the bags.
To evaluate the degree of positive, we transform the label space from discrete \( Y = \{\text{negative}(0), \text{positive}(1)\} \) into continuous space \( Y' = [0, 1] \). Let \( F' \) denote a mapping function \( F' : X \rightarrow Y' \), and \( O_i = \{o_{i1}, o_{i2}, \ldots, o_{im}\} \) is a set of outputs for each bag \( B_i \). The most positive instance of \( B_i \) is an instance \( I_{ij} \in B_i \) so that \( o_{ij} = \max(O_i) \).

The desired outputs for the “most positive” instance of positive bags and the “least negative” instance of negative bags are 1 and 0, respectively. This condition is not always achievable in reality. Therefore, the function \( F' \) is optimized in a way, so that the outputs of positive representative instances are closed to 1 for positive bags. The outputs of negative representative instances are closed to 0 for every negative bags. This can be achieved by minimizing the global error:

\[
\xi_{\text{global}} = \xi_{\text{po}} + \xi_{\text{ne}} \tag{5.2.2a}
\]

\[
\xi_{\text{po}} = \frac{1}{2} \sum_{p=1}^{P} (\max_{1 \leq j \leq m} o_{pj} - d_{\text{po}})^2 \tag{5.2.2b}
\]

\[
\xi_{\text{ne}} = \frac{1}{2} \sum_{q=1}^{Q} (\max_{1 \leq j \leq m} o_{qj} - d_{\text{ne}})^2 \tag{5.2.2c}
\]

where \( P \) and \( Q \) are the number of positive and negative bags in \( X \), \( o_{pj} \) and \( o_{qj} \) are the actual output of the positive bag \( B_p \) and negative bag \( B_q \), \( d_{\text{po}} \) is the desired output of positive bags and \( d_{\text{ne}} \) is the desired output of negative bags.

Notice that Equ. 5.2.2 can be generalized to Equ. 5.2.1, and the optimization target of \( F' \) is similar to the training objective of the BP-MIL algorithm. Thus, function \( F' \), which is implemented by BP-MIL, can provide a mapping \( X \rightarrow Y' \), so that the natural constraints for a positive degree of belonging or not belonging to a location are best preserved.

In [321], BP-MIL was originally employed as a classification algorithm with a specified decision boundary of 0.5 to predict the labels of unseen bags. In fact, BP-MIL
is a general algorithm which has not been optimized toward any data, and its performance is highly dependent on the appropriately chosen configuration. Besides which, the hyperplane of 0.5 is chosen to be a decision boundary which is the middle point in the positive degree space $Y'$, and is not optimized with respect to the representative instances. Hence, direct application of this algorithm may not be appropriate, especially when there is high overlapping between positive and negative classes. Notice that, the hyperplane of BP-MIP is a linear decision boundary in $Y'$. Therefore, the optimal decision boundary should also be linear and estimated automatically based on the representative instances. To overcome this shortcoming, SVM is employed with linear kernel as a discrimination algorithm to automatically find such a decision boundary.

Based on the output from $F'$, a set of representative instances are extracted. Only these instances are taken into account during the training process of SVM in the second step of the RIC algorithm. We denote a set of representative instances as $D_{RI} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i = I_{ij}$ is a representative instance of bag $B_i$ satisfying $F'(I_{ij}) = \max(O_i)$ and $y_i = L_i$ is its label. The task of SVM is to find a hyperplane $F(x) = x^T \beta + \beta_0 = 0$ so the margin between the training points are maximized [111].

Fig. 5.1 illustrates the solution of SVM, and the representative instances of positive and negative bags are denoted by symbols “+” and “o” respectively surrounded by squares. The hyperplane is optimized automatically on the most positive instances (most likely describing an AOI) which is more suitable for prediction purposes than a predefined one. Since SVM is a powerful discrimination function and is optimized based on the optimal set of representative instances, the combination of BP-MIL
Figure 5.1: Support Vector Classifier on representative data instances
and SVM into the RIC algorithm can provide an optimal hyperplane for the photo collection multi-instance learning problem. The effectiveness of the RIC algorithm is proven experimentally in Section 5.3 on a real world travel photo data set.

5.2.2 BA-LDA Technique

Our BA-LDA method is performed in two steps: 1) Location Concept Profile Computation, in which each AOI is represented by a concept profile that is computed based on reference photos using the LDA model; 2) Travel Diary Construction, where the location of photo chunks is computed based on the photo content and prior travel information.

5.2.2.1 Location Concept Profile Computation

The Scale-invariant feature transform (SIFT) descriptor is used to represent the visual characteristic of photo content [193]. SIFT is a suitable candidate for location recognition, and it has been used widely in computer vision area, such as scene classification [89], object recognition [182], or photo categorization [318]. The traditional bag of word approach is adopted for constructing SIFT features to represent photo contents. Visual code words were constructed by $k$-means clustering of SIFT features extracted from sample photos.

Let each visual code word represent a word $w$, each photo represents a document $d$, and each concept presents a topic $t$. Each location is referenced by a collection of geo-tagged photos taken in that area.

The LDA model is firstly applied to the photo collections of all reference locations to discover the distribution of the visual word over latent concepts and to also discover
the distribution of latent concepts in each training photo. To formalize the LDA for our case, let us restate the generative process:

- For each photo: draw a concept distribution $\theta_d \sim Dir(\alpha)$. Here, $\theta_d$ is the vector of $1 \times N$ representing the weight distribution of concept in a photo $d$, $Dir(.)$ is drawn from a uniform distribution with scaling parameter $\alpha$.

- For each visual word $w$ in the document. Draw a specific topic $z_{d,n} \sim multi(\theta_d)$, where $multi(\theta_d)$ is multi-nominal distribution. Then draw a word $w_{d,n} \sim \beta_{z_{d,n}}$.

The inferential problem of LDA is to compute posterior distribution of the hidden variable given a document:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \quad (5.2.3)$$

Since this distribution is intractable to compute, it is transformed into a variational inference problem, for each document:

$$q(\theta, z|\gamma, \varphi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\varphi_n) \quad (5.2.4)$$

where Dirichlet parameter $\gamma$ and multi nominal parameters $\varphi$ are free variational parameters. $N$ is the number of word $w$ in the document. To determine the value of variational parameters, an optimization problem is set up:

$$(\gamma^*, \varphi^*) = \arg\min_{\gamma, \varphi} D(q(\theta, z|\gamma, \varphi)||p(\theta, z|w, \alpha, \beta)) \quad (5.2.5)$$

By computing the derivatives of the KL divergence and setting them equal to zeros, we obtain the following pair of updated equations:
\[
\varphi_{ni} \propto \beta_{iw} \exp \{ E_q[\log(\theta_i) | \gamma] \} 
\]

(5.2.6)

\[
\gamma_i = \alpha + \sum_{n=1}^{N} \varphi_{ni} 
\]

(5.2.7)

Note that the optimizing parameters \( \gamma^*, \varphi^* \) are document specific. The notation \( i \) is the index of the topic \( i \in K \) number of topics. For each document, in each iteration, \( \varphi_{ni} \) parameters are first computed using the above equation, then it is normalized to sum to 1, then \( \gamma_i \) is computed. The Dirichlet parameters \( \gamma^* \) can be viewed as providing a representation of a photo in concept simplex. \( \varphi^* \) represents the weights of the concept for each visual word in a photo.

In the E-step of LDA optimization for each document, the optimizing values for variational parameters \( \gamma^*, \varphi^* \) are found. In the M-step, \( \alpha \) is computed using the Newton-Raphson method with estimated \( \gamma^* \), and \( \beta \) is computed with respect to \( \varphi^* \) and \( w \). 

\[
\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \varphi_{dni} w_{dn}^j 
\]

(5.2.8)

In this work, we are interested in computing the distribution of visual words for each concept, and the distribution of concepts for each location. Location is a collection of photos describing the visual characteristics of it. The entire data collection is fit to LDA to estimate its \( \alpha^* \) and \( \beta^* \) parameters. The estimated \( \beta^* \) value can be used for representing the distribution of words for each concept, and denoted as \( \nu_D(v) \). To compute the distribution of a topic for the location, we perform the following operation. Given \( \alpha^* \) and \( \beta^* \), the \( \gamma_i \) values is estimated using the above variational inference method. It provides a representation of a document in topic simplex for each document. In other words, \( \gamma_i \) represents the contribution of each
concept to each photo. Since a location is a set of photos, the $\gamma_i$ values of all reference photos is combined in a location and normalized to sum to 1. They are denoted as $P(v)$. In this way, $P(v)$ represents the concept distribution of each reference location. $V_D(v)$ and $P(v)$ are used in the Travel Diary Construction process as presented in the next section.

5.2.2.2 Travel Diary Construction

Suppose $img_i$ represents a photo, $S_j$ represents a location, $v(v \in S_j)$ represents concepts of location $S_j$, then the probability of $img_i$ taken in $S_j$ is:

$$P(img_i | S_j) = \sum_{v \in S_j} P(img_i | v)P(v)$$

where the probability of $img_i$ generated according to concept $v$ is measured as:

$$P(img_i | v) = \frac{1}{Dist_{K-L}(V_D(img_i), V_D(v))}$$

$Dist_{K-L}$ is the Kullback-Leibler distance, and $V_D\{\cdot\}$ is the concept distribution vector, where each entry represents a probability of a photo or view containing a specific concept. $V_D(img_i)$ is the concept distribution vector of each photo computed by a variational inference method with the trained $\alpha^*$ and $\beta^*$, followed by normalization to 1. The probability of a photo chunk taken in a location can then be measured as:

$$P(Chunk_l | S_j) = \prod_{img_i \in Chunk_l} P(img_i | S_j)$$

Here photo chunks are computed using clustering according to the temporal factor. Photos whose time taken are relatively close, are considered to be taken in the same spot and assigned to $Chunk_l$. 
The trajectory information are then integrated into the model, by adopting the Bayesian framework. Suppose $t_i$ represents a trajectory that we may predict based on the photo chunk sequence. The probability of $t_i$ based on photo chunks $D = \{\text{chunk}_1, \text{chunk}_2, \ldots\}$ is calculated as follows:

$$P(t_i|D) = P(D|t_i)P(t_i) \tag{5.2.12}$$

The prior can be measured by the product of sense transition:

$$P(t_i) = \prod_{S_i \rightarrow S_j \in t_i} P(S_i \rightarrow S_j) \tag{5.2.13}$$

The transition probability between locations should be subjected to time length. For example, if the time between the first photo chunks and second photo chunks is short, it is likely that people traveled to close by that point. Thus, we use the time-interval transition probability $A_1, A_2, \ldots$, where $A_r$ is the prior transition matrix for time interval $\Delta T$.

$P(t_i)$ is computed with respect to time length as

$$P(t_i) = \prod_{S_i \rightarrow S_j \in t_i} P_r(S_i \rightarrow S_j) \tag{5.2.14}$$

where $P_r(S_i \rightarrow S_j)$ is the entry $(i, j)$ on transition matrix $A_r$, whose time interval between photo chunks is $\Delta T = r$.

The likelihood $P(D|t_i)$ can be measured as:

$$P(D|t_i) = \prod_{\text{Chunk} \in D} P(\text{Chunk}_j|S_k) \prod_{j=1}^{k} P_r(S_{k-1} \rightarrow S_k)P(S_0) \tag{5.2.15}$$

The part $\prod_{j=1}^{k} P_r(S_{k-1} \rightarrow S_k)P(S_0)$ means that we need to multiply the probability from the first spot in the trajectory to the probability $P_r(S_{k-1} \rightarrow S_k)P(S_0)$. 
The probability of the first spot in trajectory $P(S_0)$ should be considered because we make the Markov assumption on the trajectory.

Finally, the trajectory that we prefer to give to photo chunks is:

$$\arg \max_t P(t)$$ (5.2.16)

The use of the travel diary construction framework is presented in the Section 5.3.2.

5.3 Experiment and Analysis

This section presents experiments to evaluate the performance of the proposed techniques. The first experiment is to evaluate RIC technique in the task of visited location verification. The second experiment is to evaluate BA-LDA technique in the task of travel diary construction. Details about data collection, experiment design and result analysis are presented below.

5.3.1 Visited Location Verification

5.3.1.1 Experiment Design

The data used in our experiment are photo sets downloaded from the Flickr web site. These photos were taken by tourists while traveling around a tourism destination such as Hong Kong. We randomly generated a number of photo collections to represent the travel photos for each tourist. A photo collection was labeled as positive if such collections contained one or more photos taken at a particular location such as a natural beach. A photo collection was labeled as negative if it contained no photos taken of the natural beach. It is possible that a tourist destination may have many natural
beaches opened for tourist visits in the real situation. However, we assumed there is only one area of interest under consideration. Our task was to identify if a tourist did or did not visit a natural beach using the proposed method. We also considered the fact that the number of tourists visiting a location would be much less than in other areas. Thus, the number of constructed positive collections was much smaller than the number of constructed negative collections. In detail, our data collection contained approximately 6000 photos, including 100 positive photo collections, and 300 negative collections, with each collection containing approximately 15 photos.

We used the Color Structure Descriptor (CSD) [215] and the Homogeneous Texture Descriptor (SCD) [250], which are both popular visual descriptors for describing photo characteristics. The descriptor size of CSD and the HTD default size are set to 32 and 62 respectively. Thus, each photo was represented as an instance of 94 feature vectors. Each photo collection was a bag of approximately 15 instances.

In the experiment, a 10-fold cross validation strategy was used to compare the performance of different methods. Since the data sets contained different numbers of positive and negative bags, the accuracy, recall and precision were used to measure the performance. Suppose there are $P$ positive bag and $N$ negative bag in the test data set, among which, $P_a$ positive bags are recognized correctly as visited the natural beach, and $N_a$ negative bags are recognized correctly as not visited the natural beach, while $N_r$ negative bags are recognized incorrectly as the positive bag. The accuracy, recall, precision and the F-measure are then calculated as:
\[
\text{accuracy} = \frac{P_a + N_a}{P + N} \tag{5.3.1a}
\]
\[
\text{recall} = \frac{P_a}{P} \tag{5.3.1b}
\]
\[
\text{precision} = \frac{P_a}{P_a + N_r} \tag{5.3.1c}
\]
\[
F_{\text{measure}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{5.3.1d}
\]

In order to show the overall performance of an algorithm, the averaged accuracy, recall and precision across 10-fold cross validation is reported in this experiment.

5.3.1.2 Performance Evaluation

Experiments were also performed to evaluate the performance of the proposed RIC algorithm. Since the RIC algorithm extended both the multi-instance learning (MIL) algorithm and the single-instance learning (SIL) algorithm, both approaches were considered in comparison. The included SIL algorithms were Random Forest (RF), Artificial Neural Network (ANN), and Support Vector Machine (SVM). Parameters of ANN and RF were kept by default as in Weka. SVM was tested with both linear (\textit{lin}) and non-linear polynomial (\textit{pol}) kernels. The experiment on SIL was performed in the traditional way so that all instances in positive bags were merged into a set of instances and labeled as positive, while all instances in negative bags were merged into a set of instances labeled as negative. For the MIL framework, we included BI-MIP and another standard algorithm, the Citation-KNN [294], for comparison. For RIC and BI-MIP, the desired outputs of the bags, i.e., 1 for positive and 0 for negative were replaced with 0.9 for positive and 0.1 for negative respectively, which is a trick for speeding up the training process [321]. The learning rate of the networks is set to
0.05, and running iteration was set to 500 epochs. The parameter for the Citation-KNN was set to 3 for reference and citation values, as these showed best performance on existing works. The experiments were carried out on the multi-instances data set with positive and negative bags.

Table 5.1: Experiment results.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.786</td>
<td>0.607</td>
<td>0.393</td>
<td>0.477</td>
</tr>
<tr>
<td>ANN</td>
<td>0.777</td>
<td>0.557</td>
<td>0.501</td>
<td>0.528</td>
</tr>
<tr>
<td>SVM (lin)</td>
<td>0.662</td>
<td>0.391</td>
<td>0.650</td>
<td>0.489</td>
</tr>
<tr>
<td>SVM (pol)</td>
<td>0.743</td>
<td>0.484</td>
<td>0.531</td>
<td>0.507</td>
</tr>
<tr>
<td>BP-MIL</td>
<td>0.843</td>
<td></td>
<td>0.974</td>
<td>0.380</td>
</tr>
<tr>
<td>Citation-KNN</td>
<td>0.883</td>
<td>0.827</td>
<td>0.670</td>
<td>0.740</td>
</tr>
<tr>
<td>RIC</td>
<td>0.875</td>
<td>0.692</td>
<td></td>
<td>0.783</td>
</tr>
</tbody>
</table>

In general, the overall accuracies of most algorithms were greater than 0.7. Since the provided data set was imbalanced, and the target was the identification of the AOI, the performance of classifiers must be considered in relation to other metrics (precision, recall and F-measure) as well. RF failed to detect the AOI, while the ANN and SVM (pol) showed poor performance on this task, as indicated by low recall values of around 0.5 or less. SVM (lin) had a better recall rate (0.65) but performed poorly on accuracy and precision in comparison with the others. All of these SIL algorithms achieved low performance on the F-measure of around 0.5.

The MIL framework appears to outperform other traditional algorithms using SIL, as shown with higher values for all evaluation metrics. This evidence supports the claim that MIL is more suitable than SIL in identifying the visited locations of tourists based on their photo collections. Among the MIL algorithm, BP-MIL shows the best precision, however it failed to correctly identify photo collections of interest, as shown with a low recall value (0.38). Citation-KNN shows the best accuracy (0.883) but is
was not much higher than RIC (0.875). More importantly, RIC showed best recall rate of 0.9 and F-measure (0.783). This evidence proves that the RIC algorithm can provide a better method to overcome the difficulties in location identification.

We would like to also evaluate the performance of RIC with different $\delta_{\text{pos}}$ and $\delta_{\text{neg}}$ values to see their effect, when BP-MIP is adopted into RIC. Since 0.5 is considered as the boundary between positive and negative bags in BP-MIP [321], $\delta_{\text{neg}}$ is set with different values from 0.5 to 0 with a step of 0.1. At the same time, $\delta_{\text{pos}}$ is set with different values from 0.5 to 1. The values of $\delta_{\text{neg}}$ and $\delta_{\text{pos}}$ define the decision boundary, the increase of $\delta_{\text{pos}}$ and decrease of $\delta_{\text{neg}}$ enlarge the decision boundary away from a single line of 0.5. Therefore, their variation can be considered at the same time in our experiment. The experiment results are shown in Fig. 5.2.

From Fig. 5.2, we can see that as $\delta_{\text{pos}}$ gets closer to 1 and $\delta_{\text{neg}}$ gets closer to 0, the accuracy, the precision and the F-measure steadily increased, while the recall decreases slightly. However, the recall still shows high performance as it stays above 0.85 for all cases. In consideration of the overall performance of RIC for AOI detection
application, the result is better when $\delta_{pos}$ is close to 1 and $\delta_{neg}$ is close to 0. We also notice that when the value for $\delta_{pos} = 0.9$ and $\delta_{neg} = 0.1$, a relative better performance compared to when $\delta_{pos} = 1$ and $\delta_{neg} = 0$. This is because when $\delta_{pos} = 0.9$ and $\delta_{neg} = 0.1$, it does not force the positive bag to be absolutely positive, nor the negative bag to be absolutely negative. This is more suitable for the real case when there is no definite definition of the targeted AOI, and the scenes described in the photos are varied.

Fig. 5.3 shows some sample photo collections that are labeled as positive and negative respectively. Each row represents a photo collection, which is regarded as a bag of instance. The representative instances/photos are highlighted in red and green squares. Only one photo is selected in each bag because RIC algorithm select a single most positive instance as the representative one. It can be seen that the photos in the same bags are quite different, while some photos in the positive bags are similar to those in the negative bags. The identified representative photos in the positive bags have relatively similar contents of the natural beach which does not appear in the negative bags, while the identified representative photos in the negative bags are most like the beach scene among the photos. Thus, the SVM model trained on these representative instances can better separate the photos of the natural beach in the positive bags than other most like beach photos in negative bags. A by-product of this advantage is that tourism managers can gain insight into how most tourists are viewing and taking photos at a particular AOI. They can identify the area that most interests by tourists so that appropriate plans and services can be developed.
Figure 5.3: Representative photo examples.
5.3.2 Travel Diary Construction

5.3.2.1 Experiment Design

The data used in our experiment was collected from the Flickr photo sharing database, whose GPS locations were taken in a tourist destination such as Hong Kong. The process was carried out similarly as described in Section 4.2.1. We focus on the seven popular locations in Hong Kong, as identified in Section 4.3.2.1, which include Center Mong Kok, Tsim Sha Tsui, Hong Kong Central, Times Square Towers, The Peak Tower, Tian Tan Buddha Statue and Hong Kong International Airport. Only photos taken in these areas were considered in the experiment. We removed users that had more than 100 pictures in a single location, users who visited at least 3 locations with less than 3 photos each, or users who had too many photos in total. Photos such as indoor scenes, facial photos, or photos about humans were unlikely to provide clues about the location, thus, we filtered these photos out when performing the experiment. The final data set contained approximately 10000 photos from around 500 tourists.

The photos belonging to each tourist were firstly arranged in a sequential order in a photo collection. The photo collection were clustered according to temporal information. A first photo was assigned to the first photo chunk. If a photo was taken within 10 minutes from the previous photo, it is assigned to the current photo chunk, otherwise, it was assigned to the next photo chunk. The process was repeated until all photos were assigned to a photo chunk. Thus, each photo collection from each user was transformed into a photo sequence containing many photo chunks. A photo chunk can contain one or more photos. Since, the data set has 500 users, 500
photo sequences were constructed.

In the experiment, the leave-one-out approach was adopted to evaluate the performance of BA-LDA. Namely, one photo sequence was held out as a test case, while all other photos were used as a training set. The photo content in the training set was used to construct location concept profiles, while the GPS information was used to compute the prior travel patterns of tourists. Our task was to construct the travel diary of where the tourists had visited based on the photo content and temporal information of their photo collection. Bag of word feature was used for photo content representation with the SIFT descriptor. According to [89], a small number of code words were enough for describing scene content, and w200 visual words were used in the experiment. A concept profile for each location was constructed as described in Section 5.2.2.1. We made an independent assumption on the transposition of tourists between AOIs, thus, the prior travel probability can be computed as the transition provability matrix as described in Section 4.2.3.

Despite some work on image location estimation exist, this work is the first attempt on travel diary construction based on sequence of photo trunks. We firstly demonstrated the performance of our algorithm on diary construction for several cases, and then inspected the results. It was a natural assumption that if more photos were taken in a location, there were more clues available which made it easier for estimation of the location. An experiment is further performed to examine the influence of the photo number in the photo trunks.

Suppose there are $P$ travel paths corresponding to $P$ number of photo sequences under consideration. Each photo sequence comprises a number of photo trunks, where each photo truck is corresponding to a location. The total number of locations for
all travel paths is $L$, $(L > P)$. A location is considered correctly identified by our algorithm if its estimated location matches the true location of the photo trunk. A path is considered correctly identified if and only if all of its locations are correctly identified. Let $l$ and $p$ be the numbers of locations and paths correctly identified respectively. The performance of the proposed algorithm is evaluated based on the following metrics:

\[
\text{path accuracy} = \frac{p}{P} \quad (5.3.2a)
\]
\[
\text{location accuracy} = \frac{l}{L} \quad (5.3.2b)
\]

### 5.3.2.2 Result Analysis

BA-LDA approach is applied with the leave-one-out strategy. Some representative cases with their actual and predicted travel paths along with their taken photos are shown in Fig. 5.4. The photo sequences in these cases contained three photo chunks, and each photo chunk contained four or more photos. Some sample photos in each photo chunk are shown in a row on the left hand side of the figure. Each travel path spans three different locations. The actual travel path is represented by the red line, while the estimated travel path is represented by the green line in the satellite photo. For instance, Fig. 5.4a shows the travel path from *Tsim Sha Tsui* to *Hong Kong Central* and then to *The Peak Tower* together with the photos taken in these locations in trunks 1, 2 and 3 respectively. The actual and predicted paths were relative close which indicates that our proposed algorithm was able to perform travel path estimation based on photo sequences. A similar performance was seen in Fig. 5.4b for the travel path from *Cetre Mong Kok* to *Tsim Sha Tsui* to *Hong*
Kong Central. Fig. 5.4c shows a missed prediction case of our algorithm. The tourist travels from Hong Kong International Airport to the Tian Tan Buddha Statue and to Hong Kong Central. However, our algorithm estimated the final location incorrectly as the Tsim Sha Tsui area. This is probably caused by the photos in Tsim Sha Tsui and Hong Kong Central had relatively similar visual content, while it is more likely that people travel Tian Tan Buddha Statue to Tsim Sha Tsui than travel to Hong Kong Central.

Next, the performance of BA-LDA is shown with a different number of photos in each photo chunk. The leave-one-out strategy is adopted, in which one photo sequence of an user is kept for prediction, while all others photos are used as references. The photo sequences for testing users are sampled so that only $n$ number of photos in each photo trunk is used for location prediction. $n$ is set from 1 to 10. In case the number of photos in a photo trunk is less than $n$, the results, as shown in Fig. 5.5, reflect the fact that if photo trunks contain more photos, they are more likely to be predicted accurately. The performance on the path and location are low when only 1, 2 or 3 photos are available for each trunk. It increases steadily along with an increase in the number of photos for the trunks, and becomes stable when more than 5 photos are available. The results indicate that the use of the topic model can effectively capture the visual content of photos for the location estimation task. It is interesting to see that the path accuracies are relatively smaller than the location accuracies. This is probably due to the strict condition for a path to be considered as correctly predicted only when all of its locations are correctly predicted. Many incorrectly predicted paths, in fact, have quite a number of their location correctly identified, but not all of them. In summary, the experiment has
Figure 5.4: Photo collections and travel paths.
Figure 5.5: Algorithm performance with a different numbers of photos in trunks.

demonstrated that our method is able to reconstruct the travel diary of tourists based on past travel photos and prior travel behavior.

5.4 Summary

Travel photo sharing sites have emerged as fruitful resources for effectively capturing tourists’ travel behavior and preferences. However, many of these photos are without geographical location information, which prevents tourism researchers and business managers from fully capturing and understanding the travel behavior of tourists. This chapter tackled this challenges following two innovative approaches, which were visited location verification and travel diary construction. Two techniques were proposed to accommodate this. These were the RIC and the BA-LDA. RIC is a hybrid machine learning technique, which incorporates both single instance learning
and multi-instance learning together. Experiments show that RIC achieved better performance than existing techniques on verifying visited locations based on travel photo collections. BA-LDA was proposed based on *Latent Dirichlet Allocation*, which considers both spatial, temporal and visual data. Experiments show promising performance on constructing travel diaries based on travel photo sequences. The next chapter examines the practical challenges in developing tourism applications based on GPS and satellite imaging. It presents several alternatives on satellite photo processing that allow effective displaying and processing of satellite photos on end user devices.
Travel photos with GPS information has emerged as an important resource for the analysis of travel preferences. Some attempts have been made to use the knowledge mined from these resources to develop applications to support tourist travel. Mamei et al. developed automatic intelligence services to learn from the experiences of tourists and then made recommendations on “where to go” and “what to see” to people going to visit a place for the first time [200]. Hao et al. attempted to equip tourists with knowledge mined from travelogues to facilitate the trip planning of other tourists [109]. Kurashima et al. proposed a probabilistic behavior model for recommending a travel path based on the past travel information of tourists on Flickr [162]. Okuyama and Yanai made use of actual travel paths extracted from a large number of online geo-tagged photos to develop a travel planning system [227]. Majida and colleagues proposed methods for tourism location recommendation that were relevant to user preferences and a travel context [198, 199]. Shi et al. proposed a method for personalized landmark recommendation [263] and Yin et al. facilitated tourist trip planning using a travel path search system that used geo-tagged photos [312].
These applications often involved the analysis and display of satellite images on end user devices such as smart-phones, tablets or laptop computers. The ubiquitous nature of the small display screens in such devices has reinforced the need for robust image reduction techniques. In particular, where these screens are being used to display images captured at higher resolutions than the screen is capable of displaying, image reduction is essential. Even where large screens are involved, the resolutions at which images may now be captured far exceeds the capability of many screens to display them at full resolution. It is important that the preprocessing of the image does not cause the loss of fine image details, which may convey important information relevant to the analysis of the image content.

The travel behavior analysis also involves constantly analyzing and processing a large amount of satellite images. Image reduction can reduce the number of pixels processed and can also reduce the computational complexity along with the memory and time requirements. The processing of images often requires the preservation of specific image details at the same time as removing or reducing the prevalence of unwanted artifacts, such as noise or data corruption. For instance, in satellite imagery, one may need to identify small boats at sea and discard structures such as wave crests. These are represented by nearly white pixels on a dark blue background, whereas, the spatial arrangement of these pixels are varied. However, when the number of pixels in a subregion of the image belonging to the object of interest is minor, existing techniques for filtering, smoothing or image reduction perform poorly.

*Block-based image reduction* operators based on averaging aggregation functions have been proposed to address the above problem in image reduction [236]. Recent
work has focused on the problem of the preservation of fine, pixel-scale details in images, which are represented by small, spatially coherent clusters of pixels having similar intensities. Non-monotonic averaging functions have been shown to improve the robustness of image reduction compared to monotonic averages such as the arithmetic mean or median [297]. As with convolution-based image filters in [297], spatial organisation of pixels was accounted for by using distance-based weights. This approach does not sufficiently describe the spatial structure of a set of pixels so structured image details may be preserved during the reduction.

This chapter aims to address the challenges in image processing to support the travel analysis task using satellite technology and the development of intelligence tourism applications. This chapter is devoted to developing novel methods to support image reduction, that satisfies the special requirement of satellite image processing. We are interested in distinguishing geometrically compact clusters of pixels of similar brightness (potentially part of an object or artifact) from pixels that are tonally different, and also from pixels that are tonally similar, but not geometrically related. Furthermore, we consider the possibility that such clusters may represent a minority of a local region, even when considering block sizes as small as $3 \times 3$ pixels. Pixel intensities are assumed to be corrupted by noise or undue variation due to sampling effects in discrete digital imaging arrays. To extend and improve upon previous work in image reduction, we propose that a measure of cluster compactness may be used to weight the contribution of specific pixels within a non-monotonic average, which is used to compute a representative pixel value for a given block of pixels. Novel approaches of constructing cluster compactness measures based on fuzzy measure theory are considered.
Given a subset \( A \) of pixels within an \( m \times n \) block, the following requirements for a compactness measure \( SC \) would seem to be reasonable in the image reduction context:

1. The function \( SC(A) \) must be non-decreasing in \( |A| \), as larger clusters are less likely to represent noise.

2. The function \( SC \) must be invariant with respect to translation, reflection and rotation (at least rotations by multiples of 90 degrees).

3. The function \( SC \) must discriminate between compact groups and disconnected subsets of fixed cardinality. For convenience, we require \( SC \) to be normalized.

It is noted that some simple measures of compactness exist, such as those based on inter-set and intra-set distances, however these are inadequate for our purposes as they do not satisfy the first requirement. This requirement though is precisely the monotonicity condition used in the definition of fuzzy measures and therefore it makes sense to look for a solution within the class of fuzzy measure functions.

This chapter is structured as follows. Section 6.1 formulates the image reduction task as an averaging problem over local image blocks and introduces the role of a cluster compactness measure within this context. Concepts of fuzzy measures are then provided. Section 6.2 presents details about our proposed measures based on a minimum spanning tree graph, sugeno-type fuzzy measure and a geometrical decomposition. Section 6.3 compare measures produced by the proposed methods, then present a validation of the proposed measures in the context of non-monotonic averaging for image reduction. Finally, Section 6.4 summarizes the result.
6.1 Background

6.1.1 Problem Formulation

The compactness measure problem is formulated using an example of image reduction based on a local, block based reduction operator, as shown in Fig. 6.1. An image of size $M \times N$ is subdivided into non-overlapping blocks of size $m \times n$. Each block is aggregated to generate a single value representative of the original data, which becomes the intensity of a single pixel within the reduced image. Thus, the original image is reduced to the size $M' \times N' = \frac{M}{m} \times \frac{N}{n}$.

![Figure 6.1: A scheme for image reduction within 3 × 3 blocks.](image)

Local operators based on aggregation functions have been shown to be both effective and easily admit parallel implementations [29, 237]. They require the properties of averaging functions such as means, so that the output is within the range of the intensities in the input pixels and is also idempotent [30]. This later property ensures that if the block is of uniform intensity its representative value is exactly the same
as the intensity of all input pixels. The simplest examples are the *weighted arithmetic mean* and the *median*, which play an important role in *Gaussian* and *median* filtering. All suitable averaging functions can be obtained by using a penalty-based approach [43] wherein a given penalty function for intensity deviations is minimized to obtain the aggregate value.

The most well-known averaging functions satisfy another condition, that of *monotonicity* [30]. Any increase in one or more input values does not cause a decrease in the output of the aggregation. This condition is useful when the data are noiseless, however it is not desirable when the data may be contaminated by noise, which appears as outliers within small pixel blocks. Previous work in aggregation problems has focused on the design of *non-monotonic averages* using penalty-based methods [32, 43], that have been adapted for image reduction [297]. Many of the *non-monotonic averages* were recently shown to be weakly monotone averaging functions, that have desirable properties relevant to image processing tasks such as reduction and filtering. Our current work continues to pursue the design of such functions using penalty based methods.

The intensities of a set of $p$ pixels is represented by the vector $\mathbf{x} \in \mathbb{I}^p$, where $\mathbb{I} = [a, b]$ is any closed, non-empty subset of the reals.

The average intensity of these inputs is the solution to the minimization problem

$$y = f(\mathbf{x}) = \arg \min_y \mathcal{P}(\mathbf{x}, y),$$

where $\mathcal{P} : \mathbb{I}^{p+1} \to \mathbb{R}$ is a penalty function satisfying the conditions:

1. $\mathcal{P}(\mathbf{x}, y) \geq c$ $\forall \mathbf{x} \in \mathbb{I}^p$, $y \in \mathbb{I}$;

2. $\mathcal{P}(\mathbf{x}, y) = c$ if all $x_i = y$; and,
for some constant \( c \in \mathbb{R} \) and any closed, non-empty interval \( I \).

In [297], a penalty function was proposed based on a weighted sum of intensity-based penalties, given by

\[
P(x, y) = \sum_{i=1}^{p} w_i(y) \rho(x_i, y) \tag{6.1.1}
\]

where

\[
\rho(x_i, y) = \begin{cases} 
     r(k) & r(k) < \tau, \\
     \beta r(k) & r(k) \geq \tau.
\end{cases} \tag{6.1.2}
\]

\( \tau = \alpha \max(\epsilon, r(t)) \) and \( \alpha > 0, 0 \leq \beta \leq 1, 2 \leq t \leq p. \)

Given \( r_i = \|x_i - y\| \), then \( r(k) \) denotes the \( k\)th smallest element of the set of ordered (ascending) values of \( r_i \), given the aggregate value \( y \). This function generated a mode-like non-monotonic average of the input vector \( x \), and was shown to outperform other monotonic and non-monotonic block-based reduction operators when applied to images corrupted by speckle or impulse noise.

Although, the function \( \rho \) favors compact clusters of intensity values in \( x \) by assigning smaller penalties to inputs closer to the proposed output, it does not take into account spacial information of the pixels. This was achieved by using a normalized distance for the weights \( w_i \) depending on the position of the pixels within the block [297]:

\[
w_i(y) = \frac{d(x_i, y)}{\sum_{i=1}^{p} d(x_i, y)}, \forall y = x_j \in \{x_1, ..., x_p\}. \tag{6.1.3}
\]

This function arose from the additional constraint that the average must also be an internal function (i.e., the output should be one of the inputs), which is a reasonable requirement on image reduction tasks. We wish to replace the weights \( w_i(y) \) with a function that incorporates spatial structure information and that appropriately orders candidate clusters of pixels according to their spatial organization. Consequently, we
desire a function over the power set $2^P$, where $P$ is the index set for the input vector $x$ and $w : 2^P \rightarrow [0, 1]$.

As mentioned in the introduction, we desire a monotone set function as larger subsets are favored and have smaller penalties in the expression for $\mathcal{P}$. We also impose suitable boundary conditions so that the range of $w$ is within the unit interval. Since these conditions are the same as the ones used in the definition of fuzzy measures, we consider weighting functions of the type $w(A) = C - v(A)$, where $C \geq 1$ ensures that $w$ is strictly positive, and $v$ is a fuzzy measure.

### 6.1.2 Fuzzy Measure Concepts

This section presents some concepts of fuzzy measure, which is the foundation for our approaches. Recall the fuzzy measure definition in Section 3.1.2. Let $\mathcal{N} = \{1, 2, \ldots, n\}$. A discrete fuzzy measure is a set function $v : 2^{\mathcal{N}} \rightarrow [0, 1]$ which is monotonic (i.e. $v(A) \leq v(B)$ whenever $A \subset B$) and satisfies $v(\emptyset) = 0$ and $v(\mathcal{N}) = 1$.

**Definition 2 (Submodular Fuzzy Measure).** A fuzzy measure $v$ is called submodular if for any $A, B \subseteq \mathcal{N}$

$$v(A \cup B) + v(A \cap B) \leq v(A) + v(B).$$  \hspace{1cm} (6.1.4)

It is called supermodular if

$$v(A \cup B) + v(A \cap B) \geq v(A) + v(B).$$  \hspace{1cm} (6.1.5)

**Definition 3 (Subadditive Fuzzy Measure).** A fuzzy measure $v$ is called subadditive if for any two nonintersecting subsets $A, B \subset \mathcal{N}$, $A \cap B = \emptyset$:

$$v(A \cup B) \leq v(A) + v(B).$$  \hspace{1cm} (6.1.6)
It is called *superadditive* if

\[ v(A \cup B) \geq v(A) + v(B). \]  

(6.1.7)

**Definition 4 (Symmetric Fuzzy Measure).** A fuzzy measure \( v \) is called *symmetric* if the value \( v(A) \) depends only on the cardinality of the set \( A \), i.e., for any \( A, B \subseteq \mathcal{N} \), if \( |A| = |B| \) then \( v(A) = v(B) \).

We require a *non-symmetric fuzzy measure* as we need to differentiate between compact and scattered groups of bins of the same cardinality, such as those depicted in Fig. 6.2a and Fig. 6.2b respectively. We also require *symmetry* among those subsets \( A \) that represent translation, rotation and reflection symmetry of the corresponding groups of pixels. In addition, we require a *non-additive* fuzzy measure. If we consider a compact cluster of pixels - such as the one shown in Fig. 6.2a - and add another pixel to the cluster, the measure of this larger set should depend on how close (Fig. 6.2c) or away (Fig. 6.2d) this additional pixel is to the original cluster.

![Figure 6.2: Example of different clusters.](image)

Three different approaches are considered to constructing fuzzy measures that satisfy our requirements. In the first case, fuzzy measures reflecting cluster compactness are computed directly based on Minimum Spanning Tree. In the second and third cases, we will use some extra information about the desired fuzzy measure in
terms of reference points and constraints. A reference point is a specified value of $v$ for a particular subset $\mathcal{A}$, which we believe is a reasonable choice. For example, we could specify $v(\mathcal{A}) = 1$ for all compact $\mathcal{A}$ of cardinality $|\mathcal{A}| \geq k$, for some $k < n$, and where compactness means that each pixel in $\mathcal{A}$ has a neighbor in $\mathcal{A}$ (that is, for $\forall a \in \mathcal{A} : d_H(a, \mathcal{A}) = 1$ and $d_H$ is the Hausdorff distance). Further, we will impose constraints such as $v(\mathcal{A}) \geq v(\mathcal{B}) + \delta$ for some $\delta > 0$ when we believe the measures of $\mathcal{A}$ and $\mathcal{B}$ should differ by at least $\delta$. We assume there are $r$ reference points and $c$ inequality constraints.

6.2 Methodology

This section presents three proposed approaches for constructing fuzzy measures that satisfy our preferences. The first one is based on Sugeno-type Fuzzy Measures, the second approach is based on the Minimum Spanning Tree, and the third is based on the Decomposing Fuzzy Measure.

6.2.1 Minimum Spanning Trees Approach

Minimum Spanning Trees (MST) have been used for clustering for several decades [315]. Given a connected weighted graph the MST is a subgraph (specifically a tree connecting all vertices) whose weight is the smallest. It is constructed from the adjacency matrix using Prim’s or Kruskal’s algorithms [67] and its complexity is quadratic in the number of vertices.

We shall use the weight of the MST (i.e., the sum of all edge weights in the MST), constructed from a complete graph connecting the elements of a cluster whose weights
are pairwise distances between the data. In cluster analysis, such MSTs are used to agglomerate the data, and partition it into several clusters by removing the edges of maximum weight. Here we are interested in a measure of compactness of a single cluster, hence we shall only use the weight of the MST rather than its structure.

We are interested in devising a quantity that satisfies the three criteria set out in the Introduction. We use the formula

\[ SC(A) = C - \frac{W(MST(A))}{TM} + 1, \]

where \( T = |U| \) is the cardinality of the largest cluster, \( M \) is the largest distance between the elements of a cluster, and \( C = 1 + \frac{T(M+1)-1}{TM} \). For example, for a 3 × 3 block and the Euclidean distance, \( T = 9 \) and \( M = 2\sqrt{2} \), and hence \( C = 1 + \frac{1}{9} + \frac{4}{81\sqrt{2}} \).

For brevity we will denote \( w(A) = w(MST(A)) \).

The constant \( C \) is a normalization constant which ensures \( SC(A) \in (0,1] \) (we implicitly assume that \( SC(\emptyset) = 0 \)), and \( T \) and \( M \) are scaling parameters. This ensures the function \( SC \) is a fuzzy measure, invariant with respect to translation, reflection and rotation, and discriminating compact clusters.

**Proposition 1.** The function \( SC \) in 6.2.1 is a fuzzy measure, invariant with respect to translation, reflection and rotation, and discriminating compact clusters.

**Proof.** Clearly, given the MST, pairwise distances are invariant with respect to the mentioned transformations and hence \( SC \) is also invariant. Let us show *monotonicity* and the range. Note that in our context the minimum distance between the elements of \( A \) is assumed to be one, and the largest is \( M \). We have

\[ \frac{W(A \setminus \{a\})}{M} \leq \frac{W(A) + M}{M} = \frac{W(A)}{M} + 1. \]
This implies that
\[
\frac{W(A \cup \{a\})}{|A|+1} + T \leq \frac{W(A)}{|A|+1} + T + 1 \leq \frac{W(A)}{|A|} + T.
\]
The last inequality follows from the fact that \( \frac{a+1}{b+1} \leq \frac{a}{b} \) for positive \( a, b \) such that \( a \geq b \).

We clearly have the numerator larger than the denominator as \( T \geq |A| \). The above inequality implies \( SC(A \cup \{a\}) \geq S(A) \) for every \( A \).

When \( A \) is a singleton, we have \( SC(\{a\}) = C - \frac{0+1}{1} = \frac{T(M+1)-1}{T^2 M} \). When \( A \) is the universal set, we have \( W(A) = |A|-1 \), and then
\[
SC(A) = C - \frac{T-1+1}{T} = 1.
\]
Hence we have shown \textit{monotonicity} and the range, i.e., \( SC \) is a fuzzy measure.

Now, consider two sets \( A, B \) of the same cardinality. Clearly, the less compact the set, the larger \( W(MST(A)) \) is. Since \( SC \) is anti-monotone in \( W \), more compact sets will give larger values of \( SC \). The proof is complete.

\textit{Example 2.} The numerical results illustrating the formula 6.2.1 are presented in Fig. 6.3. We observe \textit{monotonicity} with respect to set cardinality and differentiation between tightly compact and spread out sets of the same cardinality. One inconvenience of 6.2.1 is that the numerical values of \( SC(A) \) are clogged at the higher end of the spectrum near one. This means the differences between more and less compact sets of cardinality of more than three are in the second or third decimal place (see Fig. 6.3a). This can be rectified by raising \( SC(A) \) to some power (we used the fourth power) making the numerical values more evenly distributed and clearly differentiating numerical values between the more compact and less compact sets, as in Fig. 6.3b. The CPU time using this approach was negligible.
Figure 6.3: Estimated fuzzy measure using the Minimum Spanning Tree approach.
6.2.2 Sugeno-type Fuzzy Measure Approach

Our second approach is based on an analogue of Sugeno fuzzy measures, which we will call *Sugeno-type fuzzy measures*. We recall the definition of Sugeno fuzzy measures [101,130].

**Definition 5 (Sugeno \( \lambda \)-fuzzy measure).** Given a parameter \( \lambda \in ]-1, \infty[ \), a *Sugeno \( \lambda \)-fuzzy measure* is such a fuzzy measure \( v \) that for all \( A, B \subseteq \mathcal{N}, A \cap B = \emptyset \) it satisfies

\[
v(A \cup B) = v(A) + v(B) + \lambda v(A) v(B). \tag{6.2.2}
\]

Under these conditions, all the values \( v(A) \) are immediately computed from \( n \) independent values \( v(\{i\}), i = 1, \ldots, n \), by using an explicit formula

\[
v(\bigcup_{i=1}^{m} \{i\}) = \frac{1}{\lambda} \left( \prod_{i=1}^{m} \left( 1 + \lambda v(\{i\}) \right) - 1 \right), \quad \lambda \neq 0.
\]

While Sugeno fuzzy measures are popular due to their simplicity and a small number of parameters, they are too restrictive in our case as they will not differentiate between compact and non-compact groups (this information is not conveyed in the form of initial values of \( v \) at the singletons).

In our case, all \( v(\{i\}), i = 1, \ldots, n \) are the same (we can fix this common value at some small number \( \varepsilon \)). In contrast, we will use condition

\[
v(A \cup B) = v(A) + v(B) + \lambda_{AB} v(A) v(B), \tag{6.2.3}
\]

with

\[
\lambda_{AB} = \begin{cases} 
0 & \text{if } d_H(A,B) \leq 1 \\
\lambda & \text{otherwise.}
\end{cases}
\]

This means that the value of \( v(A \cup B) \) will increase more if the subsets \( A, B \) are geometrically close than when they are separated. Hence we have partial *subadditivity.*
only for separated subsets. Since a subset $C$ can be obtained by unions of different subsets $A$ and $B$, for a consistent definition we will use

$$v_{\lambda}(C) = \min_{A \cup B = C} v(A) + v(B) + \lambda_{AB} v(A) v(B).$$  \hfill (6.2.4)

Note that equation (6.2.4) implicitly accounts for translation, rotation and reflection symmetries because $d_H$ depends only on the intra-set distances and not on geometrical orientation.

The formula (6.2.4) has two parameters: the value of $\lambda \in ]-1,0]$ and a common value of $v$ at the singletons $\varepsilon > 0$. For a fixed pair $\lambda, \varepsilon$, we can compute the values of $v(A)$ for all subsets $A$. To find the suitable values of these parameters, we will fit them to the reference points and constraints we have specified.

Let us first fix some small $\varepsilon$ and fit $\lambda$, which we do by minimizing the objective

$$\text{minimize } F(\lambda) = \sum_{i=1}^{r} |v_{\lambda}(A_i) - v_i| + a \sum_{i=1}^{c} \max(0, -(v_{\lambda}(A_i) - v_{\lambda}(B_i) - \delta)), \hfill (6.2.5)$$

where $r$ is the number of reference points $(v(A_i), v_i)$ with the desired values $v_i$, $c$ is the number of constraints of the type $v(A_i) \geq v(B_i) + \delta$ and $a$ is a tradeoff parameter controlling the relative importance of fitting the reference points or constraints.

The objective $F$ is continuous as $v$ is a continuous function of $\lambda$, however it may have several local minima. We can find the global optimum numerically using the Pijavski global optimization method [241] (in particular implemented in GANSO software [31]).

Example 3. We illustrate the results of applying this approach using an example of a $3 \times 3$ block. Firstly, each pixel in the block is encoded using the index $i \in \{1, \ldots, 9\}$, from the top-left to the bottom-right corner of the block. We specify a reference point
for $v(\{1, 2, 3, 4, 5, 6\}) = 1$, and impose a constraint $v(\{1, 2\}) \geq v(\{1, 3\}) + 0.05$. The common value for $v(\{i\}), i = 1, \ldots, n$ is fixed at $\varepsilon = 0.2$.

We estimate $\lambda$ according to Equ. (6.2.5), where the tradeoff parameter $a$ is set to a default value 1. We obtained an optimal value $\lambda = -0.4792$ (which took 3 mins on a typical desktop computer). Fig. 6.4 shows the estimated values of the fuzzy measure for several representative clusters. It can be seen that for sets with the same cardinality more compact sets receive higher fuzzy measure values.

Remark 1. Since we specified a preference that $v(\{1, 2, 3, 4, 5, 6\}) = 1$, it is possible that in the fuzzy measures computed by (6.2.4) with the provided $\varepsilon$ and estimated $\lambda$, some values are greater than 1. In such cases, we convert these values to 1 by using $v = \min(v_\lambda, 1)$. Therefore, sets with cardinalities $k \geq 6$ will have fuzzy measure values of 1, as shown in the case of $v(\{1, 2, 3, 4, 5, 6, 7, 8, 9\})$. This is also done in the subsequent examples. All estimated values have a ceiling at 1.

Figure 6.4: Estimated fuzzy measure using the sub-additive Sugeno approach.
Further to fitting $\lambda$ we can also consider fitting $\varepsilon$ jointly with $\lambda$. This can be done in two ways:

1. To set up a bi-level optimization problem. At the outer level we perform a grid search for $\varepsilon$, whereas at the inner level we find the global minimizer for $\lambda$ having $\varepsilon$ fixed.

2. Solve a bi-variate global optimization problem using, for example, an Extended Cutting Angle method [22, 25, 27] (also implemented in GANSO [31]).

In both cases, however, we found the results were not very sensitive to the value of $\varepsilon$ and we can thus fix it at some “reasonable” value such as 0.2.

An alternative to using partially sub-additive measures in (6.2.3) is to use partially super-additive measures, given by (6.2.3) with

$$\lambda^+_{AB} = \begin{cases} 0 & \text{if } d_H(A, B) \geq 1 \\ \lambda > 0 & \text{otherwise.} \end{cases}$$

The fuzzy measure is constructed by using

$$v_\lambda(C) = \max_{A \cup B = C} v(A) + v(B) + \lambda^+_{AB} v(A)v(B). \quad (6.2.6)$$

In this case, rather than penalizing clusters that have separated pixels, we favor clusters which have more compact elements. Fitting the parameter $\lambda$ is done by solving problem (6.2.5) by restricting $\lambda$ to $[0, \infty)$.

**Example 4.** We illustrate the results of applying this approach using a similar setting to the partially sub-additive case: a reference point $v(\{1, 2, 3, 4, 5, 6\}) = 1$ and a constraint $v(\{1, 2\}) \geq v(\{1, 3\}) + 0.05$. For the common value for $v(\{i\}), i = 1, \ldots, n$ we fix it to $\varepsilon = 0.05$. We obtained an optimal value $\lambda = 9.6766$ (which again took
approximately three minutes of CPU time on a standard desktop PC). Fig. 6.5 shows the estimated values of the fuzzy measure for several representative clusters.

Let us examine the computational complexity of the proposed method for the solution. We note that the CPU time for solving (6.2.5) is proportional to $2^n p(r + c)$, where $p$ is the number of iterations of the Pijavski method, by which we control the quality of the minimizer $\lambda$.

Our typical CPU time (using Matlab and GANSO library) was around 3 minutes for $n = 9$. This is not excessive, as only one such fitting process is required for our application. However, for larger $n$, the CPU time may become a bottleneck. For instance for $n = 25$, it took us around 1 day to fit the fuzzy measure.

In the following sections, we explore other alternatives to constructing fuzzy measures with the desired properties, which are less CPU intensive.
6.2.3 Decomposing Fuzzy Measure Approach

This section presents another approach to constructing fuzzy measures with suitable properties based on geometrical decomposition of the sets involved. We will explicitly account for geometrical symmetries of the clusters by decomposing them into blocks. Consider Fig. 6.6 which shows examples of 4 building blocks of subsets, called the basic components. Each subset \( \mathcal{A} \) can be represented in a number of ways through the (possibly overlapping) basic components \( BC_i \). The basic components account for all geometrical symmetries and are encoded using a hash function based on the average intra-pixel distance

\[
h(A) = \frac{1}{|A|} \sum_{a \in A} d_H(a, \mathcal{A} \setminus a).
\]

The number of basic components and their shape can be specified by users according to their own preference and applications.

![Figure 6.6: Examples of basic components \( BC_1 - BC_4 \).](image)

For a given subset \( \mathcal{A} \) we identify its decompositions into one or more basic components of the same type and identify the number of possible ways to fit such a basic component into \( \mathcal{A} \). Each of the basic components \( BC_i \) is assigned a value \( u_i \). The values of \( \nu(\mathcal{A}) \) are computed as follows

\[
SC(\mathcal{A}) = \min(1, n_1 u_1 + n_2 u_2 + \ldots n_q u_q),
\]

(6.2.7)
where \( q \) is the number of basic components, \( u_i \) is the value of the \( i^{th} \) basic component, and \( n_i \) is the number of the \( i^{th} \) basic components in the decomposition of \( A \).

Suppose a subset \( A = \{1, 2, 4\} \) is a cluster with 3 connected elements in a \( 3 \times 3 \) mask. There are 3 possible ways to fit 1-element component, 2 ways to fit a 2-elements component, and 1 way to fit a 3-elements component into the provided shape. Note that overlapping is allowed. Assume that the weights of the basic components are provided as \( u = [0.1, 0.1, 0.05, 0.05] \). Thus, the value of the fuzzy measure in our example is computed as:

\[
v(A_{\text{example}}) = \min(1, 3u_1 + 2u_2 + u_3 + 0u_4) \\
= \min(1, 3 \times 0.1 + 2 \times 0.1 + 0 + 0 \times 0.05) \\
= 0.55
\]

Unlike the MST approach, the fuzzy measure decomposition approach computes the fuzzy measure based on a model with the values \( u_1, \ldots, u_q \). These values are estimated with respect to available reference points and constraints, which represent the desired numerical values and relations between the values of \( SC \). Before we proceed with formulating a fitting problem, let us demonstrate that the function \( SC \) is a fuzzy measure. Let \( U \) denote the universal set (the largest cluster – the whole block \( m \times n \)).

**Proposition 2.** The function \( SC \) in 6.2.7 is a fuzzy measure irrespective of the values \( u_i \in [0, 1]|\sum_i n_i(\mathcal{U})u_i \geq 1 \). \( SC \) discriminates between more and less compact clusters.

**Proof.** Evidently \( SC(\emptyset) = 0 \) and \( SC(\mathcal{U}) = 1 \). We need to show monotonicity. For this we note that \( A \subset B \) implies \( n_i(\mathcal{A}) \leq n_i(\mathcal{B}) \). Consequently \( \sum_i n_i(\mathcal{A})u_i \leq \sum_i n_i(\mathcal{B})u_i \). Next, as more compact subsets \( A \) allow fitting of more basic components that are larger in size, they will have larger values of \( n_i \), and hence larger values of
We now need to fit the parameters \( u_i \) to the available reference points and constraints. Because the parameters enter our expressions linearly we can set up the following mathematical programming problem

\[
\text{minimize } F(u_1, \ldots, u_q) = \sum_{i=1}^{r} |SC(A_i) - v_i| \\
\text{subject to } SC(A_i) - SC(B_l) \geq \delta_l, \ l = 1, \ldots, c, \\
SC(A_k) = \min(1, \sum_{i=1}^{q} n_i(A_k)u_i), \\
\sum_{i=1}^{q} n_i(U)u_i \geq 1, \\
u_1, \ldots, u_q \geq 0.
\]

(6.2.11)

which is subsequently converted, by introducing slack variables \( r_i^+, r_i^- \), and setting \( SC_i = SC(A_i) \), into a linear programming problem

\[
\text{minimize } \sum_{i=1}^{r} r_i^+ + r_i^- \\
\text{subject to } r_i^+ - r_i^- - SC_i = -v_i, \ i = 1, \ldots, r, \\
-SC_i + SC_k \leq -\delta_{ik}, \\
SC_k \leq \sum_{i=1}^{q} n_i(A_k)u_i, \\
SC_k \leq 1, \\
\sum_{i=1}^{q} n_i(U)u_i \geq 1, \\
r_i^+, r_i^- \geq 0, SC_i \geq 0, \\
u_1, \ldots, u_q \geq 0.
\]

(6.2.12)

The number of variables is \( 2r + c + q + t \) and the number of constraints is \( r + c + 2t + 1 \), where \( t \) is the total number of subsets engaged in all the constraints and reference

\( SC(A) \).
points. Note that this number is smaller than the number of all possible subsets $2^n$, because we care only about the values $SC_i$ for those subsets. The values $\delta_{ik} = \delta_l$ correspond to the pairs of sets $(A_l, B_l) = (A_i, A_k)$ in the $l^{th}$ inequality constraint.

The problem is solved by using the standard simplex method.

**Example 5.** We illustrate the output of this approach using an example of a $3 \times 3$ block. We provide a reference points $v(\{1, 2, 3, 4, 5, 6\}) = 1$, and impose a constraint $v(\{1, 2\}) \geq v(\{1, 3\}) + 0.05$. We use 4 basic components as in Fig. 6.6 to construct the fuzzy measure. Their weights are estimated by solving Problem 6.2.12.

We obtained the values $u = [0.0042, 0.0799, 0.0228, 0.1169]$. Fig. 6.7 shows the estimated values of the fuzzy measure for several representative clusters.

![Figure 6.7: Estimated fuzzy measure using the decomposition approach.](image)
6.3 Experiment and Analysis

This section first compares the approaches to highlight the merit of each presented approach. Their performance is then evaluated against the image reduction task. Finally, their application on satellite image reduction is also presented.

6.3.1 Comparison of the Proposed Measures

The first approach, based on Minimum Spanning Tree (MST), is a simple one, wherein fuzzy measures are computed directly from the weights of the edges without user input. However, this does not allow a user to provide reference points or constraints. Thus, it is not suggested for applications where users might want to flag clusters with specific shapes or impose some other domain dependent constraints.

The approach based on Sugeno-type fuzzy measures is more flexible than the approach based on Minimum Spanning Tree, as it allows users to specify reference points and constraints. This is due to the optimization process for estimating the parameter $\lambda$, together with the value $\varepsilon$, or with the provided value of $\varepsilon$. A shortcoming of this approach is its computational cost for generating all fuzzy measure values at every iteration of the optimization process. CPU time may become a bottleneck for a large image block.

In addition, a special feature of the MST and Sugeno approach is that the estimated fuzzy measure values are not only invariant with respect to translation, reflection and rotation, but also invariant with respect to the shape of clusters with connected elements. Fig. 6.8a shows the computed fuzzy measure for the MST approach,
computed using the default setting. Fig. 6.8b demonstrates this aspect for the super-additive Sugeno fuzzy measure approach whose values are computed based on $\varepsilon = 0.05$, $\lambda = 9.6766$. It can be observed that different cluster shapes with the same number of connected elements have the same fuzzy measure value, $v(\{1, 2, 4\}) = v(\{1, 2, 3\})$ and $v(\{1, 2, 3, 4, 5\}) = v(\{1, 2, 3, 6, 9\})$. It is suggested for use in situations where there is no need to discriminate between cluster shapes.

On the other hand, if a user wishes to impose preferences with respect to cluster shapes, the Decomposing Fuzzy Measure (DFM) is a suitable alternative due to its computational flexibility based on the basic components. For example, if we desire that a more dense cluster receives a higher value of the fuzzy measure for those clusters of connected elements with the same cardinality, we impose constraints $v(\{1, 2, 4\}) \geq v(\{1, 2, 3\}) + 0.01$ and $v(\{1, 2, 3, 4, 5\}) \geq v(\{1, 2, 3, 6, 9\}) + 0.01$. Four basic components as in Fig. 6.6 are used, and one reference point is provided, $v(\{1, 2, 3, 4, 5, 6\}) = 1$. We estimate the weights of the basic components by solving Problem 6.2.12 and obtain the values $u = [0.0106, 0.0321, 0.0621, 0.1075]$. Fig. 6.8c displays the computed values for the decomposition approach. It is obvious that all the constraints describing shape preferences are satisfied.

We further expect the behavior of the proposed methods by plotting the estimated fuzzy measure values, for the $3 \times 3$ case (see Fig. 6.9). Figures 6.9a, 6.9b, and 6.9c show the values for the MST approach with original, powered 2 and powered 4 values. Its values increase significantly along with the size of the clusters. Clusters with the same cardinality have relatively similar values, which indicate the MST approach has low discriminative power for clusters with the same cardinality but different shapes. Fig. 6.9d and Fig. 6.9e show better discriminative power for the Sugeno-type Fuzzy
Figure 6.8: Fuzzy measure of different cluster shapes of the same cardinality.

Measure approach and the Decomposing Fuzzy Measure approach, as indicated by variations of fuzzy measure values for clusters of the same cardinality.

6.3.2 Application for Image Reduction

The image in Fig. 6.10 depicts a series of concentric circles as thin curves with having a width of one pixel. This pattern contains a large variety of cluster patterns within small 3×3 blocks and in performing a reduction on this image, any operator must cope with the problem that the image detail represents a minority of pixels at all scales. This image is also easily assessed visually for global continuity of the curves, making it useful for comparative evaluation of various reduction operators. The image $I$ is constructed from the image $I = \max(C,F,B)$ where $C$ is a binary image depicting the circles, $F$ is a foreground noise field and $B$ is a background noise field. These fields were generated as uniformly distributed 8 bit pixel intensities in the ranges [0, 10] and [245, 255] respectively.

We constructed a local block-based reduction operator based on Equ. 6.1.1, by
Figure 6.9: Fuzzy measure values for cluster compactness.
replacing the individual pixel weights $w_i(y)$ with a cluster-based weight, such that

$$P(x, y) = w(A) \sum_{i=1}^{p} p(x_i, y)$$  \hspace{1cm} (6.3.1)$$

where $w(A) = 2 - v(A)^q \geq 1$, $v$ is a fuzzy measure computed using either of the aforementioned approaches and $A$ denotes the subset of pixels having intensities that satisfy $r_{(k)} < \tau$ as per Equ. 6.1.2. These pixels define a candidate cluster within the tonal space of the block for the given value $y$. This choice of penalty function means that for two candidate clusters of equal cardinality and equivalent pixel intensity differences within the cluster, the more spatially coherent cluster will have a lower penalty and thus be preferred as the significant cluster of the block. Conversely, if the spatial patterns are equivalent, the cluster with more compact tonal range will be preferred. As with the method described in [297], the candidate average values $y$ are taken from the set of input pixel values for that block so the output image is a proper subset of the input image.
We constructed versions of the penalty-based averaging function using the proposed fuzzy measures, given by equations 6.2.1 and 6.2.7, and denoted herein as MST and DFM respectively. For the MST measure, we selected the exponent value $p = 200$ to ensure the measure covered the full range $[0, 1]$. For the DFM measure, we took $p = 1$ since this function already provides a well-distributed set of measure values in $[0, 1]$. Our cluster-based mode-like average was compared against the mode-like average using distance-based weights and the Shorth function (given in [297]) as well as monotonic reduction operators using the arithmetic mean and median. The parameters for the mode-like averages were $\alpha = 12$ and $\beta = 0.3$.

Image reduction was performed over disjoint blocks of size $3 \times 3$ (producing a $\frac{1}{5}$ scale image) and the resulting images are shown in Fig. 6.11. It is apparent from these results that the median based reduction operator does not preserve the relevant curves of one pixel thickness. This is to be expected since within a $3 \times 3$ block the

Figure 6.11: Reduced circle image using cluster-based mode-like averaging functions.
high valued pixels on the curve would appear as a minority of outliers against the more common background. In the case of the arithmetic mean, while the circles are preserved, their peak intensity is diminished and they are spatially broadened, which is to be expected. While the resulting image contains the desired visual continuity of the curves, the structural detail of these curves (specifically radial gradients) has been corrupted by the reduction.

The mode-like average using distance-based weights outperforms both the mean and median operators and preserves sections of the curves. This indicates that certain cluster patterns will result in an aggregate selected from the cluster, whereas others are not significant enough to be preferred over the background pixels. In such cases, the minimization of the penalty favors reducing the number of outliers, even though they may be centrally located within the local block.

The cluster-based mode-like averaging functions, built using the various fuzzy measure construction methods, outperforms the other averaging functions on this test image, including the distance-based mode-like average, which is most similar to it. Interestingly, it is able to preserve nearly all continuous curves. The missing sections appear to be associated with sections of the original curves that were locally corrupted by noise or with contained simple linear clusters of bright pixels (such as vertical lines). In such cases the background pixels would form a larger contiguous cluster and would be preferred in the penalty minimisation, since the image detail would represent a minor cluster of outliers.
6.3.3 Satellite Image Reduction Example

To test the practicality of these reduction operators and their ability to preserve fine, pixel scale detail, the operators were applied to a real satellite image. Fig. 6.12 depicts the seaport, showing many fine detailed objects, including ships and boats with wakes, shipping containers, houses and other buildings, roads and parks.

![Figure 6.12: A satellite image of a sea port.](image)

Each of the averaging functions described above was tested using a variation on the disjoint block reduction algorithm. In this case averages were computed within overlapping 5\(\times\)5 regions, where the inner 3\(\times\)3 blocks of adjacent regions were disjoint (i.e., the overlap was a single row or column). The subsequent average was used to represent the inner block in the reduced image, which was 1/9\(^{th}\) scale. This approach was chosen as it indicated better structure preservation than using the averages of disjoint blocks.

Given the similarity of results for reducing circles using the cluster-based mode-like
average, only the MST approach is tested herein. Fig. 6.14a shows the reduction using our proposed cluster-based weights, as well as reductions performed using the mode-like average with distance-based weight, and averages computed using the Shorth, median and arithmetic mean functions. It should be noted that all images (the original, as well as the reduced versions) have been scaled for showing purposes, although this scaling has been performed using the same in-built algorithms within LaTeX. While there will be some affects on the resulting images, the differences between them are still apparent and therefore we can assess the visual quality of the local block-based reduction algorithms based on these differences.

It is apparent that the mode-like averages produce images that retain fine detail, whereas the monotonic averaging functions operate as low pass filters and smooth the image, thus removing detail. For example, the small boats in the central channel are poorly preserved in Fig. 6.13a and Fig. 6.13b. Detail such as the shipping containers (right side dock), the boats in the marina and the houses are better preserved by the mode-like averages (Figures 6.14b and 6.14b). The Shorth performs better than the monotonic averages, but not as well as the mode-like averages. Like the monotonic averages, the Shorth image is smoothed and fine detail has been lost. It is apparent from these tests that the mode-like operators not only preserve the fine detail present in each local block, but also retain information across neighboring blocks to represent larger coherent structures in the final image. Both the cluster-based and distance-based weights produce very similar results on this image.
6.4 Summary

In this work, the challenges in satellite image reduction with specific reasonable properties have been tackled to support tourism applications development and travel behavior analysis tasks. Several approaches were proposed to characterize the compactness of a single geometrical cluster of spatially organized values, using a fuzzy measure. This is necessary in image processing, where we need to distinguish between compact groups of pixels (of similar intensity) representing an object, and scattered groups likely to be noise. One approach involves a suitably scaled weight of the minimum spanning tree graph. This graph is simple to construct numerically and requires negligible computational capacity, however it is less flexible than the alternatives that involve fitting the measure to user specified reference points and constraints. The second is based on Sugeno-type fuzzy measure, which fits the fuzzy measure using the
Figure 6.14: Reduction of the Port of LA image using non-monotonic averages.
non-smooth global optimization technique. The third approach is based on the decomposing clusters into simple geometrical (basic) components. Fitting of the fuzzy measure to the power set is achieved by solving a small scale linear programming problem.

The application of our construction is within image filtering and image reduction, where mode-like non-monotonic averages have been applied to remove noise while preserving the fine details of the images, such as in the case of satellite image reduction. Initial results validating the use of a fuzzy measure based penalty weight were presented, which showed it is capable of preserving pixel-scale details in images corrupted by impulse noise and with reasonable variation in intensity values within both the background and foreground classes. This new approach was compared to previously published methods involving both monotonic and non-monotonic averaging functions and found that our new non-monotonic averaging function out-performs these previous methods on our test images.

This work contributes toward facilitating the development of tourism applications based on GPS technology and satellite imaging on end user devices, in order to enhance customer experience and provide better services to the tourism industry.
Tourism managers are interested in understanding tourist preferences with the aim of improving their strategic planning, marketing and product development. However, tourist preferences can be unstable and dynamic. Changes in preference influence the performance of tourism businesses, thus creating the need to identify and address the demands of their customers. Most existing studies focus on current demand attributes but not on emerging attributes. Tourism managers may find it difficult to make appropriate decisions in response to changes in tourists’ concerns.

Due to such demand, some data mining techniques have been proposed to study the changes in tourist’ behavior. Namely, Kim et al. utilized decision tree analysis to detect changes of customer behavior [147], and Chen et al. adopted Association Rule Mining to study changes in customer behavior in retail marketing [56]. Huang used fuzzy time-interval sequential patterns to detect changes in customer behavior in dynamic markets. Law et al. identified the changes and trends of outbound tourist behavior to Hong Kong using Contrast Positive and Negative Association Rule Mining [166]. Liu et al. used Targeted Positive and Negative Association Rule
Mining to study the changes in hotel tourists’ expectations by trip mode [189]. These techniques are effective for capturing the behavior pattern of tourists and constructing their profile. However, they are not designed in detecting emerging changes in tourist preference, where emerging changes between different data set need to be detected.

In dealing with emerging changes, Emerging Pattern Mining (EPM) appears to be a promising approach. EPM was originally proposed to capture emerging trends in time-stamped databases or to explore differentiating characteristics between groups of data [84]. An emerging pattern concept was also adopted in some work for customer behavior analysis. For instance, Kim et al. used it to develop methodology to detect the change in customer segments between time-stamped data sets [147]. Tsai and Shieh developed a method to explore emerging sequential patterns to explore the trends in consumer behaviors [281]. Shie et al. focused on discovering user behavior patterns from mobile commerce environments using mobile sequential pattern mining [264]. It would be advantageous to adopt Emerging Pattern Mining in detecting changes in tourists’ preferences. An issue with EPM is that it was originally developed to detect any emerging changes between groups of interest without indicating if such changes were increasing or decreasing change. The knowledge about increasing or deceasing change is important in application to the tourism industry. Tourism managers can make appropriate decisions to invest in whether a product receives increased attention, or direct resources away so a product receives less attention. Thus, we propose to adopt the positive and negative concepts of Association Rule Mining into Emerging Pattern Mining, as these concepts can be used to reflect the increasing or decreasing changes in customer behavior [166,189].
This chapter aims to address the challenges in emerging hotel preference identification. An effective method, *Positive and Negative Emerging Pattern Mining* (PN-EPM), is presented for detecting the emerging preferences of tourists as. The application focuses on identifying emerging tourist preferences towards tourism products such as hotel feature preference. In order to capture tourist behavior, online hotel reviews are a suitable data resource for preference analysis because they contain rich information about travel-related experiences, opinions, and concerns of tourists. The term “hotel features” includes any entity or concept that concerns tourists when reviewing a hotel.

Having introduced the background of this work, the remaining part of this chapter is organized as follows. Section 7.1 provides our problem statement for the task of emerging hotel preference identification. Section 7.2 presents detail on the PN-EPM technique together with a text processing framework for hotel feature list construction. Section 7.3 demonstrates the performance of the proposed techniques in an application of emerging hotel preference identification. Finally, Section 7.4 concludes the chapter and highlights the contributions.

### 7.1 Problem Statement

Several challenges must be addressed to analyze the changes in travelers’ concern on hotel features using online reviews. In the existing works, researchers usually have prior knowledge about what hotel features to be included for analysis. Hotel features are often represented by short-answer questions or a set of keywords. Selecting keywords that appropriately and accurately describe hotel characteristics is a challenging task. Researchers may label hotel features differently despite their shared
similarities. This is even a more challenging task in emerging hotel preference analysis, because researchers usually have no prior knowledge about what is of interests to tourists for including in the analysis. A possible solution would be to incorporate text-mining techniques into the analysis of hotel reviews. This approach can extract useful knowledge from unstructured text and then transform the information into structured data for analysis, thereby revealing relationships, patterns, or trends from textual data [265]. Therefore, we present a text processing framework that can automatically identify hotel features from review comments. In the analysis of hotel features, statistical tests can be used to examine changes and trends in preference (see, e.g., [41]). However, when a large number of variables are considered, such analysis on every available feature becomes inefficient and costly. Thus, the concept of PN-EPM is presented and described how it can be used to discover emerging hotel features.

7.2 Methodology

7.2.1 Hotel Review Processing Framework

Our review processing framework is based on GATE, a general architecture for text engineering. GATE provides a large number of utilities and resources for building an information extraction application. We make use of some of its resources, especially the English lexicon, to develop a method for hotel feature extraction.

Suppose a data set $R$ contains $m$ review comments $R = \{r_1, r_2, \ldots, r_m\}$. The process of identifying hotel features mentioned in the reviews $r_i \in R$ is carried out in two major stages, as shown in Fig. 7.1.
Figure 7.1: Hotel Features’ Identification Process.
Text processing. Transforms the unstructured text into an useful data format. In detail, each review $r_i$ is first loaded into a Text Tokenizing algorithm, where the stream of text is broken into words, phrases, symbols, or other meaningful elements called tokens. The tokens for each review is then passed through a Text Filter, where capital letters are normalized to lower case. Tokens containing symbols or numbers are removed because they are irrelevant to the hotel feature analysis. The remaining tokens are input into a Stemming process for reducing inflected words to their stem, base or root form. For instance, a stemming algorithm can reduce the words, cleans, cleaning, cleanliness and cleaned to the root word clean. This allows for the recognition of hotel features when they are mentioned using different word forms. A stemmed token list $S(i) = \{s_1^{(i)}, s_2^{(i)}, \ldots \}$ is constructed for each review $r_i$, and saved in to a processed document database. In the application of hotel feature identification, it is a natural assumption that the English vocabulary of noun type is commonly used to refer to entities such as hotel features (eg. room, location, view, service, staff). Therefore, we identify and construct a stemmed noun list $N = \{n_1, n_2, \ldots, n_o\}$ appearing in the review corpus, using the English lexicon of GATE. The lexicon resource is used as a lookup database, with approximately 63 thousand words used in English. Each word is also accompanied with a set of tags, which help determine its type, such as noun, verb, or adjective. This noun list will be used for identifying candidate hotel features in the next stage.

Hotel Feature Candidate Identification. This identification selects interesting nouns $n_i \in N$ for further analysis as potential hotel features. In detail, a binary vector $v^{(i)} = \{v_1^{(i)}, v_2^{(i)}, \ldots v_o^{(i)}\}$ is constructed for each stemmed token list $T_i$, where $v_{(j)}^{(i)}$ takes a value of 1 if $n_j \subset T_i$, or 0 otherwise. The interestingness of each noun
\( n_i \in N \) is evaluated by a support value:

\[
supp(n_j) = \frac{\text{count}(n_j)}{|R|} \quad (7.2.1)
\]

where \( \text{count}(n_j) \) is the count of vector \( v^{(i)} \) whose values \( v_j^{(i)} = 1 \), \(|R|\) is the total number of records in the data set \( R \). We use a user-specified minimum support, or \textit{support threshold} \( (\delta_s) \), to measure the significance of the nouns in the review corpus. If a noun \( n_j \) satisfies \( supp(n_j) \geq \delta_s \), it is selected into a hotel feature candidate list, otherwise, it is removed.

The advantage of this method is that users do not need to provide a set of pre-defined keywords to identify and extract hotel features for analysis. Instead, a list of hotel feature candidates is automatically constructed from the review comments. All potential features mentioned in the reviews are taken into account, and interesting candidates are returned. Here, the support threshold \( \delta_s \) eliminates insignificant features, while retaining potentially interesting ones for further analysis. The hotel review processing method is applied to an online review corpus to construct a set of features, then PN-EPM is applied to identify the emerging hotel features.

### 7.2.2 Positive and Negative Emerging Pattern Mining

Let \( F = \{f_1, f_2, \ldots, f_o\} \) be a set of items, a subset \( X \subseteq F \) is called a \textit{k-itemset}, where \( k = |X| \). Given a number of groups \( \{G_1, G_2, \ldots\} \), the supports of an item set \( X \in G_i \), denoted as \( supp(X, G_i) \), reflect how frequently \( X \) appears in this group. The change in the support of \( X \) from a group \( G_i \) to a group \( G_j \) is measured by a growth
rate metric.

\[
\text{GrowthRate}(X, G_i, G_j) = \begin{cases} 
0, & \text{if } \text{supp}(X, G_i) = 0 \text{ and } \text{supp}(X, G_j) = 0 \\
\infty, & \text{if } \text{supp}(X, G_i) = 0 \text{ and } \text{supp}(X, G_j) \neq 0 \\
\frac{\text{supp}(X, G_i)}{\text{supp}(X, G_j)}, & \text{otherwise}
\end{cases}
\]

(7.2.2)

Given \(\delta_e > 1\) as a growth rate threshold, an item set \(X\) is called an emerging pattern if it satisfies the condition:

\[
\max_{i,j}\{\text{GrowthRate}(X, G_i, G_j)\} \geq \delta_e
\]

(7.2.3)

Several issues must be considered when using EPM in hotel features analysis. Interpreting a hotel feature on its own rather than as part of an item set containing many others is easier and more meaningful. When the number of features is large, the use of item set \(X\) with one hotel feature \((k = 1)\) is suggested. Equ. 7.2.2 does not take into account the order of groups \(G_i\) and \(G_j\), although it is important in detecting changes in travelers’ concerns over time. For example, we group the customer data according to the years, therefore \(G_1\) contains records collected in the year 2012, and \(G_2\) contains records collected in the year 2013. An increase in the rate of change for product feature preference from 2012 to 2013 would occur if \(G_1\) is considered before \(G_2\), and this would mean something different from a decrease if \(G_2\) is considered before \(G_1\). Therefore, we define the concepts of positive and negative emerging patterns to distinguish between increasing and decreasing amounts of change in the concerns expressed by travelers.

Let \(G_i\) be an initial group, and \(G_j\) be a target group \((i \neq j)\). An item set \(X\) is a positive emerging pattern if it satisfies \(\max_{i,j}\{\text{GrowthRate}(X, G_i, G_j)\} \geq \delta_e\) and \(\text{supp}(X, G_i) < \text{supp}(x, G_j)\). \(\delta_e\) is a user defined emerging threshold. The growth rate
GrowthRate\((X, G_i, G_j) = \frac{\text{supp}(X, G_j)}{\text{supp}(X, G_j)}\) is called positive growth rate. On the contrary, X is a negative emerging pattern if it satisfies max\(\{\text{GrowthRate}(X, G_i, G_j)\}\) ≥ \(\delta_e\) and supp\((X, G_i) > \text{supp}(x, G_j)\). The growth rate \(\neg\text{GrowthRate}(X, G_i, G_j) = \frac{\text{supp}(X, G_i)}{\text{supp}(X, G_j)}\) is called a negative growth rate and indicated by a negative sign (\(\neg\)). Let us take a simple example to illustrate the use of EPM in applications of hotel feature preference analysis.

7.2.3 An Illustrative Example

Given a set of hotel features including price, location, room, price, service and view, a sample data set of 20 records was constructed as shown in Table 7.1. Here, each record is represented as a vector \(v(i)\), where each element \(v_{(j)}^{(i)}\) takes a value of 1 if its corresponding hotel feature is mentioned in the review \(r(i)\), and 0 otherwise. The year attribute represents the time when the reviews were created. It indicates a group which a record belongs to, either 2012 or 2013. The support of each hotel feature has been computed for two groups. Growth rates are also computed reflecting the changes in tourists’ concerns, as in Table 7.2.

In this example, we set the emerging threshold \(\delta_e = 2\). The result in Table 7.2 can be interpreted as follows. There is an emerging increase for the price aspect of hotels from the year 2012 to the year 2013, as indicated by a growth rate of 2.0. A considerable growth in tourists’ concern is also found for the room features. Hotel location have received significantly less attention from tourists as indicated with a negative growth rate of \(-2.50\). service and view are not emerging feature, because their growth rates are less than the emerging threshold.

Please note that the focus of EPM is any significant change in the support values.
Table 7.1: A sample data set.

<table>
<thead>
<tr>
<th>ID</th>
<th>price</th>
<th>service</th>
<th>location</th>
<th>view</th>
<th>room</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2012</td>
</tr>
<tr>
<td>r2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2012</td>
</tr>
<tr>
<td>r3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2012</td>
</tr>
<tr>
<td>r4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2012</td>
</tr>
<tr>
<td>r5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2012</td>
</tr>
<tr>
<td>r6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2012</td>
</tr>
<tr>
<td>r7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2012</td>
</tr>
<tr>
<td>r8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2012</td>
</tr>
<tr>
<td>r9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2012</td>
</tr>
<tr>
<td>r10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2012</td>
</tr>
<tr>
<td>r11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2013</td>
</tr>
<tr>
<td>r12</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2013</td>
</tr>
<tr>
<td>r13</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2013</td>
</tr>
<tr>
<td>r14</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2013</td>
</tr>
<tr>
<td>r15</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2013</td>
</tr>
<tr>
<td>r16</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2013</td>
</tr>
<tr>
<td>r17</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2013</td>
</tr>
<tr>
<td>r18</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2013</td>
</tr>
<tr>
<td>r19</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2013</td>
</tr>
<tr>
<td>r20</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2013</td>
</tr>
</tbody>
</table>

Table 7.2: A list of emerging hotel features.

<table>
<thead>
<tr>
<th>Features</th>
<th>Support</th>
<th>Year 2012</th>
<th>Year 2013</th>
<th>Growth Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>price</td>
<td>0.4</td>
<td>0.8</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>service</td>
<td>0.7</td>
<td>0.6</td>
<td>-1.16</td>
<td></td>
</tr>
<tr>
<td>location</td>
<td>0.4</td>
<td>0.1</td>
<td>-2.50</td>
<td></td>
</tr>
<tr>
<td>view</td>
<td>0.4</td>
<td>0.6</td>
<td>1.50</td>
<td></td>
</tr>
<tr>
<td>room</td>
<td>0.3</td>
<td>0.8</td>
<td>2.66</td>
<td></td>
</tr>
</tbody>
</table>
between groups, instead of the actual support values themselves. Standard $\chi^2$ tests can also be applied to verify the significance of the supports between the groups in real applications.

7.3 Application: Identification Emerging Hotel Feature Preference

Changes in tourists’ concerns can have a huge influence on the performance of hotel businesses. It is necessary for managers to be able to quickly and effectively identify features which are becoming important to tourists. This section presents the application of PN-EPM in addressing the challenge of emerging hotel feature preferences identification. We start with a description of our data collection from online hotel reviews. Next, a list of hotel features are constructed using the presented text processing framework. We present the analysis in results about the changes of customer preferences concerning hotel features, which is followed by a discussion on practical implications of the findings.

7.3.1 Data Collection

The data set, used in this work, is collected from TripAdvisor (www.tripadvisor.com), which is one of the most popular travel review web sites. It has been used popularly as a data resource for research on hotel selection criteria [185,189]. Review contents are extracted using Visual Web Ripper¹. The data extraction process focuses on reviews for hotels in Hong Kong, Singapore, Shanghai, Bangkok and Sydney, as these are

¹www.visualwebriper.com
major Asia Pacific tourist destinations for international tourists. Around 1740 hotels were included for review extraction, ranging from 1 star to 5 stars according to the TripAdvisor rating. The software navigated through each travel review and extracted its text comment, post date, together with demographic data about tourists such as their travel model (business, couple, family, friends or solo) and country of origin.

In tourism research, culture has been identified as an important factor influencing the behavior of people and their decision making [282], particularly in hotel evaluation [174]. It was also suggested that international tourists from different continents are likely to have different backgrounds [143]. Therefore, the location of reviewers are grouped according to continents of origin, which makes it convenient for later analysis. It is noticed that the majority of reviewers in our dataset are from North America, Europe, Asia, and Oceania, thus, only these continents are considered in this study. Most collected reviews were posted in recent years (2010 to 2013), while a small number of reviews were posted in 2009 or before. In order to make it convenient for writing, we refer to reviews in these years as belonging to the group of 2009. Records with missing attributes were removed from our data collection. This left us with a data set of a total of 118,300 review records. A detail description of the data set is presented in Table 7.3. This data set will be made available for academic use.

7.3.2 Result Analysis

**Hotel Feature List Construction.** We apply the proposed review processing method to the reviews posted in 2013. A stemmed noun list is constructed based on the English lexicon. The support threshold $\delta_s$ is then used to determine if a noun
Table 7.3: Description of the Collected Data set.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Value</th>
<th>Percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMMENT</td>
<td>Text comment of each hotel review</td>
<td>Review Text</td>
<td>100</td>
</tr>
<tr>
<td>YEAR</td>
<td>Year when the reviews were posted online</td>
<td>2013</td>
<td>26.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2012</td>
<td>39.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011</td>
<td>17.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2010</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2009 and before</td>
<td>8.39</td>
</tr>
<tr>
<td>DESTINATION</td>
<td>Location of hotels whose reviews were collected</td>
<td>Bangkok</td>
<td>23.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hong Kong</td>
<td>23.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shanghai</td>
<td>12.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Singapore</td>
<td>21.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sydney</td>
<td>18.61</td>
</tr>
<tr>
<td>ORIGIN</td>
<td>Location of reviewers according to continents</td>
<td>Asia</td>
<td>31.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Europe</td>
<td>16.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>North America</td>
<td>21.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oceania</td>
<td>30.03</td>
</tr>
<tr>
<td>GROUP</td>
<td>Travel mode of reviewers</td>
<td>Business</td>
<td>24.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Couple</td>
<td>35.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Family</td>
<td>19.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friends</td>
<td>10.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Solo</td>
<td>9.34</td>
</tr>
</tbody>
</table>

in the list is of interest for further analysis. We first examine the effect of setting different support thresholds between 0 and 0.1 to the number of hotel features returned as candidates.

Fig. 7.2 shows that the algorithm returns 5523 features with $\delta_s = 0$, which is the total number of nouns in the stemmed list. The hotel feature candidate number drops gradually to 419 when $\delta_s$ is set to 0.01, then decreases slightly with higher support thresholds. When $\delta_s = 0.1$, only 47 feature candidates are returned. Notably, the candidate generation is an automated process, thus, users may examine the output further to select their features of interest. This approach should be feasible in practice because the candidate number is usually small. The advantage of this work is that hotel features of interest are identified directly from reviews rather than from a predefined set. Hence, this condition allows for a more comprehensive and relevant
Next, we select $\delta_s = 0.05$ to generate a list of hotel feature candidates for further analysis. This is a popular value to evaluate the extent to which items in a data set are interesting [166, 177]. This analysis resulted in 111 candidates. We inspect the list and find that several popular hotel features are included (Fig. 7.3), including room, staff, location, breakfast, service, and cleanliness. This result is reasonably consistent with previous studies [51, 270], which indicates the effectiveness of our approach.

In addition, a number of other features of interest to tourists are identified. These detailed aspects of the hotel include lobby, lounge, door, coffee, tea, or the surrounding environment including road, street, park, river, and space. In particular, factors such as station, airport, taxi, train, and access are found to be significant to reviewers. Their support values are similar to or higher than some
of the popular features such as internet, club, reception, price, and bar. This is an interesting finding given that these words are about transportation which is not directly related to the hotel domain. However, they play an important role in the hotel evaluation process of international tourists. Prior work on hotel features focuses mainly on popular hotel features with less attention given to such ancillary aspects. We use the list of 39 features, as in Fig. 7.3, for further analysis.

Emerging Hotel Feature Identification. We now set out to identify features in which levels of interest among tourists have changed over the past five years. The year 2009 is set as the initial group and 2013 as the target group. Our data sets for these groups are input into the EPM algorithm. This produces a total of 16 emerging hotel features as presented in Table 7.4.

Table 7.4 shows some interesting patterns in tourists’ concerns as reflected in their reviews. It should be noted that in the context of EPM, the growth rate metric is the factor of interest for detecting change. The $\chi^2$ test results of $p$-value 0.05 demonstrate
Table 7.4: Emerging hotel features.

<table>
<thead>
<tr>
<th>Features</th>
<th>Support Values</th>
<th>GrowthRate</th>
<th>$\chi^2$</th>
<th>p-value</th>
<th>Pattern ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2009</td>
<td>2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>club</td>
<td>0.099</td>
<td>0.077</td>
<td>1.994</td>
<td>176.96</td>
<td>0.000</td>
</tr>
<tr>
<td>lounge</td>
<td>0.065</td>
<td>0.089</td>
<td>1.383</td>
<td>60.43</td>
<td>0.000</td>
</tr>
<tr>
<td>river</td>
<td>0.039</td>
<td>0.053</td>
<td>1.351</td>
<td>30.17</td>
<td>0.000</td>
</tr>
<tr>
<td>pool</td>
<td>0.186</td>
<td>0.216</td>
<td>1.159</td>
<td>39.94</td>
<td>0.000</td>
</tr>
<tr>
<td>service</td>
<td>0.306</td>
<td>0.353</td>
<td>1.155</td>
<td>39.94</td>
<td>0.000</td>
</tr>
<tr>
<td>dinner</td>
<td>0.050</td>
<td>0.058</td>
<td>1.150</td>
<td>75.39</td>
<td>0.004</td>
</tr>
<tr>
<td>food</td>
<td>0.207</td>
<td>0.230</td>
<td>1.109</td>
<td>8.16</td>
<td>0.000</td>
</tr>
<tr>
<td>view</td>
<td>0.176</td>
<td>0.194</td>
<td>1.105</td>
<td>22.30</td>
<td>0.000</td>
</tr>
<tr>
<td>price</td>
<td>0.185</td>
<td>0.122</td>
<td>-1.520</td>
<td>16.79</td>
<td>0.000</td>
</tr>
<tr>
<td>street</td>
<td>0.143</td>
<td>0.096</td>
<td>-1.495</td>
<td>256.43</td>
<td>0.000</td>
</tr>
<tr>
<td>taxi</td>
<td>0.133</td>
<td>0.094</td>
<td>-1.413</td>
<td>177.57</td>
<td>0.000</td>
</tr>
<tr>
<td>bathroom</td>
<td>0.245</td>
<td>0.176</td>
<td>-1.396</td>
<td>235.54</td>
<td>0.000</td>
</tr>
<tr>
<td>park</td>
<td>0.078</td>
<td>0.056</td>
<td>-1.383</td>
<td>60.66</td>
<td>0.000</td>
</tr>
<tr>
<td>clean</td>
<td>0.371</td>
<td>0.293</td>
<td>-1.265</td>
<td>212.05</td>
<td>0.000</td>
</tr>
<tr>
<td>bed</td>
<td>0.224</td>
<td>0.190</td>
<td>-1.174</td>
<td>52.13</td>
<td>0.000</td>
</tr>
<tr>
<td>location</td>
<td>0.479</td>
<td>0.426</td>
<td>-1.125</td>
<td>86.84</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Statistically significant changes in the support values for 2009 and 2013. Some interesting findings for the positive emerging pattern are as follows. International tourists pay more attention to the **club** feature, as shown by a growth rate of 1.994 (Pattern $P_1$). The number of tourists using this facility has nearly doubled in comparison to 2009. Some hotel areas such as the **lounge** and **pool** also received more attention than before (P’2 and P’4). **River** and **view** appear to attract more interest from tourists in 2013, with growth rates of 1.115 and 1.105 respectively ($P_3$ and $P_7$). Pattern $P_5$ shows that tourists are more concerned than before about **service**. Slightly increased attention to **dinner** and **food** provided by hotels can also be identified.

These positive patterns can draw hotel managers’ attention to features which are becoming “hot” among international tourists so they can make appropriate changes that will attract more customers. However, negative emerging patterns can also help hotel managers in terms of saving effort and concentrating investment resources away from areas which are no longer important to customers. Some popular features such as
price and cleanliness are of less concern over time, as indicated by negative growth rates in Patterns \( P_9 \) and \( P_{14} \). Interest and indoor facilities such as bathrooms and beds (\( P_{12} \) and \( P_{15} \)) or outdoor aspects such as street, taxi, and park (\( P_{10}, P_{11}, \) and \( P_{13} \)) have also decreased.

Given these emerging features, hotel managers will want to know if there are trends in the areas tourists pay attention to, for planning purposes. We therefore also compute the growth rate for each feature against each pair of years 2009-2013, with the results shown in Table 7.5. Here, the signs of the growth rates rather than the values are of interest.

Table 7.5: Emerging trends of tourists’ concerns.

<table>
<thead>
<tr>
<th>Features</th>
<th>Growth Rates</th>
<th>Trend</th>
<th>( \chi^2 )</th>
<th>p-value</th>
<th>Pattern ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>club</td>
<td>1.111</td>
<td>1.434</td>
<td>1.205</td>
<td>1.039</td>
<td>↑</td>
</tr>
<tr>
<td>lounge</td>
<td>~1.047</td>
<td>1.196</td>
<td>1.221</td>
<td>~1.009</td>
<td>~</td>
</tr>
<tr>
<td>river</td>
<td>~1.040</td>
<td>1.050</td>
<td>1.142</td>
<td>1.171</td>
<td>~</td>
</tr>
<tr>
<td>pool</td>
<td>~1.014</td>
<td>~1.003</td>
<td>1.134</td>
<td>1.039</td>
<td>~</td>
</tr>
<tr>
<td>service</td>
<td>1.072</td>
<td>1.007</td>
<td>1.074</td>
<td>~1.003</td>
<td>~</td>
</tr>
<tr>
<td>dinner</td>
<td>~1.151</td>
<td>1.139</td>
<td>1.138</td>
<td>1.020</td>
<td>~</td>
</tr>
<tr>
<td>food</td>
<td>1.041</td>
<td>1.012</td>
<td>1.049</td>
<td>1.003</td>
<td>↑</td>
</tr>
<tr>
<td>view</td>
<td>~1.142</td>
<td>1.064</td>
<td>1.103</td>
<td>1.076</td>
<td>~</td>
</tr>
<tr>
<td>price</td>
<td>~1.152</td>
<td>~1.111</td>
<td>~1.091</td>
<td>~1.090</td>
<td>~</td>
</tr>
<tr>
<td>street</td>
<td>~1.145</td>
<td>~1.117</td>
<td>~1.097</td>
<td>~1.066</td>
<td>~</td>
</tr>
<tr>
<td>taxi</td>
<td>~1.206</td>
<td>~1.030</td>
<td>~1.082</td>
<td>~1.051</td>
<td>~</td>
</tr>
<tr>
<td>bathroom</td>
<td>~1.141</td>
<td>~1.094</td>
<td>~1.041</td>
<td>~1.074</td>
<td>~</td>
</tr>
<tr>
<td>park</td>
<td>~1.072</td>
<td>~1.161</td>
<td>~1.086</td>
<td>~1.023</td>
<td>~</td>
</tr>
<tr>
<td>clean</td>
<td>~1.126</td>
<td>~1.068</td>
<td>~1.038</td>
<td>~1.012</td>
<td>~</td>
</tr>
<tr>
<td>bed</td>
<td>~1.056</td>
<td>~1.079</td>
<td>~1.002</td>
<td>~1.032</td>
<td>~</td>
</tr>
<tr>
<td>location</td>
<td>~1.065</td>
<td>~1.053</td>
<td>~1.003</td>
<td>~1.001</td>
<td>~</td>
</tr>
</tbody>
</table>

A feature has an increasing trend (↑) if the growth rates for all pairs of years are positive and a decreasing trend (↓), if its growth rates are all negative. If no clear trend is found, the symbol ~ is used. The trend in relation to club and food is increasing, as indicated by the positive growth rates across different year pairs (Trend \( T_1 \) and \( T_7 \)). Clear downward trends are found for most negative emerging
hotel features from $T_9$ to $T_{16}$, as shown by the negative growth rates over time. A slight drop for features such as lounge, dinner, and view emerges from 2009 to 2010. However, since this time, they have gained more and more attention.

### 7.3.3 Implications

The identified list of 39 hotel features are in line with those emerging from previous work. In addition, it is interesting to know that there are several outdoor features (street, park, and river), and transportation features (airport, taxi, and train) that are also very important to tourists when evaluating hotels. Hotel managers can develop more effective marketing campaigns by presenting these features in their advertisements, such as showing pictures of beautiful outdoor settings or mentioning the convenience of local transportation.

The identification of emerging hotel features also highlights changes in the topics of concern to tourists over the past five years. Hotel facilities such as club, lounge, and pool now attract significantly more attention. Managers can therefore develop investment plans focusing on those positively emerging features, while directing effort and resources away from those receiving less attention. In addition, managers' decision making can be further enhanced by trend analysis. For instance, long-term plans can be made for a clubbing facility or better food quality, because there is a clearly increasing trend of interest from tourists in these features.
7.4 Summary

As one of the fastest growing industries in the world [2], tourism offers many new opportunities and challenges. Effective planning and decision making can confer multiple benefits to the tourism industry. Hotel managers who are looking at product design and development should understand tourists’ concerns so they can enhance business performance. Managers are interested in emerging issues or trends to make appropriate adjustments to their plans, which can save internal resources and maximize returns on investment. Despite considerable efforts made from researchers, generating insights to help tourism managers address tourists’ concerns and create a competitive hotel industry remain challenging tasks. Current research has been unable to demonstrate an effective method for addressing such demands comprehensively.

To tackle such challenges, The PN-EPM technique was proposed, which can effectively identify the changes and trends in the attention of tourists. Its performance is demonstrated in an application of emerging hotel preference identification, with a comprehensive hotel feature list constructed from online review. The analysis reported in this chapter is based on a large-scale data set, which is a promising data source because the concerns expressed by tourists in these web sites closely reflect those in real-life. The introduced technique and findings are useful to support tourism managers in further planning and decision making to better accommodate the growing demand of the tourism market.
Chapter 8

Conclusion and Future Work

Data mining provides a number of popular techniques for analyzing customer behavior, which is the key to effective strategic planning and decision making for businesses. Increasing demand for better insight into customer behavior has made existing data mining techniques insufficient to accommodate different situations. Using the tourism industry as a testbed, this thesis studied and developed advanced data mining techniques for the efficient mining of tourist preferences and for analyzing their behavior.

The research was carried out following three areas:

1. Decision making modeling focuses on developing better approaches for modeling the complex decision making process of tourists, so deeper insights into their behavior can be discovered.

2. Travel behavior discovery focuses on developing new techniques for processing and analyzing geo-tagged photos, and then for exploring the travel behavior of tourists.

3. Emerging preference identification focuses on developing new techniques for identifying emerging preferences of tourists so new interesting knowledge about
their demands and preferences can be obtained.

8.1 Contributions

The theoretical and experimental results have led to the conclusion and the main contributions of this thesis:

- **Decision Making Modeling:** Effective modeling of customer decision-making process should consider multiple criteria simultaneously and account for all possible interaction between criteria. These complex requirements are unable to be satisfied by traditional approaches. As such, deep insight into the decision making process was not achieved. This thesis aims to address the problem of decision making modeling by proposing a *fuzzy measure based preference mining* technique (Chapter 3). The Shapley and the Interaction Index computed from the fuzzy measure are able to provide insight into the preferences of customers, along with the interaction among criteria. A new R package, the Rfmtool package, was developed specifically for customer preferences analysis. The practical advantages of the proposed technique and R package have been demonstrated in a case study to model the behavior of international travelers in the hotel selection process. The proposed approach outperforms existing techniques in decision making modeling. It has significant potential in supporting researchers and industry practitioners in dealing with the decision making process and developing strategic plans.

- **Geo-tagged Photos Analysis:** Online travel photos with GPS information have presented a new way for studying travel behavior with great potential.
Unfortunately, no standard framework has been available to effectively exploit such data. Limited tools support the analysis process and application development based on these data resources. Aiming to address this limitation, this thesis proposed a travel behavior mining framework for geo-tagged photos, based on Density Clustering and the Markov Chain (Chapter 4). This technique is efficient in processing and mining geo-tagged photos to identify tourism attractions, and to reveal travel the flow of tourists between visited locations. Practical capability of this technique was demonstrated in an application of Hong Kong inbound tourists, which supports destination development, transportation planning, and impact management.

- **Travel Diary Construction:** Travel behavior analysis from geo-tagged photos is an emerging research direction. There is a massive amount of travel photos available online, whereas only a small number of them are tagged with geographical data. This prevents tourism researchers and business managers from fully capturing and understanding the travel behavior of tourists. Aiming to address such challenges, this thesis incorporates travel photos without geographical data in travel analysis following two approaches (Chapter 5): 1) visited location verification: aiming to determine if a photo was actually taken at a particular location or not; and 2) travel diary construction: aiming to reconstruct a travel diary for travel photo collections taken by tourists. The former approach is performed using a newly proposed machine learning technique, named Representative Instance Classification (RIC). The latter approach is carried out using a new technique, named Bayesian Latent Dirichlet Allocation (BA-LDA). Experiment results showed that RIC achieves the best performance in visited
location verification, while, BA-LDA has promising results for travel diary construction.

- **Satellite Image Reduction**: There is an increasing interest in developing applications using geo-tagged photos and GPS technology to support tourist travel. Those applications often involve the analysis and display of high resolution satellite images, on small display screens of user devices. This has reinforced the need for robust image reduction techniques that can preserve specific image details while removing or reducing the prevalence of unwanted artifacts, such as noise or data corruption. *Block-based image reduction* operators are potential candidates, however existing approaches sufficiently describe the spatial structure of a set of pixels so that structured image details may be preserved during the reduction. To facilitate the development of tourism applications, this thesis proposed three alternative weighting approaches to improve the robustness of *Block-based image reduction* operators for satellite image reduction (Chapter 6). These approaches are based on fuzzy measure theory, including the *Minimum-Spanning Tree approach* (MST), *Sugeno-type Fuzzy Measure approach* (SFM) and the *Decomposing Fuzzy Measure Approach* (DFM). They are effective in reducing image size, while preserving special characteristics and satisfying special requirements of satellite images.

- **Positive and Negative Emerging Pattern Mining**: Tourism managers are interested in identifying emerging changes in the preferences of tourists, to improve their strategic planning, marketing, and product development. The challenge in *emerging* changes identification is that researchers usually have no prior knowledge about what should be included in the analysis. None of the
existing techniques are effective in detecting the *emerging* changes in the preferences of tourists. Aiming to address these challenges, this thesis presented a hotel review processing framework and *Positive and Negative Emerging Pattern Mining* (PN-EPM) (Chapter 7). The online review processing framework allows for automatic identification of tourist’s preferred features, which are expressed in online review. Then, PN-EPM can effectively capture emerging changes, in customer preference. The performance of the proposed technique is demonstrated in an application of emerging hotel preference identification.

The practical applications of the proposed data mining techniques are demonstrated in applications of discovering tourists’ behavior and preferences so they can be used in tourism management. The mined knowledge helps to completely understand tourist behavior from *pre-trip accommodation planning, in-trip travel activities*, to *emerging changes in hotel preferences*. The techniques presented in this thesis and the discovered knowledge have the potential to benefit tourism practitioners worldwide to better understand traveler behavior and develop sustainable tourism industries.

### 8.2 Future Work

Although a number of preference mining techniques were proposed to address the critical challenges in tourism management, there remains further research and applications of data mining for customer behavior analysis.

Chapter 3 introduced a decision making modeling approach based on fuzzy measure, which considers the interaction between all possible combinations of criteria. It
should be noted that the concept of $k$-additivity controls the trade off between model complexity and modeling capability. When the complexity reduction is desired, small $k$ values such as 2 or 3 are preferred. Interaction of more than 3 criteria would be hard to predict. In contrast, there may be interactions between large groups of criteria in real situations. It would be critical to determine if it is worth increasing the value of $k$, especially when there are a large number of criteria under consideration. Besides, we are aware of the fact that customer decision making are in fact often carried out by groups of people. It would be beneficial to propose methods which can model group based decision making in the future research.

Chapter 4 presented the use of GPS information from travel photos for analyzing the travel behavior of tourists. Besides the spatial and temporal information, the online travel photos can provide textual and visual information from photo tags and visual content. The information would have great potential for providing insight into tourists’ behavior, which can then be further explored in further research. More effort can be spent on investigating different kinds of techniques in text mining or computer vision, so that effective methods can be developed to mine the knowledge hidden inside such data sources. Besides, it would be interesting to extend the work in Chapters 5 and 6 and construct a more detailed travel diary from ungeotagged travel photos or develop an intelligence system for travel diary recommendation. The performance of RIC algorithm and BA-LDA can be further evaluated on more different data sets.

The PN-EPM method and text processing framework, presented in Chapter 7, are general techniques that can identify specific features from online reviews. This means they can be used to pinpoint travelers’ concerns in other tourism contexts, such as airlines, restaurants, or other attractions. The support thresholds used for
constructing the features list and generating emerging patterns are pre-determined by end users. More experiments can be carried out in different applications to determine the threshold values for producing the best result. The analysis of hotel features of concern was performed solely on the review text using text processing approach. It would be advantageous to consider other factors such as hotel rating, feature rating or other meta data associated with reviews. Sentiment analysis can also be applied together with text processing to provide deeper understanding into tourists’ behavior.

While this thesis provides a thorough report of our research in data mining techniques for customer behavior analysis, there are interesting and promising issues that remain unexplored. I would like to continue studying and proposing useful techniques, that bring insightful benefits to researchers in customer behavior analysis, especially in the context of tourism and hospitality.
Bibliography


[303] C. L. Wu and D. Carson. Spatial and temporal tourist dispersal analysis in


[305] J. Xia, F. Evans, K. Spilsbury, V. Ciesielski, C. Arrowsmith, and G. Wright.
Market segments based on the dominant movement patterns of tourist. *Tourism

[306] J. Xia, P. Zeephongsekul, and C. Arrowsmith. Modelling spatio-temporal move-
ment of tourists using finite markov chains. *Mathematics and Computers in

[307] F. Xu, M. Morgan, and M. Moital. Cross-cultural segments in international
student travel: An analysis of British and Chinese market. *Tourism Analysis*,

[308] R. R. Yager. On ordered weighted averaging aggregation operators in multi-
criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*,

based methodology for mapping consumer preferences and setting product tar-

[310] O. H. M. Yaua and C. F. Chan. Hong Kong as a travel destination in South-
1990.


