Mining People's Semantic Trajectory Behaviours from Geotagged Photographs

Guochen Cai

A thesis presented for the degree of
Doctor of Philosophy

College of Business, Law and Governance
James Cook University
Australia

September 20, 2017

Declaration of originality

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

(Signed) ______________________________

Guochen Cai

Acknowledgements

I would first like to express my sincere gratitude to my supervisors Dr.Kyungmi (Joanne) Lee and Professor Ickjai Lee. This thesis was completed under the guidance of my supervisors. Thanks very much for Joanne and Ickjai's great time and effort in my study and research life over the past four years! Whether it is the design of the research topic, experimental guidance, or paper writing, Joanne and Jai devoted a lot of effort. Every one of my papers was carefully revised by my supervisors many times. They took the trouble to teach me how to use scientific methods to undertake research, and unreservedly shared their rich life and scientific research experience with me, which is a very valuable asset for me. They are conscientious, responsible and patient supervisors. I could not have imagined having a better advisor and mentor for my PhD study.

I would like to thank James Cook University for providing a high quality study and research environment. I enjoyed every second that I spent in this university. A very special gratitude goes out to the university for helping and providing the funding for my study and the journey to academic conferences.

I would also like to thank all my friends. You have made my life in Australia great and unique.

Finally but most importantly, I deeply and forever thank my wife, Jiaoyan, my children Runxin and Linfeng, and my parent. My special thanks to Jiaoyan for her understanding, encouragement and inspiration. Another special thanks to my parent for their endless material, financial and spiritual

supports. I dedicate this thesis to these people with love and gratitude.

Statement of the contribution of others

Research funding

- JCUPRS Scholarship

Thesis committee

- Dr Kyungmi Lee, Information Technology Academy, James Cook University Australia
- Professor Ickjai Lee, Information Technology Academy, James Cook University Australia
- Dr Insu Song, Information Technology Academy, James Cook University Singapore
- Dr Yok-Yen Nguwi, Information Technology Academy, James Cook University Singapore

Editorial support

- Dr Kyungmi Lee
- Professor Ickjai Lee

- Helena Bond
- Anonymous reviewers of the published papers derived from this thesis

Publications arising from this thesis

Conference papers:

1. Guochen Cai, Kyungmi Lee, and Ickjai Lee. "Mining Semantic Sequential Patterns from Geo-Tagged Photos". In: *Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 2187-2196. IEEE Computer Society, 2016.

2. Guochen Cai, Kyungmi Lee, and Ickjai Lee. "Discovering Common Semantic Trajectories from Geo-tagged Social Media". In: *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE) 2016*, pp. 320-332. Springer Lecture Notes in Computer Science, 2016.

3. Guochen Cai, Kyungmi Lee, and Ickjai Lee. "A Framework for Mining Semantic-Level Tourist Movement Behaviours from Geo-tagged Photos". In: *Australasian Joint Conference on Artificial Intelligence (AusAI)*, pp. 519-524. Springer Lecture Notes in Computer Science, 2016.

Submitted Journal Articles:

1. Guochen Cai, Kyungmi Lee, and Ickjai Lee. "Mining Mobility Patterns from Geo-tagged Photos through Semantic Trajectory Clustering". Submitted to *Cybernetics and Systems*. Submitted on 8/12/2016.

2. Guochen Cai, Kyungmi Lee, and Ickjai Lee. "Mining Semantic Trajectory Patterns from Geo-tagged Photos". Submitted to *Journal of Computer Science and Technology*. Submitted on 31/3/2017.
3. Guochen Cai, Kyungmi Lee, and Ickjai Lee. "Itinerary Recommender System with Semantic Trajectory Pattern Mining from Geo-tagged Photos". Submitted to *Expert Systems with Applications*. Submitted on 24/4/2017.

Abstract

Web 2.0 technology has changed the way users use the web. In Web 2.0, users can create their own data and upload to the web. The new web technology promotes the evolution of social media applications built on Web 2.0. Social media allow the creation and exchange of user-generated content. One main type of social media is photo-sharing websites where the main objective is the sharing of photo media content between users. Through these websites users store and manage their photos, and share and communicate with friends, families and colleagues.

The development of mobile devices facilitates people's easy usage of the Internet. This has lead to dramatic growth in the number of user-generated photos in the website. This massive collection of photo data may enclose people's movement behaviours, which are useful to domain experts and areas such as traffic management and tourism. However, this large and complex dataset requires advanced techniques to extract the hidden useful knowledge from the big data.

Some previous studies have been conducted, and various approaches have been proposed to extract people's movement behaviours from online geo-tagged photos. These studies are mainly about three topics. The first topic is to reveal the spatial behaviours of people that the approaches detect the spatial locations that people prefer to visit (Kisilevich, Mansmann, and Keim 2010; Lee, Cai, and Lee 2013). The second topic is to find out the spatial

place association rules that determine the sets of places visited together in people's movements. The third topic is to discover people's dynamic spatio-temporal movement behaviours including the spatio-temporal traffic flow (Girardin et al. 2008b) and frequent spatio-temporal movement patterns (Zheng, Zha, and Chua 2012; Cai et al. 2014; Bermingham and Lee 2014).

However, previous approaches lack consideration of the additional aspatial semantics information of trajectories. They are traditional geometric-feature analyses. The main drawback of previous methods is that their result patterns contain only pure geometric data, without meaningful semantics information about the mobility. Most applications analyses require complementing trajectory with additional information from the application context. The contextual information provides useful knowledge about moving behaviours with richer and more meaningful semantic information and the semantic-level patterns.

This thesis aims to develop a systematic framework for extraction of people's movement patterns with meaningful and understandable semantics information. We add the aspatial semantics annotations to trajectories and analyse trajectories with spatial, temporal and aspatial features together. We aim to find the semantics-enhanced movement patterns, including semantic sequential patterns, semantic common patterns and semantic trajectory patterns. Finally, this thesis also aims to build an itinerary recommender system based on the extracted trajectory patterns.

In this thesis, we propose a systematic framework for discovery of people's semantic mobility patterns from geotagged photos. The framework has four main functions for extraction of the three semantic patterns and for building the recommender system, respectively. At the first step, the framework builds spatio-temporal trajectory data from the geotagged photos. Then, we add background geographic information, place type annotation and multiple environmental contextual data to the raw trajectories to generate people's semantic trajectories.

From the semantic trajectories, the framework's first main function is to find out the frequent semantic sequential patterns. This thesis proposes a sequential pattern mining method to extract semantic sequential patterns, which are sequences of stops that frequently occur in people's trajectories. This method can deal with multi-dimensional semantic trajectories. The extracted groups of patterns include not only the basic patterns, which contain geographic place category information only, but also the multi-dimensional semantic patterns, which are associated with flexible combinations of frequent environmental contextual information.

The framework's second main function is to reveal the semantic common patterns. This thesis proposes a semantic trajectory clustering approach to find semantic common patterns in the semantic trajectories. The common pattern shows the common track drawn from many people having similar trajectories. A distance function is designed and proposed for the multi-dimensional semantic trajectories.

The third main function of the framework is to extract the semantic trajectory patterns. This thesis presents a semantic trajectory pattern mining method to find frequent trajectory patterns from semantic trajectories. A semantic trajectory pattern demonstrates a frequently visited sequence of stops with typical transition time information. The transition time shows the time interval between two stops that indicates temporal behaviour of people's mobility.

Finally, this framework builds a recommendation system based on the extracted semantically enhanced movement patterns. The system provides users with suggestions about travel itineraries including travel route and time interval information between two stops. The system is semantic-aware, allowing users to customise sets of place types that they want to visit in the trip and to set up travel duration.

We conduct experiments to evaluate proposed methods using real photo

dataset from Flickr[1]. The experimental results prove the effectiveness of our framework. The results show that the proposed semantics added trajectory analysis methods can extract detailed and semantically enhanced semantic patterns that not only show people's semantic-level mobility patterns, but also provide rich meaningful information and better understanding of people's movements. The results also demonstrate that our recommender system effectively generates a set of customised and targeted semantic-level itineraries that meet the user-specified constraints and with an efficiency itinerary generation property. In addition, our system produces higher place type-layer itineraries with richer meaningful information about travel contexts.

[1]Flickr: https://www.flickr.com/

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Massive user-generated content is publicly available online with the advance of Web 2.0 technology. Web 2.0 is a platform in which applications deliver software as a continually updated service (O'reilly 2005). Data on Web 2.0 is from multiple sources, including not only the service providers, but also individual users who provide their own data. This is compared to Web 1.0, where online data is mainly created, uploaded and published by website providers. In Web 1.0, the vast majority of users simply act as consumers of content; in Web 2.0, all users can create their own data and upload online.

There are now thousands of websites and applications built on Web 2.0. Especially, Web 2.0 is the platform for the evolution of social media – such as Wikipedia, Facebook, YouTube – that allow the creation and exchange of user-generated content (Kaplan and Haenlein 2010). Tens of millions of users are providing data every day. These data are about users' experiences, ideas, events, activities and life, and are used to share and communicate with their friends, communities and families. Meanwhile, network-enabled mobile devices, such as smart phones, make it convenient for users to visit social media services and generate content. The enormous amount of online data provides a potential data-rich environment for the discovery of people's behaviours (Pang and Lee 2008; Papadopoulos et al. 2009; Agarwal et al. 2011; Zafarani, Abbasi, and Liu 2014) .

One important type of social media is the photo-sharing website. The main objective of this type of website is the sharing of photo media content between users. Photo-sharing sites provide users with a large space for storage of photo collections, and services to organise and manage their photos and share with their groups. As an example, Instagram [1] and Flickr are two of the most popular photo-sharing sites. These sites are a great resource for photography enthusiasts, and increasingly for travellers. Following the increasing number of photos that are manually geotagged by users, these photo-sharing sites have recently launched their own services for adding latitude and longitude information to a photo. Flickr provides a tool that allows a user to select the location on a map in which a picture was taken, and then the corresponding latitude and longitude information is added as metadata to the picture. In addition, many photos are geotagged automatically using global positioning system (GPS) logs or location-aware devices. Therefore, the location and time data associated with photos and other related text tags can be considered as useful geographically annotated materials on the web. Millions of photo data are created and uploaded to photo-sharing websites. These user-generated photo data may contain useful knowledge about people's behaviours, which are valuable and useful to domain experts and applications.

The data on the photo-sharing websites are large, unstructured and complex. As a result, domain experts and decision makers cannot directly find useful information and knowledge from the big data by using traditional analysis methods. It requires advanced analysis tools and techniques to deal with big data. "Data mining" encompasses the processes and technologies developed to extract the valuable knowledge embedded in the vast amounts

[1] Instagram: https://www.instagram.com/

of data. Data mining uses a series of techniques to discover interesting patterns and knowledge from large amounts of data (Han, Pei, and Kamber 2011). Data mining focuses on the properties of methods for handling large datasets, including accuracy, efficiency and scalability, as well as on ways to handle complex types of data. There are several data mining functionalities and tasks used to specify the kinds of patterns that can be mined, including characterisation and discrimination, the mining of frequent patterns, associations, and correlations, classification and regression, cluster analysis and outlier analysis (Han, Pei, and Kamber 2011). Data mining techniques are popular tools used to discover useful patterns and knowledge from the massive social media data (Thelwall, Wilkinson, and Uppal 2010; Barbier and Liu 2011; Jin et al. 2011; Chen, Vorvoreanu, and Madhavan 2014) and specific online photo data (Kennedy et al. 2007; Girardin et al. 2008a; Papadopoulos et al. 2011; Zheng, Zha, and Chua 2012; Lee, Cai, and Lee 2014).

1.2 Motivations

Making sense of large online geotagged photo datasets is of significant importance to derive more thorough movement behaviour information, particularly of informative patterns which are valuable to various domains, such as city planning, traffic management and tourism. However, there is no existing systematic data mining framework for extracting people's movement behavioural patterns with meaningful and understandable semantic information from the geotagged photos. This research aims to fill that gap.

Many studies were interested in analysing geotagged photos and some approaches have been proposed to discover knowledge about movement behaviours from the photo data. Most of the studies cover the following three topics:

- spatial interest of places and events;

- place association rules;
- spatio-temporal trajectory behaviours.

The first topic is discovering the spatial behaviour of people, that is, detecting the spatial locations, places and events they find of interest: the places where people prefer to gather. Another topic is to find out the association rules of spatial places – a rule indicates certain association relationships among a set of spatial places. The proposed approaches find out sets of places that frequently occur together in people's movements.

The third topic is to extract the spatio-temporal movement behaviours of people. Though these previous methods were useful and recognisable in terms of spatial locations and location association rules, there is a lack of studies on discovering people's dynamic spatio-temporal movement patterns. Few recent studies have attempted to discover dynamic spatio-temporal trajectory behaviours from geotagged photos (Girardin et al. 2008b; Zheng, Zha, and Chua 2012; Cai et al. 2014). Instead, the existing studies use spatial features of trajectories to find movement knowledge, in terms of geographical spatial shapes and patterns. Their results show purely geometric information about movement.

However, these previous traditional spatial-geometric-feature-only trajectory analysis approaches are insufficient. Many specific applications require richer application-related semantic information and meanings in people's mobility. In many application domains, useful knowledge about moving behaviour or moving patterns can only be extracted from trajectories if the background geographic information where trajectories are located is considered (Alvares et al. 2007b). Thus, trajectory analysis needs to be integrated with aspatial semantics information (Parent et al. 2013). Aspatial semantics information is mobility-related background contextual data in which movement takes place.

Integrating trajectory analysis with aspatial semantics data is important because it provides applications with richer and more meaningful knowledge about movement and indicates semantic-level patterns. Detailed and semantically enhanced patterns provide better understanding of people's movement. The importance of semantics in trajectory data mining has received attention in recent years as in GPS data (Alvares et al. 2007b) and in geotagged photos (Kisilevich et al. 2013). However, there has been little research with geotagged photo with semantics until very recently.

In summary, massive online collections of shared photos with geographic data indicate people's movement trajectories and contain potential behavioural movement patterns that are valuable for various domains. However, previous approaches have not been able to extract the semantic mobility patterns in the trajectories. A few past studies analyse trajectories with traditional geometric-feature-only methods, but they lack of consideration of the important aspatial semantic features. The traditional analysis could miss the interesting and meaningful semantically enhanced patterns. Some recent studies have paid attentions on analysing trajectory data incorporating semantics. However, there has been little research with geotagged photo with semantics. These motivate this research to discover people's movement behavioural patterns with the meaningful aspatial application of semantic information from online geotagged photos.

1.3 Research aim

The aim of this research is to develop a systematic data mining framework to extract semantic trajectory behavioural patterns from geotagged photos and to build an itinerary recommender system based on the patterns. Specifically, this research aims to find out four kinds of patterns that scope the four aims of this research, which are:

1. To find meaningful and understandable semantic sequential patterns from geotagged photos;

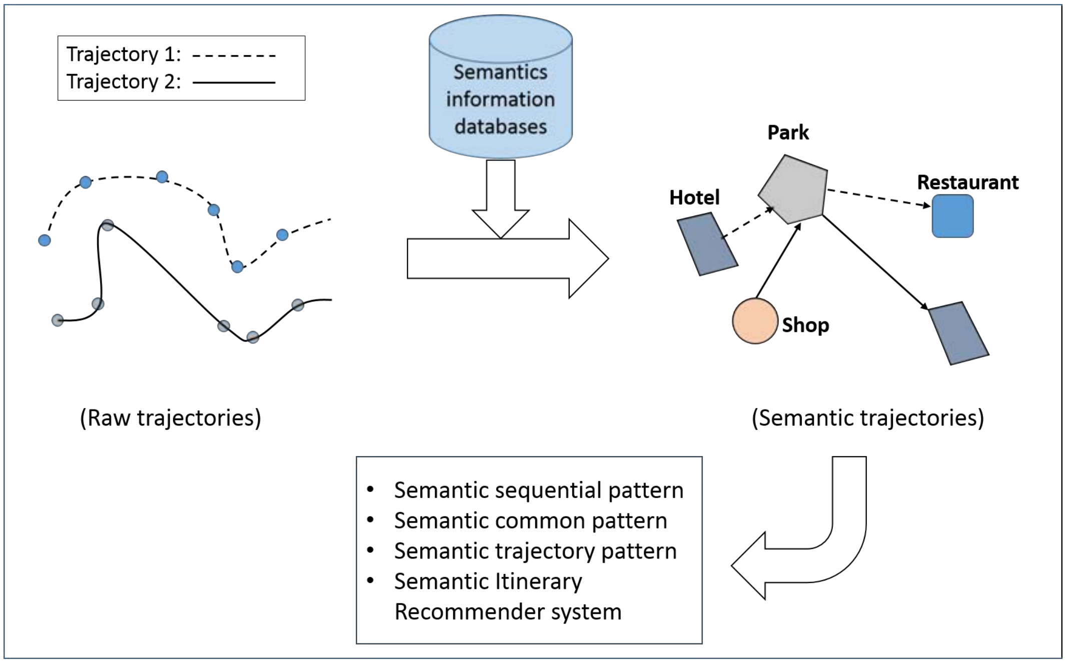


Figure 1.1: Overall research of this thesis.

2. To find meaningful and understandable semantic common patterns from geotagged photos;

3. To find meaningful and understandable semantic trajectory patterns from geotagged photos;

4. To build a semantic itinerary recommender system using the extracted trajectory patterns from geotagged photos.

Figure 1.1 briefly shows the overall work of this thesis. There is a more detailed description of our work in Chapter 3. This study is to add aspatial semantics information to raw trajectories formed from the geotagged photos, and then to extract three semantic mobility patterns – semantic sequential patterns, semantic common patterns and semantic trajectory patterns – and finally to build an itinerary recommender system based on trajectory patterns. The sequential pattern indicates a frequent sequence of stops in the trajectories. The common pattern shows the common track of people's trajectories. The trajectory pattern presents the frequent sequence of stops with transit time information. There are many semantics information databases, but in this study, we use certain semantics information only, including a geographic information database, which is used to add basic place type annotation and city area information to geo-objects, and a weather observation database, which is used to add contextual weather conditions to the trajectories. Note that users can add more aspatial and semantic databases to our framework with ease for domain specific applications.

A semantic trajectory is a raw trajectory with semantic annotations. In particular, in this study, it is with place-type and weather semantic annotations. More details will be covered in Section 3.4. Given semantic trajectories, the definitions of the three kinds of patterns are as follows:

1. Semantic sequential pattern is a sequence of stops that frequently occurs in semantic trajectories;

2. Semantic common pattern is a clustered pattern that is common in semantic trajectories;

3. Semantic trajectory pattern is a sequence of stops with relevant time interval that frequently occurs in semantic trajectories.

Figure 1.2 illustrates the differences between the three kinds of patterns. Semantic sequential pattern shows a sequence of stops frequently occurring in semantic trajectories indicating a sequential order of frequent movements. Semantic common pattern refers to a cluster of similar patterns in semantic trajectories. Semantic trajectory pattern indicates a sequence of frequent stops with relevant time interval.

This research is based on four hypotheses:

- Semantic trajectories from geotagged photos indicate meaningful and understandable semantic sequential patterns;

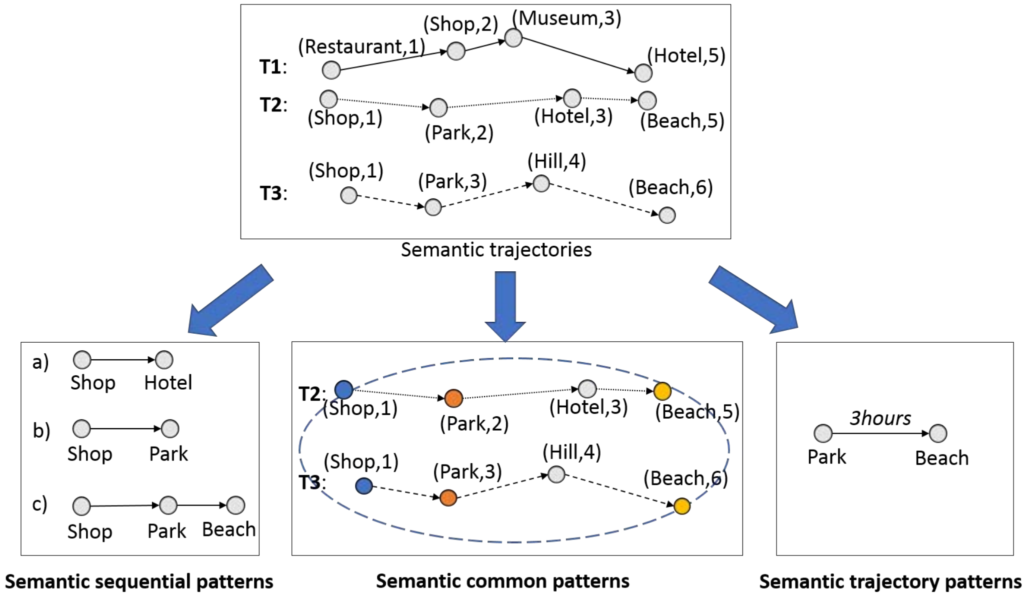


Figure 1.2: Differences of semantic sequential pattern, semantic common pattern and semantic trajectory pattern.

- Semantic trajectories from geotagged photos indicate meaningful and understandable semantic common patterns;
- Semantic trajectories from geotagged photos indicate meaningful and understandable semantic trajectory patterns;
- Semantically enhanced patterns from geotagged photos are good indications for travel itinerary recommendations.

Based on the hypotheses, this research seeks to answer four questions:

- How can we detect more meaningful and understandable sequential patterns from geotagged photos-based semantic trajectories?
- How can we discover more meaningful and understandable common patterns from geotagged photos-based semantic trajectories?

- How can we find more meaningful and understandable common patterns from geotagged photos-based semantic trajectories?

- How can we use the semantically enhanced patterns from geotagged photos to provide users with travel itinerary recommendations?

Table 1.1 presents four specific research aims and the related objectives. These four aims will answer the four research questions.

This thesis aims to develop a systematic framework to extract people's semantic trajectory behavioural patterns from geotagged photos. Overall, the framework generates trajectories from geotagged photos, creates semantic trajectories, extracts three kinds of mobility patterns from trajectories and builds an itinerary recommender system based on the extracted patterns. At the beginning, this framework first constructs people' raw movement trajectories from their geotagged photos. A raw trajectory is a sequence of geographical points with time stamps. Then the framework generates people's semantic trajectories. A semantic trajectory is a sequence of stops with background geographic information semantics annotations, using place category in this study. This research uses region of interest (RoI) as the stops and this study proposes a semantic RoI mining algorithm to detect RoIs with semantics annotation from raw trajectories. This study also adds multiple environmental context semantics data to trajectories including city name, day type, day time and weather conditions in which the movement takes place.

From the semantic trajectories, the first function of the framework is mining semantic sequential patterns. A semantic sequential pattern mining approach is proposed to find out frequent sub-sequences of stops in semantic trajectories based on the projection-based PrefixSpan algorithm (Pei et al. 2001). These frequent sub-sequences are semantic sequential patterns in people's trajectories. The proposed approach is able to deal with multi-dimensional semantic trajectories: the results are sets of semantic sequential

Table 1.1: Research aims and related objectives.

Aims	**Objectives**
Aim 1: to find out meaningful and understandable semantic sequential patterns from geotagged photos	Objective 1: reconstructing trajectories from geotagged photo data Objective 2: building semantic trajectories from geotagged photos Objective 3: mining semantic sequential patterns from semantic trajectories
Aim 2: to find out meaningful and understandable semantic common patterns from geotagged photos	Objective 4: mining semantic common patterns from semantic trajectories
Aim 3: to find out more meaningful and understandable semantic trajectory patterns from geotagged photos	Objective 5: mining semantic trajectory patterns from semantic trajectories
Aim 4: to build a semantic itinerary recommender system using the extracted trajectory patterns from geotagged photos	Objective 6: building itinerary recommender system using semantic trajectory patterns

patterns and each pattern is a sequence of stops integrated with a frequent combination of multiple dimensions where a combination could contain a subset or all of the original environmental context data.

The second function of the framework is extracting semantic common patterns from semantic trajectories. This study presents a semantic trajectory clustering method to group similar semantic trajectories and each cluster indicates a common track of people: the common pattern. The proposed clustering method adopts a density-based OPTICS algorithm scheme (Ankerst et al. 1999). It produces an ordering list of objects based on distance and extracts clusters from the ordering list. To deal with multi-dimensional semantic trajectories, this study introduces a distance function with a strategy of assigning different weight values to multiple dimensions. Distance function first finds out the commonality of two trajectories by using the Longest Common Sub-Sequence (LCSS) algorithm, then calculates the similarity score between two trajectories, and at last utilises the dissimilarity score as the distance between two trajectories.

The third function of the framework is finding out frequent semantic trajectory patterns in people's semantic trajectories. A trajectory pattern is a frequent sequence of stops with time intervals between two stops. The frequently repeated time intervals show temporal relations between stops. A semantic-trajectory-pattern-mining algorithm is introduced to extract the patterns based on the temporally annotated sequential pattern $\mathcal{TAS}$ algorithm (Giannotti, Nanni, and Pedreschi 2006). The proposed method can find basic semantic patterns, which are sequences of basic geographic semantics only; it is also able to find multi-dimensional semantic trajectory patterns, which are basic geographic semantic patterns with additional semantic annotations. These additional annotations could be combinations of arbitrary subset of the initial semantics.

Finally, the framework builds a semantic itinerary recommender system based on semantic trajectory patterns to provide users with suggestions for travel itineraries. The system generates a set of customised and targeted semantic-level itineraries that meet the user-specified constraints. The recommender system is an offline–online architecture. In the offline component, the system extracts people's previous semantic trajectory patterns from geotagged photos and stores the patterns into a database. In the online component, the system receives users' queries, verifies the queries, searches appropriate patterns from the pattern database and returns the final patterns to generate appropriate travel itineraries, including travel route sequences and transition time information. The user query includes a set of user-customised place types and travel duration. The system generates appropriate itineraries that satisfy the user's request.

1.4 Contributions of study

This study contributes to the literature of data mining and analysis of online geographic referenced photo data, by investigating the analysis of the trajectory feature of geotagged photos and the extraction of trajectory behavioural patterns. Moreover, this study considers the aspatial semantics features in the analysis of trajectory data to learn about meaningful and understandable semantically enhanced movement patterns. This study proposes a systematic data mining framework with a series of techniques and approaches for extracting people's semantic trajectory patterns from geotagged photos and building a semantic itinerary recommender system. The following is a summarised list of contributions made in this thesis.

Table 1.2: Contributions of this thesis.

Chapter	Publications	Contributions
3		1. Analysis of trajectories with consideration of spatial, temporal and aspatial semantics features. This research is a novel investigation into trajectory analysis using additional important aspatial semantics features to find semantic-level trajectory behaviours. We find semantic-level trajectory behavioural patterns from geotagged photos by considering multiple semantic annotations in addition to raw trajectories. Trajectories are enriched with multiple background geographic information and environmental context data that provide richer meanings to understandings of people's mobility;

4	• "Mining Semantic Sequential Patterns from Geo-tagged Photos". *HICSS 2016*	2. Semantic RoI detection. The semantic RoI mining method is proposed to detect stops in raw trajectories. This method uses spatial and aspatial semantics features together that show a natural process for generating semantically enhanced stops. This method can find fine and accurate semantic RoIs; 3. Semantic sequential pattern mining. This thesis proposes a semantic sequential pattern mining method to find frequent semantic sequential patterns in trajectories. The method has the ability to deal with multi-dimensional sequences. The method produces frequent patterns with flexible combinations of various frequent dimensions, including basic sequential patterns with only basic geographic data and multi-dimensional patterns with sets of additional semantics information;

5	• "Discovering Common Semantic Trajectories from Geo-tagged Social Media". *IEA/AIE 2016* • "Mining Mobility Patterns from Geotagged Photos through Semantic Trajectory Clustering". Submitted to *Journal: Cybernetics and Systems*	4. Semantic common pattern mining. This research proposes a semantic trajectory clustering method for finding the common semantic trajectories. This study presents a novel similarity measure for multi-dimensional semantic trajectories for semantic trajectory clustering. The proposed semantic trajectory clustering is exploratory, allowing users to explore diverse combinations of semantic dimensions. This flexibility enables users to refine patterns, and supports what-if analysis;
6	• "A Framework for Mining Semantic-Level Tourist Movement Behaviours from Geo-tagged Photos". *AusAI 2016* • "Mining Semantic Trajectory Patterns from Geo-tagged Photos". Submitted to *Journal of Computer Science and Technology*	5. Semantic trajectory pattern mining. This thesis proposes a semantic trajectory pattern mining algorithm for discovering frequent semantic trajectory patterns in semantic trajectories. The algorithm generates frequent sub-sequences with frequently occurring time intervals in people's trajectories. It also produces basic patterns and multi-dimensional patterns with diverse combinations of frequent dimensions;

7	• "Itinerary Recommender System with Semantic Trajectory Pattern Mining from Geo-tagged Photos". Submitted to *Journal: Expert Systems with Applications*	6. Semantic itinerary recommendation system. This research presents a semantic-aware offline–online travel itinerary recommender system. The proposed method generates these semantic itineraries from historic people's movements by mining frequent travel patterns from geotagged photos. It effectively generates a set of customised and targeted semantic-level itineraries that meet the user-specified constraints. The system generates appropriate higher place type-layer itineraries with rich, meaningful information about travel contexts and with an efficient itinerary generation property.

1.5 Organisation of the thesis

This thesis has eight chapters including this introduction. The other seven chapters of this thesis are described briefly as follows.

- Chapter 2 presents the background and related work. It first illustrates typical studies on mining knowledge of people's behaviours from online social media data. Then it presents previous work on mining and analysing geotagged photo data. In the next part, it describes previous approaches and studies on pattern mining from trajectories for geotagged photo data and GPS data. Specifically, it reviews and discusses previous studies on sequential patterns, trajectory clustering and trajectory patterns. Last, it presents a review of the studies on building travel recommendation systems using online geotagged photos.

- Chapter 3 describes the overall framework of this thesis. It briefly summarises the framework and describes each component and function of the framework. Next, the data collection and apatial semantics information databases used in this thesis are introduced. In this chapter, we also present detailed definitions and terminologies used in this thesis.

- Chapter 4 presents the study of extracting semantic sequential pattern, which is the first research aim. In this chapter, we describe the basic steps and approaches for building people's trajectories from geotagged photos and generating semantic trajectories with a proposed semantic RoI mining algorithm. Then, from the semantic trajectories, this study finds frequent sequences of stops with meaningful semantics information that show semantic sequential patterns by using the proposed semantic sequential pattern mining method.

- Chapter 5 demonstrates the study of mining semantic common patterns, which is the second research aim. This study proposes a semantic trajectory clustering method for finding people's common tracks in trajectories. A distance function is also presented for multi-dimensional semantic trajectories.

- Chapter 6 illustrates the study of discovering semantic trajectory patterns, which is the third research aim. This study is about finding out the frequent sequences of stops with typical transition time between stops in people's semantic trajectories. We propose a semantic trajectory pattern mining method to reveal the frequent patterns.

- Chapter 7 shows the study of building a semantic itinerary recommender system, which is the fourth research aim. The system is developed as offline–online architecture. It generates itineraries by using the extracted semantic trajectory pattern. The system is semantics-aware – allowing users to customise a set of preferred place types and travel duration. The system produces sets of appropriate travel itineraries including sequences of stops with transition time between stops and rich information on travel environmental contexts.

- Chapter 8 concludes this thesis by summarising the content and contributions obtained in this research. It also summarises potential future work that could be explored following on from the research work in this thesis.

Chapter 2

Literature review

This chapter surveys previous studies on analysing online social media data and reviews previous research relating to extraction of people's mobility patterns and building itinerary recommender systems from geotagged photo data. The abundance of online user-generated content provides a data-rich environment for many areas and applications. In Chapter 2.1, we briefly introduce diverse studies on using, analysing and mining online social media data. Chapter 2.2 reviews past studies exploring online geotagged social data and mining people's movement behaviour. Then, in Chapter 2.3, we present existing work on extracting people's dynamic mobility behavioural patterns from geotagged photos. We describe work related to each objective and task of our study. Chapter 2.4 describes previous research on building travel itinerary recommender systems using geotagged photo data. Finally, Chapter 2.5 summarises the literature review in relation to our research topic and work.

2.1 Social media data mining

Social media data mining extracts useful information, patterns, rules and trends from large-scale online user-generated social media data. There are many categories of social media and various types of social data including records of events, social activities, and people's presentations of ideas and daily life. Online social media data is growing fast with the pervasive use of social media. Consequently, the enormous amount of online data attracts various research communities' investigations into valuable information and knowledge about human behaviours for specific applications and aims (Tang and Liu 2010; Pang and Lee 2008; Rattenbury, Good, and Naaman 2007; Lee, Cai, and Lee 2014).

One of the most popular research areas is using social data entities to discover social structure, relationships and transformation of information in the online virtual world. This information helps our understandings of people's social activities, relations and behaviours in the social networking sites. Zafarani, Abbasi, and Liu (2014) extracted information about interactions between individuals and entities that shows how individuals interact with others about the contents (Tang, Wang, and Liu 2009). Kempe, Kleinberg, and Tardos (2003) investigated influence modelling to understand the process of influence or information diffusion. Tang and Liu (2010) studied the formation and detection of community and users' behaviours in social media websites, showing that individuals in the group interact with each other more frequently than with those people outside the group. Community detection is an important tool for the analysis of complex networks by enabling the study of microscopic structures that are often associated with organisational and functional characteristics of the underlying networks (Papadopoulos et al. 2012). Papadopoulos et al. (2009) proposed a method to detect communities from complex networks. Figure 2.1, cited from (Papadopoulos et al. 2009), presents the communities around the tags "computers" from a collaborative question/answering application, LYCOS iQ.

Another active research area is to learn and obtain people's ideas, opinions and sentiments about events, goods, products and services, using online user-generated content. The rich online public data is a significant data source for research communities on sentiment analysis and opinion mining. Senti-

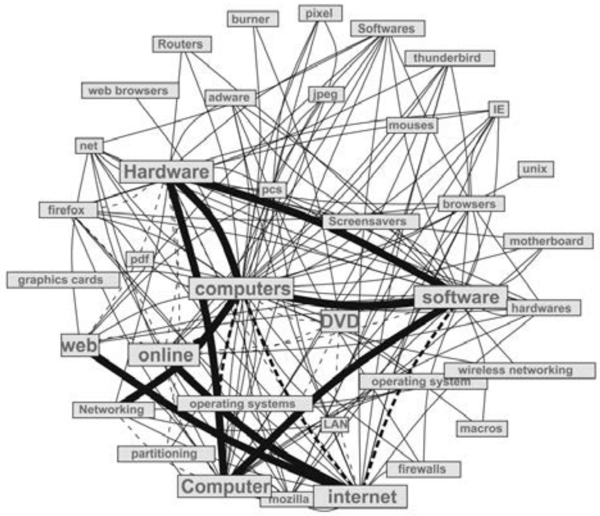


Figure 2.1: Communities around "computer" (Papadopoulos et al. 2009).

ment analysis (Pang and Lee 2008) extracts people's opinions on events and topics, as expressed in user-generated contents such as blogs. Consequently, various applications and services are built based on the opinions extracted. Ye, Zhang, and Law (2009) proposed a method of sentiment classification for online reviews to travel destinations that assigns sentiment class labels to the unlabelled reviews. Agarwal et al. (2011) classified overall sentiment of Twitter data. Sobkowicz, Kaschesky, and Bouchard (2012) mined citizens' opinions on elections. Liang and Dai (2013) determined people sentiment directions from micro-blogs data in Twitter. Figure 2.2, cited from (Sobkowicz, Kaschesky, and Bouchard 2012), presents the two online opposing communication networks (Opensource logic vs. Proprietary logic) opinions on the issue of governance of the Java software standard in 2002, showing that Java Opensource software was an issue that attracted some interest.

Folksonomies (Trant 2009) have become another rich data repository for researchers to mine useful information. Tagging services in social networking platforms allow users to add tags that enhance description of online content.

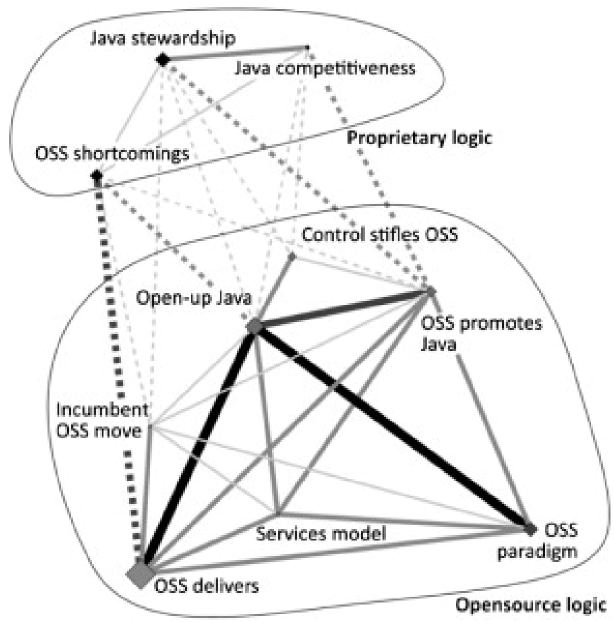


Figure 2.2: Topics, centrality, momentum and cross-references of important issues (Sobkowicz, Kaschesky, and Bouchard 2012).

Folksonomies are the bodies of tagged data. The information retrieval research community studied folksonomy from the perspective of its ability to support information retrieval (Mathes 2004; Hotho et al. 2006; Zhou et al. 2008). Kennedy et al. (2007) used the concept of representative tags and a tag-driven approach to extract place and event semantics, and to retrieve representative images using visual features. Rattenbury, Good, and Naaman (2007) conducted research investigating ways to extract place and event semantics from folksonomies. Another analysis of folksonomies seeks to build a tag recommender application based on the collective tags to find good tags for any resource as it is uploaded by the user (Jäschke et al. 2007; Sigurbjörnsson and Van Zwol 2008). Figure 2.3, cited from (Kennedy et al. 2007), shows a sample set of representative tags for San Francisco using Flickr data.

However, the above popular studies do not focus on the use and analysis of the abundant online geographic data associated with social media data

Figure 2.3: Representative tags for San Francisco (Kennedy et al. 2007).

entities. They mainly focus on the behaviours of social networks and relationships, sentiment and opinion on events and products using online textual tags, content entities and social networking data. Social media websites host sheer volume of geographically referenced data including documents, photos and videos (Zheng, Zha, and Chua 2011). Geographic-data-enriched online resources open up a new world of opportunities to discover the geographic-related knowledge and information of human society. These georeferenced data provide the information about events, activities and people that help understand geographic knowledge of objects, and behaviours of people. This study analyses the geotagged data to explore people's geographic-related behaviours.

2.2 Geotagged social media data mining

A recent novel research area is learning knowledge about physical geographic objects from online user-generated geographic referenced data. Online social media data contain massive amounts of geographic information benefiting from geo-tagging services. Geo-tagging enables users to add geographical identification metadata to media data. Moreover, location-based services enabled in cell phones or geo-tagging services on the sharing websites let users annotate geographic information to media data for sharing personal location information. Thus, the volume of geotagged social data provides a potential data-rich environment for geography (Sui and Goodchild 2011). These massive amounts of volunteered geographic information data are important repositories and play a critical role in understanding of geographic objects in the world (Goodchild 2007). Mining and analysing geotagged social media data aims to extract useful knowledge about geographic objects and to learn and understand geographic objects.

A recent useful analysis is using online geotagged data as immediate real-time geographical location information for emergency management. The geographic data is used to map online geographic data to the physical world for various areas. In particular, user-generated geographic information is crucial to fields of natural disaster and crisis management. Real-time geotagged social media data provide immediate location and diverse geographic information that are important and valuable in disaster management (Haworth and Bruce 2015). Goodchild and Glennon (2010) presented an early frontier study of using online user-contributed geographic information to assist official organisations with the response management for wildfire in Santa Barbara, US. In the same year, Zook et al. (2010) published a study on using volunteered geographic information for disaster mapping in the response to the 2010 Haiti earthquake. McDougall (2011) used the online geographic information from users' social media data to assist in mapping the flood extents in regions where there was little or no mapping available. Figure 2.4, cited from (McDougall 2011), shows a crowdmap of the Queensland floods based on the Ushabidi platform in January 2011 using individuals' generated geographic information data. These studies use rich online user-generated geographic information as extra volunteered data to learn about physical geographic objects and events. In contrast, this research study focuses on discovery of knowledge about people's geographical movement behaviours.

Figure 2.4: Map of Queensland flood (McDougall 2011).

Some studies exploit geotagged social media data to extract information on spatial locations to assist in exploration and understanding of spatial areas and objects. Ahern et al. (2007) used Flickr photo data to build a World Explorer tool to visualise tags representative of the spatial areas. Xie et al. (2013) detected the events that happened in a specific location. Other studies have a different application: building useful online applications to help users organise their content. Crandall et al. (2009) and Serdyukov, Murdock, and Van Zwol (2009) built a map model of users' content, predicted the location of users' photos according to the tags of photos, and helped place photos on the map. Some other interesting studies investigated people's social relationships in relation to physical locations. Crandall et al. (2010) learnt the social ties and Cranshaw et al. (2010) learnt the social networks between people from their physical geographic location of posted social media data.

Discovery of people's movement behaviours is another popular topic in mining online geotagged data. Geotagged data are people' footprints that point out the locations visited. These data are records of people's historical movement that provide opportunities to learn people's movement behaviours. Some studies have been conducted to extract the attractive and functional locations in urban areas from geotagged data, and people's preferences for spatial locations. Kisilevich, Mansmann, and Keim (2010) extracted the attractive areas that are characterised by high photo activity in a specific area from geotagged photos. Lee, Cai, and Lee (2014) presented a study identifying points of interest (PoI) where a great number of people gathered, from mining geotagged photo data, and further found out the association rules of visiting PoIs, to show which PoIs are visited together. Similarly, Hu et al. (2015) extracted urban areas of interest by clustering geotagged photos. Shirai et al. (2013) detected the areas of interest and shooting hotspots from geotagged photos by spatial clustering. An area of interest is tourist spots, whilst a shooting hotspot is a location where people take photos related to an event. Figure 2.5, cited from (Lee, Cai, and Lee 2013), presents a sample set of PoIs mined from Flickr geotagged photos. Figure 2.6, cited from (Lee, Cai, and Lee 2014), shows a sample association rule about two PoIs.

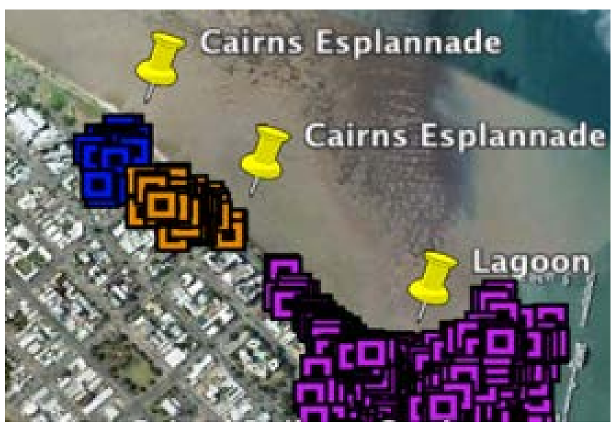


Figure 2.5: Sample set of point of interest (Lee, Cai, and Lee 2013).

However, previous studies on mining online geotagged data mainly focus on the extraction of spatial information and knowledge; there is a lack of studies analysing people's dynamic movement behaviour. The geotagged social data contains people's mobility data, including dynamic spatio-temporal trajectories. A geotagged entry is a person's footprint in the physical world. Consequently, a series of geotagged data, connected in time order, indicate

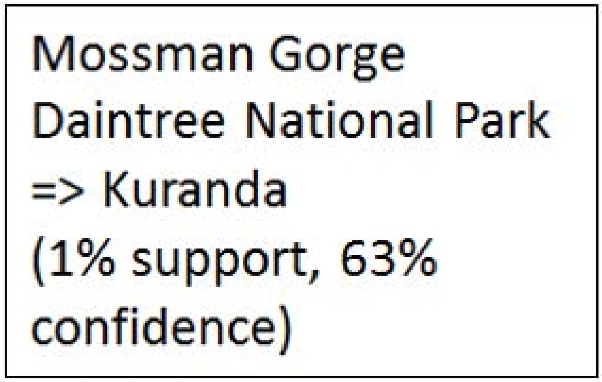


Figure 2.6: Association rule for PoIs (Lee, Cai, and Lee 2014).

a user's spatial trajectory during a certain time period. The vast amount of online geotagged data provides a potential data repository of people's trajectories, which may contain useful knowledge of people's mobility behaviours and patterns that are useful to various specific applications.

Our study aims to extract dynamic and mobile information from trajectories of geotagged photo data to help learn and understand human trajectory phenomena. Photo data is a rich online geotagged social data resource. Abundant photos are publicly available on the Internet with the advance of Web 2.0. Flickr is one of the most popular photo-sharing sites. Flickr allows people to upload and manage their photos, and communicate with others. It also provides rich Application Program Interfaces (APIs) to users that allow users to collect and explore photo data from their site. This study uses geotagged photos as examples of geotagged social media data to extract people's dynamic trajectory behavioural patterns. Figure 2.7 shows the geo-tagging interface in Flickr that allows users to drag and drop photos to a position on the map to geo-tag them. Figure 2.8, cited from (Zheng, Zha, and Chua 2012), shows the travel movement trajectories generated from geotagged photos in London.

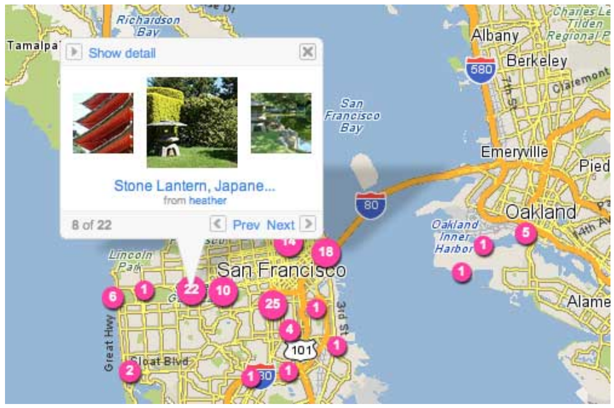


Figure 2.7: Geo-tagging in Flickr.

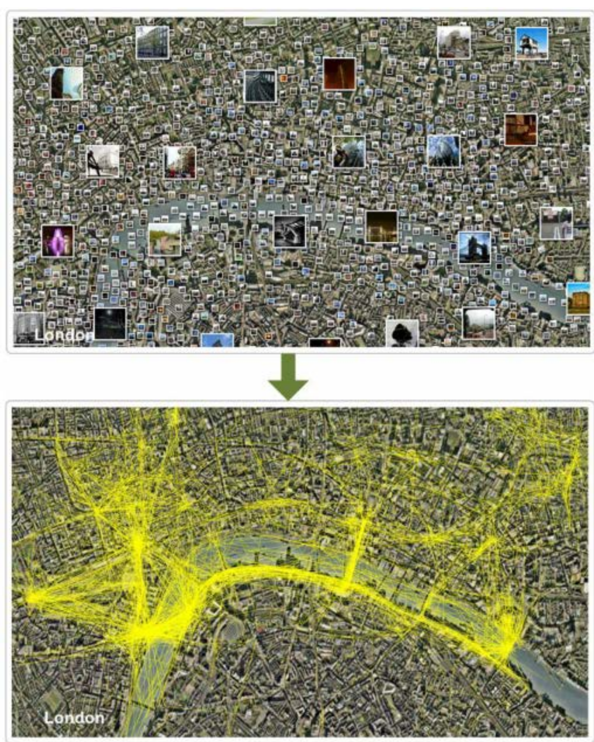


Figure 2.8: Moving trajectories generated from geotagged photos (Zheng, Zha, and Chua 2012).

2.3 Trajectory data mining

Few studies have investigated methods for extracting people's dynamic movement behaviours from geotagged photos. Girardin et al. (2008b) and Girardin et al. (2008a) presented early novel studies of explicit spatio-temporal data mined from online user-generated content to provide insights into human dynamics that illustrate understandings of visitors' dynamic movement flows in an urban space. Vu et al. (2015) investigated tourists' historical movement data from geotagged photo data to discover travel flows and directions for different tourist groups. Gao et al. (2014) detected people's regional spatio-temporal original-destination mobility flows from individual geotagged tweets. These studies investigated extraction and understanding of people's dynamic movement flow and structure from a potential trajectory repository generated from online geotagged social media data. Figure 2.9, cited from (Girardin et al. 2008b), shows the flow of visitors among main areas of tourist activity in Florence, Italy.

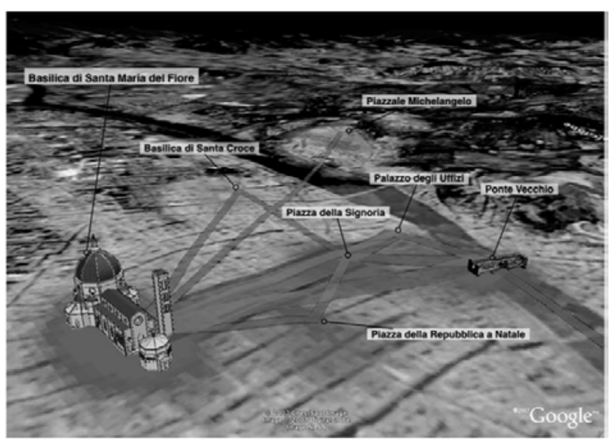


Figure 2.9: Trajectory flow of visitors (Girardin et al. 2008b).

However, they are not able to find out the frequent collective patterns over multiple people's moving trajectories. The aggregate patterns describe a group of moving objects sharing similar movement patterns. Successfully mining movement patterns has many significant applications in human mobility understanding, urban planning and transportation (Li 2014). Especially, this research discovers sequential patterns, frequent transition sequences of stops; common patterns, common paths shared by trajectories; trajectory patterns, and frequent sequences of stops with the typical transition time between stops.

2.3.1 Sequential trajectory pattern mining

Sequential pattern mining (SPM) from trajectories aims to find out sequences of locations that appear frequently in trajectories (Cao, Mamoulis, and Cheung 2005). A sequential pattern is formed by frequently occurring sequential movements from one location to the next location.

A sequential pattern mining problem is initially introduced to discover sub-sequences of transaction data that appear frequently in market transaction data (Agrawal and Srikant 1995), and then has been extended to explore sequential relationships from other sequence data. In this context, "frequent" means that the number of occurrences of a sequential pattern in the trajectory database is no smaller than the user-defined support threshold *minSup*. A transaction data represents a collection of item sets. Many efficient algorithms have been proposed to find sequential patterns from datasets, including GSP algorithm (Srikant and Agrawal 1996), Spade algorithm (Zaki 2001), FreeSpan algorithm (Han et al. 2000) and PrefixSpan algorithm (Pei et al. 2001).

Tsoukatos and Gunopulos (2001) first proposed a study on mining spatio-temporal patterns to find sequences of events that occur frequently in spatio-temporal datasets. They found ordered sequences of events that occur in the spatial locations. Similarly, Huang, Zhang, and Zhang (2008) discovered frequent sequence patterns of events. Most previous sequential pattern studies mainly aimed to find out frequent sequences of visit locations from trajectories. This study is different because it focuses on mining semantic people's trajectories and is interested in extracting people's semantic movement patterns.

Traditional sequential trajectory pattern mining

A trajectory shows a spatial mobility. It contains a spatial feature, which is spatial position, and a temporal feature, which is the time stamp when the spatial position is recorded. The spatial feature plays a principal role in the analysis of trajectories. Traditional studies on mining sequential patterns from trajectories are divided into two categories: those that consider spatial features only in 2-D spatial space; and those that consider both spatial and temporal features in 3-D spatio-temporal space.

The main group of studies uses spatial geometric features only of trajectories, while the time feature is used to order the spatial positions in the computing of trajectories. The final result is a sequence of spatial locations. As location coordinates do not typically match exactly at the same geographic coordinates in a pattern instance, most of these studies need a spatial neighbour function to determine whether two spatial positions are in the same meaningful spatial area or location. Cao, Mamoulis, and Cheung (2005) extracted frequent periodic sequential patterns from a long individual trajectory that appear repeatedly and periodically in the long trajectory. Using online geotagged photo data, Kisilevich, Keim, and Rokach (2010) generated people PoI-based travel sequences and found out tourists' frequent PoI sequential patterns. Majid et al. (2015) mined frequent visit sequences of locations from geotagged photos.

Another group of trajectory analyses applies spatial and temporal attributes. They consider the time dimension of spatial trajectories to compute the trajectories in a spatio-temporal 3-D space. The aim of considering the temporal dimension is to determine whether the spatial sequential patterns occur within the same specified time range. Bermingham and Lee (2014) found people's spatio-temporal sequential patterns from geotagged photos for tourism.

However, the main drawback of previous studies on mining sequential patterns is lack of aspatial semantics information, that is, they focus on traditional spatial geometric-only trajectory analysis. Trajectories can be enriched with additional application background and contextual semantics representing useful information about movement, and enhance some novel unknown semantic-level behaviours whose predicates bear on contextual semantics (Parent et al. 2013).

Semantic sequential pattern mining

Semantic trajectory aims to provide applications with semantic contextual knowledge about the movement. A few recent studies focus on mining of semantic trajectories (Alvares et al. 2007b; Zhang et al. 2014; Chen and Chiang 2016; Chakri, Raghay, and El Hadaj 2017). Alvares et al. (2007a) tried to enrich trajectories with semantic geographical information called semantic trajectories. They define the semantic trajectory as a sequence of stops with semantics annotations. Specifically, Alvares et al. (2007b) first found a set of stops in geometric trajectories, then matched the stops to semantic places, and finally mined frequent place sequences as sequential patterns. These semantic sequential patterns provide meaningful semantics information to better understand people's mobility. Zhang et al. (2014) added place category features to people's trajectories to extract frequently occurring sequences of visited place categories. These category-based sequential patterns show people's behaviours of visiting and moving among place categories.

However, semantic trajectories in previous studies only consider basic place type semantics of spatial location. There has been no similar study dealing with multi-dimensional semantics until now. This study concerns multi-dimensional semantic trajectories. Specifically, we add basic place type semantics and several additional environmental semantics to trajectories. The aim of this research is to find finer patterns and provide more meaningful semantics information. Previous methods for computing single dimensional semantic trajectory are not suitable for our multi-dimensional semantic trajectories.

Multi-dimensional sequential pattern mining has been studied for traditional market transaction data. Pinto et al. (2001) first tried dealing with several dimensions in the framework of sequential patterns. Different from traditional transaction data, in (Pinto et al. 2001) multiple customer information, such as category and age, is added to a sequence of purchased items. They focus on frequent sequences containing the additional profile information of customer and other information about the purchase. They use a combination of methods including PrefixSpan algorithm (Han et al. 2001), which is for finding frequent item-purchased sequences, and Bottom-up Cube computation (BUC) algorithm (Beyer and Ramakrishnan 1999), which is for computing frequent multi-dimensional value combinations, to find the multi-dimensional patterns. While the item of sequences used in (Pinto et al. 2001) is dealt with only in a single dimension, our study uses a multi-dimensional sequence where we consider several dimensions combined over time. Plantevit et al. (2005) displayed an approach for multi-dimensional sequences. Their study introduces a generalised multi-dimensional sequential pattern called jokerised patterns in which some dimension values may not be instantiated. But they focus on finding sequential patterns in inter-reference dimensions that describe the whole transaction sequence, while in our work, every stop entry is trajectory is enriched with several semantic annotations. We concentrate on the multi-dimensional trajectory.

2.3.2 Common trajectory pattern mining

Common trajectory pattern mining (CTP) aims to find the common trajectory tracks of any moving objects that have similar trajectories. A common trajectory pattern shows popular movements of people. A common trajectory pattern is indicated and triggered by a group of similar trajectories. That is, to find out the common trajectory patterns in a trajectories database, we need to gather similar trajectories into clusters, and then a group of similar trajectories reveals one common movement track.

Clustering is the task of grouping objects so that objects similar to each other are in the same cluster but dissimilar objects are in different clusters (Han, Pei, and Kamber 2011). Clustering is one of the most popular data mining techniques, which has been widely used in various applications, such as geographic data (Miller and Han 2009). There exist many classic clustering algorithms, including partitioning-based k-means and k-medoids, hierarchical-based BIRCH (Zhang, Ramakrishnan, and Livny 1996) and CURE (Guha, Rastogi, and Shim 1998), density-based DBSCAN (Ester et al. 1996) and OPTICS (Ankerst et al. 1999). Several traditional clustering methods have been applied to geographic spatial points data for geographic data analysis in various specific applications, including crime hot-spot analysis (Estivill-Castro and Lee 2000).

Traditional common pattern mining

Several studies have been undertaken on trajectory clustering for extracting common patterns of moving object in different applications. Many clustering methods for trajectory data have been proposed. In the previous studies, trajectory data are modelled in varied representations by different methods for specific purposes. There are two groups of trajectory clustering analysis: whole trajectory or sub-trajectory. In whole trajectory analysis the whole trajectory is the analysed object. Such work clusters whole trajectories. Gaffney and Smyth (1999) proposed a model-based method to cluster whole trajectories. An individual trajectory was represented as mixtures of regression models and then the EM algorithm applied to cluster trajectories. Nanni and Pedreschi (2006) proposed a density-based T-OPTICS algorithm, adopting the OPTICS clustering algorithm, for trajectory clustering. A trajectory is represented as a sequence of locations. The distance between two trajectories used in the method is defined as the average distance between objects that is based on the Euclidean distance between spatial points. Another group is to find local trajectory patterns. Lee, Han, and Whang (2007) claim that some portions of trajectories show a common behaviour. This kind of work focuses on clustering similar sub-trajectories. Lee, Han, and Whang (2007) proposed the TRACLUS method for detecting similar portions of trajectories. The method represents each trajectory as a line and clusters the spatially nearby line segments with similar shape using DBSCAN algorithm. Overall, these methods use different similarity measure approaches that are based on spatial or spatio-temporal and other features of trajectory data. Most previous studies employ traditional spatial trajectory analysis. This analysis method groups trajectories into a cluster using the basic spatial geometric attribute, so that trajectories that have similar geometrics are grouped together.

Some studies take into account the important temporal dimension in their trajectory clustering, especially for finding local spatio-temporal common patterns. The temporal feature of trajectory is the time stamp of spatial points in the trajectory. The time stamps are used to define a local common pattern, moving together in a specific consecutive time duration, from trajectories. Gudmundsson, Kreveld, and Speckmann (2004) used the absolute time condition to find trajectory flock patterns. A flock pattern refers to a group of a frequent number of trajectories moving spatially close for a specific consecutive timestamp. Using a loose requirement of consecutive timestamps, Jeung et al. (2008) extracted the convoy patterns from trajectories. Moreover, Li et al. (2010) detected the swarm patterns from trajectories without requirement of consecutive timestamps. To cluster trajectories only in meaningful time intervals, Nanni and Pedreschi (2006) proposed the TF-OPTICS for temporal focusing trajectory clustering. It finds the moving clusters in a set of objects that move close to each other for a long time interval.

However, previous traditional spatial feature-focused common pattern mining lacks the aspatial semantics features of trajectories. As mentioned above, the semantics feature provides opportunities to reveal potential semantic-level common movements of people, which are useful to specific applications. Moreover, these aspatial semantics provide trajectories with meaningful background contextual information that is helpful to understanding people's trajectory behaviours. This research is discovering the semantic common patterns from trajectories by using spatial, temporal and aspatial semantics features.

Semantic common pattern mining

Recently, a few studies have presented attempts to consider aspatial semantics information for trajectory data mining reflecting the importance of aspatial semantic information (Parent et al. 2013). But there is no such study for mining semantic common patterns from trajectories with consideration of additional aspatial semantics features. Bermingham and Lee (2015) develop an ND-TRACLUS method for multiple-dimension trajectory clustering. It is an extension of TRACLUS method (Lee, Han, and Whang 2007), which is a well-known spatial trajectory clustering method, to multi-dimensional trajectories. ND-TRACLUS has an ability to uncover new, previously unknown, higher dimensional trajectory patterns. This method considers additional speed and direction semantics features of the geographic trajectory. However, it is originally designed for GPS trajectories. The GPS trajectory data is different from the trajectory generated from geotagged photos as the latter trajectory data is neither continuous nor regular. Similar to the original method, ND-TRACLUS groups similar spatial trajectories according to a spatio-temporal geographic proximity measure but takes additional speed and direction aspects into account. It is spatial-trajectory oriented, whilst our study is both spatial- and aspatial- oriented.

2.3.3 Trajectory pattern mining

Trajectory patterns consider the typical time information in people's mobility. A trajectory pattern is a frequent sequence of locations with typical interval time information between two locations (Giannotti et al. 2007). The trajectory pattern not only indicates the sequential features of the visited locations, but also discloses important information about time relations between locations. The time information in the trajectory pattern shows a typical time interval from one location to another that frequently occurs in individual trajectories. As mentioned in (Yoshida et al. 2000), the interval time provides people with information of the "specific time after" (a period) whilst sequential pattern without interval time only tells "after" (a succession in time), and the order information of locations in the sequence. This interval time included patterns useful to understanding of human movement. It provides the important time information about when the locations are visited.

Mining frequent sequential patterns with interval time annotations has been studied for transactional sequence dataset in the recent decades. Yoshida et al. (2000) presented an early investigation in data mining frequent sequential patterns including time intervals, named delta patterns, from market purchasing data. A delta pattern is an ordered list of item sets with the time intervals between two neighbouring item sets. Each time interval is a range including several values. Vautier, Cordier, and Quiniou (2005) also extracted sequential patterns with time intervals from events sequences. They call the interval times "chronicles". Unlike (Yoshida et al. 2000), the chronicles can be intervals with negative bounds. Similarly, Chen, Chiang, and Ko (2003) propose a method to extract time-interval sequential patterns from purchasing datasets. These patterns reveal not only the order of items purchased, but also the time intervals between successive items. Hirate and Yamana (2006) discovered sequential patterns with time intervals from real earthquake events, demonstrating that the time-integrated patterns are more useful than conventional sequential patterns without time information. Giannotti, Nanni, and Pedreschi (2006) proposed a prefix-projection-based algorithm to mine the temporally annotated sequences, giving a sequential pattern with typical transition time.

Traditional trajectory pattern mining

Several investigations have been undertaken into mining the time-annotated trajectory patterns from geographic spatio-temporal trajectories. Giannotti et al. (2007) first extracted spatial location sequential patterns with temporal relations between locations from trajectory data. They proposed the trajectory pattern mining (TPM) method for mining trajectory patterns. Later, Lee, Chen, and Ip (2009) proposed a graph-based method to find the frequent trajectory patterns from trajectories. The method refers spatial points to square spatial space and builds graph for trajectories with vertex as the label of spatial location. The trajectory patterns – sequences of locations with time spans – are then generated with a condition of maximum time-span threshold. Focusing on a different temporal aspect, Kang and Yong (2010) extracted frequent spatio-temporal patterns with a duration time spent in the spatial location from trajectories. The method first finds spatio-temporal regions in a spatio-temporal 3-D space, and then mines frequent spatio-temporal patterns based on a prefix-projection approach from the sequences of these regions.

Few studies have found the time-interval-enhanced trajectory patterns from trajectories of online geotagged photo data. Cai et al. (2014) extracted people's trajectory patterns from geotagged photos by using the TPM method with an advanced RoI mining method. An RoI shows a dense region that many trajectories visit. The extracted trajectory patterns show how people collectively visit sequences of RoIs, and the popular transition times between two neighbouring RoIs. Arase et al. (2010) detected people's trips from geotagged photos. They model people's trip sequences based on geographic information from photos and then mine the frequent visit sequences of cities and the typical duration between two consecutive cities.

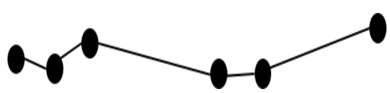
These previous studies explore traditional trajectory pattern mining that mainly focuses on spatial features of trajectories. The extracted spatial trajectory patterns provide order information of visited locations in the sequence, the interval time information and spatial information of movement. However, these spatial patterns are not able to provide meaningful semantics information. Parent et al. (2013) emphasized the importance of aspatial semantic information in trajectory data mining. The aspatial semantics information was recently considered for trajectory analysis to reflect the importance of aspatial semantic information (Ying et al. 2011; Wang et al. 2013). For trajectory patterns, the traditional spatial feature-only methods could not trigger the application background semantic-level behaviour of people's movement patterns. And these spatial-geometric methods are not specific for semantic trajectory patterns. This study, however, considers the aspatial semantics feature of trajectory to extract patterns with both semantic and temporal features.

Semantic trajectory pattern mining

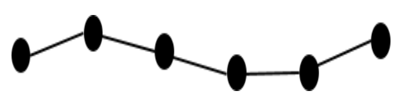
The main aim of semantic trajectory mining is to provide applications with semantic knowledge about movement, going beyond geographic-feature-only trajectories. Recently, some research (Chen, Kuo, and Peng 2015; Chen and Chiang 2016) attempts to extend PrefixSpan to incorporate semantics and time information by transforming trajectory sequences into symbolised sequences before using PrefixSpan. However, the transformation of spatio-temporal trajectories into symbolised sequences can mask off important spatio-temporal trajectory patterns, and these studies do not consider various spatial and aspatial semantic databases as we investigate in this study. The task of semantic trajectory pattern mining takes into account time information to extract semantic patterns, which are sequences of semantic stops with typical interval time between stops.

2.3.4 Comparisons of geotagged photo-based trajectories and GPS-based trajectories

In the trajectory data mining area, many research studies have been done for GPS-related trajectory data, due to the increasing prevalence of GPS devices and fast growing availability of trajectory data. Some of these studies have been discussed above. In particularly, Zheng et al. (2009) mined the interesting locations and travel sequences from GPS trajectories. Lee, Han, and Whang (2007) developed a partition-and-group framework in which trajectories are partitioned into a set of quasi-linear segments and DBSCAN algorithm is applied to cluster these segments to find the common sub-trajectories. Giannotti et al. (2007) presented a trajectory pattern mining algorithm to extract frequent movement patterns, which are the sequences of places visited by different objects with similar time intervals. Kang and Yong (2010) proposed a method to find spatio-temporal trajectory patterns by partitioning trajectories into segments and clustering the segments.



(a) Geotagged photo trajectory



(b) GPS trajectory

Figure 2.10: Examples of geotagged photo-based trajectory data and GPS-based trajectory data.

Though various approaches have been proposed and applied on GPS-related data to extract trajectory patterns, we cannot directly adopt those methodologies to analyse the trajectory data of online geotagged photo data, due to two main reasons. The first reason is the significant difference in features between GPS log data and geotagged photo data. GPS logging

Table 2.1: Comparisons of characteristics for geotagged photos data and GPS data.

	Geotagged photos	**GPS**
Characteristics	• Manual disclosure of location • Irregular recording	• Automatic capturing of trace • Regular recording
Characteristics of trajectory data	1. Trajectory in geotagged photos can only show the order of positions visited 2. Spatial information between two connected points in individual trajectory can be non-closed 3. Temporal information is irregular	1. Trajectory in GPS can illustrate the detail of route the object is taking 2. Spatial information between two connected points in individual trajectory is closed 3. Temporal information is regular

devices can record the movement continuously, while trajectories generated from geotagged photo data only contain sporadic spatial and temporal information. Figure 2.10 shows examples of trajectory data generated from geotagged photos and GPS-related data in Figure 2.10 (a) and Figure 2.10 (b), respectively. Table 2.1 summarises the main differences between the spatial features, temporal features, size of points and techniques for processing trajectories from the two data sources. The second reason is the difference in dimensions and features of trajectory that the methods focus on. Previous methods are designed for traditional spatial geometric feature trajectory analysis whilst our study considers additional aspatial semantics features of trajectory. Therefore, those techniques originally designed for and used in GPS log data mining cannot be directly applied to this research project. For these two reasons, the initial goal of this project aims to propose a data mining framework to analyse geotagged photo-generated trajectories and to extract semantic trajectory behavioural patterns.

2.4 Travel itinerary recommender system

Building travel recommender systems using the vast online geotagged datasets is another popular research area. Recommendation systems aim to assist people with their travel planning. Online social media data plays an increasingly important role as information sources for travellers (Xiang and Gretzel 2010). Especially, the geographic information-annotated data contains rich experience data about destinations. The previous trajectory data formed from geotagged data also provided information on movement during travel. Using online data to build applications for travel information search and retrieval is one important research area (Xiang and Gretzel 2010). These applications make people search for travel information from the huge amount of online travel experiences of people who have already visited the locations. The recommender system research community focuses on building a useful recommendation tool and application that aims to suggest products and provide users with information to facilitate their decision-making processes (Schafer, Konstan, and Riedl 2001). A popular travel recommender system gives location-focused recommendations that provide people with assistance on ways to travel to the destination, including location recommendations, route recommendations and itinerary recommendations. These recommendations suggest where to go and what to visit, offering the sequential routes and useful time information.

Location recommender systems suggest to travellers the best locations in the destinations. Personalised recommendation considers user's preferences and recommends locations that match user's interests learned from their online travel history. Popescu and Grefenstette (2011) recommended landmarks to users based on destinations that had been visited by similar users, who are measured based on the landmark histories. Yamasaki, Gallagher, and Chen (2013) recommended personalised landmarks to users in inter-city. Shi et al. (2011) measured the similarity between users by using an additional category of landmarks. Chen, Cheng, and Hsu (2013) recommended personalised next destinations to users based on their gender, age, and travel group types, which are detected from the photo image features. Majid et al. (2013), and Memon et al. (2015) recommended personalised tourist locations, which are relevant to the temporal and weather context environments. There are several other services considered in various location recommender systems. One function is context-aware recommendation. This kind of system generates more appropriate locations that are relevant to the environmental context of the travel (Adomavicius and Tuzhilin 2015), like spatial, temporal and weather conditions. Temporal context of movements is about the time the user visits a destination. Some popular temporal contexts used are monthly, weekly and daily. Van Canneyt et al. (2011) recommended popular POIs in the specific time context. Similarity, Bhargava et al. (2015) considered the time context in addition to recommendations of location and activities.

Yamasaki, Gallagher, and Chen (2013) considered the seasonal and temporal contexts to recommend the next landmarks in intra-city travel. Majid et al. (2013), and Memon et al. (2015) recommended personalised tourist locations which are relevant to the temporal and weather conditions. Another function is considering semantics information of locations, especially the type or category of location. Shi et al. (2011) measured the similarity between users by using an additional category of landmarks.

A good route recommender system suggests to people specific travel routes for appropriate locations. The travel route and visit sequence of locations provide useful information about visit order of these places and an appropriate integrated route. Okuyama and Yanai (2013) recommended travel routes to the target destinations that were reconstructed from trajectory data formed from geotagged photos. Sun et al. (2015) recommended most popular landmarks, with the best travel routings between the landmarks based on road network. However, these route recommendations lack time information. Time is significant information for travellers, telling people how much time they will spend. It will provide better advice to people who have predefined travel duration time and let them plan their travel itinerary better. This study is to recommend routes with interval time information between stops: the itinerary recommendation.

2.4.1 Traditional itinerary recommender system

Itinerary recommender systems recommend movement routes with important associated travel time information to help people plan a travel itinerary. These recommender systems provide appropriate itineraries, in which the total cost time of the itinerary fits the time budget constraint that users pre-define. One kind of time information is the typical transit time between neighbourhood locations of the route. Kurashima et al. (2013) generated sequences of locations with both stay time and transit time. Lu et al. (2010) and Lim et al. (2015) generated recommendations of sequences of locations with stay times at each location. Both methods built a graph-based travel sequence model and then generated appropriate candidate itineraries from the model. De Choudhury et al. (2010) generated sequence of locations with both stay time and transit time.

These previous studies consider various users' travel requirements and constraints, like travel duration, distance, and recommend sequence of specific geographic spatial locations. But, they are lack of function for dealing with requirement of semantic type of place that users want to visit. Rather than specific geographical locations, travellers who plan to visit an unfamiliar area where they do not know any specific places, are still likely to want to visit some place types in certain weather conditions, and to customise these requirements as a constraint, like beach in good weather, restaurant in any weather, and cultural park in wet weather. Previous itinerary systems are lack of consideration of the semantic query. Symeonidis, Ntempos, and Manolopoulos (2014) considered an additional semantic category of landmark, but they focused on landmark recommendation. This study considers users' semantic type-of-place query for itinerary recommendations.

2.4.2 Semantic itinerary recommender system

A higher semantic-level itinerary is also important for users' travel planning. Chen et al. (2011) and Gionis et al. (2014) let users customise the category visit sequence they preferred, and then generated specific geographic routes that matched the sequence and were appropriate to users' actual situation, including their actual position. These studies prove that the higher semantic-level itinerary is useful. But there is no such work to recommend a higher semantic-level itinerary to users. Previous itinerary recommender systems produce final specific geographic itineraries that could not solve the semantic-level itinerary problem. This study focuses on semantic-level itinerary recommendations. Moreover, our system produces an itinerary with additional useful recommendations for rich and meaningful contextual information.

2.5 Summary

An enormous amount of online social media data provides a data-rich environment available to various research communities for specific research purposes and applications. In particular, the online user-generated geographic information, geotagged social data, becomes a potential geographic data repository for understanding of objects in the physical geographic world. Several studies have analysed the online geographic information referenced social data to learn geographic objects and people's behaviours, including reconstructions of maps, spatial hotspot detection, and detection of PoIs. Unlike past work, this study aims to extract people's dynamic semantic trajectory patterns from the geotagged photo data.

Some studies have discovered people's dynamic movement behaviours and patterns from geotagged photo data, such as people's trajectory flow in cities (Girardin et al. 2008a). As listed in Table 2.2, this study will extract three types of mobility patterns from geotagged photos: sequential patterns, common patterns and trajectory patterns. SPM is to extract the visit sequences of stops that frequently occur in people's trajectories. CTP is to mine the common trajectory tracks of a group of people who have similar trajectories. TPM is to find out the frequent visit sequences of stops with important interval time information between consecutive stops. These patterns help domain experts understand people's movement behaviours including common path and time information. As the online data contains rich travel experiences, travel recommender system communities try to use these repositories to suggest itineraries to users and assist users with travel planning. An itinerary is a visit sequence of stops with transition time information between two stops. IRS provides useful suggestions to travellers about travel routes and time planning.

However, there are two important problems motivating this study. One motivation is a comprehensive framework for a set of important processes.

Table 2.2: Comparison of literatures.

	Important features			**Important functions**			
	spatial	temporal	aspatial	SPM	CTP	TPM	IRS
Traditional SPM	yes	yes		yes			
Semantic SPM	yes	yes	yes	yes			
Traditional CTP	yes	yes			yes		
Semantic CTP	yes	yes	yes		yes		
Traditional TPM	yes	yes				yes	
Semantic TPM	yes	yes	yes			yes	
Traditional IRS	yes	yes					yes
Semantic IRS	yes	yes	yes				yes
Our proposed framework	yes	yes	yes	yes	yes	yes	yes

There are four tasks, including analysis of three kinds of mobility patterns and construction of the itinerary recommender system. There were studies for each of these tasks, but no framework proposed to undertaken all four tasks. The other motivation is consideration of three important features for geotagged photos: spatial (since they are geotagged), temporal (since they are time stamped) and aspatial (semantic metadata to add additional information to trajectories). Whilst there were some studies considering spatio-temporal aspects, there was nothing considering these three together for geotagged photos. Aspatial semantics features are important in trajectory analysis for many applications. Table 2.2 summarises and highlights the differences between our study and past studies in considerations of these important three features and functions. In summary, there were some studies but none satisfies the three important features, and four important functions. The aim of this thesis is to propose a semantic trajectory framework for geotagged photos considering three important features – spatial, temporal and aspatial – and also providing four important functions for decision-making.

Chapter 3

Overall framework

This chapter introduces the overall framework proposed in this thesis. It first summarises the framework for mining semantic trajectory behavioural patterns from online geotagged photos in Chapter 3.1. The photo dataset collection used in this study is introduced in Chapter 3.2. This research uses two external aspatial semantics information databases for semantics information enrichment, which are described in Chapter 3.3. Finally, Chapter 3.4 lists and introduces the general definitions and terms used in this thesis.

3.1 Framework

This thesis proposes a systematic data mining framework for extracting people's dynamic trajectory behavioural patterns from geotagged photos. Enormous amounts of publicly available online photo data, which are already referenced with geographic information, provide a movement data repository of great potential to various research communities and applications. This research study aims to mine people's movement patterns from the geotagged photos, as shown in Figure 1.1 in Chapter 1.3. A methodical framework is proposed in this thesis to achieve the aim. Overall, the framework, shown in Figure 3.1, extracts people's collective semantic trajectory behavioural patterns that show their mobility behaviours in semantic level and are enriched with meaningful semantics information. This data is then used to build a system to recommend travel itineraries to users based on the extracted trajectory patterns. This framework contains four main functions that solve the four specific research aims: mining meaningful semantic sequential patterns, discovering understandable semantic common patterns, finding meaningful semantic trajectory patterns and building the itinerary recommender system using extracted patterns.

Figure 3.1: Overall framework of proposed semantic data mining.

Figure 3.2 presents the detailed main framework proposed in this thesis. The main framework contains six modules linking to each of six objectives: (1) building raw trajectories from photo data; (2) building semantic trajectories; (3) mining semantic sequential patterns; (4) discovering semantic common patterns; (5) extracting semantic trajectory patterns; and (6) building an itinerary recommender system. In summary, this research collects photo data from Flickr website, which is a popular photo-sharing application. From the photo data, we build people's geographic trajectories and then generate semantic trajectories. The semantic trajectories are generated using an external aspatial semantics database to enhance the trajectories with application contextual aspatial semantics. We then extract three kinds of trajectory patterns from the semantic trajectories: semantic sequential patterns, semantic common patterns and semantic trajectory patterns. Finally, we build a semantic itinerary recommender system based on the extracted semantic trajectory patterns. A more detailed summary of each module follows.

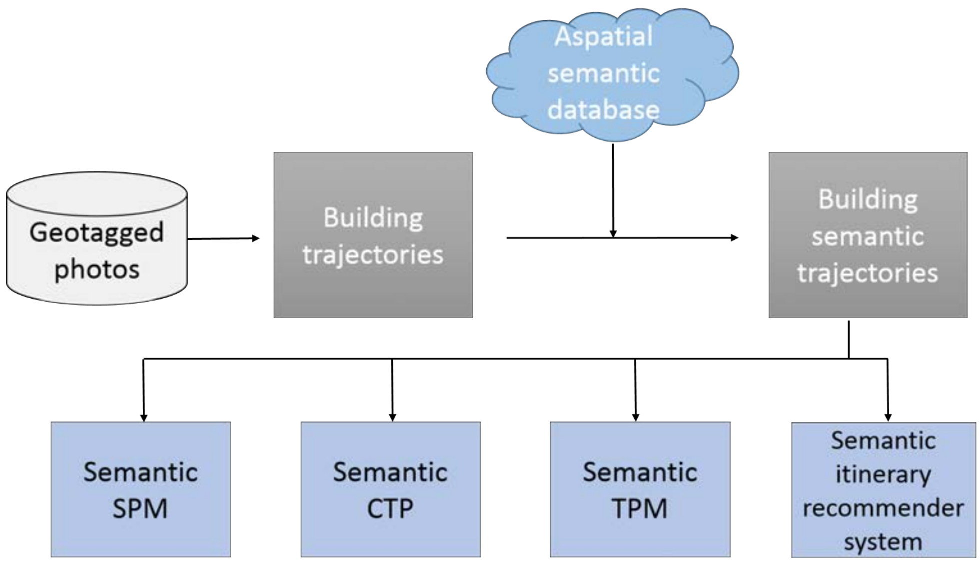


Figure 3.2: Main framework of proposed semantic data mining.

3.1.1 Modules of framework

Building trajectories

The first module in the framework is to build people's trajectories from geotagged photos. The trajectory data of geotagged photos are the focused data in this research study. We extract people's mobility patterns from the trajectory data. A geotagged photo indicates the spatial point and time stamp of a visit. Given a set of geotagged photos from a single photo-taker, we identify a raw geographic trajectory. Each raw trajectory is represented as a sequence of spatial points with time stamps that show the dynamic spatio-temporal mobility. The detailed process of building trajectory data is presented in Chapter 4.

Building semantic trajectories

The second module is to build people's semantic trajectories. The raw geographic trajectory data contains only spatial and temporal information. In this study, we focus on mining people's meaningful semantically annotated mobility patterns. To find such patterns, we enrich raw spatial trajectories with additional contextual aspatial semantics information. Later, we use this trajectory data with spatial, temporal and aspatial semantics features to discover the semantic patterns. We generate semantic trajectories from raw trajectories using external aspatial semantics databases for semantic enrichment. A semantic trajectory is defined as a sequence of stops with contextual aspatial semantic annotations. In this study, a stop is a RoI that many trajectories passed through. We annotate each RoI with basic background contextual aspatial semantics, which is the type-of-place annotation of geo-object in the RoI in this study. In addition, we enrich stops with multiple extra contextual environmental semantics including spatial, temporal and weather conditions. Finally, the format of the semantic trajectory is a sequence of type-of-place annotations with a set of additional environmental semantics. We generate semantic trajectories from raw trajectories by using the proposed semantic RoI mining algorithm and a further semantics enrichment process. Detailed information about generating the semantic trajectories is set out in Chapter 4.

Semantic sequential pattern mining

Semantic sequential pattern mining is the first main function of the framework. The function is to extract the semantic sequential patterns from the semantic trajectories. A sequential pattern is a frequent sequence of visited stops that is shared by a group of people. At the basic semantic level, a semantic sequential pattern is a sequence of basic type-of-place semantics annotations. This basic pattern shows people's sequential trajectory behaviour in visiting types of places. We find not only basic patterns, which contain only type-of-place semantics, but also multi-dimensional sequential patterns, which are basic patterns with sets of additional semantics. The study of semantic sequential pattern mining is further introduced in Chapter 4.

Semantic common pattern mining

The second main function of the framework is semantic common pattern mining. This task is to discover, from the semantic trajectories, people's similar trajectories that show a common track pattern. At the semantic level, a semantic common pattern is a semantic trajectory compiled from a group of similar trajectories undertaken by different people. A common pattern refers to a cluster of similar trajectories. This study proposes a semantic trajectory clustering approach to disclose all the semantic common patterns in people's semantic trajectories. Specific content on mining semantic common patterns is presented in Chapter 5.

Semantic trajectory pattern mining

The third task is mining frequent semantic trajectory patterns. A set of trajectories may share the property of visiting the same sequence of stops with similar interval times. This property shows a frequent trajectory pattern of people. Trajectory patterns consider the important interval time knowledge that indicates time relations between stops. A basic semantic trajectory pattern is a sequence of stops integrated with type-of-place annotations with interval time between two stops. We find out both basic semantic trajectory patterns and multi-dimensional semantic trajectory patterns, which are basic semantic trajectory patterns with sets of additional semantics. Chapter 6 describes thoroughly the study of semantic trajectory pattern mining.

Semantic itinerary recommender system

Finally, the framework builds a semantic itinerary recommender system to assist people with travel planning based on extracted semantic trajectory patterns. The system recommends travel itineraries as a route of stops with typical interval time information. The recommender system is developed as an offline–online structure. In the offline component, frequent semantic trajectory patterns in previous photo-takers are extracted. The trajectory patterns represent the frequent stop visit sequences and the typical interval time between stops. In the online component, the system receives users' queries and generates recommendations using the trajectory patterns. This itinerary recommender system also enables users to customise a set of preferred types of place and the total travel duration constraint. When someone plans to visit one destination, given a set of required types of places and travel duration, this system produces and lists possible itineraries with multiple additional valuable contextual environmental information. Details about building the semantic itinerary recommender system are presented in Chapter 7.

3.2 Photo dataset collection

This research study uses geotagged photo data collected from Flickr photo-sharing platform. Flickr is one of the best-known photo-hosting websites. Users share their stories with photos, and add other relevant information using tags (folksonomy). In particular, Flickr has been a popular way of communication for tourists to share their travelling stories and moments with their personal links and social networks. Flickr also works with online maps such as OpenStreetMap to provide geo-spatial references. Michel (2017) reported in January 2017 that by 2016 Flickr had a total 5.87 billion public photos and 1.68 million public photos were uploaded daily, on average, in 2016.

We collect photo data using the Flickr API. The Flickr API allows developers to access and collect Flickr data, and is available for non-commercial use by outside developers. In particular, we use flickrpy API kit[1], a Python library for the Flickr API, to collect photo data and the metadata associated. We use the 'photo.search' API method to query photo data. In this study, we set up parameter "bbox", a comma-delimited list of four values defining the bounding box of the area that will be searched, and the parameters "min taken data" and "max taken data", which are minimum and maximum dates when photos were taken, This search method returns a list of photos in the search area and time period. Returned photo data contains some metadata, including upload date, taken date, latitude, longitude, owner id, owner name, place id, tags, title, and description, etc.

This study uses the data of user id, photo id, latitude, longitude and taken-time of photos. The format of photo data used in this study is shown in Table 3.1 with a sample photo data. A user id is a user ID of the photo owner assigned by Flickr. A photo id is the id of a photo. Latitude and longitude are the geographic coordinates a photo referenced. Time is the time stamp of when a photo was taken.

Table 3.1: Format of photo data for this study.

user_id	photo_id	latitude	longitude	time
67908321@N02	8703168606	-25.286436	152.883521	2014-05-13 17:20:35

The photo taken time information used in this study is captured directly from the Flickr photo metadata. The time information is set by either the photo taking device's time or the photo upload time. Users' photo taking devices, such as cameras or mobile phones, could be set to their own time zones where the users are from, and the time on their photo taking devices may not be adjusted correctly to local time. That is, the time of photo could

[1] https://code.google.com/archive/p/flickrpy/

not be the correct local time of the visit. Since this study mainly focuses on the pattern mining framework and methods for finding semantic patterns, our pre-processing phase does not correct the time information in different time zones. However, we leave this one as one of our future work.

This study collects photos in Queensland, Australia, from 1 April 2014 to 30 March 2015. Australia is one of the biggest tourist destination countries in the world. Its beautiful scenery and competitive industries attract numerous people to travel there. Queensland is a state of many landscapes that range from sunny tropical coastal areas, through lush rainforests to dry inland areas. Queensland is blessed by natural beauty, especially along the eastern coastline and it is home to many well-known amusement parks. The tourism industry plays a key role in the economies of regional areas and supports thousands of small businesses. This study analyses the photo data to mine people's trajectory patterns in Queensland, Australia, providing information on people's movement behaviours to tourism-related organisations.

As the raw Flickr geotagged photo dataset contains noise and redundancy, data cleaning is required to remove faults and redundancy from the dataset. Since the temporal dimension is an important factor for trajectory, it is crucial to ensure that the data has the correct time attached, thus stripping incorrect time annotations from the photo dataset is an important cleaning process. After pre-processing, we have a total of 64,733 cleaned photo records. Figure 3.3 presents the photo points on NASA earth. Most of the photos are located on the coastlines, where the big cities are.

We classify the photo-takers into tourists and non-tourists based on the time span of trajectory. A geotagged photo is associated with geographic location and time information, and all photos when connected chronologically result in a spatio-temporal trajectory. The time span of a trajectory is calculated by using the time gap between the last photo and the first photo. We consider a photo-taker as a tourist if the time span of trajectory is less than 31 days, otherwise, this photo-taker is defined as a non-tourist. However,

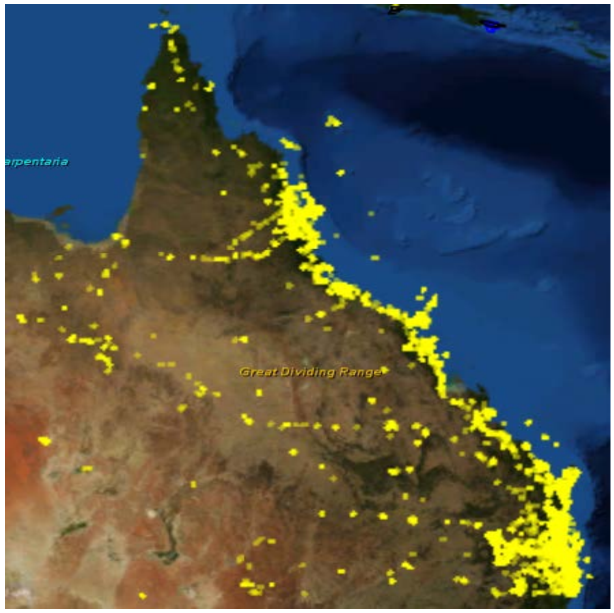


Figure 3.3: Flickr photo points in study region.

users are able to reset this value for exploratory data analysis.

3.3 Aspatial semantics databases

This study requires external aspatial semantics information databases to enrich trajectories with contextual semantic annotations. One semantics information database is the geographic information database. This database is used to annotate the type-of-place semantics to a spatial place. Another database is weather observation information database. It is used to annotate the weather condition environmental semantics in which each visit to a place occurred.

Background geographic information database is collected from GeoNames[2]. The GeoNames geographical database covers all countries and contains over ten million place names that are available for download freely. It

[2]Geonames: http://www.geonames.org/

consists of over 9 million unique features, and all features are categorised into one of nine feature classes and further sub-categorised into one of 645 feature codes. This study uses the feature codes as the type-of-place semantics information for trajectories. In particular, this work uses the GeoNames Australia Gazetteer database. We use the main 'geoname' table, including fields of latitude, longitude and feature code. The GeoNames Australia geographical database is used to do the reverse geocoding that receives the feature code of the nearest place based on queried geographic coordinates.

Reverse geocoding technique is the process of back coding of a point location (latitude, longitude) to a physical place. However, there are some possible limitations of the proposed approach using reverse geocoding. One limitation is the multiple geo-objects retrieval in crowded urban areas. That is, several places could have the same nearest distance to a point location, and the reverse geocoding technique could return all these places. Another possible limitation is the uncertain geo-objects retrieval due to spatial uncertainty (Goodchild 2008; Zheng 2015). These two are inherent limitations with reverse coding, and in this study we utilise the first type of place found from databases in order to overcome these limitations as in other studies (Cao, Cong, and Jensen 2010).

The weather information database is collected from Bureau of Meteorology Australia[3]. This database contains both an observation stations database and a daily weather observations database. This study uses only the stations in Queensland, using data from the stationID, stationName, latitude and longitude attributes. This study collects the daily weather observations data for Queensland from 1 April 2014 to 30 March 2015. At last, 121 records are stored for the station list for Queensland, and 47,949 records of daily weather observations are collected and stored for Queensland, Australia.

[3]http://www.bom.gov.au/climate

3.4 Preliminaries

This section introduces some basic definitions and terms used in this thesis. This study uses geotagged photo data, and a geotagged photo is associated with geographic location information and a time stamp. As described in Chapter 3.2, this study uses the data of user id, photo id, latitude, longitude and timestamp of photos. A photo data is represented in Definition 3.1.

Definition 3.1 *A geotagged photo data is represented as a record of (photo_id, owner_id, latitude, longitude, time), where owner_id is the id number of the photo owner, latitude and longitude shows spatial geographic coordinates position and time shows the timestamp of when the photo was taken.*

A photo data points out the location the photo-taker visited. Consequently, a series of photos, connected in time order, show dynamic changes of location that reflect a person's spatio-temporal trajectory. A trajectory is a sequence of geographic points as defined in Definition 3.2.

Definition 3.2 *A trajectory, raw geographic trajectory, is a sequence of geographic points with time information* $T = \langle (lat_1, lon_1, t_1), (lat_2, lon_2, t_2), \cdots, (lat_n, lon_n, t_n) \rangle$*, where* (lat_n, lon_n) *is the geographical coordinates of a photo* n *that shows the visited geographic position, and* t_n *is the corresponding time.*

A raw geographic trajectory shows the distribution of geographic sampling points. In the specific geographic space or applications, a trajectory may pass through a set of specific targeted spatial areas, destinations or geo-objects. These episodes are considered as the stops in the trajectory. Thus, the raw trajectory can be transformed into a sequence of stops that provide a meaningful expression of the movement in specific scenes and applications. In this study, a stop is defined by a RoI, which is a spatial area that a number of trajectories visit. An RoI consists of several square grid cells, which compose a RoI with an approximate and arbitrary shape. So a structured trajectory is a sequence of RoIs that people have visited.

Definition 3.3 *A RoI is a spatial area, defined as (R_id, $CELLS$) where R_id is the id number of RoI, and $CELLS$ is a set of square cells c and each cell c has a record of geographic coordinates that points out the specific spatial position.*

Definition 3.4 *A structured-trajectory is a sequence of RoIs $Structured-T = \langle$ (R_id, t_1), (R_id, t_2), $\cdots$, (R_id, t_m)$\rangle$, where R_id is the id number of a visited RoI and t_k is the visit time of the k_{th} visited RoI.*

A RoI has specific spatial location data, but can also be enriched with application contextual information, such as data related to the geo-object in the spatial region. A geo-object is located in the spatial RoI, such as a PoI, a Chinese restaurant or a gift shop. The type of place presents a basic category information of the geo-object, like restaurant or shop. The type of place provides the contextual semantics information in which a visit takes place. This study defines a semantics-enhanced RoI, named semantic RoI, as a spatial RoI with basic type-of-place annotation.

Definition 3.5 *A semantic RoI (SemRoI) is a spatial RoI with basic type-of-place annotation of geo-objects $SemRoI = (R\_id, CELLS, Type)$, $Type$ is the type of place reference to the geo-object in the spatial area.*

A trajectory can be enriched with application contextual data and annotations in which such a movement takes place. The semantics-enhanced trajectory provides more meaningful application-dependent semantics information. In the case of adding basic type-of-place annotation, a semantic trajectory, in this study, is a structured trajectory with basic type-of-place annotations where spatial data (RoI) is replaced by basic geographic object information annotations.

Definition 3.6 *Basic semantic trajectory is a sequence of place type annotations. $basicSemT = \langle(type_1, t_1), (type_2, t_2), \cdots, (type_n, t_n)\rangle$, where each $type_k$ is a place type annotation indicating the category of RoI_k in the trajectory.*

A trajectory can be further enriched with multiple semantic annotations. Specifically, besides the place type, a visited RoI can also be tagged with additional contextual environmental information in which a visit takes place. For instance, we may enrich RoI with annotations of city name where the RoI is located, days of the week, and the time period of a day when the visit occurred. These multiple annotations provide richer semantic descriptions and meanings about the movement trajectory.

Definition 3.7 *Multi-dimensional semantic trajectory (SemT): a sequence of place type annotations with a set of additional semantics* $SemT = \langle (SemA_0, t_0), \cdots, (SemA_n, t_n) \rangle$, *where* $SemA$ *is semantic annotations of RoI* $SemA = (e, V)$ *where* e *is the basic semantics,* V *is a set of additional semantic annotations, and* t *is the time stamp.*

This study utilises the multi-dimensional semantic trajectory. We enrich trajectories with additional city name, day of week, time of day and weather condition annotations. For brevity, we refer to the final multi-dimensional semantic trajectory as the "semantic trajectory" in the rest of the thesis.

Chapter 4

Semantic sequential pattern mining

This chapter presents the study of semantic sequential pattern mining from geotagged photos. Chapter 4.1 introduces the semantic sequential pattern mining. In Chapter 4.2, we summarise literature related to our study. We present previous studies on mining people's sequential patterns from photo data, and existing work about modelling and generating semantic trajectories. Chapter 4.3 describes the problem statement of our semantic sequential pattern mining. In Chapter 4.4, we provide details of the proposed framework and method for mining people's semantic sequential patterns. We develop a RoI mining method and a sequential pattern mining algorithm to mine sequential patterns from multi-dimensional semantic trajectories. Chapter 4.5 shows the experimental evaluations of our proposed method. The conclusion of this chapter is in Chapter 4.6.

4.1 Introduction

Sequential movement patterns represent the movement sequence of visited stops (Cao, Mamoulis, and Cheung 2005) that exist in a significant number of people's trajectories. A trajectory presents a time-ordered visit sequence of locations. It delivers sequential features of a list of visited locations that provide the temporal order relations between objects. In people's trajectories, a frequently occurring sequence of locations provides one frequent trajectory pattern comprised of a set of locations visited together and the time at which they were visited, thus showing the order of relations between locations. Massive repositories of online geotagged photos provide opportunities to mine knowledge about people's trajectory behaviours. Sequential behaviour of mobility is one type of information that can be obtained from trajectories. Several studies have investigated the extraction of frequent sequential mobility patterns (Kisilevich, Keim, and Rokach 2010; Bermingham and Lee 2014) from geotagged photos as discussed in Chapter 2.3.1. People's trajectories are built and then their sequential trajectory patterns are mined, including frequent sequential PoI sequences (Kisilevich, Keim, and Rokach 2010) and frequent spatio-temporal RoI sequences (Bermingham and Lee 2014).

However, previous studies of geometric feature only trajectory analysis are insufficient. Most applications require more semantics information on people's mobility patterns, and past studies lack semantics information. Adding the type-of-place data, for instance, at the type-of-place semantics level, reveals a sequential pattern of visiting the beach first and then going to restaurants. This provides useful knowledge of people's frequent mobility among different types of place. This type-of-place semantic-level sequential pattern is more useful to applications that care about human semantic-level movement behaviour than the pattern with only geometric information. Past studies of mining sequential patterns from geotagged photos lack consideration of aspatial semantics of trajectory data. Annotating semantics to trajectories can provide better understanding and more useful information to the results and enhance some novel unknown knowledge of people's semantic-level trajectory behaviours (Parent et al. 2013). This study will analyse aspatial semantically enhanced trajectories to find people's frequent sequential mobility pattern at a semantic level and with meaningful information.

In this chapter, we propose a framework to find semantic sequential patterns from geotagged photos. We use spatial, temporal and aspatial semantics features of trajectories and mine semantic patterns from the aspatial semantically enhanced trajectories. To generate each semantic trajectory, we propose a semantic RoI mining method to detect semantic RoIs with basic place type annotation from raw trajectories in which each semantic RoI is considered as a stop in the semantic trajectories. We build people's semantic trajectories as a sequence of semantic RoIs with sets of additional environmental semantics. We develop a sequential pattern mining method to deal with multi-dimensional semantic trajectories to find frequent semantic sequential patterns. The semantic sequential pattern result of the proposed method provides more information for understanding human behaviours that are valuable to various domains, for instance, tourism industry understands tourists moving among place types at different weather conditions and day time periods. Interesting RoIs with fine, precise, place type semantics are found, and sets of useful semantic trajectory patterns are extracted.

4.2 Related work

4.2.1 Sequential pattern mining on trajectory data for geotagged photos

The standard sequential pattern mining problem is to find all the frequent sub-sequences whose occurrence frequency in the sequence database is no less than the given minimum support (Agrawal and Srikant 1995). Cao, Mamoulis, and Cheung (2005) mined sequential patterns from spatio-temporal trajectories, that is, to extract frequent sequences of stops from a set of trajectory data. A stop is a spatial region, area or a specific geo-object in the area.

Several pioneering studies have investigated mining people's sequential mobility patterns from massive online collections of photo data that already include geographic information (Bermingham and Lee 2014; Kisilevich, Keim, and Rokach 2010). Knowledge of human frequent sequential patterns is beneficial to various applications especially tourism (Bermingham and Lee 2014). Kisilevich, Keim, and Rokach (2010) extracted tourists' frequent visit sequence of PoIs. Bermingham and Lee (2014) added a temporal dimension into trajectories to extract spatio-temporal trajectory patterns for tourism science. All previous studies focus on the analysis of geometric-feature-only trajectories for spatial-level patterns without considering contextual aspatial semantics information in the trajectory analysis and patterns. However, most application analyses require raw data complemented with additional information from the application context (Parent et al. 2013). This study aims to find out people's semantic-level sequential patterns by investigating aspatial semantics-enhanced trajectories.

4.2.2 Analysis of semantic trajectory for geotagged photos

Alvares et al. (2007a) added semantic geographical information to trajectories for analysis of mobility pattern with meaningful semantics knowledge. The semantic trajectory is modelled as a sequence of semantic annotated stops. Their method for building semantic trajectory was to find spatial stops first and then annotate semantics to the stops. It is a kind of post-process method that divides detection of semantic stops into two separate steps. The main drawback of the post-process method is that the results contain some false areas thus these areas are actually different place types. This is mainly due to considering only spatial features without semantics.

To detect spatial stops from trajectories, Giannotti et al. (2007) proposed a grid-based RoI mining algorithm. The main strength of grid-based RoI mining algorithm is its time efficiency. It requires $O(m)$ time where m is the number of grid cells. But the RoIs found are always rectangular shapes that could contain false positives. To improve the method, Hio et al. (2013) proposed a more effective hybrid grid-based RoI mining algorithm that is able to detect arbitrary shapes of RoIs with almost the same time efficiency. This study defines a semantic trajectory as a series of semantic RoIs. We adopt the idea of hybrid grid-based RoI mining algorithm (Hio et al. 2013) to find spatial RoIs while we apply a new strategy to generate semantic RoIs. This new strategy is different from the post-process method proposed in (Alvares et al. 2007a). The proposed algorithm can find semantic RoIs with fine accurate semantics.

Recently, a few novel studies attempted to investigate mining sequential patterns from semantic trajectories (Alvares et al. 2007b; Zhang et al. 2014). Alvares et al. (2007b) discovered place-level stop-move sequence patterns. More spatial compactness and temporal continuity features are considered and the fine-grained sequential patterns in semantic trajectories are extracted in (Zhang et al. 2014). Semantic trajectories in these studies only consider a single dimension (the spatial dimension) and the methods could not deal with the multi-dimensional semantic trajectories, in which each RoI has multiple semantics.

4.2.3 Multi-dimensional sequential pattern mining for geotagged photos

Multi-dimensional analysis can extract patterns with more information that are more useful. Pinto et al. (2001) tried a frontier investigation on mining sequential patterns in multi-dimensional circumstances. They use a single dimension for item of sequence whilst our study uses multiple dimensions for item of trajectories. Another kind of study on multi-dimensional sequential pattern mining was proposed by (Plantevit et al. 2005). This study introduces jokerised patterns that are not fully instantiated multi dimensional patterns.

4.3 Problem statement

This section presents the terminologies and problem statement of sequential pattern mining. As explained in Definition 3.2 in Chapter 3.4, a trajectory is represented as a sequence of time-ordered geographic points. It shows people's movements in a geographic area.

The raw geographic trajectories are then enriched with application-dependent contextual data to build meaningful, and semantically enhanced trajectories. A semantic trajectory is a sequence of stops with contextual environmental annotations. The semantic trajectory indicates people's movement at the semantic-level. A detailed definition of semantic trajectory is presented in Definition 3.7 in Chapter 3.4.

From the semantic trajectories, we aim to find out the sequence of semantic stops that frequently occurred in a density of individual trajectories. This frequent visit sequence is named semantic sequential pattern. A *semantic sequential pattern* contains a sequence of semantic elements.

Definition 4.1 *A semantic sequential pattern (SemSPattern) is a sequence of place type annotations with set of additional semantics* $SemSPattern = \langle (E_0), \cdots, (E_n) \rangle$*, where* E *is a multi-dimensional element* $E = (e, V)$ *where* e *is basic semantics, and* V *is a set of additional semantic annotations.*

When an element contains the basic place type annotations only, SemSPattern will be called *basic SemSPattern*; when the element is associated with multiple semantics, SemSPattern will be called *multi-dimensional SemSPattern.*

Figure 4.1 shows an example of a semantic sequential pattern. It shows mobility at the basic type-of-place semantics layer. The movement pattern starts at a hotel, then goes to a shop, visits a restaurant in the next step and moves to a park at the end. Figure 4.2 shows an example of a multi-dimensional semantic sequential pattern. It integrates a basic type-of-place semantic pattern with additional information of weather and temporal day type and day time.

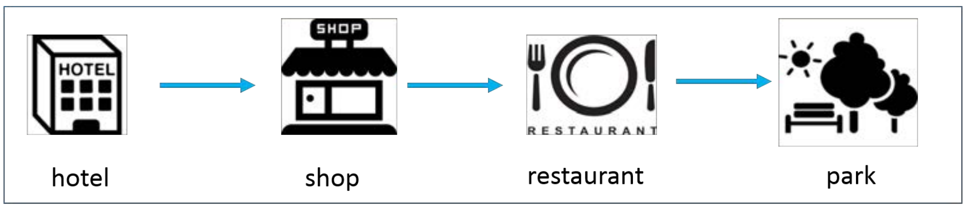


Figure 4.1: Example of basic SemSPattern.

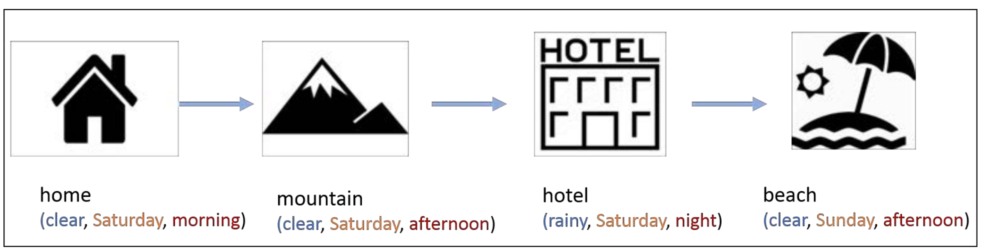


Figure 4.2: Example of multi-dimensional SemSPattern.

The determination of a semantic sequential pattern is based on the frequency and *support* of such patterns, that is, the number of trajectories that contain the pattern.

In standard sequence mining, a key step is the equality test for matching two elements of two compared sequences. In our study, it requires matching the basic type-of-place annotation label and tests for containment of multiple additional semantics annotations. Specifically, each element of semantic trajectory is multi-dimensional consisting of a place type annotation and a set of additional annotations. Thus, for two compared elements, the matching test is not to check the matching. In summary, for semantic trajectories, the *containment* requires that RoIs are matching in dimensions of semantics that the set of additional semantics of a RoI is full or partial match the set of semantics of the other RoI, defined as follows:

Definition 4.2 *Dimensional containment ($\preceq_d$): Given two semantic trajectories $SemT_1 = \langle (e'_1, V'_1),\ \cdots,\ (e'_m, V'_m) \rangle$, and $SemT_2 = \langle (e_1, V_1),\ \cdots,\ (e_n, V_n) \rangle$, $m \leq n$ we say that $SemT_1$ is contained in $SemT_2$, denoted as $SemT_1 \preceq_d SemT_2$, if $\forall_{0 \leq k \leq m}, e'_k = e_i$, and $V'_k \subseteq V_i$.*

Definition 4.3 *Support of a semantic sequential pattern SemSPattern as*

$$supp(SemSPattern) = \frac{|SemT^* \in D | SemSPattern \preceq_d SemT^*|}{|D|}.$$

Definition 4.4 *Semantic sequential pattern mining: Given a database of input semantic trajectories D, and a minimum support threshold minSup, the semantic sequential pattern mining problem is to find all frequent semantic sequential patterns whose support is no less than minSup.*

In this study, the problem is to find out all of frequent semantic sequential patterns from semantic trajectories which are built from a given database of geotagged photos.

4.4 Semantic SPM framework and methods

Our proposed framework for extracting human semantic sequential patterns from georeferenced photos, shown in Figure 4.3, consists of three main steps. First, raw geotagged photos are pre-processed, including data cleaning, and reconstruction of raw trajectories. The next step is to generate semantic trajectories from raw trajectories.

Step 2 – the semantic trajectories generating stage – includes two substeps: first semantic RoI mining, and then trajectory transformation and additional semantics enrichment. Spatial RoIs with place type semantics

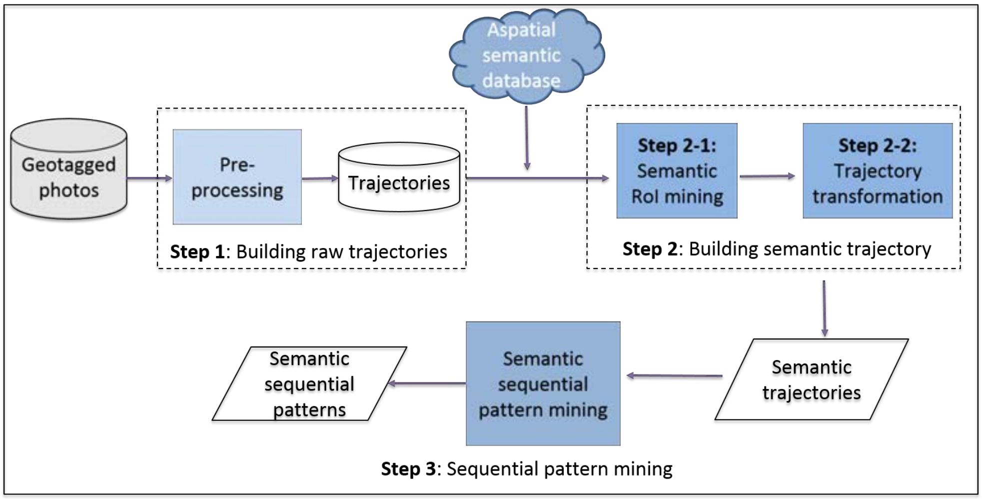


Figure 4.3: Framework for semantic sequential pattern mining.

are extracted from the raw trajectories during the semantic RoI mining sub-step. Based on these extracted semantic RoIs, those raw trajectories are then transformed into RoI-based structured trajectories. A structured trajectory is a sequence of RoIs. During this process, place type semantic RoIs are first used as basic contextual semantics, and then further environmental semantics, including temporal and weather condition information, are added to generate multi-dimensional semantic trajectories. To finish, the semantic sequential pattern mining algorithm is applied to the semantic trajectories to find frequent semantic sequential patterns.

4.4.1 Building trajectories

In the first step, we reconstruct people's raw trajectories from geotagged photos. As presented in Chapter 3.2, photo data are cleaned and redundancies are removed in the pre-processing process. A record of photo data includes these fields: owner id, latitude, longitude and time. We group photos by photo owner id as each photo-taker has several photos. For each photo-taker, geotagged photos are ordered by timestamp and then connected to form a trajectory. Each element in the trajectory is a photo data record including latitude and longitude, which refer to the spatial location, and timestamp of the photo, which refers to time information. Trajectories with more than two points are considered useful in this research.

4.4.2 Semantic RoI mining method

We then generate people's semantic trajectories. This process includes mining semantic RoIs from trajectories, and additional semantics enrichment using external aspatial semantics databases. We describe semantic RoI mining methods in this part.

The core concept behind the semantic RoI mining is that areas with place type information that contain a high density of moving entities are both interesting and significant. We apply the hybrid grid-based RoI mining algorithm to calculate dense spatial cells and generates RoI (Cai et al. 2014). To obtain semantic RoIs, we enrich these spatial areas with place type contextual semantics. Two methods are proposed to find the semantic RoIs: a post-process method (Figure 4.4) and an inter-process method (Figure 4.5). The post-process method first finds spatial RoIs and then annotates them with place type. The inter-process method works by generating semantic RoIs directly from the semantic dense cell. This post-process method generates loosely coupled semantic RoIs, and also could produce false positive RoIs whilst the inter-process method generates tightly coupled semantic RoIs.

Post-process method

The post-process method contains two main steps. At the first step, the hybrid grid-based RoI mining algorithm is applied to raw trajectory datasets to find spatial RoIs. At the next step, spatial RoIs are assigned with place type semantics by using the place type of its nearest PoI searched from a geographic information database. A detected spatial RoI has a record of all the points within it. At first, for each point, we annotate it with place name and place type. To do the annotating work, we search for the nearest place to the point from a background geographic information database. The PoI geographic information database is used in this research. And, the reverse geocoding technique is used to search for the nearest place that corresponds to the place information for a given latitude and longitude. After all points inside the spatial RoI have been annotated with place type, the spatial RoI has a list of annotations of place types. Next, we calculate the most frequent annotation, of place type for the spatial RoI by applying Term Frequency-Inverse Document Frequency (TF-IDF) method to the list of annotations. The most frequent place type is considered the place type of the RoI. RoIs with place type annotation are semantic RoIs.

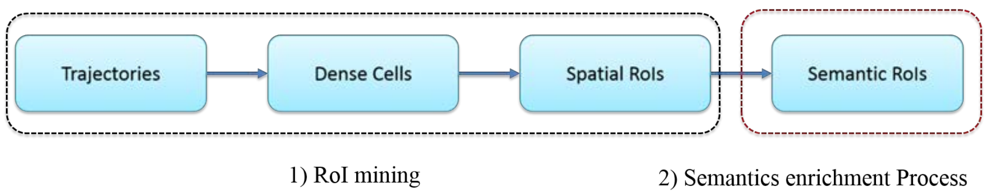


Figure 4.4: Post-process semantic RoI mining method.

The post-process semantic RoI mining algorithm extracts semantic RoIs from raw trajectories. However, the post-processed semantic RoIs may contain some false areas that are different place types. In particular, considering only spatial features, some neighbouring dense spatial cells are merged into a spatial RoI and assigned the same place type at the post-processing step. This becomes invalid when those dense cells are actually different place types. Moreover, these different place type cells need to be considered as different semantic RoIs, but the post-process method can only find a single big RoI from neighbouring dense cells, even if they are all different place types.

Inter-process method

The inter-process method fixes the false area problem that post-process method has. Compared to the post-process method, this approach integrates the semantics discovery process into the intermediate part of the grid-based RoI mining algorithm. This approach is a derivative algorithm from the hybrid grid-based RoI mining algorithm. It includes three steps. First, dense cells are computed. Second, using the same technique that is used to find semantic RoIs in the post-process method, we calculate semantics for the dense cells. Third, we merge neighbouring dense semantic cells to construct semantic RoIs. From this process, different place type areas can be found as several independent semantic RoIs.

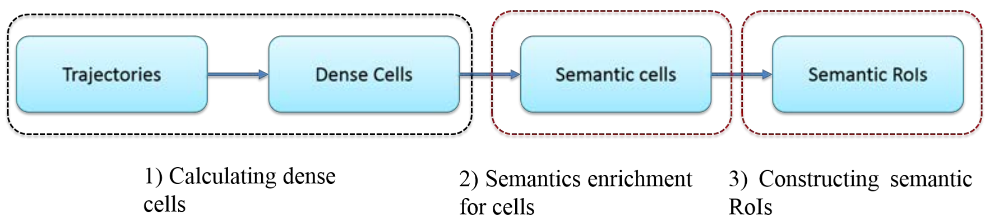


Figure 4.5: Inter-process semantic RoI mining method.

4.4.3 Additional semantics enrichment

After semantic RoI mining, raw trajectories are then transformed into multi-dimensional semantic trajectories. The transformation is completed in two steps which are generating single-dimensional semantic trajectories and annotating multiple additional semantics. Firstly, based on the semantic RoIs, the result of the inter-process method, a raw trajectory is transformed into a sequence of semantic RoIs called a single-dimensional semantic trajectory. These semantic RoIs contain only place type semantics. Then, in the second step, RoIs are enriched with temporal and weather condition, and become multi-dimensional RoIs. At last, semantic trajectories become multi-dimensional.

First, RoIs are enriched with two temporal features (day type and day time) based on their timestamp values. For day type, one RoI's visit time is converted into the day of week, and it is then grouped into weekday type or weekend type. For day time, based on the RoI's timestamp, the hour time is found, and it is can then be grouped into different temporal concepts as defined in Table 4.1

Table 4.1: Time of day concepts for this research.

Time period (12h)	**Context concept**
2:00am - 6:59am	Dawn
7:00am - 11:59am	Morning
12:00pm - 5:59pm	Afternoon
6:00pm - 9:59pm	Evening
10:00pm - 1:59am	Midnight

Then RoIs are enriched with weather condition features based on their geographical location and timestamp values. We use daily weather observation databases to query climate information. At first, based on the RoI's geographical location value, we find the nearest observation station. Using the observation station found and the RoI's timestamp value, we query the rainfall and temperature values from the daily weather observation database. Based on the rainfall and temperature, the weather condition is calculated and added to the RoIs.

At the end of this process, raw trajectories are transformed into semantic trajectories that include basic type-of-place semantics and additional temporal and weather condition information.

4.4.4 Semantic sequential pattern mining

This step finds semantic sequential patterns from the multi-dimensional semantic trajectories. After the semantic trajectories have been converted from raw trajectories, our sequential pattern mining method is then applied to the semantic trajectories. The main method used for the sequential pattern mining in this step is based on the PrefixSpan algorithm (Han et al. 2001), in conjunction with the BUC algorithm (Beyer and Ramakrishnan 1999).

The PrefixSpan algorithm is used to find frequent single-dimensional items and frequent single–dimensional item sequences. The BUC algorithm is used to find frequent multi-dimensional value combinations for each frequent single item. During the PrefixSpan algorithm process, when a frequent single-dimensional item is found, values of its multiple dimensions are collected as a value tuple. For all the sequences containing the item, all the item value tuples are obtained and used to create a dimension value database. Then, the BUC algorithm is applied to the dimension value database to find frequent value combinations. Each value combination and the item composes a frequent multi-dimensional item. Each frequent multi-dimensional item is a 1-length prefix in this current recursion. Based on the dimension combination, these new 1-length prefixes are added to the previous prefixes, which are received from the last recursion, to generate new multi-dimensional prefixes. These new multi-dimensional prefixes are then sent to the next recursion as parameters. As the PrefixSpan algorithm runs recursively, multi-dimensional sequences are extended continuously. At last, all the different length, frequent multi-dimensional sequences are found. Benefits of the BUC algorithm and semantic sequential pattern mining algorithm include their ability to handle sequences with multi-dimensional items and to find frequent multi-dimensional sequential patterns.

Example 1 *A level 1 projection p1. Basic semantics is type of place, and 3 additional semantics are day time, day type and weather condition. Existing* *prefix is: Restaurant; three multi-dimensional prefixes are* $Restaurant_{[afternoon]}$, $Restaurant_{[weekday]}$ *and* $Restaurant_{[morning][rainy]}$.

1. There is a newly found frequent basic item: Park. There are three sequences that contain item Park and the three involved item are
$I_1 : Park_{[morning][weekend][clear]}$, $I_2 : \ Park_{[morning][weekday][rainy]}$ *and*
$I_3 : \ Park_{[afternoon][weekend][clear]}$ *respectively.*

2. Start to find frequent multi-dimensional item for Park.

1) Collecting initial semantics values and creating value matrix.

a. morning, weekend, clear;

b. morning, weekday, rainy;

c. afternoon, weekend, clear.

2) Applying the BUC algorithm on value matrix. The results are four frequent combinations of values: $\{[morning]\}$,$\{[weekend]\}$, $\{[clear]\}$, *and* $\{[weekend][clear]\}$.

3) frequent multi-dimensional items are: $Park_{[morning]}$, $Park_{[weekend]}$, $Park_{[clear]}$, $Park_{[weekend][clear]}$.

3. Generating new extended level 2 projection p2 with basic prefix: $Restaurant \ \rightarrow \ Park$. *There are two multi-dimensional prefixes:*

a)(dimension: day time): $Restaurant_{[afternoon]} \ \rightarrow \ Park_{[morning]}$;

b)(dimension: day type): $Restaurant_{[weekday]} \ \rightarrow \ Park_{[weekend]}$.

Example 1 shows the process of generating projection steps. For a 1st level projection database, a basic prefix is Restaurant and the other three multi-dimensional prefixes with frequent dimensions day time, day type and weather condition, respectively. In the first step, we found a frequent item Park. In the second step, we are going to find multi-dimensional items. By applying the BUC algorithm to the initial value set of three additional semantics, we obtain four frequent values of dimension combinations: value morning for day time dimension; value weekend for day type dimension; value clear for weather condition dimension; and value weekend and clear for combination of day type and weather condition dimensions. In the third step, based on the consistency of dimension combinations, we generate two 2-length multi-dimensional prefixes.

A Pseudocode of the semantic sequential pattern mining algorithm is shown in Algorithm 4.1. First step is to find all the frequent single-dimensional items (Line 2). Line 3 iterates through frequent single-dimensional items. Each single-dimensional item holds a record of belonging sequences that contain the item. Line 5 scans through these belonging sequences, and finds the first such item (Line 8). Values of all dimensions of such item are obtained and made into values tuples (Line 9). After these sequences have been scanned, all the dimension values become the database for the BUC algorithm. The BUC algorithm is then applied to the dimension values database (Line 11) to generate the frequent dimension values combinations. Single-dimensional items with frequent dimension values combinations become a frequent 1-length multi-dimensional sequence (Line 14). These 1-length sequences are added to previous sequences to generate new and longer frequent sequences. The combination of dimensions is used as the determination condition for the sequences prolonging process (Lines 16–21). If the number of the same sequences meets the minimum support threshold, this new 1-length sequence is added to the previous sequence to make a new and longer sequence (Line 21) and saved (Lines 22–23). Finally, for this single-dimension item, a projected database is built from the initial sequence database (Line 24). The next recursion of the semantic sequential pattern mining algorithm starts with the projected database as its initial sequence database and all the new longer sequences as previous sequences (Line 25).

Algorithm 4.1 Semantic_Sequential_Pattern_Mining_algorithm

Input: A set of semantic trajectories D_t, a minimum sup $minSup$;

Output: Semantic sequential patterns;

```
P_md ← ∅, P_sd ← ∅;
Find out all unique items Items;
for all item ∈ Items do
    if item.support ≥ minSup then
        P_sd ← P_sd ∪ item, D_buc ← ∅;
        for all sequence ∈ D_t do
            if sequence contains item then
                fitem = sequence.getFirstItem(item.basicSemantics);
                Tuple ← fitem.additionalSemantics;
                D_buc ← D_buc ∪ Tuple;
        freDims = BUC-Algorithm(D_buc);
        newP_md ← ∅;
        for all cdimvalue ∈ freDims do
            curP_md ← item + cdimvalue.values;
            dims ← cdimvalue.dimensions;
            if P_md contains dims then
                preP_md ← P_md.get(dims);
                for all pattern p ∈ preP_md do
                    sameSeqs ← P.sequences ∧ curP_md.sequences;
                    if sameSeqs ≥ minSup then
                        P ← append(P, curP_md);
                        newP_md ← newP_md ∪ P;
        Save newP_md;
        D_proj = projectedDatabase(item,D_t);
        Semantic_Sequential_Pattern_Mining_algorithm(D_proj, newP_md,
        P_sd, minSup);
```

4.5 Experiments

We conducted experiments to validate the proposed method in terms of: (1) effectiveness of proposed semantic RoI mining method; and (2) effectiveness of semantic sequential pattern mining method to find interesting semantic-level sequential patterns.

4.5.1 Dataset

This study collected geotagged photos taken in Queensland, Australia, from 1 April 2014 to 30 March 2015. After pre-processing, we have 64,733 cleaned photo records, and 1,404 trajectories including, 61,322 points, as shown in Figure 4.6

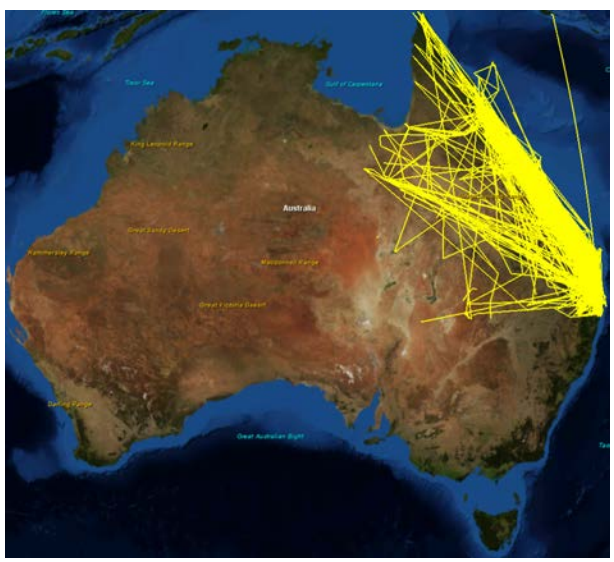


Figure 4.6: Raw trajectories in Queensland ($|n|$=1,404).

4.5.2 Parameters selection

The hybrid grid-based RoI mining algorithm and sequential pattern mining algorithm are sensitive to parameter selections. Both algorithms rely heavily on the minimum support (*minSup*) value for a cell to become a RoI and also on the size of cell (*CellSize*) that is used to partition the study region. It is a non-trivial problem to choose the best values of parameters to produce meaningful and insightful RoIs and patterns. Thus, the approach adopted during experimentation was a systematic trial and error approach, in which the *minSup*, *CellSize* and resultant number of RoI were recorded.

Table 4.2: Values of parameters tested for SPM.

Parameter	**Values**
MinSup	0.01, 0.008, 0.005
CellSize	0.008, 0.005, 0.003

To estimate best values for these two parameters, several experiments have been performed and parameter values chosen for this study are shown in Table 4.2. For parameter *CellSize*, value 0.005 means 0.5km whilst for parameter *minSup* 0.005 equals to 0.5%.

4.5.3 Semantic RoIs

The two semantic RoI mining algorithms work in different ways after the dense spatial cells have been calculated. Using 0.005 for parameter *minSup* and 0.005 for parameter *CellSize*, a total 173 cells are dense.

Post-process semantic RoI mining method

For the post-process semantic RoI mining algorithm, spatial RoIs were generated. The number of spatial RoIs we found is 89. Then these spatial RoIs were annotated with place type semantics. Finally, 89 semantic RoIs were constructed. Among the 89 semantic RoIs, there were 30 different place types. Table 4.3 shows the five most frequent place types. There are detailed descriptions for codes HTL, PPL and PIER in Table 4.4. Feature code BDG is for bridge, and RCH is for reach: that is, a straight section of a navigable stream or channel between two bends.

Table 4.3: Five most frequent type of place semantic RoIs from the post-process method.

Type of place	**Support**
HTL	1,484
PPL	213
BDG	200
PIER	181
RCH	137

Figure 4.7 displays a set of spatial cells in Cairns city area. Minimum support is 8, giving a total of 1,404 trajectories. Using the post-process RoI mining algorithm, spatial RoIs are then found as shown in Figure 4.8. Then these RoIs are enriched with place type, as shown in Figure 4.9, based on the GeoNames feature codes description. Place types shown in Figure 4.9 are explained in Table 4.4.

Inter-process semantic RoI mining method

After calculation of spatially dense cells, the inter-process semantic RoI mining algorithm first enriches place type annotations to generate semantic cells, and then merges neighbouring dense semantic cells to generate semantic RoIs. Applying 0.005 for parameter minSup and 0.005 for parameter CellSize, 173 spatially dense cells are calculated. Then semantic cells are produced. Finally, we found 136 semantic RoIs containing 36 different place types. Table

Figure 4.7: Spatial cells in Cairns area (*minSup* = 0.005, *CellSize* = 0.005).

4.5 displays the five most frequent place types. Codes HTL, PPL and RSTN are described in Table 4.4. BDG is for bridge and BCH means beach.

Figure 4.7 shows spatially dense cells in Cairns city area, and the semantic cells are displayed in Figure 4.10. Each semantic cell is shown with support number and place type annotation. Figure 4.11 presents the semantic RoIs generated from these semantic cells by merging neighbouring cells containing same place type. Table 4.4 displays the description of feature codes shown in Figure 4.11.

Discussions

The post-process method and the inter-process method extract semantic RoIs in different ways. Both methods can find semantic RoIs from trajectories by cooperating with a contextual geographic information database. However, experimental results show that the inter-process method produces better results in terms of quality, and also quantity than the post-process method. In

Figure 4.8: Post-process method: spatial RoIs in Cairns city area (*minSup* = 0.005, *CellSize* = 0.005).

detail, for quality, the inter-process technique produces more accurate semantic RoIs. For example, in Figure 4.11, bottom semantic RoIs have different place types generated from the inter-process technique that are well separated and grouped, but in Figure 4.9 semantic RoIs generated from the post-process technique are grouped as one bigger semantic RoI, which is incorrect. The reason for different quality is explained in Chapter 2.3: the post-process method generates RoIs containing false areas, but the inter-process method can fix this problem. As a result of the increased accuracy, for quantity, the inter-process method finds more semantic RoIs. The inter-process method found 136 semantic RoIs, while the post-process method found 89 semantic RoIs.

Using the same parameter values, 0.005 for parameter *minSup* and 0.005 for parameter *CellSize*, the two methods cost similar running time. The post-process method costs 742,092ms to find semantic RoIs, and the inter-process method costs 738,398ms. Both running times include the query time from

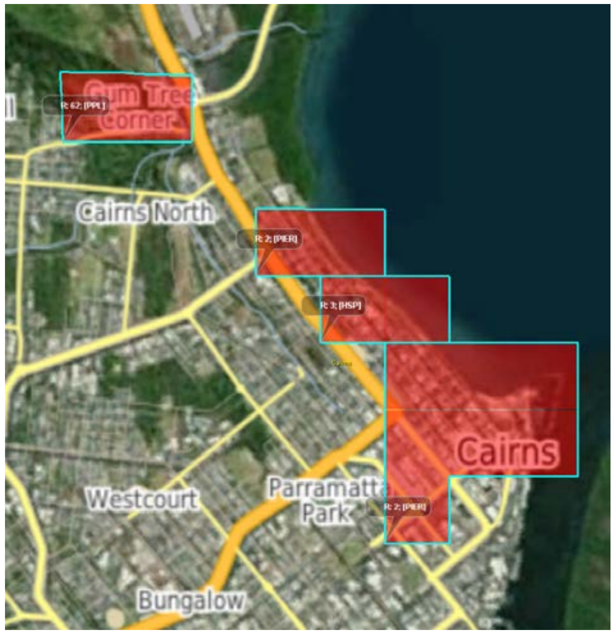


Figure 4.9: Post-process method: semantic RoIs in Cairns city area (*minSup* = 0.005, *CellSize* = 0.005).

the geographical database.

4.5.4 Semantic sequential patterns

We use semantic RoIs results of the inter-process method to transform raw trajectories into semantic trajectories. There are 1,404 raw trajectories, and 136 semantic RoIs. At last, we have 916 multi-dimensional semantic trajectories. Example 2 shows a 2-length semantic trajectory, beginning with the trajectory id (567) and next the description of length (2), meaning 2 RoIs in the trajectory. This trajectory shows that the photo-taker went to a beach on a rainy weekday evening and then went to an island on a weekday clear morning.

Table 4.4: Description of feature codes for SPM study.

Feature code	**Description**
HTL	Hotel
PPL	Populated place
BDG	Bridge
PIER	Pier (a structure built out into navigable water on piles providing berthing for ships and recreation)
RCH	Reach (a straight section of a navigable stream or channel between two bends)
PT	Point (a tapering piece of land projecting into a body of water, less prominent than a cape)
RSTN	Railroad station

Example 2 *A 2-length semantic trajectory*

$$567 : 2 < (BCH_{[weekday][Lightrain][evening]})$$
$$(ISL_{[weekday][Clear][morning]}) > .$$

Using parameter values defined in Chapter 4.5.2, we finally obtained 52,163 frequent semantic patterns including 1-length patterns. These results contain sequences with only place type semantics and sequences with any combination of other temporal and weather semantic dimensions. The longest pattern is 16 in length. The two longest patterns have only the place type semantics, and both end with a 14-length sequence moving among hotels only. Some other long patterns also have sub-sequences with only hotel place types. These could be caused by those raw trajectories that each trajectory includes some hotel stops, and because of using only place type semantics. For 2-length patterns, the most frequent pattern is from hotel to hotel with support 229. With only basic place-type semantics, 229 trajectories contain sub-sequences moving from hotel to hotel. There are also patterns from

Table 4.5: Five most frequent type of place semantic RoIs from the inter-process method.

Type of place	**Support**
HTL	1,026
PPL	308
BDG	240
BCH	212
RSTN	208

hotel to hotel with many frequent value combinations. Table 4.6 shows a pattern, hotel to hotel with dimension combinations of weather condition and day time. As each dimension has several values, there are 7 patterns with different value combinations.

With only place type semantics, the second most frequent 2-length pattern is from hotel to bridge. This pattern also has many further multi-dimensional patterns. Table 4.7 shows the number of patterns of every frequent dimension combination for patterns from hotel to bridge. There are 4 frequent dimension value combinations for patterns with the day type dimension and 3 patterns for both day time and weather condition dimensions. For the combination of all 3 dimensions, 2 patterns are listed in Table 4.8.

Similarly, various numbers of frequent patterns various dimension combinations can also be found for other length patterns.

Our method can find interesting semantic-level sequential patterns. Also, our patterns provide rich and meaningful semantics information. Experimental results demonstrate that our proposed method is able to find semantic sequential trajectory patterns from geotagged photos. From the results, we find that trajectory analysis integrated with semantic information provides understanding of people's semantic behaviours. For example, a frequent place type movement is hotel to bridge, as listed in Table 4.7 that is followed by

Table 4.6: 2-length SemSPatterns from hotel to hotel (dimensions: weather condition and day time).

Pattern	Support
$\langle(HTL_{[Clear][midnight]}) \rightarrow (HTL_{[Clear][midnight]})\rangle$	56
$\langle(HTL_{[Clear][dawn]}) \rightarrow (HTL_{[Clear][dawn]})\rangle$	44
$\langle(HTL_{[Clear][morning]}) \rightarrow (HTL_{[Clear][morning]})\rangle$	23
$\langle(HTL_{[Clear][afternoon]}) \rightarrow (HTL_{[Clear][afternoon]})\rangle$	18
$\langle(HTL_{[Clear][evening]}) \rightarrow (HTL_{[Clear][evening]})\rangle$	28
$\langle(HTL_{[Lightrain][midnight]}) \rightarrow (HTL_{[Lightrain][midnight]})\rangle$	19
$\langle(HTL_{[Lightrain][dawn]}) \rightarrow (HTL_{[Lightrain][dawn]})\rangle$	14

Table 4.7: Number of SemSPatterns for each dimension combination (pattern: hotel to bridge).

Dimension combination	Number of patterns
Day_Type	4
Weather	3
Day_Time	3
Day_Type + Weather	2
Day_Type + Day_Time	2
Weather + Day_Time	3
Day_Type + Day_Time + Weather	2

Table 4.8: SemSPatterns for combination of 3 dimensions Day_Type, Weather and Day_Time (pattern: hotel to bridge).

Pattern	Support
$\langle(HTL_{[weekday][Clear][midnight]}) \rightarrow (BDG_{[weekday][Clear][midnight]})\rangle$	13
$\langle(HTL_{[weekday][Clear][dawn]}) \rightarrow (BDG_{[weekday][Clear][dawn]})\rangle$	8

Figure 4.10: Inter-process method: semantic cells in Cairns city area (*minSup* = 0.005, *Cellsize* = 0.005).

a number of photo-takers. Such patterns of tourist behaviour are valuable to various domains, including tourism. Compared to place-level, which is presented in (Cai et al. 2014), these semantic patterns can provide more meaningful behaviours and understanding of the movement.

Another important finding from the outcomes of experiments is that by adding more semantics, more novel patterns can be found. By adding place type, weather condition and temporal semantics, our method finds frequent movement of place types with more information. Moreover, all the experimental results show that our method can process multi-dimensional semantic trajectories that find patterns with various combinations. As listed in Table 4.8, several patterns of hotel to bridge with various dimension combinations are found. If all dimensions appear together, when the support threshold is high, we could not obtain patterns. But with a flexible combination strategy, we can derive useful patterns with subsets of dimensions.

In summary, experimental results demonstrate that semantic trajectory

Figure 4.11: Inter-process method: semantic RoIs in Cairns city area (*minSup* = 0.005, *Cellsize* = 0.005).

patterns can be mined from geotagged photo data by applying our method. Moreover, adding multiple semantics provides rich information with good understanding of trajectory behaviours. Results also show that our method is able to deal with multi-dimensional trajectories and can find patterns with a flexible combination of dimensions.

The extracted semantic sequential pattern results can be used in tourism. For example, a pattern of "from hotel to Harbour bridge" indicates one frequent travel pattern from previous tourists visiting Harbour bridge in Sydney from their hotels. In fact, Harbour Bridge is one of the most popular destinations that attracts numerous people to visit and cross it. It is located in the heart of Sydney with Opera house, and there are many hotels around. This travel pattern can be used as a suggestion for potential tourists who plan to travel to Sydney and book nearby hotels.

4.6 Conclusion

This chapter presents the study of converting geotagged photos into semantic trajectories for providing information about people's movement especially their sequential trajectory patterns on semantic level. We propose a multi-dimensional sequential pattern mining method to extract frequent semantic patterns, and introduce an inter-process semantic RoI mining method to find semantic RoIs. Experimental results show that the inter-process method is able to find RoIs with finer accuracy on place type semantics, when compared with a post-process method. We found numerous people's sequential trajectory patterns on a semantic level. Such semantic patterns provide more meaningful knowledge and understanding of human mobility behaviours that are valuable to tourism industry (Alvares et al. 2007b) than those have been previously available. By adding multiple semantics to trajectories, including weather condition and temporal information, we found more novel knowledge about trajectory patterns with several contextual semantics. Our method also generated patterns with various combinations of dimensions from multi-dimensional trajectories, demonstrating the ability of the proposed method to deal with multi-dimensional semantic trajectories.

Chapter 5

Semantic common pattern mining

This chapter describes the study of semantic common pattern mining from geotagged photos. The introduction to this study is presented in Chapter 5.1. Chapter 5.2 surveys previous studies related to common trajectory pattern mining and briefly reviews existing techniques for trajectory clustering and methods for measuring similarity of trajectories. Then in Chapter 5.3 we describe the research problem. Chapter 5.4 illustrates the overall framework and presents the proposed semantic trajectory clustering method. The details of experimental design are explained in Chapter 5.5, and the experimental results are presented and discussed in Chapter 5.6. Finally, the conclusions of this chapter are set out in Chapter 5.7.

5.1 Introduction

Common trajectory pattern is the approximate track shown in many objects' historical movements. It shows collective common mobility behaviours (Lee, Han, and Whang 2007). With the advances of Web 2.0 and geo-tagging technologies, georeferenced photos containing location information, have become a potential repository for discovering people's common trajectories (Zheng, Zha, and Chua 2012). Tourists' common popular travel movements are useful to tourism domains and urban management. Zheng, Zha, and Chua (2012) extracted people's spatial or spatio-temporal common trajectory patterns.

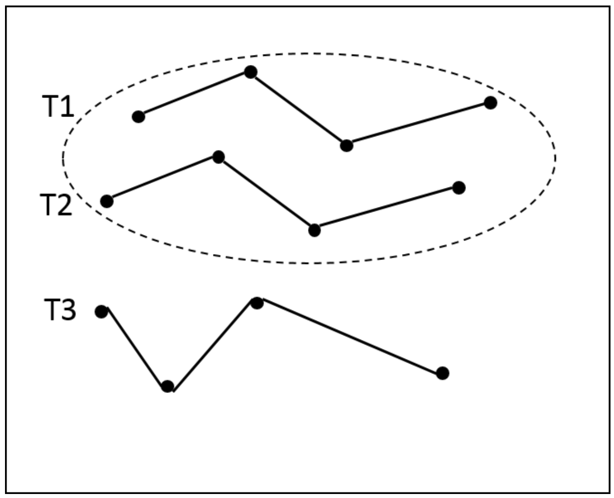


Figure 5.1: Geometric-feature-based trajectory clustering.

However, past studies use only spatial geographic features (location information) in analysis of mobility trajectories, while ignoring aspatial semantics features (Alvares et al. 2007b). But some specific applications require more meaningful aspatial semantic information, which these trajectories cannot provide. This chapter proposes a common trajectory pattern mining method that is based on both geographic and aspatial semantic features. Figure 5.1 and Figure 5.2 present an example of clustering based on different features of trajectories to highlight the limitation of the geographic-feature-only approach. Specifically, using the geographic-feature-only approach shown in Figure 5.1, trajectories $T1$ and $T2$, having similar geographic information, are grouped into the same cluster, whilst these two trajectories are not considered to be similar when the type-of-place semantic feature is considered, as shown in Figure 5.2. Considering the type-of-place semantic feature, trajectories $T2$ and $T3$ are grouped into the same cluster since they exhibit the same type-of-place visitation sequence. Semantically enhanced trajectories reveal more detailed and meaningful patterns for specific applications.

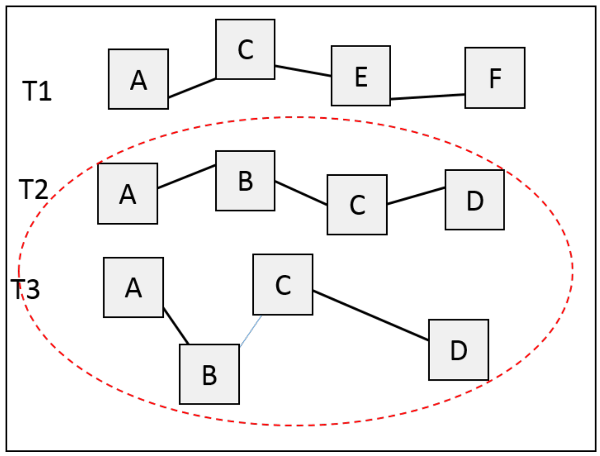


Figure 5.2: Semantic-feature-based trajectory clustering.

This study aims to extract semantic common trajectory behaviours from geotagged social media data by considering both geographic and aspatial features. We exploit semantic trajectories to uncover more meaningful semantic behaviours. First we enrich raw geographic trajectories with application related semantic annotations to generate semantic trajectories. We use type-of-place information of specific geographic object and multiple environmental annotations. Then, from the semantic trajectories, we extract semantic common trajectory patterns by using a proposed clustering technique. For the clustering method, we propose a similarity measure method for multi-dimensional semantic trajectories and present an OPTICS-based algorithm to group similar semantic trajectories. Using real geotagged photo data, we were able to discover interesting semantic common trajectory patterns that traditional geographic-feature-only approaches cannot.

5.2 Related work

5.2.1 Trajectory clustering for geotagged photos

Trajectory clustering means partitioning a set of trajectories into clusters in which each cluster contains similar trajectories according to a certain similarity measure. Each cluster represents a common trajectory behaviour (Lee, Han, and Whang 2007; Gaffney and Smyth 1999). Trajectory clustering has been a popular research area for other trajectory sources, such as GPS data. As early as 1999, Gaffney and Smyth (1999) presented a study on clustering trajectories and proposed a model-based trajectory clustering method. Nanni and Pedreschi (2006) proposed a density-based method for temporal-focused trajectory clustering. Several studies focused on clustering sub-trajectories (Lee, Han, and Whang 2007; Gudmundsson, Kreveld, and Speckmann 2004; Jeung et al. 2008; Li et al. 2010). They extracted clusters of common local parts of whole trajectories. Massive repositories of online geotagged photos provide a rich collection of people's trajectory data that has attracted several recent studies on extracting people's common mobility patterns. Zheng, Zha, and Chua (2012) mined tourists' popular travel routes from geotagged photos using trajectory clustering. However, a common drawback of previous studies is that they focus only on geometric features of trajectory data in the common trajectory analysis, with the result that the extracted patterns lack of meaningful semantics information and have low understandability.

Recently, Bermingham and Lee (2015) proposed a method for multi-dimensional trajectory clustering that includes consideration of speed and direction features. It adopts the TRACLUS method, which partitions trajectories into line segments. We claim that their approach also mainly focuses on the spatial geometric feature analysis. Their method is not suitable for our sequential feature focused analysis of trajectories formed by geotagged photos due to the differences between GPS-based trajectories and geotagged photo-based trajectories which are presented in Chapter 2.3.4 .

5.2.2 Similarity of trajectories for geotagged photos

In trajectory clustering method, the distance between trajectories is defined as the dissimilarity or similarity of trajectories. The sequence of elements is the most widely used form for representing a trajectory and it is used as a dissimilarity measure. Each element is considered as a multi-dimensional feature vector containing spatial and temporal values. For the dissimilarity measure, one common method is the Euclidean distance. It sums up distances between two ordered pairs of elements for sequences. However, it is unable to handle two unequal length trajectories, and it uses real values that are sensitive to noise points containing extreme values. A distance function Dynamic Time Warping (DTW) was proposed by Berndt and Clifford (1994) in order to relax the restriction of time dimension when comparing trajectories. Both distances are based on the sum of distances between two elements. Another type of distance is based on common parts of sequences. Vlachos, Kollios, and Gunopulos (2002) applied the LCSS to measure the similarity for trajectories. This distance lies on the length of the common sub-sequence of two trajectories. It reduces the effect of real values. Edit distance, ERP (Chen and Ng 2004), is another popular method used for trajectory data. This distance is the number of edits required to convert one sequence into another. Edit Distance on Real sequence (EDR) distance function (Chen, Özsu, and Oria 2005) quantifies the distance between values 0 and 1 to remove the noise effect. Furtado et al. (2016) proposed a similarity measure method for multi-dimensional trajectories. They apply weight strategy to a multi-dimensional similarity problem that assigns different weight to each dimension such that the sum of all weights is equal to 1. These weights of multiple dimensions are used to calculate a matching score between two elements. The final similarity score between trajectories is defined as the average parity, where the parity is the sum of the highest matching score of all elements of one trajectory compared with all elements of the other trajectory.

5.2.3 Density-based clustering methods for geotagged photos

For clustering, a density-based technique is one of the most popular methods applied in previous studies for discovering common trajectories (Nanni and Pedreschi 2006; Lee, Han, and Whang 2007). This is because the density-based method is efficient for finding noise and detecting outliers. Also, it is able to detect clusters of arbitrary shapes. Density-based clustering is based on data intensity. It defines a cluster as a set of objects within an area with high density. DBSCAN (Ester et al. 1996) is one classic density-based clustering method. DBSCAN requires two parameters Eps and $MinPts$ to search for areas of high density. Areas of given neighbourhood (Eps) containing at least a minimum number of points ($MinPts$) are reported as a part of a cluster. OPTICS (Ankerst et al. 1999) is a well-known density-based approach, an extended version of basic DBSCAN. Unlike DBSCAN, OPTICS algorithm does not generate object clusters explicitly, but instead, orders objects of the dataset. It reduces the sensitivity of parameter Eps. In OPTICS, an object is a core object if there exist at least $MinPts$ objects within the area of Eps distance. And a core distance of object is the distance to the $MinPts$th closest point. The reachability-distance of another object from a core object is the biggest distance between these two objects. The outcome of OPTICS algorithm is a reachability distance ordering of objects. In the ordering, objects which are closest become neighbours. Later, these ordered objects can be grouped into clusters based on additionally set radius and maximum distance considered. Object clusters of varying densities can be obtained by choosing different radius thresholds.

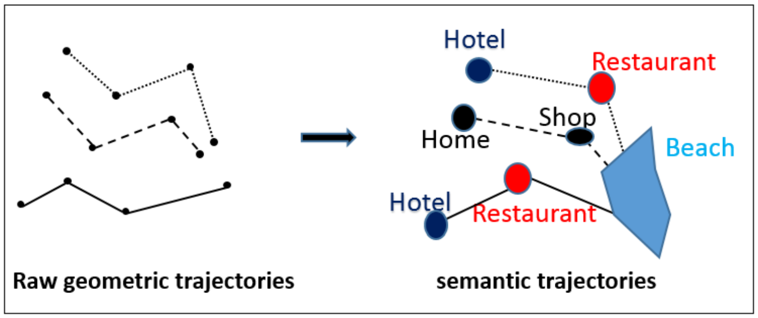


Figure 5.3: Example of semantic trajectories generated from raw geographic trajectories.

5.3 Problem statement

This study aims to mine people's common semantic mobility behaviours from geotagged photos. Given a set of geographic-information-referenced photo data, the research problem is to extract people's semantic-level common trajectory patterns. A *common trajectory pattern* shows a general semantic trajectory for which a group of people's trajectories have similar mobility at the semantic level.

The semantic trajectories, as detailed in Definition 3.7 in Chapter 3.4, provide richer semantic descriptions and meanings about the trajectory. An example of semantic trajectories transferred from raw geographic trajectories is shown in Figure 5.3. Note that each sample semantic trajectory visits several kinds of places.

A semantic common trajectory pattern refers to a group of trajectories have similar mobility. This task becomes a problem of semantic trajectory clustering. Given a set of semantic trajectories, the problem is to find a set of clusters where each cluster contains similar trajectories according to a given dissimilarity measure. Each cluster of semantic trajectories represents a person's semantic common trajectory pattern drawn from several people's trajectories.

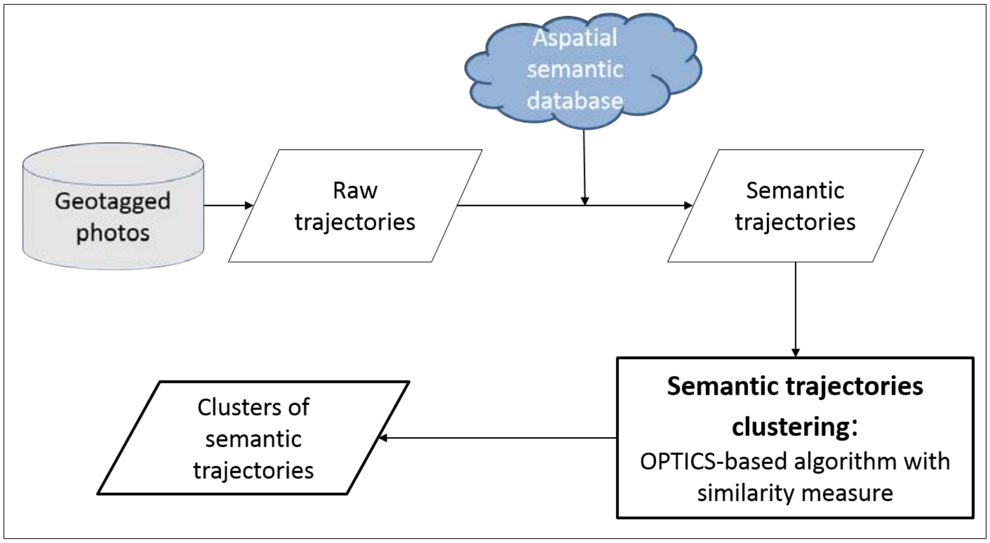


Figure 5.4: Framework for semantic common trajectory mining.

Definition 5.1 *A semantic common pattern refers to a group of trajectories which are semantically similar in terms of the visit sequences of types of places with a given set of additional contextual semantics.*

5.4 Semantic CTP Framework

Figure 5.4 shows our framework for discovering semantic common trajectory patterns from geotagged social media data. We create raw geographic trajectories of people from geotagged photos. Then, we generate people's semantic trajectories enriched with application-dependent aspatial contextual semantics information. From the semantic trajectories, we extract the semantic common patterns. This study proposes a semantic trajectory clustering approach, extending the OPTICS algorithm with a new similarity measure for semantic trajectories. We finally obtain some clusters of semantic trajectories in which each cluster represents a semantic common trajectory pattern. Details of methods for building raw trajectories and building semantic trajectories are presented in Chapter 4.4. In the rest of this section, we describe the details of our similarity method and clustering algorithm for semantic trajectories.

5.4.1 Semantic trajectory clustering

Once semantic trajectories obtained as detailed in Chapter 4.4.2, semantic trajectory clustering is in place to find common trajectory patterns. In this section, we illustrate the semantic trajectory clustering method. The method adopts the density-based clustering scheme, extended OPTICS algorithm (Ankerst et al. 1999), with a novel similarity measure for semantic trajectory data. First, the similarity function is described in Chapter 5.4.1, and then we illustrate the details of the clustering algorithm in Chapter 5.4.1.

Semantic trajectory similarity function

Given two multi-dimensional sequences, in order to compute a similarity between them, we need to figure out how many commonalities they have according to the intuition of similarity (Lin 1998). We apply the LCSS algorithm to find the common sub-sequence of two trajectories. Our similarity function includes two parts: finding the longest common sub-sequence (LCS) of trajectories and calculating a similarity score for two trajectories. Our semantic trajectories are enriched with multi-dimensional semantic annotations, which could have different degrees of importance in different applications. Specifically, in the measure of similarity among trajectories, some dimensions are required to be more important than others. In order to reflect this application-dependent and context-sensitive importance, this study adopts two classifications of semantic dimensions: compulsory dimensions and optional dimensions. Compulsory dimensions are ones whose values must match in the comparison of two RoIs using the LCSS algorithm, whilst values of optional dimensions do not have to match. This flexibility is of particular use where some dimensions must be matched but other dimensions are assigned different weights and used to calculate a matching score. We use a similar idea of weight strategy as in (Furtado et al. 2016) for multiple dimensions, especially for optional dimensions in order to reflect the different level of importance. Trajectories considered to be similar must have the same value in the compulsory dimensions, while the optional dimensions can be used as loose requirements. Various combinations of compulsory and optional dimensions would produce application-dependent and context-sensitive patterns.

Finding the longest common sub-sequence Finding the common subsequence of two semantic trajectories is the first step to process. We apply LCSS algorithm to do this task. Finding matched RoIs is a key step in the discovery of the LCS in LCSS algorithm. As each RoI is multi-dimensional, we use a weighting strategy to reflect context-sensitive weights for different dimensions. To compute a matching score of two RoIs, we consider compulsory and optional dimensions separately. For compulsory dimensions, two elements must have an exact match. Then the optional dimensions, with their associated weights, are used to calculate a matching score between two elements. For each dimension, when two elements have the same value for the dimension, we set "1" for matching score of the dimension, otherwise we set "0". As shown in Equation 5.1, the element-matching score is the sum of the matching score of every optional dimension times its associated weight. For each optional dimension k, $match_k(mdEle1, mdEle2)$ is the matching score and $weight_k$ is its associated weight value. A threshold is used to compare the matching score to determine whether two elements match or not. At the end of this process, we obtain the LCS of two trajectories.

$$MScore\left(mdEle_1, mdEle_2\right) = \sum_{k=1}^{d}\left(match_k\left(mdEle_1, mdEle_2\right) \times weight_k\right). \quad (5.1)$$

Calculating similarity score for semantic trajectories The LCSS algorithm finds the longest common sub-sequence of two semantic trajectories. To compute a similarity between trajectories, first we need to ensure that a commonality is present in most parts of both trajectories. This is measured by checking the ratio of LCS to the length of the whole trajectory. When both ratios are valid, they are used to compute the similarity score between two trajectories using the average ratio as a similarity score. Equation 5.2 shows the calculation of the ratio of LCS to the length of one multi-dimensional trajectory, where $mdst_1$ and $mdst_2$ denote two semantic trajectories, respectively. The average similarity function is in Equation 5.3. Finally, the similarity score between two given semantic trajectories is obtained. A higher similarity score value means more similar. Finally, the distance between two trajectories is the dissimilarity score calculated by using "1 - similarity" score as in Equation 5.4.

$$ratio\,(mdst_1) = \frac{|LCS\,(mdst_1, mdst_2)|}{|mdst_1|}, \tag{5.2}$$

$$sim(mdst_1, mdst_2) = \frac{ratio(mdst_1) + ratio(mdst_2)}{2}, \tag{5.3}$$

$$distance(mdst_1, mdst_2) = 1 - sim(mdst_1, mdst_2). \tag{5.4}$$

Semantic trajectory clustering algorithm

Our semantic trajectory clustering method is based on the OPTICS algorithm scheme (Ankerst et al. 1999) . We extend it to semantic trajectory-data type, by proposing a similarity measure method for semantic trajectory, which is described Chapter 5.4.1. We call this extended algorithm the SemTra-OPTICS algorithm. We also apply the ExtractDBSCANClustering method (Ankerst et al. 1999) to generate clusters from the ordering results. We integrate TB-tree (Pfoser, Jensen, and Theodoridis 2000) as an index structure in OPTICS algorithm, in order to facilitate efficient neighbourhood queries.

TB-tree structure is designed strictly for trajectory data. Each trajectory is represented as a sequence of segments, and a leaf node of TB-tree stores several segment Minimum Bounding Boxes (MBBs). Each leaf node keeps only those segments that belong to the same trajectory. A semantic trajectory is a sequence of multi-dimensional textual semantics annotations enriched stops. A stop is as a textual episode which contains a vector of textual word. The original MBB model is not suitable for textual episodes containing only text. We modified the original structure to adapt to semantic trajectory data. We transfer the model of MBB of TB-tree into Minimum Term Bag (MTB) for multi-dimensional semantic RoIs. A MTB stores episodes that build a minimum cover of terms for each dimension of episodes.

The overall process of our method for extracting common semantic trajectory patterns is illustrated in Algorithm 5.1. For a given dataset of raw geographic trajectories, we first build semantic trajectories from raw trajectories in Step 1. Then we apply SemTra-OPTICS algorithm to generate a reachability distance-ordered list of the semantic trajectories (Step 3). In Step 4, the ExtractDBSCAN-clustering method is used to generate clusters of trajectories from the ordered list. ExtractDBSCAN-clustering visits every object of the ordered list. It creates a cluster for a set of continuing objects until the terminal semantic trajectory, whose reachability distance is greater than a given epsilon threshold.

Our SemTra-OPTICS algorithm is demonstrated in detail in Algorithms 5.2 and 5.3. At the beginning, we build a SemTB-tree for semantic trajectories in Step 1 and use our semantic trajectory distance function in the SemTB-tree in Step 2. The expected ordered list is initialised in Step 3. We traverse unvisited semantic trajectories and process each of them in Steps 4-6. For an unvisited trajectory, Step 6 expands it to its neighbours. Specifically, in Algorithm 5.3, we mark the unvisited semantic trajectory as visited

Algorithm 5.1 ExtractSemanticCommonTrajectoryPattern

Input: A dataset of raw geographic trajectory T;

Output: A set of clusters of semantic trajectories;

```
semT ← generatingSemanticTrajectory(semT);
clustersList ← ∅;
orderedFile ← SemTra-OPTICS(semT,);
clustersList ← extractDBSCAN(orderedFile);
return clusterList
```

Algorithm 5.2 SemTra-OPTICS

Input: A semantic trajectory dataset T, epsilon e, and minimum points $minPts$;

Output: A list of reachability distance ordered semantic trajectories $orderedFile$;

```
build SemTB-tree;
set semanticTraDistance as the distance function of SemTB-tree;
orderedFile ← ∅;
for all semantic trajectory t ∈ T do
    if t is unvisited then
        expandClusterOrder(t, orderedFile, e, minPts);
return orderedFile
```

in Step 1, calculate and set its core distance and reachability distance in Steps 3–4 and add it into the ordered list in Step 5. We then further process its neighbourhood trajectories. In Step 8, the unvisited neighbourhood trajectories are put in a priority queue and sorted in ascending order based on their reachability distance. We handle the neighbourhood trajectories of the queue from beginning to end. For each actual neighbourhood trajectory, we progressively find its neighbourhood trajectories, add these trajectories into the priority queue and update the queue in Steps 9–16. In this processing method, semantic trajectories with their neighbourhood trajectories are stored in a final ordered list where the order the neighbourhood trajectories is based on the reachability distance. The ordered list is the outcome of SemTra-OPTICS algorithm.

5.5 Experimental setup

Our study focuses on the extraction of semantic behaviours. We propose a clustering method to find common trajectories, and conduct experiments with real geotagged photos to present semantic patterns. As mentioned in Chapter 5.4.1, our method provides flexibility to choose compulsory and optional dimensions to reflect context-sensitive and application-dependent scenarios. We design a set of experiments to explore this flexibility with various dimension combinations.

5.5.1 Baseline methods

We benchmark our approach with two popular methods: popular travel route (PTR) and ND-TRACLUS. The first method for extracting popular travel routes was proposed by Zheng, Zha, and Chua (2012). We call it the PTR method for short in this study. It is one of the state-of-art studies for mining common trajectories – popular travel routes – from geotagged social media

Algorithm 5.3 ExpandClusterOrder

Input: A semantic trajectory t, a list of distance ordered semantic trajectories $orderFile$, epsilon e, and minimum points $minPts$;

Output: A list of reachability distance ordered semantic trajectories $orderedFile$;

```
mark t as visited;
neighbors ← semTBtree.getNeighbors(t,e);
t.reachabilityDistance ← UNDEFINED;
t.setCoreDistance(neighbors, e, minPts);
add t into orderedFile;
if t.coredistance ≠ UNDEFINED then
    orderseeds ← ∅;
    update(neighbors,t,orderseeds);
    while orderseeds is not empty do
        get first trajectory actualTra in orderseeds;
        actualneighbors ← semTBtree.getNeighbors(actualtra,e);
        mark actualtra as visited;
        actualtra.setCoreDist(actualneighbors, e, minPts);
        add actualtra into orderedFile;
        if actualtra.coredistance ≠ UNDEFINED then
            update(actualneighbors,actualtra,orderseeds);
```

data. A travel route is a sequence of places visited. Zheng, Zha, and Chua (2012) first transfer raw trajectories, generated from geotagged photos, into sequences of spatial places. Then they group similar sequences by applying a clustering technique. Each resulting cluster is considered a potential popular travel route. Similar to our work, it is a trajectory clustering task. However, this method considers only the geographical spatial feature of trajectory data and lacks semantic features. Specifically, the researchers use specific geographic places without adding any further semantic annotations except the specific place entity of the spatial region. The results are sequences of places that show people's spatial behaviours only.

The other method is ND-TRACLUS, developed by Bermingham and Lee (2015). It is an extension of the well-known spatial trajectory clustering method TRACLUS (Lee, Han, and Whang 2007) to multiple dimensions. TRACLUS is proposed for grouping geographic trajectory segments. It partitions trajectories into line segments and then groups similar line segments into clusters to find common sub-trajectories. ND-TRACLUS extends it to n-dimensions. ND-TRACLUS is able to uncover new, previously unknown, higher dimensional trajectory patterns. This method considers speed and direction features of geographic trajectory. However, it is originally designed for continual GPS trajectories. This is a different type of trajectory from those generated from geotagged social media, which are neither continuous nor regular. Like the original method, ND-TRACLUS groups similar spatial trajectories according to a spatio-temporal geographic proximity measure, but takes additional speed and direction aspects into account. It is spatial-trajectory oriented, whilst our method is both spatial- and aspatial-oriented.

Table 5.1 shows major differences between our method and the baseline methods. PTR deals with geotagged social media data, but does not consider semantics; ND-TRACLUS deals with semantics used as additional features of geographic trajectory, but is oriented towards traditional GPS trajectories (regular and continuous). Our method (irregular and discrete) deals with

Table 5.1: Comparison of baseline methods for semantic common pattern mining.

	Geotagged data	**Semantics**
ND-TRACLUS	NO	YES (speed and direction of geographic trajectory)
PTR method	YES	NO (geographical spatial places)
SemTra-OPTICS	YES	YES (type of place, city, temporal and weather)

both geotagged social media data and semantics. Importantly, our method targets semantic mobility behaviours.

5.5.2 Dataset

We use the same dataset as in Figure 4.6 in Chapter 4.5 for this semantic common pattern mining. Most of the trajectories lie in coastal areas, where Queensland's major cities are located. For semantic information, we use the geographic information database of GeoNames to obtain the type of place and city information for trajectories and the weather information database from Bureau of Meteorology for weather semantics.

5.6 Results and discussions

In this section, we present two types of experimental results. First we present experimental results using our proposed method. Note that, our method has two important features: enriched semantic-level trajectories and the flexibility of compulsory and optional dimensions. We summarize experimental results with various compulsory–optional dimension combinations and present the interesting common semantic trajectory patterns extracted in

Chapter 5.6.1. Second we present comparative results with the PTR and ND-TRACLUS methods in Chapter 5.6.2. These comparisons focus on the types of common trajectory patterns.

5.6.1 Results of semantic trajectory clustering

After several rounds of parameter tuning with the RoI mining method, we were able to obtain 737 semantic trajectories having at least 2 stops with *minSup* of 0.002 and *cellSize* of 0.003.

Parameter selection

The proposed framework facilitates an exploratory data mining approach enabling users to explore datasets with different parameter values. Like other exploratory data mining tools (Zheng, Zha, and Chua 2012), there are several parameters to tune in our method. Most importantly, our framework enables users to select a set of semantic trajectory dimensions to start with. That is, the framework assists users to choose domain-specific and application-dependent semantics. In addition, the framework assists users to select compulsory (important and crucial) and optional dimensions in the calculation of semantic trajectories. In this study, there are five dimensions defined, which are PLACE_TYPE, CITY, DAY_TYPE, DAY_TIME and WEATHER, for the semantics annotations place type, city, day type, day time and weather condition, respectively. The dimension (PLACE_TYPE) is considered to be the basic semantic annotation of the trajectory and it is set as a default compulsory dimension. The optional dimensions require weight values to be set to reflect their respective importance. In this study, we take all the dimensions into consideration for the calculation of semantic trajectories. For the compulsory–optional dimension combination, we conduct two case studies with different combinations to show the effect of combination on the common semantic trajectory patterns. The first study uses dimension PLACE_TYPE as the only compulsory dimension and the other four dimensions, CITY, DAY_TYPE, DAY_TIME and WEATHER, as optional dimensions. The second study uses PLACE_TYPE and CITY as compulsory dimensions while the other three as optional dimensions. For simplicity, we set an equal weight value for each optional dimension in both two cases, that is, 0.25 for 4 optional dimensions and 0.33 for 3. Results of these two cases experiments are shown in Chapter 5.6.1.

We also need to set a threshold value for the parameter element matching score ($elematThreshold = 0.3$) and a ratio threshold ($rThreshold = 0.3$) for our similarity method. We use parameter values $minPts = 5$ and $epsilon = 1$ for OPTICS algorithm, and $epsilon = 0.5$ for extractDBSCAN-clustering method. These default values are chosen after several rounds of exploratory tests for this particular study.

Experiments with different dimensional combinations

Case 1: four optional dimensions In this experiment, we set PLACE_TYPE as the only compulsory dimension and the other four dimensions (CITY, DAY_TIME, DAY_TYPE and WEATHER) as optional dimensions. Figure 5.5 shows the reachability ordering plot outcome of the OPTICS algorithm for Case study 1. As shown in the plot, many semantic trajectories' reachability distances are "UNDEFINED" which means their similarity values are too small based on our distance function.

Figure 5.5 shows that there are 2 clusters referring to two concave areas below 0.5 reachability distance. The first cluster contains 141 semantic trajectories and the second contains 5 trajectories. For simplicity of presentation, we list members of the smaller cluster in Table 5.2. These semantic trajectories are similar in the sense that the commonality takes up at least 30% of both trajectories and each element matching score is no less than 0.3. In particular, the first trajectory is computed as a centre object that has 4 neighbours in the search range and the distance between each of the

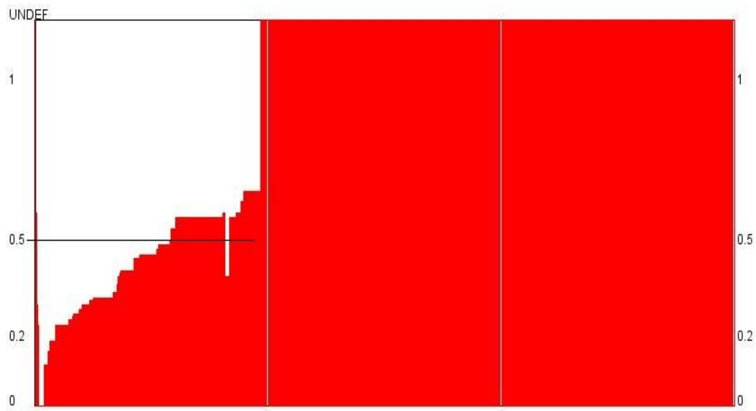


Figure 5.5: Plot chart of ordered list of semantic trajectories for Case study 1.

four trajectories and the first trajectory is 0.4 indicating that each pair of trajectories has 60% commonality.

Focusing on people's common semantic mobility behaviour, we can see from Table 5.2 that the common mobility is in Cairns area. People have similar moving patterns from hotels to a ship berth (fleets departing place to the Great Barrier Reef) and then to hotels. Also, note that, they move mostly on clear days. Even though clusters have slightly different visiting times and weather conditions, their trajectories show similar patterns. This common trajectory is well-supported by the fact that the Great Barrier Reef is one of the most famous daily travel destinations in Cairns, attracting millions of visits each year. This background knowledge supports validation of the extracted behaviour of the common semantic trajectory. Note that traditional geographic-feature-only approaches are unable to detect these detailed semantic patterns outlining place types, times, and weather conditions information.

Table 5.2: Semantic trajectory clusters having 5 members for Case study 1.

Tra_*id* & length	**Semantic trajectory**
407 : $len - 5$	$[(\mathrm{HTL}_{[weekday][Cairns][Clear][morning]})$, $(\mathrm{HTL}_{[weekday][Cairns][Clear][morning]})$, $(\mathrm{PIER}_{[weekday][Cairns][Heavy\,rain][morning]})$, $(\mathrm{HTL}_{[weekday][Cairns][Heavy\,rain][morning]})$, $(\mathrm{HTL}_{[weekday][Cairns][Clear][evening]})]$
49 : $len - 5$	$[(\mathrm{HTL}_{[weekday][Cairns][Light\,rain][morning]})$, $(\mathrm{PIER}_{[weekday][Cairns][Light\,rain][morning]})$, $(\mathrm{HTL}_{[weekday][Cairns][Light\,rain][evening]})$, $(\mathrm{PPL}_{[weekday][Cairns][Clear][dawn]})$, $(\mathrm{HTL}_{[weekday][Cairns][Clear][evening]})]$
646 : $len - 3$	$[(\mathrm{HTL}_{[weekday][Cairns][Clear][evening]})$, $(\mathrm{HTL}_{[weekday][Cairns][Clear][dawn]})$, $(\mathrm{HTL}_{[weekday][Cairns][Clear][midnight]})]$
961 : $len - 3$	$[(\mathrm{HTL}_{[weekday][Cairns][Light\,rain][evening]})$, $(\mathrm{HTL}_{[weekend][Cairns][Clear][dawn]})$, $(\mathrm{HTL}_{[weekday][Cairns][Clear][evening]})]$
704 : $len - 5$	$[(\mathrm{HTL}_{[weekday][Cairns][Clear][evening]})$, $(\mathrm{PIER}_{[weekday][Cairns][Clear][evening]})$, $(\mathrm{PIER}_{[weekday][Cairns][Clear][evening]})$, $(\mathrm{PIER}_{[weekend][Cairns][Clear][midnight]})$, $(\mathrm{HTL}_{[weekday][Cairns][Light\,rain][evening]})]$

Table 5.3: Description of feature codes for the dataset used in Fig. 5.4.

Feature code	**Description**
HTL	Hotel
ISL	Island
HSP	Hospital
DEVH	Housing development
PIER	Structure built out into navigable water providing berthing
PT	Point (a tapering piece of land projecting into a body of water)
RSTN	Railroad station
PPLX	Section of populated place: a city, town, village, or other agglomeration of buildings where people live and work

Case 2: three optional dimensions In this case, we set PLACE_TYPE and CITY (spatial semantics) as compulsory dimensions and the other three dimensions, DAY_TIME, DAY_TYPE and WEATHER as optional. This dimension combination directly affects the similarity calculation of semantic trajectories. The weight values of optional dimensions are set to 1/3 to make the total sum equal to 1. The reachability ordering plot of this case study is shown in Figure 5.6. As expected, using a more strict similarity measure condition, more semantic trajectories have "UNDEFINED" reachability which means they are not similar to each other. Interestingly, adding CITY to the compulsory dimensions, the shape of the plot chart is significantly different from the plot in Figure 5.5. We detect three clusters in this case, and they correspond to three concave areas below the 0.5 reachability distance as shown in Figure 5.6. The first cluster has 53 semantic trajectories, the second cluster has 7, and the third cluster has 9. The total number of

clustered trajectories is smaller than that of Case study 1.

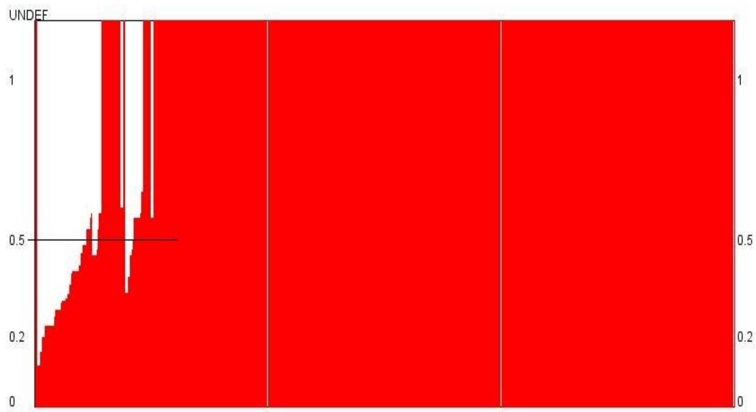


Figure 5.6: Plot chart of ordered list of semantic trajectories for Case study 2.

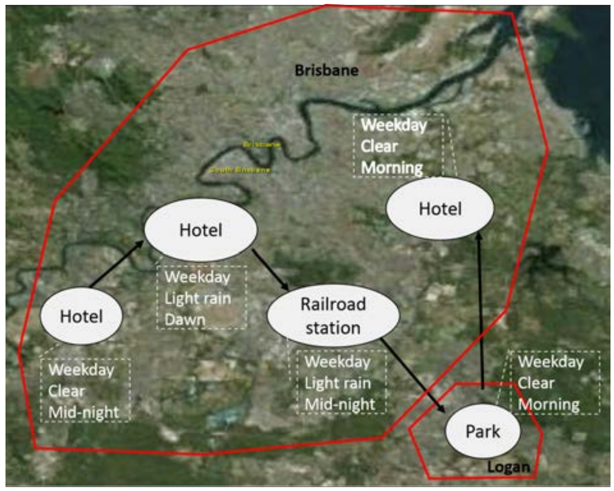
Table 5.4 lists the three common semantic trajectories of people in this case study. Figure 5.7 illustrates semantic common trajectory patterns for cluster 2 and 3. For cluster 1, common mobility is a movement between hotels in Gold Coast city on clear days. Other trajectories in this cluster have similar movements. Based on the commonality of movement between hotels, these trajectories are grouped into the same cluster. Note that Gold Coast is renowned not only for its beautiful beaches, but also for various theme parks. This cluster pattern is explained by popular daily trips to beaches and various amusement parks in Gold Coast. The second cluster presents a different common trajectory that starts from a hotel in Brisbane city on a clear morning, then passes through a rail station in Brisbane, visits a park in Logan in the morning, and goes back to the hotel. Most RoIs of this mobility are in Brisbane area. This cluster pattern explains a typical movement pattern in big cities. The third common trajectory is in Cairns, gateway to the Great Barrier Reef, and the third most popular tourist destination after Brisbane and Gold Coast. People have similar movement patterns from hotels to the ship berth where fleets depart to the reef, and then go back to hotels. This supports the popular day excursions to the Great Barrier Reef in Cairns.

Table 5.4: Representative common semantic trajectory for each cluster for Case study 2.

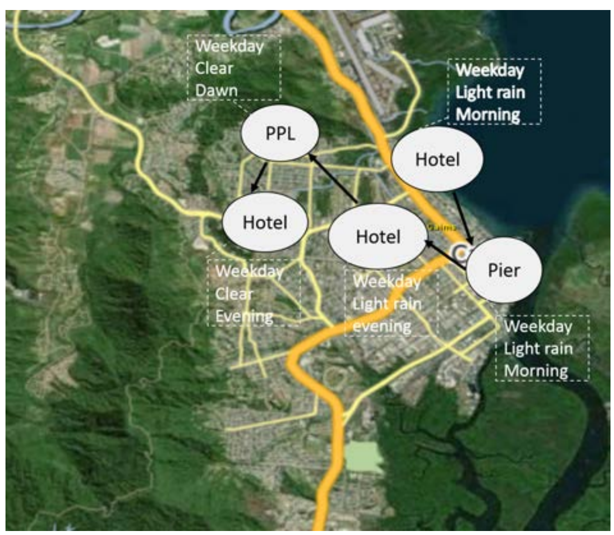
Cluster	Common semantic trajectory
cluster 1	$[(\text{HTL}_{[weekday][Clear][Gold\,Coast][midnight]}) \rightarrow (\text{HTL}_{[weekday][Clear][Gold\,Coast][midnight]}) \rightarrow (\text{HTL}_{[weekday][Clear][Gold\,Coast][dawn]}) \rightarrow (\text{HTL}_{[weekend][Clear][Gold\,Coast][dawn]})]$
cluster 2	$[(\text{HTL}_{[weekday][Clear][Brisbane][morning]}) \rightarrow (\text{HTL}_{[weekday][Light\,rain][Brisbane][dawn]}) \rightarrow (\text{RSTN}_{[weekday][Light\,rain][Brisbane][midnight]}) \rightarrow (\text{PRK}_{[weekday][Clear][Logan][morning]}) \rightarrow (\text{HTL}_{[weekday][Clear][Brisbane][morning]})]$
cluster 3	$[(\text{HTL}_{[weekday][Light\,rain][Cairns][morning]}) \rightarrow (\text{PIER}_{[weekday][Light\,rain][Cairns][morning]}) \rightarrow (\text{HTL}_{[weekday][Light\,rain][Cairns][evening]}) \rightarrow (\text{PPL}_{[weekday][Clear][Cairns][dawn]}) \rightarrow (\text{HTL}_{[weekday][Clear][Cairns][evening]})]$

5.6.2 Comparative results

In this section, we present comparative results using the PTR and ND-TRACLUS methods. We use our raw trajectory datasets extracted from Flickr for both methods, as they require geographic-feature-only trajectory datasets. We first illustrate common trajectory patterns from these three methods, and then we compare and discuss differences among them. Comparisons will focus on the types of extracted trajectory behaviours. Figure



(a) Cluster 2



(b) Cluster 3

Figure 5.7: Semantic common trajectory patterns for cluster 2 and cluster 3 shown in Table 5.4.

5.8 presents common trajectory patterns generated from the two baseline methods.

Results of PTR

Several clusters of trajectories have been extracted by the PTR method, where each cluster represents a popular tour route. We list two popular tour routes in Table 5.5. The extracted popular tour routes are sequences of spatial regions of area (RoAs). Every RoA is named by the most frequent title of all photos inside. And each RoA indicates a specific spatial region that a density of photos are taken in the region. One of the popular routes is visiting Palm Cove and then going to Agincourt Ribbon Reef. Palm Cove is a beautiful beach community with palm trees lining the beach and Agincourt Ribbon Reef is a PoI at the outer edge of the Barrier Reef renowned for snorkelling and diving. Another popular route is from Bargara Beach (near Bundaberg) to Greenmount Beach (Gold Coast). Even though these extracted routes indicate popular rough and crude tour sequences, they are unable to provide the weather-specific, place type-specific, or time-specific semantic information that our proposed method does.

Table 5.5: Popular tour routes in Queensland extracted by the PTR method.

	Popular tour route
1	Palm Cove → Agincourt Ribbon Reef
2	Bargara Beach→ Greenmount Beach

Results of ND-TRACLUS

For this particular study, we choose the following parameters: *spatial epsilon* of 500 meters, *temporal epsilon* 10 days, *directional epsilon* of $\sin 30°$, *speed epsilon* of $1 \times 10^{-8} m/s$ and used the average representative method with parameter *min lines* of 10. Figure 5.8(b) shows the outcome of ND-TRACLUS,

(a) PTR

(b) ND-TRACULS

Figure 5.8: Comparison of common trajectory patterns from three methods.

which is a representative geographic trajectory. The extracted common mobility shows a moving trace on the geographical spatial surface exhibiting a spatial trajectory behaviour. The temporal feature is represented as a vertical perspective. Geographic line segments can be extended with speed and direction features providing additional semantics to geographic trajectories.

Comparison and discussion

Figure 5.7 and 5.8 display and compare outcomes from the three methods for the same dataset. It shows that our proposed method reveals more detailed patterns than traditional geographic-feature-only approaches. We can summarise into two findings. Our first finding is that our common trajectory patterns are in semantic-level behaviours whilst traditional trajectory patterns are spatial-level behaviours. Our second finding is that our results produce different patterns from traditional approaches. Namely, our approach produces local and condition-specific patterns whilst traditional geographic-feature-only approaches produce global patterns disregarding local and condition specific semantics. Patterns from our approach are spatial and aspatial whilst patterns from traditional approaches are only spatial. Note that our approach can produce the same spatial-only patterns when the same aspatial semantics is given. Our approach produces much more detailed and semantically enhanced patterns that traditional approaches are not able to detect. The proposed method is flexible enabling users to choose a set of semantic annotations based on domain-specific applications, and also allows users to explore the effect of compulsory and optional dimensions. Table 5.6 summarises the results and mobility behaviours for the three methods under consideration.

At last, understanding tourists' common patterns has implications for tourism. For instance, a pattern of "from hotel to pier" shows one of the common routes tourists approximately take in Cairns. Cairns is the gateway to the Great Barrier Reef. There are several islands and reef platforms.

Table 5.6: Comparisons of three common pattern mining methods.

	Result	**Behaviour**
ND-TRACLUS	Geographic trajectories	Spatio-temporal behaviours
PTR	Sequences of geographic spatial places	Spatial behaviours
Our method	Sequences of place types with weather, temporal and city information	Semantic (spatial and aspatial) behaviours

This popular travel route shows tourists' movements that can provide useful advices for planning and managing the transport between hotels and piers. It also helps travel agencies arrange tours for tourists, and manage trips and ships between piers to islands and reef platforms.

5.7 Conclusion

In this study, we investigate the extraction of semantic common trajectory patterns from geotagged social media data. These patterns are useful for understanding people's mobility behaviours on a semantic level. Aspatial semantic information is as important as spatial information in spatial data mining and geographic data analysis (Miller and Han 2009). Traditional approaches consider only spatial information while disregarding aspatial semantic information. This chapter introduces a semantic trajectory mining method for common trajectory patterns. The proposed approach first builds raw trajectories from geotagged photos, and enriches them with various semantic annotations. We propose an extended OPTICS clustering algorithm to handle semantics enriched trajectories, and introduce a new similarity measure for semantic trajectories. Experimental results with real datasets reveal that our approach produces detailed spatial and aspatial patterns that traditional approaches fail to identify.

Chapter 6

Semantic trajectory pattern mining

This chapter describes the study of mining people semantic trajectory patterns from geotagged photos. Chapter 6.2 covers background techniques for spatio-temporal patterns, trajectory pattern mining and semantic sequential pattern mining, and reviews related work. Chapter 6.3 defines the semantic trajectory pattern mining problem. Then, in Chapter 6.4, we introduce the framework for extracting semantic trajectory patterns from geotagged photos and describe the proposed semantic trajectory pattern mining algorithm in detail. Experimental results are presented and discussed in Chapter 6.5. We conclude our work in Chapter 6.6.

6.1 Introduction

A trajectory pattern represents a moving sequence of places associated with transit time annotations (Giannotti et al. 2007). The transit time annotations indicate frequent time intervals between adjacent places in the sequence. A trajectory pattern is the spatio-temporal dynamic mobility that frequently occurs in movements. It reflects regular behaviours of moving objects, and delivers both the specific spatial location information and the sequential features. It also provides important frequent interval time information for the sequence and shows temporal relations between adjacent locations. Knowledge of human spatio-temporal behaviours is useful for various areas, such as city planning and tourism. The ever-increasing volume of user-generated geotagged photos provides a valuable repository of people's trajectories. These trajectories contain rich information about people's mobility behaviours, which are potentially useful and valuable to domain experts. Recently, Cai et al. (2014) presented an investigation into extraction of trajectory patterns from geotagged photos. Their results reveal patterns about people's frequent movements among spatial regions with annotated transit time information.

Previous studies dealing with geographic-feature-only trajectories have proven insufficient for many applications. There has been a great deal of research aimed at incorporating additional aspatial contextual semantic information into trajectory data mining (Ying et al. 2011; Parent et al. 2013; Zhang et al. 2014). Trajectory patterns enriched with semantic meanings are referred to as *semantic trajectory patterns*. For example, a trajectory pattern of moving among different place types (hotels, restaurants, etc.) can reveal people's mobility behaviours with respect to place categories. For an application where the type-of-place information plays an important role, semantic trajectory patterns are more relevant, focused and valuable than those without semantics.

This study investigates the extraction of semantic-level trajectory patterns. The following example shows a semantic trajectory pattern that is a frequent moving sequence of "going to a hotel and then going to a park after 2 hours on a rainy weekday and visiting a beach on a clear weekend 2 days later". This is a much more detailed and meaningful pattern than "going to Place A and then to Place B", which is what you get from a traditional geographic-feature-only trajectory. Mining semantic trajectory patterns requires a new technical development to handle both spatial and aspatial information.

$$Hotel_{[weekday][Rainy]} \xrightarrow{2hours} Park_{[weekday][Rainy]} \xrightarrow{2days} Beach_{[weekend][Clear]}.$$

This chapter proposes a semantic trajectory pattern mining algorithm to generate semantic trajectory patterns from the semantically enhanced trajectories. Our method can not only find *basic semantic patterns* which are the sequence of basic geographic semantics only, but also find *multi-dimensional semantic trajectory patterns* which are basic geographic semantic patterns with additional aspatial semantic annotations. These additional annotations could be arbitrary combinations of the initial multiple semantics. We conduct experiments using real geotagged photos to find semantic trajectory patterns, and undertake comparative experiments with the traditional geographic-feature-only trajectory pattern mining method. The results show that our method can find richer semantically meaningful and finer trajectory patterns.

6.2 Related work

In this section, we review background techniques including spatio-temporal patterns, trajectory pattern mining, semantic sequential pattern mining, and related terminologies. Some previous work related to trajectory pattern mining from geotagged social media data is also presented here.

6.2.1 Spatio-temporal patterns for geotagged photos

The massive online geotagged photos, containing a large number of people's movement trajectories, have recently attracted some investigations into extraction people spatio-temporal trajectory pattern (Cai et al. 2014; Bermingham and Lee 2014). A spatio-temporal pattern not only shows the spatial aspect of a frequent movement sequence, but also includes important temporal knowledge about mobility. Bermingham and Lee (2014) found collective spatio-temporal patterns by using spatial and temporal dimensions (3D) of trajectory. Their patterns include the information of specific frequent occurrence time of visited locations. Spatio-temporal pattern mining from trajectories has been studied in other areas and sources, like GPS trajectory data. Cao, Mamoulis, and Cheung (2007) extracted periodic patterns from a long trajectory. Time was used to decide the time period in which a movement sequence frequently occur. Kang and Yong (2010) mined spatio-temporal patterns with spatial and temporal dimensions and also considered the duration time of visited regions. None of these approaches, however, are able to consider the interval time information between locations in the movement sequence.

6.2.2 Trajectory pattern mining

Trajectory pattern mining (Giannotti et al. 2007) aims to identify the moving sequences of places with time interval annotations, the *trajectory patterns*, that frequently occur in people's trajectories. Specifically, the sequence of places shows the mobility, and the time interval annotations indicate the typical transit time between adjacent places of the mobility. Following the same spirit of temporally annotated sequences introduced by Giannotti, Nanni, and Pedreschi (2006), a *trajectory pattern* (*T-pattern*) has the following form:

$$\textit{T-pattern} = (x_0, y_0) \xrightarrow{\alpha_1} (x_1, y_2) \xrightarrow{\alpha_2} \cdots \xrightarrow{\alpha_k} (x_k, y_k).$$

This can also be represented as a couple $T = (S, A)$ of sequence $S = <(X_0, Y_0), \cdots, (X_k, Y_k)>$ with temporal annotations $A = <\alpha_1, \cdots, \alpha_k>$. Especially, any single transit time annotation α in T-pattern refers to a time span with a form of $[t_1, t_2]$. A *T-pattern* is frequent in a group of trajectories with a loose similar transit time where a time span annotation of *T-pattern* indicates an frequent area where all the similar time falls in.

The concept of the *frequent* of T-pattern is based on the notion of *support* for the T-pattern, which is defined as the number of trajectories that contain the T-pattern. For this spatio-temporal trajectory, *containment* of a *T-pattern* takes place when both spatial positions and transition times of the pattern approximately correspond to those found in input trajectories. This *spatio-temporal containment* requires that the two spatial locations are an approximate match, with some error tolerance, that the pair of spatial positions are neighbouring, and also that the two time intervals are similar, such the tolerance τ is within the temporal constraint. Refer to Giannotti's work for more details (Giannotti et al. 2007; Giannotti, Nanni, and Pedreschi 2006). Furthermore, for spatial containment, Giannotti et al. (2007) proposed a grid-based RoI mining method to determine a spatial match for spatial points. This method generates spatial RoIs, that a density of trajectories passes through, from input trajectories. Spatial points located in the same RoI are considered as neighbours. These neighbouring spatial regions are represented as a RoI. A trajectory pattern is consequently represented as a sequence of spatial regions with time intervals. For temporal containment, Giannotti et al. (2007) adopt the τ-containment introduced by Giannotti, Nanni, and Pedreschi (2006), as defined in Definition 6.1.

Definition 6.1 *(τ-containment($\preceq_\tau$)) Given a time threshold τ, a sequence $T = s_0 \xrightarrow{\alpha_1} \cdots \xrightarrow{\alpha_n} (s_n)$ is τ-contained in an input sequence $I =< (I_0, t_0)$, $\cdots, (I_m, t_m) >$, denoted as $T \preceq_\tau I$, if and only if there exists a sequence of integers $0 \leq i_0 \leq \cdots \leq i_n \leq m$ such that:*
1. $\forall_{1\leq k\leq n}.|\alpha_k - \alpha'_k| \leq \tau$ where $\forall_{1\leq k\leq n}.\alpha'_k = t_{i_k} - t_{i_{k-1}}$.

Cai et al. (2014) extracted trajectory patterns from geotagged photos by applying TPM (Giannotti et al. 2007). Various interesting trajectory patterns, moving among spatial regions with transit times, were found.

However, previous trajectory pattern mining work focused on geographic-feature-only trajectories. Specifically, the analysis of trajectories is based on the measurement of geographical information of entities, and the trajectory patterns are about movements on the spatial level. Recently, several studies (Chen, Kuo, and Peng 2015; Chen and Chiang 2016) analysed trajectories incorporating semantics and time information by transforming trajectory sequences into symbolised sequences before using PrefixSpan. However, they fail to consider multiple spatial and aspatial semantic information.

Unlike previous work, our study aims to find semantic trajectory patterns whose predicate bears on both spatial and aspatial semantic contextual data. One example of semantic trajectory pattern could be mobility among some types of places in certain weather conditions with frequent interval time information when focusing on a place type and weather semantic context. We attempt to find frequent trajectory patterns on a contextual semantic level and obtain semantically meaningful patterns on mobility.

6.3 Problem statement

From semantic trajectories, we aim to find frequent sequences of semantic elements with transit times that are frequent from trajectories in this section. These mobility behaviours are named semantic trajectory patterns. A *semantic trajectory pattern* contains a sequence of semantic elements and a sequence of transit times where each demonstrates a frequent time interval α between two consecutive elements. Adopting the spirit of trajectory patterns (Giannotti et al. 2007), we represent semantic trajectory patterns (SemT-pattern) as a pair of sequences of semantic elements and time annotations. When an element is the basic geographic semantic annotation only, SemT-pattern is called *basic SemT-patterns*; when the element is associated with multiple other semantics, SemT-pattern will be called *multi-dimensional SemT-patterns*.

Definition 6.2 *(SemT-pattern) A semantic Trajectory Pattern is a pair* $(SemS, A)$*, where* $SemS = \langle (SemA_0),\ \cdots,\ (SemA_n) \rangle$ *is a sequence of se-*

mantic elements, an element $SemA = (e, V)$ *where* e *is the basic semantic annotation and* V *is a set of additional semantic annotations, and* $A = \langle\ \alpha_1, \cdots, \alpha_n\ \rangle$ *describes the (temporal) annotations of the sequence.*

$$SemT - pattern = SemA_0 \xrightarrow{\alpha_1} SemA_1 \xrightarrow{\alpha_2} \cdots \xrightarrow{\alpha_n} SemA_n.$$

Our task is to find all frequent SemT-patterns such that the number of occurrences in trajectories given *support* of SemT-pattern is greater than a pre-defined minimum support threshold. An occurrence of SemT-pattern is that there is a trajectory containing the SemT-pattern. In this study, using multi-dimensional semantic trajectories, a containment of pattern occurs when both semantic elements and time intervals of the pattern approximately match those found in a trajectory. Specifically, we consider that a semantic element corresponds to another when the basic semantics of two elements are the same and additional semantics of an element partially match or fully match the additional semantics of the other element. This definition of match of elements adopts the dimensional containment specifically described in Definition 4.2 in Chapter 4.3. For the match of two time intervals, the gap between two time intervals is smaller than a given tolerance threshold.

Definition 6.3 *Dimensional and* τ*-containment* ($\preceq_{d,\tau}$)*. Given a semantic trajectory* $SemT = \langle (SemA_0, t_0), \cdots, (SemA_n, t_n) \rangle$*, time tolerance* τ*, and a SemT-pattern* $(SemS, A) = SemA_0 \xrightarrow{\alpha_1} \cdots \xrightarrow{\alpha_n} SemA_k$*, we say that* $(SemS, A)$ *is contained in* $SemT$ *(*$(SemS, A) \preceq_{d,\tau} SemT$*), if and only if there exists a sub-sequence* $SemT'$ *of* $SemT$*,* $SemT' = \langle (SemA'_0, t'_0), \cdots, (SemA'_k, t'_k) \rangle$ *such that:*

1. $SemS \preceq_d SemT'.sequence\ of\ SemA;\ \forall_{0 \le j \le k}, e_j = e'_j,\ and\ V_j \subseteq V'_j;$

2. $\forall_{1 \le j \le k}.|\alpha_j - \alpha'_j| \le \tau;$ *where* $\forall_{1 \le k \le n}.\alpha'_j = t'_j - t'_{j-1}.$

Definition 6.4 *support of a SemT-pattern as*

$$supp(SemT\text{-}pattern) = \frac{|SemT^* \in D | SemT\text{-}pattern \preceq_{N,\tau} SemT^*|}{|D|}.$$

Definition 6.5 *Semantic trajectory pattern mining: Given a database of input trajectories D, a time tolerance τ and a minimum support threshold $minSup$, the semantic trajectory pattern mining problem is to find all frequent SemT-patterns whose support is no less than $minSup$. Support of a SemT-pattern is the number of trajectories $SemT \in D$ such that SemT-pattern $\preceq_{d,\tau} T$.*

6.4 Semantic TPM framework

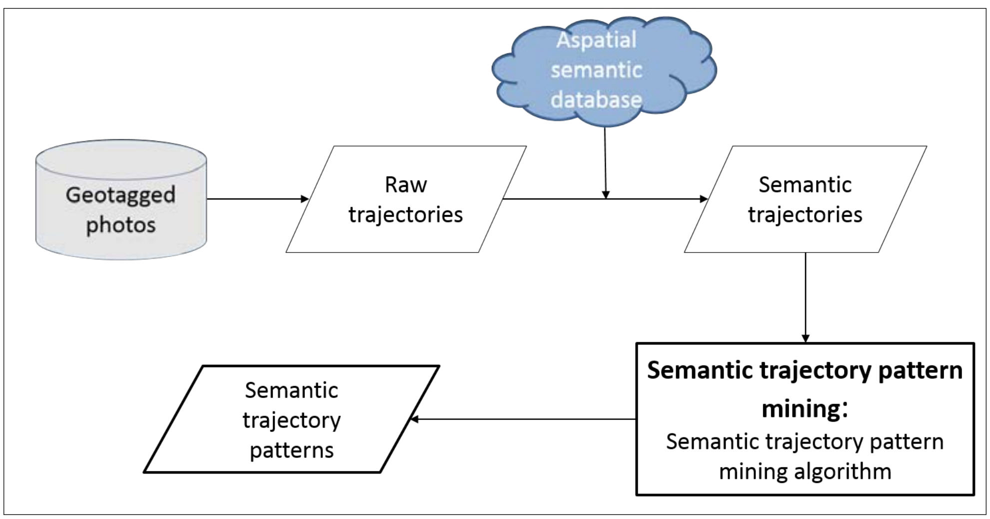


Figure 6.1: Framework of semantic trajectory pattern mining.

Figure 6.1 demonstrates our framework for finding semantic trajectory patterns from user-generated geotagged contents. The framework contains three main steps: (1) generation of raw geographic-feature only trajectories; (2) generation of semantic trajectories from raw trajectories; and (3) extraction of semantic trajectory patterns from those semantic trajectories. The first two steps are common processes with the main research framework shown in Figure 3.2 and other tasks. First, we create raw geographic trajectories from geotagged photos. Then, we generate semantic trajectories. In the third step, we apply our semantic trajectory pattern mining algorithm to find frequent semantic patterns from semantic trajectories.

6.4.1 Generating semantic trajectories

The first two steps are creating raw trajectories from geotagged photos and building semantic trajectories. Detailed information on these two common steps is presented in Chapters 4.4.1 and 4.4.2. Each extracted semantic RoI, particularly the place type semantic annotation, is considered to be an interesting stop of the trajectory. Last, the semantic trajectories are built as sequences of semantic stops with multiple contextual environmental semantics.

6.4.2 Semantic trajectory pattern mining

We propose a semantic trajectory pattern mining algorithm to find SemT-patterns from the generated semantic trajectories. Our algorithm is developed based on $\mathcal{TAS}$ algorithm scheme (Giannotti, Nanni, and Pedreschi 2006) which extends the PrefixSpan (Pei et al. 2001) projection-based method. PrefixSpan algorithm is a classic method designed for frequent sequential pattern mining. It finds frequent items and adds each frequent item to the existing item sequence to make a new longer item sequence. An item sequence is defined as the prefix and the prefixes are considered to be sequential patterns. The algorithm also creates a projection of sequences for each new prefix. The projection stores sequences that contain the latest found frequent item. For each involved sequence, all items that precede the latest found frequent item are removed. PrefixSpan recursively generates longer patterns and computes projection of sequences.

Based on PrefixSpan, Giannotti, Nanni, and Pedreschi (2006) proposed the $\mathcal{TAS}$ mining algorithm to find temporally annotated sequential patterns which are sequential patterns with additional time annotations. The time annotation indicates the transition times between each two adjacent elements of the sequence. It is different from normal projected sequence data, in that $\mathcal{TAS}$ algorithm uses T-sequence data type. A projection is composed of T-sequences. A T-sequence is a projected sequence enriched with an annotation sequence. An annotation sequence includes records of occurrences of the prefix in the original sequence. The occurrence record contains a sequence of time stamps of such occurrence and a pointer to the element of sequence where the occurrence terminates. In the $\mathcal{TAS}$ algorithm, prefixes are generated in the normal projection-based way, and the occurrences are used to calculate the frequent transition time annotations for the prefixes. Integrating the prefix and a frequent time annotation makes temporally annotated sequential patterns. In each successive projection extension, T-sequences, including annotation sequence, are maintained for the computation in the next level.

Adopting $\mathcal{TAS}$ algorithm scheme, our semantic trajectory pattern mining method finds semantic trajectory patterns from semantic trajectories. Our algorithm includes a new technique of handling multiple semantic dimensions, technique for generating multi-dimensional patterns and a modified approach of calculating frequent time interval annotations. For multiple dimensions issues, it includes computing frequent combinations of dimensions and creating multi-dimensional prefixes. The algorithm computes frequent interval sequences for SemT-pattern in a progressively increasing way such that length of frequent interval sequence is incremented along the deeper level projection extension.

Algorithm 6.1 shows the procedure of the semantic trajectory pattern mining algorithm. For an actual projection, the algorithm extracts frequent time interval sequences in Step 7, and generates semantic trajectory patterns by integrating prefixes and frequent interval sequences in Step 8. Step 10 removes the occurrences of prefixes that do not contribute to the frequent interval sequences. Steps 11-13 extend actual projection that generates sub-

Algorithm 6.1 SemanticTrajectoryPatternMining

Input: A set of semantic trajectories T, a min sup *minSup*, a temporal threshold *tau*;

Output: A set of semantic trajectory patterns (SemT-patterns);

$L \leftarrow 0$;
$P_0 \leftarrow \{T \times \{ \langle \rangle \} \}$;
while $P_L \neq \emptyset$ **do**
    $P_{L+1} \leftarrow \emptyset$;
    for all $P \in P_L$ **do**
        if P.prefix ≥ 2 **then**
            ExtractFrequentIntervalAnnotations(P);
            patterns $\leftarrow$ GeneratingTrajectoryPatterns(P);
            Output(*patterns*);
            $P \leftarrow$ PruneAnnotations(P,*Intervals*);
        for all element $e \in P$ **do**
            if support(e) $\geq$ *minSup* **then**
                $P_{L+1} \leftarrow P_{L+1} \cup \{$ExtendProjection($P$,$e$)$\}$;
    $L++$;

projection for each newly extracted frequent item of actual projection. This algorithm progressively finds longer patterns.

Multi-dimensional sequence projection

Besides the generation of projected sequences and up-to-date annotations in T-sequences, our method requires extra steps to generate multi-dimensional prefixes that will be used to make multi-dimensional SemT-patterns. Specifically, we extend the basic prefix, the sequence of basic geographic semantic annotations, in the usual way finding a frequent element and adding it to the prefix of the actual projection to make new longer prefixes. Also, we generate the projected sequences by selecting the sub-sequence of the sequence that starts at the next element of the frequent element. For those up-to-date annotation sequences in T-sequence, each annotation will be extended with an occurrence of the projected element successive to the entry-point of the former, as described in (Giannotti, Nanni, and Pedreschi 2006). These basic prefixes will be used to generate basic SemT-patterns. In this study, we also produce multi-dimensional SemT-patterns based on the multi-dimensional prefixes. In the next step, our method also needs to extend each multi-dimensional prefix, which is a basic prefix with additional frequent semantics. The key point is to find multi-dimensional elements. The task is to find the combination of multiple additional semantics that is frequent for each frequent basic element. To do this, we apply the BUC algorithm (Beyer and Ramakrishnan 1999) to all initial additional semantics for every frequent element. The results are arbitrary combinations of partial semantics or all those multiple semantics that are frequent. Frequent basic elements with these extracted combinations (that are frequent) become multi-dimensional elements. And, based on consistent combinations, these multi-dimensional elements are added to the multi-dimensional prefixes of the actual projection to make new longer multi-dimensional prefixes.

Algorithm 6.2 illustrates a procedure for extending the projection method. For multi-dimensional prefixes we need to store values of dimensions for every frequent item in Step 7. The values of dimensions are then calculated by the BUC algorithm as in Step 9. The results of the BUC algorithm are sets of frequent values of arbitrary dimensional combinations. This approach solves the issue of arbitrary combination of dimensions. As we did with the extension of basic prefixes, we add frequent values of dimensions to existing multi-dimensional prefixes to create expected new multi-dimensional prefixes in Steps 11-15. To keep consistent combination of dimensions, we connect new frequent values to multi-dimensional prefixes that both have the same dimensions in Step 12. We also need to ensure that both frequent value and

Algorithm 6.2 ExtendProjection

Input: A projection P and an element ele;

Output: A projection of P w.r.t ele;

1: $P' \leftarrow \emptyset$, $D_{dims} \leftarrow \emptyset$;
2: **for all** T-sequence t=$(S,A) \in P$, $e \in t$ **do**
3:     $S' \leftarrow S|_{ele}$ and $A' \leftarrow \langle\rangle$;
4:     **for all** annotation$(a,e) \in A$ **do**
5:         **for all** $(s,t) \in S$, $ele \in s \wedge t > e$ **do**
6:             $A' \leftarrow append(A', (append(a,t), \rightarrow t))$;
7:             D_{dims}.add($t.ele.dimValues$);
8:     $P' \leftarrow P' \cup \{(S', A')\}$;
9: $freVals \leftarrow$ computeFrequentValues(D_{dims});
10: Remove infrequent values base on number of unique sequences;
11: **for all** $mdPrefix \in P.mdPrefixes$ **do**
12:     $fredimValues \leftarrow$ freVals.values(mdPrefix.dimensions);
13:     **for all** $freVal \in fredimValues$ **do**
14:         **if** $(mdPrefix.sequences \cap freValue.sequences) \geq minSup$ **then**
15:             $P'.mdPrefix \leftarrow P'.mdPrefix$
            $\cup\ (dim, append(mdPrefix, freValue))$;
16: **return** P'

the extended prefixes belong to the same sequence in Step 14.

Finding frequent time interval annotations

Another process is to calculate time interval annotations for trajectory patterns. In projections, the type of T-sequence is used as a projected sequence. A T-sequence contains an annotation sequence that stores several records of occurrences of prefixes in the sequence. An occurrence includes a sequence of timestamps of such occurrences. The algorithm uses these timestamps to find the time interval between elements, and calculates frequent intervals. We generate frequent interval sequences in a progressively increasing way. Specifically, we first calculate the frequent intervals for the last two elements of the prefix, and then add the frequent interval into the interval sequence to make a longer interval sequence for actual projection. These new, longer interval sequences will be integrated with prefixes of actual projection to generate SemT-patterns.

The procedure for finding frequent interval annotations is presented in Algorithm 6.3. For an actual projection, the algorithm first collects all the time blocks of two elements from occurrence sequences in Steps 1–6. Then, Steps 7–15 calculate frequent time intervals. Last, this algorithm generates new, longer interval annotations in Steps 15–20. In projection, a prefix may occur several times at different positions in a sequence where several occurrences are stored. To find a frequent interval between the last element and its former element of the basic prefix, we need to use all transit times of all occurrences in Step 4. A time block, an interval, is then created for each time by making a time range of 2τ in Step 5. An interval has two boundaries, the lowest time value and the highest value. Some intervals probably have intersecting ranges. Areas of intersections are various. To make the calculation of frequent intervals easier, we create some basic interval cells based on unique values of boundaries of all intervals in Steps 7–9. So, each interval will take place at several basic interval cells and conversely each basic interval cell is covered by some intervals. The density of a basic cell is the number of intervals that covers the basic cell in Steps 10–14. We remove invalid basic cells whose density is less than the frequency threshold. In the next step, this algorithm merges the neighbouring basic interval cells to make final frequent intervals in Step 15. This strategy is to reduce the number of SemT-patterns. Finally, we append frequent intervals to interval sequences to make new, longer frequent interval sequences according to that both belonging to the same occurrence in Step 16. The new interval sequences, which occur in

Algorithm 6.3 ExtractFrequentIntervalAnnotations

Input: A Projection database P;

Output: A set of extended frequent interval sequences;

```
intervals ← ∅;
for all T-sequence t=(S,A) ∈ P do
    for all annotation(a,e) ∈ A do
        time ← a.lastEle.time - a.SecondlastEle.time;
        interval with center time and edge 2τ;
        intervals.add(interval);
basicIntervals ← ∅, timeBoundaries ← ∅;
for all interval ∈ intervals do
    timeBoundaries = timeBoundaries ∪ interval.timeBoundaries;
Sort timeBoundaries;
Build basicIntervals based on timeBoundaries;
for all interval ∈ intervals do
    involvedBasicIntervals ← interval ∩ basicIntervals;
    For each basicInterval in involvedBasicIntervals, increment basicInterval.density;
Remove sparse basic intervals from basicIntervals;
frequentIntervals ← mergeNeighborhood(basicIntervals);
for all sequence ∈ P.lastLevelIntervalSequence do
    for all interval ∈ frequentIntervals do
        if (sequence.occurrences∩interval.occurrences) ≥ minSup then
            P.intervalSequences ← P.intervalSequences
            ∪ append(sequence, interval);
```

a density of sequences, are stored for actual projection in Steps 17–20 that will be used to produce semantic trajectory patterns.

Once frequent interval annotations have been extracted, we generate semantic trajectory patterns from actual projections in Step 8 of Algorithm 6.1. Frequent interval annotations are integrated with the basic prefix and multi-dimensional prefixes to make basic SemT-patterns and multi-dimensional SemT-patterns, respectively. During the process, we need to check the number of unique sequences that contain the patterns to ensure if it is frequent.

6.5 Experiments

We conducted experiments to show that our method has the ability to find SemT-patterns: both basic patterns and multi-dimensional patterns with an arbitrary combination of dimensions. We also executed experiments to compare our method with the traditional geographic-feature based TPM method (Cai et al. 2014). Experimental results demonstrate that the SemT-patterns discovered provide more semantically meaningful information than the results of the TPM method and demonstrate people's semantic-level mobility behaviours. Moreover, our method finds a greater number of SemT-patterns, while the TPM method extracts fewer geographic trajectory patterns.

6.5.1 Parameters

We use the same dataset used in Figure 4.6 in Chapter 4.5. Our method requires three parameters: minimum support (*minSup*) for a cell to become a RoI and trajectory to become SemT-pattern, size of geographical grid cell (*cellSize*) which is used to partition the study region and time tolerance (*tau*) which is the acceptable range for a time interval. In this study, semantic trajectories that have fewer than 30 elements will be used because the few long trajectories, containing many identical place type elements, will produce a huge number of occurrences that cost expensive running time. Moreover, small values of parameter *minSup* lead to high consumption of memory as the PrefixSpan algorithm quickly generates sequence dataset and holds the candidate sequence patterns as well as the sequence dataset. As reported by Cai et al. (2014), increasing the value of *tau* generates more frequent patterns, the valid range of time interval becomes wider, and more intersections of time intervals occur. Similarly, an increase of *cellSize* values will produce more RoIs. As a result, more valid points and trajectories will be considered in the calculation of patterns, and thus will result in more potential patterns. In this experiment, we select a value 0.008 (0.8%) for parameter *minSup*, a value 0.0015 (150 meters) for parameter *cellSize*, and a value of 2 days for *tau*.

6.5.2 Semantic trajectory patterns

We found 65 basic semantic trajectory patterns, of which the bulk were 2- and 3-length patterns (29 and 28 respectively). The remainder were 4-length patterns, and only one 5-length pattern. We also obtained 2,124 multi-dimensional semantic trajectory patterns. We demonstrate several typical results including basic place type semantic trajectory patterns, patterns associated with multiple additional dimensions, arbitrary combinations of dimensions, and patterns with various frequent time intervals.

Table 6.1 lists some basic semantic trajectory patterns for lengths of 2 to 5. Every pattern shows a frequent trajectory moving from a type of place to others type of places, with frequent time interval information. Each element of the pattern is a feature code from the GeoNames database used to categorise places. The descriptions of feature codes are listed in Table 6.2. These patterns provide us with meaningful information on mobilities among types of places and transit times. Specifically, for 2-length patterns, one frequent pattern is “going to a hotel and then going to a rail station with a time interval range of 0 to 5 days". Another 2-length pattern is "from a park to a populated place with an interval time in 0 to 2 days". A third pattern is "moving from a hotel to a bridge after 0 to 3 days". For other patterns with a longer length involve more types of places and show diverse mobility. As shown in the example patterns, the place type of "hotel" occurs in many patterns and occupies most elements in long patterns; in particular, the 5-length pattern is about movements among hotels only. This is because many initial semantic trajectories have several hotel elements. However, the SemT-patterns generated from our method help users understand people's frequent mobility behaviours on the geographic semantic level.

Table 6.1: Samples of basic SemT-patterns.

Length	**Basic semantic trajectory pattern**
2	$\text{HTL} \xrightarrow{[0,5]} \text{RSTN}$, $\text{PRK} \xrightarrow{[0,2]} \text{PPLX}$, $\text{HTL} \xrightarrow{[0,3]} \text{BDG}$
3	$\text{HTL} \xrightarrow{[0,3]} \text{PPLA} \xrightarrow{[0,2]} \text{HTL}$, $\text{HTL} \xrightarrow{[0,35]} \text{HTL} \xrightarrow{[0,2]} \text{BDG}$, $\text{RSTN} \xrightarrow{[0,8]} \text{HTL} \xrightarrow{[0,2]} \text{HTL}$
4	$\text{HTL} \xrightarrow{[0,35]} \text{HTL} \xrightarrow{[0,15]} \text{HTL} \xrightarrow{[0,2]} \text{PPLA}$
5	$\text{HTL} \xrightarrow{[0,35]} \text{HTL} \xrightarrow{[0,15]} \text{HTL} \xrightarrow{[0,4]} \text{HTL} \xrightarrow{[0,2]} \text{HTL}$

Our method also generates semantic trajectory patterns with multiple additional semantics. Table 6.3 lists some multi-dimensional semantic patterns belonging to the group of basic pattern of hotels to populated places. These additional semantics provide much richer information about people's frequent mobilities. For a combination of day type and city dimensions, we find that a pattern "visiting hotel on a weekday in Brisbane first and then moving to a place in an administrative division on weekday after 0 to 3 days".

Table 6.2: Descriptions of feature codes for semantic trajectory pattern mining.

Feature Code	**Description**
HTL	Hotel
PRK	Park
BDG	Bridge
PPLA	Seat of a first-order administrative division
RSTN	Railroad station
PPLX	Section of populated place: a city, town, village, or other agglomeration of buildings where people live and work

This pattern indicates the day type and city information of the mobility. For dimensions of day type, city and weather together, a finer SemT-pattern is found that a frequent mobility occurs on a clear day and in a weekday in Brisbane that contains the additional weather condition information besides the day type and city information.

Another important feature of our results is the ability to have arbitrary combinations of multiple dimensions. We add four additional dimensions into the semantic trajectories. The multi-dimensional SemT-patterns we obtained are basic patterns with different combinations of partial dimensions or all four additional dimensions. As shown in Table 6.3, we find patterns with a combination of day type and city dimensions, patterns with a combination of day type and weather dimensions, and patterns with all three of day type, city and weather dimensions. The benefit of this feature is that we can find some interesting semantic trajectory patterns with partial additional dimensions when all four dimensions together are calculated as infrequent.

Benefiting from the $\mathcal{TAS}$ algorithm, the same sequence of place types can have various time interval annotations that can generate various SemT-

Table 6.3: Samples of multi-dimensional SemT-patterns for a pattern: HTL $\xrightarrow{[0,3]}$ PPLA.

Combination of dimensions	**Semantic pattern**
DAY TYPE, CITY	$\text{HTL}_{[weekday][Brisbane]} \xrightarrow{[0,3]} \text{PPLA}_{[weekday][Brisbane]}$
DAY TYPE, WEATHER	1) $\text{HTL}_{[weekend][clear]} \xrightarrow{[0,3]} \text{PPLA}_{[weekday][clear]}$, 2) $\text{HTL}_{[weekday][clear]} \xrightarrow{[0,3]} \text{PPLA}_{[weekday][clear]}$
DAY TYPE, CITY, WEATHER	$\text{HTL}_{[weekday][Brisbane][clear]} \xrightarrow{[0,3]} \text{PPLA}_{[weekday][Brisbane][clear]}$

patterns. These various time annotations provide people with more knowledge about transit time between types of places. As shown in Table 6.4, for the basic pattern of visiting two hotels followed by a railroad station, there are two different frequent transit times. One interval group is spending a range of 0 to 35 days between the first two hotels, and 0 to 3 days from the second hotel to a railroad station. The other is 39 to 59 days interval between two hotels. Two multi-dimensional patterns with weather and city semantics have the same time interval group with the two basic patterns, respectively.

6.5.3 Comparison with TPM method

T-patterns from TPM

In this section, we compare our method with the TPM method (Cai et al. 2014), which already outperforms the original TPM (Giannotti et al. 2007). We focus on comparisons of the type of extracted trajectory patterns, and comparisons of the number of SemT-patterns and T-patterns. Using the same values of parameters with *minSup*=0.008, *cellsize*=0.0015, *tau*=2, the TPM method generated 33 spatial RoIs and found 25 patterns, mostly 2-length

Table 6.4: Samples of SemT-patterns with various time intervals: HTL$\rightarrow$ HTL $\rightarrow$ RSTN.

Type of pattern	**Combination of dimensions**	**Semantic pattern**
Basic patterns		1)HTL $\xrightarrow{[0,35]}$ HTL $\xrightarrow{[0,3]}$ RSTN, 2)HTL $\xrightarrow{[39,59]}$ HTL $\xrightarrow{[0,3]}$ RSTN
Multi-dimensional patterns	WEATHER, CITY	1)HTL$_{[Clear][Brisbane]}$ $\xrightarrow{[0,35]}$ HTL$_{[Clear][Brisbane]}$ $\xrightarrow{[0,3]}$ RSTN$_{[Clear][Brisbane]}$, 2)HTL$_{[Clear][Brisbane]}$ $\xrightarrow{[39,59]}$ HTL$_{[Clear][Brisbane]}$ $\xrightarrow{[0,3]}$ RSTN$_{[Clear][Brisbane]}$

patterns (24), with just one 3-length pattern.

Table 6.5: Samples of T-patterns from the TPM method.

Length	**Trajectory pattern**
2	$R_1 \xrightarrow{[0,2]} R_{35}$ $R_8 \xrightarrow{[0,4]} R_{45}$, $R_{24} \xrightarrow{[11,12]} R_{67}$
3	$R_{27} \xrightarrow{[0,2]} R_{85} \xrightarrow{[0,2]} R_{134}$

Table 6.5 lists four examples of T-patterns. These T-patterns are sequences of spatial RoI labels with time intervals. Each spatial RoI is composed of several neighbouring spatial cells represented as bounding boxes of geographical coordinates. One 2-length T-pattern is “visiting region 1 first and then going to region 35 after 0 to 2 days”. By visualising these four spatial T-patterns on NASA earth shown in Figure 6.2, we can see where the

spatial RoIs and T-patterns are located. From the top left picture (Figure 6.2(a)), we can find that this 2-length pattern is located in Brisbane and its two spatial regions are geographically close. The other two 2-length patterns, Figure 6.2 (b) and (c), are located in Cairns city and Brisbane city, respectively. The 3-length pattern, Figure 6.2 (d) is in Brisbane moving between two regions where one region has two different labels representing two different visits. One major issue with the traditional approach is that we need a map overlay or map-matching to make sense of those detected RoIs.

(a) R1 to R35

(b) R8 to R45

(c) R24 to R67

(d) R27 to R85 to R134

Figure 6.2: Samples of T-patterns plotted on NASA earth.

Table 6.6: Samples of 2-length SemT-patterns.

Length	**Semantic trajectory pattern**
2	HTL $\xrightarrow{[0,5]}$ RSTN PARK $\xrightarrow{[0,2]}$ PPLX $\text{HTL}_{[Cairns]} \xrightarrow{[0,2days]}$ Pier $\text{HTL}_{[weekday][BNE][clear]} \xrightarrow{[0,3]} \text{PPLA}_{[weekday][BNE][clear]}$

Comparison

Semantic trajectory patterns provide richer semantically meaningful information and semantic-level behaviours than the geographic T-patterns. Table 6.6 lists some 2-length SemT-patterns similar to the 2-length T-patterns shown in Table 6.5. Note that, the main differences between SemT-patterns and T-patterns is that SemT-patterns present frequent movement patterns between types of places while T-patterns show frequent movement patterns between spatial regions labelled with *id*. Obviously, the type of place provides more meaningful and readable semantic information than the abstraction of a spatial region with an identification number. Moreover, there are several additional pieces of semantic information added to the SemT-patterns. Though we can obtain semantic knowledge of T-patterns through post-processing methods such as map-matching or map overlap, there are still some drawbacks. First, post-processing is not a natural way of producing semantic patterns. Second, the TPM method misses some potential semantic-level patterns.

Figure 6.3 visualises a 2-length SemT-pattern ($\text{HTL}_{[Cairns]} \xrightarrow{[0,2days]}$ Pier) corresponding to the 2-length T-pattern ($R_8 \xrightarrow{[0,4]} R_{45}$) shown in Figure 6.2(b). Note that in Cairns, the Great Barrier Reef is one of the most famous daily travel destinations which attracts millions of people to visit. Obviously, Figure 6.3 is easier to understand than Figure 6.2(b). This SemT-pattern shows a mobility behaviour of hotel to the fleet station. In fact, there are some other reef tour routes from spatially different piers to different islands or reef platforms in Cairns. These routes fail to be triggered as patterns because the number of involved trajectories is less than the minimum support threshold.

Figure 6.3: A 2-length SemT-pattern: $\text{Hotel}_{[Cairns]} \xrightarrow{[0,2days]} \text{Pier}_{[Cairns]}$.

Obviously, our SemT-patterns provide richer information and higher semantic-level behaviours in both behaviour and information levels. Differences between SemT-patterns and spatial T-patterns can be summarised as follows:

1. Behaviour level:

    - SemT-patterns: (a) (basic patterns) movement on the type-of-place semantic level; (b) (finer patterns) movement on the type-of-place + subset of additional weather, temporal and city dimensions.
    - T-patterns: (spatial level) movement on the spatial level.

2. Information level:

    - SemT-patterns: (place type) hotel, park, rail station etc; (weather) clear, rainy, overcast etc; (temporal) day, weekday, weekend, morning, daytime, etc; (city type): Brisbane, Cairns, Gold Coast etc.

- T-patterns: spatial RoIs with *id*.

SemT-patterns show people's movement behaviours on a high semantic level whilst spatial T-patterns depict those on a low geographic spatial level. SemT-patterns reveal **how** people move on the type-of-place semantic level, whilst T-patterns show spatial positions of patterns. For applications that require advanced knowledge about people's movement behaviours at the high semantic level, SemT-patterns are more understandable, readable, readily usable, useful and valuable than spatial T-patterns.

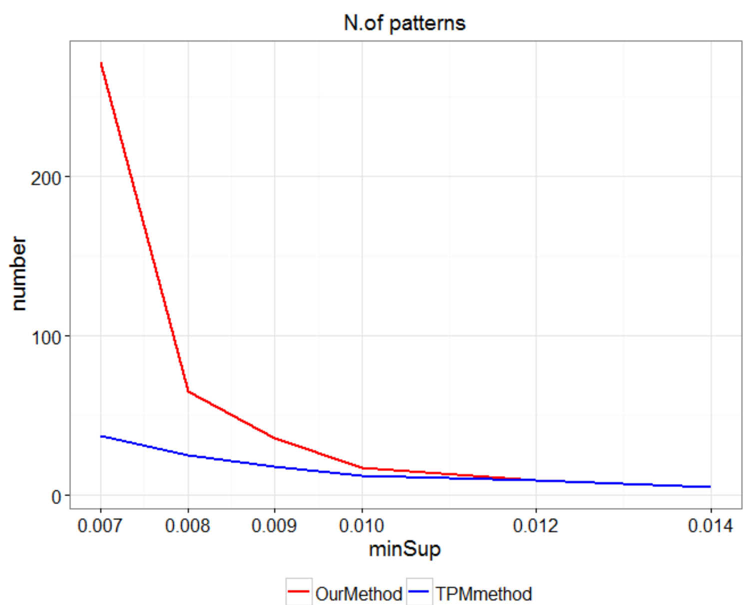


Figure 6.4: Comparison of number of patterns for SemT-patterns and T-patterns.

In addition, our method finds more trajectory patterns than the TPM method does. Figure 6.4 compares the number of basic patterns found by our method, and the numbers of patterns found by the TPM method. As we mentioned an increase in the value of parameter τ produces more patterns for both methods, we use $cellSize$=0.0015 and τ=2, and test values of $minSup$ from 0.007 to 0.014. Using bigger values, both methods find few patterns, specifically, for a value of 0.014, our method finds 5 basic patterns and TPM finds 2 patterns. In contrast, using small values, our method finds many more basic patterns than the TPM method. As discussed above, more patterns are trigged when using semantic features however they become infrequent with the geographic-feature-only approach. Moreover, our method can find many multi-dimensional semantic patterns that are also useful to the understanding of mobility, which is not possible with traditional TPM.

The two methods require a similar memory requirement due to the rapidly growing number of projections. Both methods cost a similar running time as the value of $minSup$ increases. Our method requires slightly more time with smaller $minSup$ values as our approach generates a greater number of patterns. Note that our method needs extra steps to process additional projections for multi-dimensional prefixes and extension of multi-dimensional patterns.

6.5.4 Discussions

There are some obvious findings from the experimental results. First, our method is able to generate semantic-level trajectory patterns that provide semantically meaningful information and knowledge about trajectory patterns. We are able to find people's frequent mobility patterns among different place type semantics. Second, using additional semantic dimensions, we are able to find patterns associated with additional semantics that provide richer information about mobility. The patterns are associated with city, day type, day time, and weather condition semantics. Third, our method finds potential semantic trajectory patterns, which the previous geographic-feature-only method was unable to find. Last, another important benefit of our method is an automatic extraction of various combinations of semantics dimensions.

These semantic trajectory patterns with arbitrary combinations of dimensions also provide useful insight into people's mobility.

This paragraph illustrates an example of using trajectory patterns for tourism. These temporally annotated trajectory patterns show travel route behaviours and corresponding interval time information. These trajectory patterns have implications for tourism. Specifically, these patterns help tourists plan a travel itinerary based on the travel duration, and provide transport departments or travel agencies with useful advices for designing tours and transports between places. For example, a pattern of visiting from "hotel to pier with about 0 to 2 days in Cairns" tells tourists how long they can spend in Cairns to visit the Great Barrier Reef. This pattern also provides transport service suppliers with information about the time tourists use to transfer between two stops which could assist them in planning transport and arranging public transport vehicles.

6.6 Conclusion

This study is an investigation into analysing georeferenced social media data to find people's semantic trajectory patterns which are in the place type semantic level, frequent moving sequences with interval times. We propose a semantic trajectory pattern mining method to find semantic trajectory patterns from actual semantic trajectories. Using real geotagged photos, we find many interesting semantic trajectory patterns. These patterns show frequent mobility among types of places along with transit times between entities. Experimental results also show that our method is able to find trajectory patterns with various additional semantics. These semantic trajectory patterns provide rich semantic information about people's mobility behaviours. SemT-patterns are more readable, potentially useful, readily usable, interpretable, and valuable than any patterns generated from this type of data before.

Chapter 7

Semantic itinerary recommender system

This chapter describes the semantic itinerary recommender system. Our system uses data from geotagged photos to provide users with suggested travel itineraries containing a sequence of places with transit time information. Chapter 7.1 introduces the motivations of our study. Chapter 7.2 reviews current studies in itinerary recommender systems. Chapter 7.3 formulates problems and provides the problem statements for the semantic itinerary recommender. Chapter 7.4 introduces the framework of our proposed itinerary recommender system based on trajectory pattern mining from geotagged photos. Chapter 7.5 outlines the experimental design and datasets used. Chapter 7.6 illustrates experimental results to demonstrate the effectiveness and efficiency of our framework over traditional approaches. Finally, Chapter 7.7 draws conclusion for this chapter

7.1 Introduction

Travel itinerary recommender systems attempt to assist users with travel planning (Yoon et al. 2012). They provide useful suggestions about popular places to visit and ideas on travel routes and corresponding stay times for users who travel to an unfamiliar destination. An itinerary is a detailed trip plan made up of a travel route associated with stay time information, where the travel route is a sequence of places. The enormous amount of online photo data has become a potential data repository for discovering useful travel information and building travel recommender systems (Beel et al. 2016; Bobadilla et al. 2013), such as location recommendation (Popescu and Grefenstette 2011; Waga, Tabarcea, and Fränti 2012; Yamasaki, Gallagher, and Chen 2013) and travel route recommendation (Okuyama and Yanai 2013).

Existing itinerary recommender systems generate specific itineraries with geographic location information from available geotagged photos. Typically, they generate popular PoIs where many photos have been taken, and map-match PoIs with specific geographic place types to construct a suggested itinerary. However, traditional approaches share a common major drawback: they are mainly based on geographic spatial information when they recommend an itinerary. That is, they do not take any aspatial semantic information into account. In many real-world scenarios, a user wants to visit a certain place type, for instance "zoo", in a given trip. This specific, semantically enhanced request is not considered at all in the traditional approaches. Instead, they accommodate this aspatial semantic information as a post-processing stage for their spatial-information-only recommender systems. Therefore, the traditional recommender systems are not able to accommodate users' semantically enhanced requests to generate meaningful and semantically enhanced itineraries.

The semantic place type request is an important feature in the user's travel planning. For users who are unfamiliar with specific geographic locations and PoIs in a certain destination, they prefer to list some place types (categories) they would like to visit (Gionis et al. 2014). For instance, a user may want to visit "Great Barrier Reef", "rain forest" and "cultural Aboriginal park" in a trip to Cairns. In addition, the user may want to visit "Great Barrier Reef" on a clear day to enjoy swimming with fish and exploring the beauty of reefs, whilst "rain forest" is okay for a rainy day. Any recommended itineraries are expected to contain at least one of each of these requested place types. However, existing itinerary recommender systems fail to consider the user's actual constraints, because they have insufficient semantic information in the recommender system. This study presents an itinerary recommender system that considers users' predefined semantic spatial and aspatial constraints on place types, weather conditions and travel duration time.

A semantic-level itinerary offers detailed journey planning with semantic spatial and aspatial information incorporated. It is more detailed and specific than a general spatial-location-only itinerary. It shows a sequence of movements among different place types with certain weather conditions and certain stay times. This semantic-level itinerary provides users with flexible choices (rain forest in Kuranda or rain forest in Port Douglas) of specific geographic-level routes that satisfy their actual requests (Chen et al. 2011). To the best of our knowledge, there is no other itinerary recommender system that produces a semantic-level itinerary with a set of spatial and aspatial user-specified constraints.

This study develops a semantic-level itinerary recommendation system from geotagged photos. This system considers users' semantic spatial and aspatial requests, and travel duration constraints, and generates semantic-level itineraries that meet the user constraints. The proposed semantic-level itinerary recommender system aims to provide users with higher level advice on place types, weather conditions and stay times. We generate itineraries based on mining semantic trajectory patterns from geotagged photos. We test our algorithm with real datasets from Flickr against traditional spatial-only recommender systems. The experimental results support the effectiveness and efficiency of our system.

7.2 Related work

Online user generated databases, comprised of varying numbers of photos from a wide range of individuals, have become massive, and now offer a potentially useful resource for the tourism-related research community to build collective intelligence and to generate collectively filtered and recommended travel itineraries (De Choudhury et al. 2010). Itinerary recommender systems provide people with advices on travel itinerary with travel routes and time constraints. Kurashima et al. (2013) proposed a system to to generate travel itinerary containing a sequence of locations and the transit time information between two locations. Different from the transit time information, Lu et al. (2010) and Lim et al. (2015) produced the stay time information that tourist could spend at each location of travel route. De Choudhury et al. (2010) provided suggestions for both stay time and transit time for the recommended travel route. However, there are two main differences between previous itinerary recommender systems and our itinerary recommender system. The first difference is that previous systems are unable to deal with semantic-level requirements, such as place types. Users may also want the recommended itineraries to contain some required place types. Symeonidis, Ntempos, and Manolopoulos (2014) play attention to the query with containment of semantic categories, but their system generates landmark-only recommendations. Our system considers users' queries with required containment of place types. The second difference is that our system focuses on the place type layer itinerary recommendations, while previous systems produce geographic object-itinerary recommendations. Travel itinerary in the place type layer is another kind of important suggestion for users' travel planning that provides useful information for users who are unfamiliar with travel destinations and who have no idea what sequence of place types is popular. Our recommender system produce semantic-level itinerary recommendations. Moreover, our system generates additional richer and more meaningful contextual information than any previous itinerary recommender system.

As for the itinerary generation, previous methods use original travel sequences formed from photo data to build people's travel sequences as a probabilistic model (Kurashima et al. 2013) and graph-model (Lu et al. 2010; Quercia, Schifanella, and Aiello 2014; De Choudhury et al. 2010; Lim et al. 2015), and then they reconstruct itineraries from the travel models based on various criteria, such as popularity maximisation. However, the main drawback is that there is no guarantee these itineraries have ever been taken by people. Unlike past methods, this study directly extracts people's "frequent" movement patterns from their historic data, and uses these extracted patterns to provide recommendation to users. Being "frequent" means that the itinerary is regularly occurring in and supported by a certain number of people's travel movements. These frequent movement patterns in people's trajectories guarantee the validity and trustworthiness of recommended itineraries.

7.3 Problem statement

This study aims to build a semantic itinerary recommender system using data mined from massive repositories of online geotagged photos. The problem could be described as follows: given a query including a set of required place types and a travel duration, $query = \{< Types >, Duration\}$, our system replies a list of semantic-level candidate itineraries. An itinerary is a sequence of place types with interval time information, that contains some of the required place types and satisfies the travel duration. Each place type stop is associated with several pieces of additional semantic contextual information: day time, day type and weather condition. An itinerary is expressed as shown below:

$$semantic\ itinerary = Stop_0 \xrightarrow{\alpha_1} Stop_1 \xrightarrow{\alpha_2} \cdots \xrightarrow{\alpha_n} Stop_n.$$

We generate itineraries based on previous people's semantic trajectory patterns, as explained in Chapter 6. These patterns are extracted from historical trajectories formed from geotagged photos. We create raw trajectories from photo data, then we build semantic trajectories that contain application-dependent contextual aspatial semantics, as defined in Definition 3.7. From these semantic trajectories, we find frequent sequences of semantic elements with transit times that are frequent from semantic trajectories. These mobility behaviours are semantic trajectory patterns, as defined in Definition 6.2. Using basic place type semantics, a semantic trajectory pattern is a sequence of visited place type stops with an interval time between two stops.

7.4 Semantic itinerary recommender system and methods

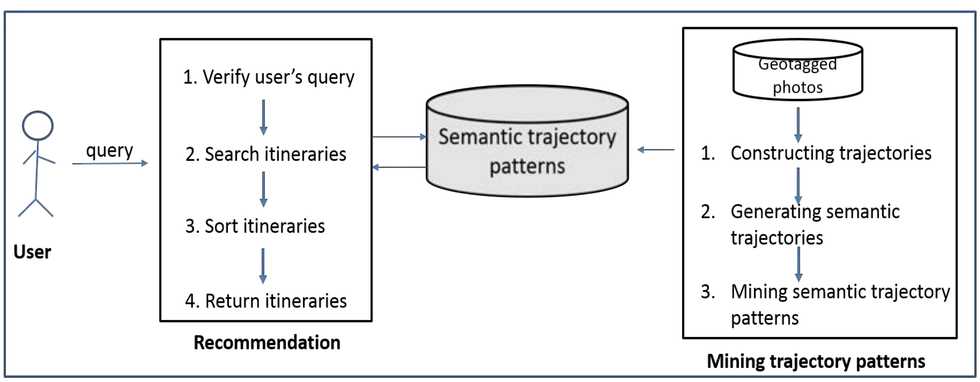


Figure 7.1: Framework of semantic itinerary recommender system.

Figure 7.1 shows our architecture for the semantic itinerary recommender system. The framework includes two main components: offline semantic trajectory pattern mining from geotagged photos; and online itinerary recommendation. In the offline component, we reconstruct people's trajectories from geotagged photos, generate semantic trajectories associated with basic place type semantics and additional contextual semantics, and extract previous users' semantic trajectory patterns. This part contains three steps: (1) constructing basic trajectories from geotagged photos; (2) generating enhanced semantic trajectories using additional spatial and aspatial databases; and (3) mining semantic trajectory patterns. These patterns are used for itinerary recommendations. In the online component, our system verifies the user's query, searches for related candidate itineraries, and sorts and displays them. The remainder of this section describes the methods in detail, step by step. We first present a brief description of methods for extracting semantic trajectory patterns in the offline part. The semantic trajectory pattern mining method is described in detail in Chapter 6.4 in Chapter 6. Then, we illustrate methods for recommending itineraries in the online part.

7.4.1 Mining semantic trajectory pattern

We utilise a semantic trajectory pattern mining algorithm proposed in Chapter 6.4. Given a dataset of semantic trajectories, this algorithm finds all semantic trajectory patterns. These results contain two kinds of semantic trajectory patterns (SemT-patterns). One is a basic semantic pattern, which is a sequence of basic semantics, whilst the other is multi-dimensional semantic pattern, which is a sequence of basic semantics with several additional semantics.

7.4.2 Semantic itinerary recommendation

In the online itinerary recommendation component, for a given user query, the proposed system recommends appropriate semantic itineraries through a series of processes including verifying query, searching and filtering candidate itineraries from our semantic trajectory pattern database, and ranking itineraries. At last, a set of semantic itineraries is extracted and displayed.

The first step is to verify a user's query to check its correctness and validity. A query should include a set of place types and travel duration. Each place type is a word describing a place category, such as hotel, beach, park, etc. We check to ensure that the query contains correct and valid category words. Travel duration indicates the number of days a user will spend travelling. We check to ensure the number is a valid positive integer.

The second step is to search and filter itineraries from the semantic trajectory pattern database. We search every semantic trajectory pattern: a candidate pattern must contain some of required place types, and have a total duration no greater than the user-specified time constraint. Once found, the computed candidate itineraries are stored based on their degree of satisfaction of the user-specified constraints. If there is no pattern containing any of the required types, we choose a set of long patterns that matches the travel duration.

The last step is to sort candidate itineraries. We need to place the candidate itinerary that satisfies the most user constraints at the top. Satisfaction is defined based on the number of required place types a candidate itinerary meets and contains. We sort candidate itineraries based on the number of required types they meet and contain. The final output of the itinerary recommender system is a list of sorted candidate itineraries.

7.5 Experimental setup

We conducted experiments to evaluate the efficiency and effectiveness of the proposed recommender system. The first experiment mainly focuses on the effectiveness of our system. Specifically, we validate the performance of recommended semantic itineraries, that is, to check the number of user's requests the recommended itineraries contain. The second experiment evaluates the informativeness of recommendations. We explain the additional useful information our recommendations can provide, with a comparison to previous traditional methods.

7.5.1 Baseline methods

We chose two previous popular traditional methods – popularity-based method (De Choudhury et al. 2010; Lim 2015) and random-based method (Lim 2015) – as baseline comparative studies. They reconstruct itineraries from people's historical travel routes generated from geotagged photos. A travel route is a sequence of PoIs with transit time information. A PoI is a geographical location that a great number of users visits. In the implementation of baseline methods in this experiment, PoIs are extracted from people's geotagged photos by clustering photo points. We store PoIs and create a PoI database. Then, we generate people's travel routes using these PoIs. For each sequence, we collect transit time between each pair of sequential PoIs, and compute an average transit time database. Two baseline methods use the PoI database and average transit time database to construct itineraries. Each itinerary is a sequence of geographic PoIs with transit time between two PoIs.

- **Random selection method: (Lim 2015)** This method randomly selects a PoI from the PoI database as the next destination, and finds out the average time between these two PoIs from the transit time database as the recommended interval time. This process continues until the total duration time of recommended itinerary reaches the user specified travel duration constraint.

- **Popularity-based method: (De Choudhury et al. 2010; Lim 2015)** This method aims to recommend itineraries with maximum popularity. The popularity of a PoI is the number of people who visit and take a photo at this PoI. This is one of the most popular approaches used in previous recommendation systems, using the basic assumption that a route with maximum popularity will be preferred by users. From the PoI database and the average transit time database, this method finds all potential itineraries that have a total duration not greater than the user's queried travel duration constraint. Then, these potential itineraries are sorted in a descending order based on the total popularity. Finally, a list of top itineraries is recommended to users.

As baseline methods produce specific spatial itineraries without any basic place type semantics information, we conduct an extra post-processing step in order to add place types to the spatial itineraries for the baseline methods. We find a place type for every PoI to transform each spatial only itinerary into a place type semantic-level itinerary. We record statistics about place type in the recommended itineraries for a comparison with our itinerary results.

7.5.2 Evaluation approaches

Here we introduce the metrics used to measure the effectiveness of the itinerary recommender's results. We measure how well the recommendations satisfy the user's query. Specifically, when a user searches for a travel itinerary with a set of customised place types and a travel duration, each system generates multiple candidate itineraries. Each candidate itinerary may potentially be selected by users. To evaluate the effectiveness of recommender systems, we measure the following aspects:

1. Given a user query q, a system generates a list of n candidate itineraries $I = \{i_i, i_2, \ldots, i_n\}$. If a candidate itinerary $i_k \in I$ contains some of user's queried types (user's customised constraints $C = \{c_1, c_2 \ldots, c_m\}$), it is called **positive**. For n candidates, there are $n+1$ situations from where no candidate contains any queried type (0 positive) to where all candidates contain some queried types (n positive). Higher positive values mean better performance. That is, it is formally defined as: $|I'|$ for $I' \subset I$ where $i' \in I'$, $\exists\ c \in C$ such that i' contains c. $|I'|$ varies between 0 and n. Given a set of itinerary recommender systems

$R = \{R_1, R_2, \ldots, R_l\}$, R_i is said to be better than R_j in performance, for R_i and $R_j \in R$ iff $|I'_{R_i}| > |I'_{R_j}|$.

2. Let us assume there are n candidature itineraries I for a given query q. For a candidate itinerary $i_k \in I$, let us assume that it has x unique types (number of unique values of place type in a candidate itinerary), and contains y common types (number of place types both user query and a candidate itinerary contain). The **percentage** of common types in the candidate i_k is $per_k = y/x$. The higher percentage, the better performance. For all n candidates, we calculate the average percentage for I, that is $aver\_per = (per_1 + per_2 + \cdots + per_n)/n$. For a given set of m testing routes, we record the average percentage for each route. Then we count the number of testing routes in each average percentage range: that is the metric we use to measure the degree of concentration. The higher the number of routes with higher average percentage range, the better the quality of candidate itineraries.

3. The efficiency of each system, that is, how fast the system generates a list of recommendations.

We also compare the degree of **additional meaningful semantics information** in the itinerary recommendation results. This comparison measures information richness. Specifically, the comparison focuses on what information our approach and the baseline systems recommend. We conduct these evaluation experiments by simulating user queries for test travel routes.

7.5.3 Datasets for experiments

Training dataset

The dataset used is the same as the one in Figure 4.6 in Chapter 4.5. The statistics about the training dataset are as follows.

- For our recommender system, we generate 1,404 trajectories, find 49 semantic RoIs with 12 place types, and extract 65 basic semantic trajectory patterns (each one has a set of multi-dimensional semantic patterns). Most of the patterns have a length of 2 to 4.

- For two baseline methods, we apply the DBSCAN clustering method (Ester et al. 1996) to photo data and validate final PoIs based on the number of users. Finally, we obtain 46 PoIs with 17 unique place types.

Testing dataset

We used simulated travel routes as our testing dataset, constructed from a combination of unique place types from our method and unique place types from the baseline methods (12 and 17, respectively). The final dataset for the simulation included 22 unique place types. We then generated travel routes by randomly selecting place types from the place type set and randomly generating an interval time between two types. We kept the total duration of each route at 16 days or less, since the popularity-based baseline method cost significantly more time to construct itineraries when a query travel duration constraint is more than 16 days. Finally, we created a testing dataset with the following statistics:

- 300 travel routes;

- containing total 22 unique place types;

- nearly half the routes are 2-length, while containing 2 unique place types;

- number of common types that training dataset also contains (our system: 12, baseline 17);

- higher diversity than real Flickr travel route testing dataset.

7.6 Results and discussions

7.6.1 Effectiveness of recommendation results

We used testing route as simulated user query input. A query includes a set of place types and a travel duration constraint. In evaluation experiments for the effectiveness of itinerary recommendations, we chose the top five candidate itineraries as final recommendations, because the top five candidates provide a temperate diversity and number of recommendations for all three methods. When using more than five candidates, the redundancy of itineraries increases.

The first experimental result is how many candidate itineraries contain the user-queried place types in the recommendation results. If a candidate itinerary contains any user-required type, it is noted as positive. When all five candidate itineraries contain any of the required types, it becomes 5-positive. Figure 7.2 shows the performance of itinerary recommendation results from three systems. For 300 testing queries, the figure presents the distribution of queries for each positive situation.

Obviously, the random selection method generates the worst quality of recommendations. It is not able to generate 5-positive for any testing route, whilst the popularity-based method and our method can produce 5-positive for more than half the testing routes. The number of routes for which the popularity-based method generates a high positive is a little more than that of our method. The reason is that the extracted patterns from our system are mostly in short length. As a result, our recommended itineraries are short, whereas the popularity-based method constructs itineraries by finding PoIs and connecting PoIs as long as possible until the total time duration reaches the user-defined duration constraint. A long itinerary has a higher probability of containing the user-queried place types. However, the difference is not significant. A positive value indicates the system's ability to generate valuable and useful candidates: the higher the positive value, the higher the performance. Both our method and the popularity-based method are able to produce high positive itineraries.

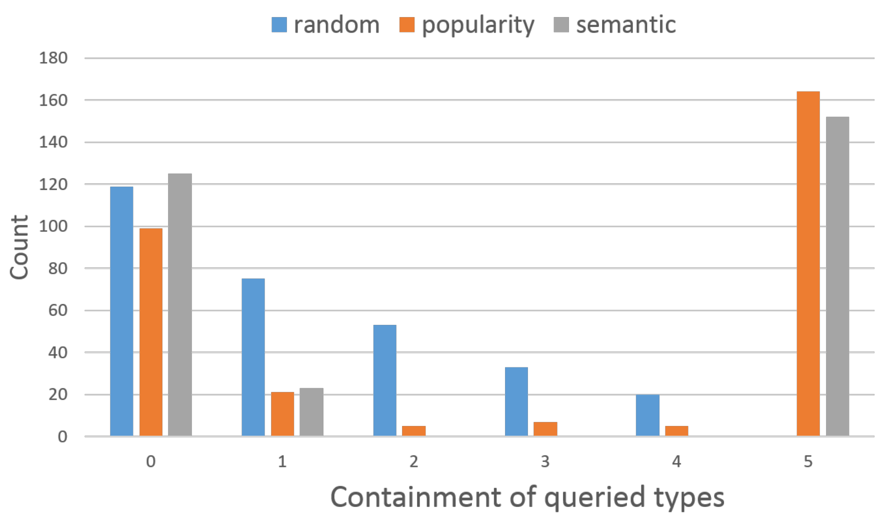


Figure 7.2: Number of routes for each different containment for user-queried types.

The second result is the average percentage of common place types in the recommendations. A candidate containing a high percentage of user-queried place types exhibits high concentration on queried types and thus it is useful to users. The average percentage indicates the quality and degree of concentration of whole recommendation results. Figure 7.3 shows the distribution of routes with different average percentage ranges. Our system produces better itinerary recommendations since it is able to generate higher concentration recommendation results for more than half the queries. In contrast, the two baseline methods can only generate low concentration itineraries.

The third experiment is to measure the running time of the system. The total duration time for testing the route is used as a queried travel duration constraint. Figure 7.4 shows running time requirements for the three systems. As shown in Figure 7.4, the popularity-based method exhibits the

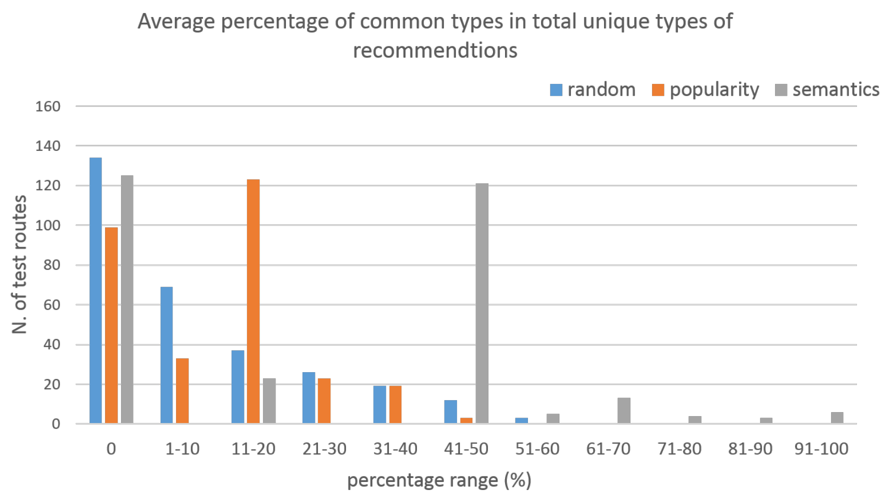


Figure 7.3: Average percentage ($N.comtypes/N.uniqueTypes$).

worst efficiency, that is not scalable to large datasets. For larger travel duration queries, the popularity-based method costs much more time to generate recommendations than the other two. The random-based method and our method cost much less time, and are scalable to large datasets. However, note that our semantic method requires consistently less time for all travel duration queries and exhibits the best efficiency performance.

Table 7.1: Performance comparison of recommendation results.

	Positiveness	**Concentration**	**Efficiency**
Random-based	Bad	Bad	Good
Popularity-based	Good	Bad	Bad
Semantic-based	Good	Good	Good

In summary, Table 7.1 presents the performance comparison of the three

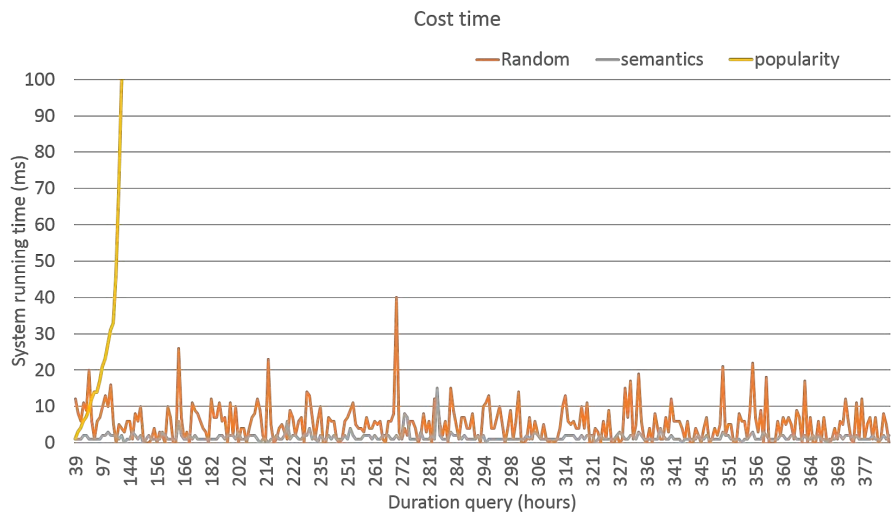


Figure 7.4: Specific running time of random-based and semantic-based recommendation systems.

itinerary recommender systems under study. The random-based method is fairly efficient, costing less time to generate candidate itineraries, but ineffective: the quality of candidates is poor since it contains fewer user required place types. The popularity-based method, on the other hand, is able to generate good positive candidate itineraries containing the user-required place types. However, this method costs much more time to generate candidate itineraries, and the percentage of required types in the candidates is small. Our proposed semantic-based approach exhibits effective performance in positiveness and concentration, as well as efficiency in time costs.

7.6.2 Comparative results of information

Higher layer semantic-level itinerary *vs.* Lower layer geographic-level itinerary

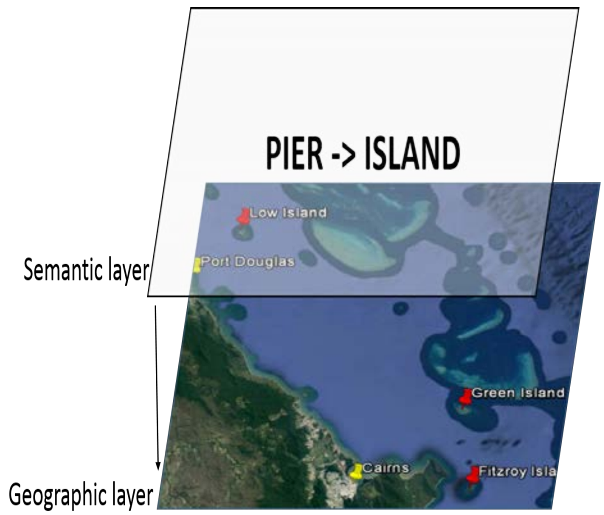
Our system produces semantic-level itineraries including basic semantic itineraries, and itineraries with additional semantics. The baseline methods produce geographic itineraries with specific geographic places. Table 7.2 lists one example of typical itinerary recommendations from the three systems, for query: set of types = ISL, travel duration = 15 days. This sample selects a route from PoI Cairns pier to PoI Green Island in Cairns area.

Table 7.2: Example of itinerary recommendations.

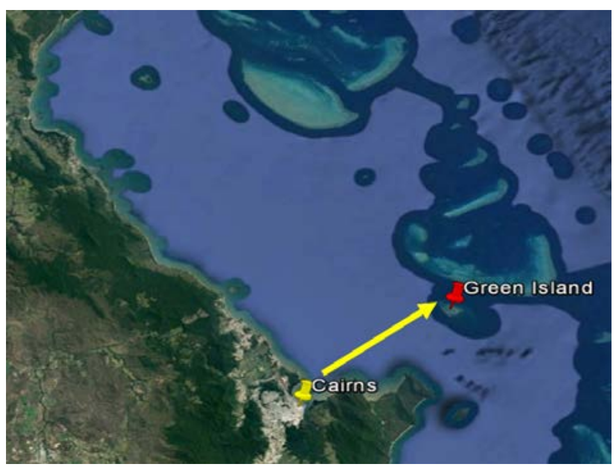
Method	**Itinerary recommendation**
Our system	route: PIER – [0 to 2 days] – ISL — additional info — ∗ PIER[weekday][Cairns] – [0 to 2 days] – ISL[weekday][Cairns] ∗ PIER[Clear][Cairns] – [0 to 2 days] – ISL[Clear][Cairns]
Random-based & Popularity-based	route : P45 – 4days – P31

One main difference between itineraries of our system and itineraries of the baseline systems is in the layer of the recommended itinerary. For the sample shown in Table 7.2, the baseline systems recommend a user a specific geographic layer travel route which is from Cairns pier to Green Island. It is one of the popular travel routes for visiting the Great Barrier Reef. Our system recommends a user a higher layer semantic route which is from pier to the island. Figure 7.5 shows the semantic-level itinerary and geographic-level itinerary.

The higher semantic-level itinerary is better than the lower specific geographic route. The semantic-level itinerary enables users to freely choose various optional geographic routes according to their preferences and actual environments, such as the geographic location where they are, whilst the baseline systems recommend specific routes, and users receive these routes only but are unable to obtain other optional routes. For the semantic-level



(a) Semantic itinerary



(b) Geographic itinerary

Figure 7.5: Visualisation of semantic-level itinerary and geographic-level itinerary.

route, pier to island, there are many piers and islands. We find several specific piers and islands in Cairns shown in Figure 7.5(a). Users can choose a route from Port Douglas pier to Low Island, or a route from Cairns pier to either Green Island or Fitzroy Island. All three of these are popular travel routes for visiting the Great Barrier Reef. In comparison, the baseline systems recommend a specific geographic route, which limits the selection of other popular routes. The semantic-level itinerary provides a higher layer perspective and guidance to users on how to travel, and also they can select a specific geographic route of their choice.

Other additional semantics information

Table 7.3: Information from itinerary recommendation results.

	Basic information	**Other information**
Itineraries of baseline systems	Interval time for a route of specific PoIs	N/A
Itineraries of our system	A route with place types; Interval time	Recommended contextual information: temporal (day time, day type), weather condition

Our system recommends more useful travel context information than the baseline systems. Table 7.3 lists specific information the recommended itinerary results can provide. The baseline systems supply basic information on travel routes among specific geographic PoIs, including interval time. Our system can provide basic information on frequent travel routes among place types and interval time. Moreover, our results can provide additional and useful information about travel context environment including temporal information about day time and day type, and weather condition. In particular, our system recommends itineraries based on previous people's frequent trajectory patterns mined from their historic trajectories. The trajectory patterns are associated with additional semantic information. The additional information shows the frequent contextual environment in which frequent trajectories occur. As a result, this information in the itinerary recommendation results provides users with useful advice on environmental contexts in which people travel the destination. Moreover, this useful information can be potentially further used for context-aware recommendation services that recommend itineraries to users based on their contextual environment like day time and weather condition.

7.7 Conclusion

In this study, we present an itinerary recommender system using online geotagged photos. Our system allows users to customise a set of place types and an overall travel duration in the query. The system generates itinerary recommendations based on previous people's semantic trajectory patterns extracted from their historical photo data. Experimental results show that our system is able to produce itinerary recommendations that satisfy user's predefined requirements. Our system recommends semantic-level itineraries to users that show more place type layer route suggestions compared to specific geographic-level ones. The higher layer routes provide users with more flexible selections of potential spatial routes. Moreover, our system generates itinerary recommendations with additional and useful environmental semantics information.

Chapter 8

Conclusions

This chapter summarises the study of this thesis. Then, some potential future work is presented.

8.1 Summary of study

Online user-generated and shared photo data is already a massive resource, and is growing fast. These photos, together with geographic information, timestamp and other tag annotations, are a rich potential data repository for extracting people's movement behaviours. This research aimed to extract semantically enhanced trajectory behavioural patterns from geotagged photos and then to build an itinerary recommender system using the extracted patterns. These semantic patterns are useful for understanding people's mobility. This thesis has proposed a systematic framework tool to extract dynamic movement patterns. The framework contains four main functions, for finding three kinds of patterns and building an itinerary recommender system. In particular, this project proposed four trajectory data mining approaches for the four tasks. Overall, this study has proposed a semantic RoI mining method for detecting stops from raw trajectories. The stops are used to build semantic trajectories. Then, a semantic sequential pattern mining method has also been proposed to find the frequent semantic sequential patterns from semantic trajectories. In addition, a semantic trajectory clustering method has been introduced to discover the semantic common patterns in the semantic trajectories. A semantic trajectory pattern mining method has been presented to extract semantic trajectory patterns from semantic trajectories. Finally, a semantic-aware itinerary recommender system has been built using the semantic trajectory patterns to provide users with suggestions of travel itineraries with routes and typical transition time information.

This research analyses the trajectory data of geotagged photos and explores people's dynamic mobility behaviours and movement patterns. The previous traditional spatial geometric-feature-only approaches consider only spatial data, while disregarding aspatial contextual semantics information. Aspatial semantics information is as important as spatial information in spatial data mining and geographic data analysis (Miller and Han 2009). Our research focused on the analysis of semantically enriched trajectories using spatial, temporal and aspatial semantics features. First, people's spatio-temporal trajectory data are reconstructed from geotagged photos, then they are enriched with multiple background geographical information and environmental data. This research proposed a semantic RoI mining algorithm to detect stops from raw trajectories that are then used to build stop-based semantic trajectories. The stop is a geo-object with a place type annotation. The proposed approach is able to find RoIs with fine and accurate place type semantics. The final semantic trajectory is a sequence of stops with basic geographical place type annotations and a set of environmental context data including city name, day type, day time and weather condition.

From people's semantic trajectories, this study first found frequent sequential patterns. A sequential pattern shows a frequent visit sequence of stops that indicates a set of frequently visited stops with certain time order behaviour. The frequent sequential patterns reveal people's semantic-level movement behaviours. This study proposed a semantic sequential pattern mining method to find the semantic sequential patterns. The method found numerous people's sequential trajectory patterns on a semantic level. The semantic patterns provide more meaningful knowledge and understanding of human mobility behaviours than those having previously been available, and which are valuable to the tourism industry. Moreover, the proposed method can deal with multi-dimensional semantic trajectories. That is, by adding multiple semantics to trajectories, the method generates patterns with various combinations of dimensions. From the multi-dimensional semantic trajectories, the proposed method found novel knowledge about semantic-level trajectory patterns with several contextual semantics.

As the second aim, this study extracted semantic common patterns from semantic trajectories. A common pattern refers to a cluster of similar semantic trajectories. This study introduced a semantic trajectory clustering method for finding common patterns in the trajectory dataset. The proposed method is an extension of the OPTICS clustering algorithm to handle semantically enriched trajectories with a new similarity measure method for multi-dimensional trajectories. The experimental results reveal that the proposed approach has the ability to produce much more detailed spatial and aspatial patterns that traditional geographic-feature-only approaches fail to identify. These semantic common patterns are useful for understanding people's mobility behaviours on the semantic level.

Another aim of this work was discovering the frequent semantic trajectory patterns in semantic trajectories. A trajectory pattern indicates the frequent visit sequence of stops with frequent transition time data between stops. It shows frequent movement sequences and time relations between stops in each sequence. This study proposed a semantic trajectory pattern mining method to find the semantic trajectory patterns from semantic trajectories. The experimental results show that the proposed method can find many interesting place type semantic-level trajectory patterns of people. Experimental results also show that the proposed method is able to find trajectory patterns with various additional semantics, while previous methods are not able to find understandable semantic patterns. These semantic trajectory patterns provide richer semantic information about people's mobility behaviours than those having previously been available. In addition, the proposed method generates more potential, detailed semantic trajectory patterns that more number of patterns provide benefits to have less chance to miss positive patterns.

Last, this study built a semantic-aware itinerary recommender system based on semantic trajectory pattern mining from geotagged photos. The trajectory patterns show real-world people's movement behaviour, including frequent visit sequences and time information which are good indications to travel itineraries. The proposed recommender system receives users' queries including a set of preferred place types and travel duration and returns appropriate travel itineraries with route and transition time to users. Experimental results show that the proposed system, considering semantics queries, is able to generate highly concentrated itineraries in which most of the stops in the itinerary match user-queried place types. In terms of efficiency, the proposed system costs the least time of the three systems tested to generate candidature itineraries. Moreover, the itineraries recommended by the proposed system contain rich, meaningful and understandable information about travel routes and movement environmental contexts including time information and weather condition information that are useful to users.

8.2 Future work

There are several areas where future work could be undertaken, as listed below:

1. An immediate area for future study is to test our approaches with more social media datasets, including GPS-logged and sensor-tagged datasets. In this research, the experiments used Flickr photos to evaluate the proposed approaches. In the future, more datasets will be

tested. A comprehensive set of experiments will further support the usefulness and richness of our approaches.

2. Another future work is to extend the current framework to include more contextual semantics information. This research analysed multi-dimensional semantic trajectories using five pieces of semantics information: type of place, city, day type, day time and weather. In the future, more semantic information, including social, economic and environmental information could be added and considered. This will require a flexible framework to easily add additional semantic annotations into the framework. Note that, our proposed framework is flexible and designed to accommodate more dimensions with ease.

3. Post-processing of detected semantic patterns is another interesting target, to find some positive associations and cause-effect patterns. Trajectories indicate movement that shows the sequential relations of visited stops. This research extracted three types of semantic movement patterns. Further work could be conducted on patterns to discover the associations and cause-effect sequential patterns of visited stops in trajectories.

4. Pre-processing of time information from original geotagged photos is another future work. The taken time information from photos could be in different time zones that is not the correct local time in the study area. This time information directly affects semantic patterns in this study, and an appropriate pre-processing approach could be developed to fix this local time in order to produce more meaningful semantic patterns.

5. Applicability is the next aim in future work. This study developed a framework to extract semantic mobility patterns. We plan to conduct a case study to observe before and after scenarios with the tourism in-

dustry to observe what benefits these semantically enhanced movement patterns provide in real life.

6. For the itinerary recommender system, a future study is needed on further processing of recommended itineraries to produce more diverse semantic-level routes with suggestions of specific geographic places and locations. This research built an itinerary recommender system using the extracted semantic trajectory patterns. The recommended itineraries are about travel routes between place type that provide advice on higher place type layer movement and users can choose specific places according to their actual location. In future, more appropriate specific places could be extracted and suggested in the recommendation results. Moreover, post-processing of recommended itineraries could be undertaken to generate diverse and longer itineraries.

7. Visualisation is another main improvement for future work. This study analysed semantic trajectories and extracted semantics patterns, adding extra semantics information. A better visualisation of the semantic trajectory, semantic patterns and additional semantic information will be investigated to provide more effective presentation of the information, and easier understanding of movement behaviour for end users. Moreover, for the itinerary recommender system, an easy-to-use interface needs to be built, and should include the visualisation of recommended travel itineraries on a map.

Bibliography

Adomavicius, Gediminas and Alexander Tuzhilin (2015). "Context-aware recommender systems". In: *Recommender systems handbook*. Springer, pp. 191–226.

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau (2011). "Sentiment analysis of twitter data". In: *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, pp. 30–38.

Agrawal, Rakesh and Ramakrishnan Srikant (1995). "Mining sequential patterns". In: *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*. IEEE, pp. 3–14.

Ahern, Shane, Mor Naaman, Rahul Nair, and Jeannie Hui-I Yang (2007). "World explorer: visualizing aggregate data from unstructured text in georeferenced collections". In: *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, pp. 1–10.

Alvares, Luis Otavio, Vania Bogorny, Bart Kuijpers, Jose Antonio Fernandes de Macedo, Bart Moelans, and Alejandro Vaisman (2007a). "A Model for Enriching Trajectories with Semantic Geographical Information". In: *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*. GIS '07. ACM, pp. 22–30.

Alvares, Luis Otavio, Vania Bogorny, Bart Kuijpers, JAF de Macelo, B Moelans, and Andrey Tietbohl Palma (2007b). *Towards semantic trajectory* *knowledge discovery*. Technical Report. Hasselt University, Hasselt, Belgium.

Ankerst, Mihael, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander (1999). "OPTICS: Ordering Points to Identify the Clustering Structure". In: *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. SIGMOD '99. ACM, pp. 49–60.

Arase, Yuki, Xing Xie, Takahiro Hara, and Shojiro Nishio (2010). "Mining people's trips from large scale geo-tagged photos". In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM, pp. 133–142.

Barbier, Geoffrey and Huan Liu (2011). "Data mining in social media". In: *Social network data analytics*, pp. 327–352.

Beel, Joeran, Bela Gipp, Stefan Langer, and Corinna Breitinger (2016). "Research-paper recommender systems: a literature survey". In: *International Journal on Digital Libraries* Vol. 17. No. 4, pp. 305–338.

Bermingham, Luke and Ickjai Lee (2014). "Spatio-temporal sequential pattern mining for tourism sciences". In: *Procedia Computer Science* Vol. 29, pp. 379–389.

Bermingham, Luke and Ickjai Lee (2015). "A general methodology for n-dimensional trajectory clustering". In: *Expert Systems with Applications* Vol. 42. No. 21, pp. 7573–7581.

Berndt, Donald J and James Clifford (1994). "Using Dynamic Time Warping to Find Patterns in Time Series." In: *KDD workshop*. Vol. 10. No. 16, pp. 359–370.

Beyer, Kevin and Raghu Ramakrishnan (1999). "Bottom-up computation of sparse and iceberg cube". In: *ACM SIGMOD Record*. Vol. 28. No. 2. ACM, pp. 359–370.

Bhargava, Preeti, Thomas Phan, Jiayu Zhou, and Juhan Lee (2015). "Who, what, when, and where: Multi-dimensional collaborative recommendations using tensor factorization on sparse user-generated data". In: *Pro-*

ceedings of the 24th International Conference on World Wide Web. ACM, pp. 130–140.

Bobadilla, Jesús, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez (2013). "Recommender systems survey". In: *Knowledge-based systems* Vol. 46, pp. 109–132.

Cai, Guochen, Chihiro Hio, Luke Bermingham, Kyungmi Lee, and Ickjai Lee (2014). "Sequential Pattern Mining of Geo-tagged Photos with an Arbitrary Regions-of-interest Detection Method". In: *Expert Systems with Applications* Vol. 41. No. 7, pp. 3514–3526.

Cao, Huiping, Nikos Mamoulis, and David W. Cheung (2005). "Mining Frequent Spatio-Temporal Sequential Patterns". In: *Proceedings of the Fifth IEEE International Conference on Data Mining*. ICDM '05. Washington, DC, USA: IEEE Computer Society, pp. 82–89.

Cao, Huiping, Nikos Mamoulis, and David W. Cheung (2007). "Discovery of Periodic Patterns in Spatiotemporal Sequences". In: *IEEE Trans. on Knowl. and Data Eng.* Vol. 19. No. 4, pp. 453–467.

Cao, Xin, Gao Cong, and Christian S. Jensen (2010). "Mining Significant Semantic Locations from GPS Data". In: *Proc. VLDB Endow.* Vol. 3. No. 1-2, pp. 1009–1020. ISSN: 2150-8097.

Chakri, Sana, Said Raghay, and Salah El Hadaj (2017). "Semantic Trajectory Knowledge Discovery: A Promising Way to Extract Meaningful Patterns from Spatiotemporal Data". In: *International Journal of Software Engineering and Knowledge Engineering* Vol. 27. No. 03, pp. 399–421.

Chen, Chien-Cheng and Meng-Fen Chiang (2016). "Trajectory pattern mining: Exploring semantic and time information". In: *Technologies and Applications of Artificial Intelligence (TAAI), 2016 Conference on*. IEEE, pp. 130–137.

Chen, Chien-Cheng, Chia-Hsiang Kuo, and Wen-Chih Peng (2015). "Mining spatial-temporal semantic trajectory patterns from raw trajectories". In:

Data Mining Workshop (ICDMW), 2015 IEEE International Conference on. IEEE, pp. 1019–1024.

Chen, Haiquan, Wei-Shinn Ku, Min-Te Sun, and Roger Zimmermann (2011). "The partial sequenced route query with traveling rules in road networks". In: *GeoInformatica* Vol. 15. No. 3, pp. 541–569.

Chen, Lei and R Ng (2004). "On the Marriage of Lp-norms and Edit Distance". In: *Proceedings of the Thirtieth International Conference on Very Large Data Bases*. Vol. 30. VLDB '04. VLDB Endowment, pp. 792–803.

Chen, Lei, M Tamer Özsu, and Vincent Oria (2005). "Robust and fast similarity search for moving object trajectories". In: *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, pp. 491–502.

Chen, Xin, Mihaela Vorvoreanu, and Krishna Madhavan (2014). "Mining social media data for understanding students' learning experiences". In: *IEEE Transactions on Learning Technologies* Vol. 7. No. 3, pp. 246–259.

Chen, Yan-Ying, An-Jung Cheng, and Winston H Hsu (2013). "Travel recommendation by mining people attributes and travel group types from community-contributed photos". In: *IEEE Transactions on Multimedia* Vol. 15. No. 6, pp. 1283–1295.

Chen, Yen-Liang, Mei-Ching Chiang, and Ming-Tat Ko (2003). "Discovering time-interval sequential patterns in sequence databases". In: *Expert Systems with Applications* Vol. 25. No. 3, pp. 343–354.

Crandall, David J, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg (2009). "Mapping the world's photos". In: *Proceedings of the 18th international conference on World wide web*. ACM, pp. 761–770.

Crandall, David J, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg (2010). "Inferring social ties from geographic coincidences". In: *Proceedings of the National Academy of Sciences* Vol. 107. No. 52, pp. 22436–22441.

Cranshaw, Justin, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh (2010). "Bridging the gap between physical location and online social networks". In: *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, pp. 119–128.

De Choudhury, Munmun, Moran Feldman, Sihem Amer-Yahia, Nadav Golbandi, Ronny Lempel, and Cong Yu (2010). "Automatic construction of travel itineraries using social breadcrumbs". In: *Proceedings of the 21st ACM conference on Hypertext and hypermedia*. ACM, pp. 35–44.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu (1996). "A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. AAAI Press, pp. 226–231.

Estivill-Castro, Vladimir and Ickjai Lee (2000). "Amoeba: Hierarchical clustering based on spatial proximity using delaunay diagram". In: *Proceedings of the 9th International Symposium on Spatial Data Handling. Beijing, China*, pp. 7–26.

Furtado, Andre Salvaro, Despina Kopanaki, Luis Otavio Alvares, and Vania Bogorny (2016). "Multidimensional Similarity Measuring for Semantic Trajectories". In: *Transactions in GIS* Vol. 20. No. 2, pp. 280–298.

Gaffney, Scott and Padhraic Smyth (1999). "Trajectory clustering with mixtures of regression models". In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '99. ACM, pp. 63–72.

Gao, Song, Jiue-An Yang, Bo Yan, Yingjie Hu, Krzysztof Janowicz, and Grant McKenzie (2014). "Detecting origin-destination mobility flows from geotagged Tweets in greater Los Angeles area". In: *Eighth International Conference on Geographic Information Science (GIScience'14)*.

Giannotti, Fosca, Mirco Nanni, and Dino Pedreschi (2006). "Efficient Mining of Temporally Annotated Sequences". In: *SDM*. SIAM, pp. 348–359.

Giannotti, Fosca, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi (2007). "Trajectory pattern mining". In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 330–339.

Gionis, Aristides, Theodoros Lappas, Konstantinos Pelechrinis, and Evimaria Terzi (2014). "Customized tour recommendations in urban areas". In: *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, pp. 313–322.

Girardin, Fabien, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat (2008a). "Digital footprinting: Uncovering tourists with user-generated content". In: *IEEE Pervasive computing* Vol. 7. No. 4, pp. 36–43.

Girardin, Fabien, Filippo Dal Fiore, Carlo Ratti, and Josep Blat (2008b). "Leveraging explicitly disclosed location information to understand tourist dynamics: a case study". In: *Journal of Location Based Services* Vol. 2. No. 1, pp. 41–56.

Goodchild, Michael F (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* Vol. 69. No. 4, pp. 211–221.

Goodchild, Michael F. (2008). "Imprecision and Spatial Uncertainty". In: *Encyclopedia of GIS*. Boston, MA: Springer US, pp. 480–483.

Goodchild, Michael F and J Alan Glennon (2010). "Crowdsourcing geographic information for disaster response: a research frontier". In: *International Journal of Digital Earth* Vol. 3. No. 3, pp. 231–241.

Gudmundsson, Joachim, Marc van Kreveld, and Bettina Speckmann (2004). "Efficient detection of motion patterns in spatio-temporal data sets". In: *Proceedings of the 12th annual ACM international workshop on Geographic information systems*. ACM, pp. 250–257.

Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim (1998). "CURE: an efficient clustering algorithm for large databases". In: *ACM Sigmod Record*. Vol. 27. No. 2. ACM, pp. 73–84.

Han, Jiawei, Jian Pei, and Micheline Kamber (2011). *Data mining: concepts and techniques*. 3rd. Morgan Kaufmann Publishers Inc.

Han, Jiawei, Jian Pei, Behzad Mortazavi-Asl, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu (2000). “FreeSpan: frequent pattern-projected sequential pattern mining”. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 355–359.

Han, Jiawei, Jian Pei, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and MC Hsu (2001). “Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth”. In: *proceedings of the 17th international conference on data engineering*, pp. 215–224.

Haworth, Billy and Eleanor Bruce (2015). “A review of volunteered geographic information for disaster management”. In: *Geography Compass* Vol. 9. No. 5, pp. 237–250.

Hio, Chihiro, Luke Bermingham, Guochen Cai, Kyungmi Lee, and Ickjai Lee (2013). “A Hybrid Grid-based Method for Mining Arbitrary Regions-of-Interest from Trajectories”. In: *Proceedings of Workshop on Machine Learning for Sensory Data Analysis*. MLSDA ’13. ACM, pp. 5–12.

Hirate, Yu and Hayato Yamana (2006). “Sequential pattern mining with time intervals”. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 775–779.

Hotho, Andreas, Robert Jäschke, Christoph Schmitz, and Gerd Stumme (2006). “Information retrieval in folksonomies: Search and ranking”. In: *European Semantic Web conference*. Springer, pp. 411–426.

Hu, Yingjie, Song Gao, Krzysztof Janowicz, Bailang Yu, Wenwen Li, and Sathya Prasad (2015). “Extracting and understanding urban areas of interest using geotagged photos”. In: *Computers, Environment and Urban Systems* Vol. 54, pp. 240–254.

Huang, Yan, Liqin Zhang, and Pusheng Zhang (2008). “A framework for mining sequential patterns from spatio-temporal event data sets”. In:

IEEE Transactions on Knowledge and data engineering Vol. 20. No. 4, pp. 433–448.

Jäschke, Robert, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme (2007). "Tag recommendations in folksonomies". In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, pp. 506–514.

Jeung, Hoyoung, Man Lung Yiu, Xiaofang Zhou, Christian S Jensen, and Heng Tao Shen (2008). "Discovery of convoys in trajectory databases". In: *Proceedings of the VLDB Endowment* Vol. 1. No. 1, pp. 1068–1080.

Jin, Xin, C Lin, Jiebo Luo, and Jiawei Han (2011). "A data mining-based spam detection system for social media networks". In: *Proceedings of the VLDB Endowment* Vol. 4. No. 12, pp. 1458–1461.

Kang, Ju-Young and Hwan-Seung Yong (2010). "Mining spatio-temporal patterns in trajectory data". In: *Journal of Information Processing Systems* Vol. 6. No. 4, pp. 521–536.

Kaplan, Andreas M and Michael Haenlein (2010). "Users of the world, unite! The challenges and opportunities of Social Media". In: *Business horizons* Vol. 53. No. 1, pp. 59–68.

Kempe, David, Jon Kleinberg, and Éva Tardos (2003). "Maximizing the spread of influence through a social network". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 137–146.

Kennedy, Lyndon, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury (2007). "How flickr helps us make sense of the world: context and content in community-contributed media collections". In: *Proceedings of the 15th ACM international conference on Multimedia*. ACM, pp. 631–640.

Kisilevich, Slava, Daniel Keim, and Lior Rokach (2010). *A novel approach to mining travel sequences using collections of geotagged photos*. Springer Berlin Heidelberg, pp. 163–186.

Kisilevich, Slava, Florian Mansmann, and Daniel Keim (2010). "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos". In: *Proceedings of the 1st international conference and exhibition on computing for geospatial research & application*. ACM, pp. 38–41.

Kisilevich, Slava, Daniel Keim, Natalia Andrienko, and Gennady Andrienko (2013). "Towards acquisition of semantics of places and events by multi-perspective analysis of geotagged photo collections". In: *Geospatial visualisation*. Springer, pp. 211–233.

Kurashima, Takeshi, Tomoharu Iwata, Go Irie, and Ko Fujimura (2013). "Travel route recommendation using geotagged photos". In: *Knowledge and information systems* Vol. 37. No. 1, pp. 37–60.

Lee, Anthony JT, Yi-An Chen, and Weng-Chong Ip (2009). "Mining frequent trajectory patterns in spatial–temporal databases". In: *Information Sciences* Vol. 179. No. 13, pp. 2218–2231.

Lee, Ickjai, Guochen Cai, and Kyungmi Lee (2013). "Points-of-interest mining from people's photo-taking behavior". In: *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. IEEE, pp. 3129–3136.

Lee, Ickjai, Guochen Cai, and Kyungmi Lee (2014). "Exploration of geotagged photos through data mining approaches". In: *Expert Systems with Applications* Vol. 41. No. 2, pp. 397–405.

Lee, Jae-Gil, Jiawei Han, and Kyu-Young Whang (2007). "Trajectory clustering: a partition-and-group framework". In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, pp. 593–604.

Li, Zhenhui (2014). "Spatiotemporal pattern mining: algorithms and applications". In: *Frequent Pattern Mining*. Springer, pp. 283–306.

Li, Zhenhui, Bolin Ding, Jiawei Han, and Roland Kays (2010). "Swarm: Mining relaxed temporal moving object clusters". In: *Proceedings of the VLDB Endowment* Vol. 3. No. 1-2, pp. 723–734.

Liang, Po-Wei and Bi-Ru Dai (2013). “Opinion mining on social media data”. In: *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*. Vol. 2. IEEE, pp. 91–96.

Lim, Kwan Hui (2015). “Recommending tours and places-of-interest based on user interests from geo-tagged photos”. In: *Proceedings of the 2015 ACM SIGMOD on PhD Symposium*. ACM, pp. 33–38.

Lim, Kwan Hui, Jeffrey Chan, Christopher Leckie, and Shanika Karunasekera (2015). “Personalized Tour Recommendation Based on User Interests and Points of Interest Visit Durations”. In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. AAAI Press, pp. 1778–1784.

Lin, Dekang (1998). “An information-theoretic definition of similarity.” In: *ICML*. Vol. 98, pp. 296–304.

Lu, Xin, Changhu Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang (2010). “Photo2trip: generating travel routes from geo-tagged photos for trip planning”. In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM, pp. 143–152.

Majid, Abdul, Ling Chen, Gencai Chen, Hamid Turab Mirza, Ibrar Hussain, and John Woodward (2013). “A context-aware personalized travel recommendation system based on geotagged social media data mining”. In: *International Journal of Geographical Information Science* Vol. 27. No. 4, pp. 662–684.

Majid, Abdul, Ling Chen, Hamid Turab Mirza, Ibrar Hussain, and Gencai Chen (2015). “A system for mining interesting tourist locations and travel sequences from public geo-tagged photos”. In: *Data & Knowledge Engineering* Vol. 95, pp. 66–86.

Mathes, Adam (2004). *Folksonomies-cooperative classification and communication through shared metadata*. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`. Accessed: 2017-02-10.

McDougall, Kevin (2011). “Using volunteered information to map the Queensland floods”. In: *Proceedings of the 2011 Surveying and Spatial Sciences Conference: Innovation in Action: Working Smarter (SSSC 2011)*. Surveying and Spatial Sciences Institute, pp. 13–23.

Memon, Imran, Ling Chen, Abdul Majid, Mingqi Lv, Ibrar Hussain, and Gencai Chen (2015). “Travel recommendation using geo-tagged photos in social media for tourist”. In: *Wireless Personal Communications* Vol. 80. No. 4, pp. 1347–1362.

Michel, Franck (2017). *How many public photos are uploaded to Flickr every day, month, year?* `https://www.flickr.com/photos/franckmichel/6855169886/in/dateposted/`. Accessed: 2017-04-10.

Miller, Harvey J and Jiawei Han (2009). *Geographic data mining and knowledge discovery*. 2nd. CRC Press.

Nanni, Mirco and Dino Pedreschi (2006). “Time-focused clustering of trajectories of moving objects”. In: *Journal of Intelligent Information Systems* Vol. 27. No. 3, pp. 267–289.

Okuyama, Kohya and Keiji Yanai (2013). “A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the web”. In: *The era of interactive media*. Springer, pp. 657–670.

O’reilly, Tim (2005). *What is Web 2.0: Design patterns and business models for the next generation of software*. `http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html`. Accessed: 2016-11-15.

Pang, Bo, Lillian Lee, et al. (2008). “Opinion mining and sentiment analysis”. In: *Foundations and Trends in Information Retrieval* Vol. 2. No. 1–2, pp. 1–135.

Papadopoulos, Symeon, Andre Skusa, Athena Vakali, Yiannis Kompatsiaris, and Nadine Wagner (2009). *Bridge bounding: A local approach for efficient community discovery in complex networks*. Technical Report.

Papadopoulos, Symeon, Christos Zigkolis, Yiannis Kompatsiaris, and Athena Vakali (2011). "Cluster-based landmark and event detection for tagged photo collections". In: *IEEE MultiMedia* Vol. 18. No. 1, pp. 52–63.

Papadopoulos, Symeon, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos (2012). "Community detection in social media". In: *Data Mining and Knowledge Discovery* Vol. 24. No. 3, pp. 515–554.

Parent, Christine, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, et al. (2013). "Semantic trajectories modeling and analysis". In: *ACM Computing Surveys (CSUR)* Vol. 45. No. 4, pp. 42–74.

Pei, Jian, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu (2001). "Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth". In: *Proceedings of the 17th International Conference on Data Engineering*. ICDE '01. IEEE Computer Society, pp. 215–224.

Pfoser, Dieter, Christian S Jensen, Yannis Theodoridis, et al. (2000). "Novel approaches to the indexing of moving object trajectories". In: *Proceedings of VLDB*, pp. 395–406.

Pinto, Helen, Jiawei Han, Jian Pei, Ke Wang, Qiming Chen, and Umeshwar Dayal (2001). "Multi-dimensional sequential pattern mining". In: *Proceedings of the tenth international conference on Information and knowledge management*. ACM, pp. 81–88.

Plantevit, Marc, Yeow Wei Choong, Anne Laurent, Dominique Laurent, and Maguelonne Teisseire (2005). "M2SP: Mining sequential patterns among several dimensions". In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, pp. 205–216.

Popescu, Adrian and Gregory Grefenstette (2011). "Mining Social Media to Create Personalized Recommendations for Tourist Visits". In: *Proceed-*

ings of the 2nd International Conference on Computing for Geospatial Research & Applications. COM.Geo '11. ACM, pp. 37–42.

Quercia, Daniele, Rossano Schifanella, and Luca Maria Aiello (2014). "The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city". In: *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, pp. 116–125.

Rattenbury, Tye, Nathaniel Good, and Mor Naaman (2007). "Towards automatic extraction of event and place semantics from flickr tags". In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 103–110.

Schafer, J Ben, Joseph A Konstan, and John Riedl (2001). "E-commerce recommendation applications". In: *Applications of Data Mining to Electronic Commerce*. Springer, pp. 115–153.

Serdyukov, Pavel, Vanessa Murdock, and Roelof Van Zwol (2009). "Placing flickr photos on a map". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 484–491.

Shi, Yue, Pavel Serdyukov, Alan Hanjalic, and Martha Larson (2011). "Personalized Landmark Recommendation Based on Geotags from Photo Sharing Sites." In: *ICWSM* Vol. 11, pp. 622–625.

Shirai, Motohiro, Masaharu Hirota, Hiroshi Ishikawa, and Shohei Yokoyama (2013). "A method of Area of Interest and Shooting Spot Detection using Geo-tagged Photographs". In: *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*. COMP '13. ACM, pp. 34–41.

Sigurbjörnsson, Börkur and Roelof Van Zwol (2008). "Flickr tag recommendation based on collective knowledge". In: *Proceedings of the 17th international conference on World Wide Web*. ACM, pp. 327–336.

Sobkowicz, Pawel, Michael Kaschesky, and Guillaume Bouchard (2012). "Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web". In: *Government Information Quarterly* Vol. 29. No. 4, pp. 470–479.

Srikant, Ramakrishnan and Rakesh Agrawal (1996). "Mining sequential patterns: Generalizations and performance improvements". In: *International Conference on Extending Database Technology*. Springer, pp. 1–17.

Sui, Daniel and Michael Goodchild (2011). "The convergence of GIS and social media: challenges for GIScience". In: *International Journal of Geographical Information Science* Vol. 25. No. 11, pp. 1737–1748.

Sun, Yeran, Hongchao Fan, Mohamed Bakillah, and Alexander Zipf (2015). "Road-based travel recommendation using geo-tagged images". In: *Computers, Environment and Urban Systems* Vol. 53, pp. 110–122.

Symeonidis, Panagiotis, Dimitrios Ntempos, and Yannis Manolopoulos (2014). *Recommender Systems for Location-based Social Networks*. 1st. Springer-Verlag New York.

Tang, Lei and Huan Liu (2010). "Community detection and mining in social media". In: *Synthesis Lectures on Data Mining and Knowledge Discovery* Vol. 2. No. 1, pp. 1–137.

Tang, Lei, Xufei Wang, and Huan Liu (2009). "Uncoverning groups via heterogeneous interaction analysis". In: *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, pp. 503–512.

Thelwall, Mike, David Wilkinson, and Sukhvinder Uppal (2010). "Data mining emotion in social network communication: Gender differences in MySpace". In: *Journal of the Association for Information Science and Technology* Vol. 61. No. 1, pp. 190–199.

Trant, Jennifer (2009). "Studying social tagging and folksonomy: A review and framework". In: *Journal of Digital Information* Vol. 10. No. 1. Retrieved 2017-02-09, from `https://journals.tdl.org/jodi/index.php/jodi/article/view/269/278`.

Tsoukatos, IIias and Dimitrios Gunopulos (2001). "Efficient mining of spatiotemporal patterns". In: *International Symposium on Spatial and Temporal Databases*. Springer, pp. 425–442.

Van Canneyt, Steven, Steven Schockaert, Olivier Van Laere, and Bart Dhoedt (2011). "Time-dependent recommendation of tourist attractions using Flickr". In: *Proceedings of the 23rd Benelux conference on artificial intelligence (BNAIC 2011)*, pp. 255–262.

Vautier, Alexandre, Marie-Odile Cordier, and René Quiniou (2005). "An inductive database for mining temporal patterns in event sequences". In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05. Morgan Kaufmann Publishers Inc., pp. 1640–1641.

Vlachos, Michail, George Kollios, and Dimitrios Gunopulos (2002). "Discovering similar multidimensional trajectories". In: *Data Engineering, 2002. Proceedings. 18th International Conference on*. IEEE, pp. 673–684.

Vu, Huy Quan, Gang Li, Rob Law, and Ben Haobin Ye (2015). "Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos". In: *Tourism Management* Vol. 46, pp. 222–232.

Waga, Karol, Andrei Tabarcea, and Pasi Fränti (2012). "Recommendation of points of interest from user generated data collection". In: *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on*. IEEE, pp. 550–555.

Wang, Xiaofeng, Gang Li, Guang Jiang, and Zhongzhi Shi (2013). "Semantic trajectory-based event detection and event pattern mining". In: *Knowledge and information systems* Vol. 37. No. 2, pp. 305–329.

Xiang, Zheng and Ulrike Gretzel (2010). "Role of social media in online travel information search". In: *Tourism management* Vol. 31. No. 2, pp. 179–188.

Xie, Ke, Chaolun Xia, Nir Grinberg, Raz Schwartz, and Mor Naaman (2013). "Robust Detection of Hyper-local Events from Geotagged Social Media

Data". In: *Proceedings of the Thirteenth International Workshop on Multimedia Data Mining*. MDMKDD '13. ACM, pp. 2–11.

Yamasaki, Toshihiko, Andrew Gallagher, and Tsuhan Chen (2013). "Personalized intra-and inter-city travel recommendation using large-scale geotags". In: *Proceedings of the 2nd ACM international workshop on Geotagging and its applications in multimedia*. ACM, pp. 25–30.

Ye, Qiang, Ziqiong Zhang, and Rob Law (2009). "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches". In: *Expert Systems with Applications* Vol. 36. No. 3, pp. 6527–6535.

Ying, Josh Jia-Ching, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S. Tseng (2011). "Semantic trajectory mining for location prediction". In: *19th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2011, November 1-4, 2011, Chicago, IL, USA, Proceedings*, pp. 34–43.

Yoon, Hyoseok, Yu Zheng, Xing Xie, and Woontack Woo (2012). "Social itinerary recommendation from user-generated digital trails". In: *Personal and Ubiquitous Computing* Vol. 16. No. 5, pp. 469–484.

Yoshida, Mariko, Tetsuya Iizuka, Hisako Shiohara, and Masanori Ishiguro (2000). "Mining sequential patterns including time intervals". In: *AeroSense 2000*. International Society for Optics and Photonics, pp. 213–220.

Zafarani, Reza, Mohammad Ali Abbasi, and Huan Liu (2014). *Social media mining: an introduction*. Cambridge University Press.

Zaki, Mohammed J (2001). "SPADE: An efficient algorithm for mining frequent sequences". In: *Machine learning* Vol. 42. No. 1-2, pp. 31–60.

Zhang, Chao, Jiawei Han, Lidan Shou, Jiajun Lu, and Thomas La Porta (2014). "Splitter: Mining fine-grained sequential patterns in semantic trajectories". In: *Proceedings of the VLDB Endowment* Vol. 7. No. 9, pp. 769–780.

Zhang, Tian, Raghu Ramakrishnan, and Miron Livny (1996). "BIRCH: an efficient data clustering method for very large databases". In: *ACM Sigmod Record*. Vol. 25. No. 2. ACM, pp. 103–114.

Zheng, Yan-Tao, Zheng-Jun Zha, and Tat-Seng Chua (2011). "Research and applications on georeferenced multimedia: a survey". In: *Multimedia Tools and Applications* Vol. 51. No. 1, pp. 77–98.

Zheng, Yan-Tao, Zheng-Jun Zha, and Tat-Seng Chua (2012). "Mining travel patterns from geotagged photos". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* Vol. 3. No. 3, pp. 56–74.

Zheng, Yu (2015). "Trajectory Data Mining: An Overview". In: *ACM Trans. Intell. Syst. Technol.* Vol. 6. No. 3, 29:1–29:41. ISSN: 2157-6904.

Zheng, Yu, Lizhu Zhang, Xing Xie, and Wei-Ying Ma (2009). "Mining interesting locations and travel sequences from GPS trajectories". In: *Proceedings of the 18th international conference on World wide web*. ACM, pp. 791–800.

Zhou, Ding, Jiang Bian, Shuyi Zheng, Hongyuan Zha, and C Lee Giles (2008). "Exploring social annotations for information retrieval". In: *Proceedings of the 17th international conference on World Wide Web*. ACM, pp. 715–724.

Zook, Matthew, Mark Graham, Taylor Shelton, and Sean Gorman (2010). "Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake". In: *World Medical & Health Policy* Vol. 2. No. 2, pp. 7–33.