UNIVERSITAT ROVIRA I VIRGILI

DEVELOPMENT OF CONTEXT-AWARE RECOMMENDERS OF SEQUENCES OF TOURISTIC ACTIVITIES

Ayebakuro Jonathan Orama

Development of Context-Aware Recommenders of Sequences of Touristic Activities

UNIVERSITAT ROVIRA i VIRGILI

Jonathan Ayebakuro Orama

Supervised by

Prof. Dr. Antonio Moreno and Dr. Joan Borràs
Universitat Rovira i Virgili

This dissertation is submitted for the degree of

Doctor of Philosophy

Tarragona, November 2022

UNIVERSITAT
ROVIRA i VIRGILI

Departament d'Enginyeria Informàtica i
Matemàtiques

Av. Països Catalans, 26 Campus
Sescelades 43007 Tarragona Spain
Tel. (+34) 977 559 703
Tel. (+34) 977 558 512
Fax (+34) 977 559 710

eurecat
Centre Tecnològic de Catalunya

Data Science And Big Data
Analytics

C/ Joanot Martorell,15 43480 Vila-seca
Tel: (+34) 977 39 48 71

I STATE that the present study, entitled "Development of Context-Aware Recommenders of Sequences of Touristic Activities", presented by Jonathan Ayebakuro Orama for the award of the degree of Doctor, has been carried out under our supervision at the Department d'Enginyeria Informàtica i Matemàtiques of this university, and at Eurecat, Technology Centre of Catalonia.

Tarragona, November 2022

Doctoral Thesis Supervisors:

Prof. Dr. Antonio Moreno

Dr. Joan Borràs

Declaration

I declare that the contents of this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree or qualification in this, or any other University. Except where states otherwise by reference or acknowledgment, this work is entirely my own, and not done in collaboration.

Jonathan Ayebakuro Orama

Tarragona 2022

Acknowledgements

This dissertation has been supported by EURECAT, Technology Centre of Catalonia, under the Vicente López doctoral fellowship for which I express my gratitude to the IT & Media Scientific Director, Eugenia Fuenmayor Garcia, and all other Eurecat staff for their help and support.

I would like to acknowledge the directors of this thesis work, Prof. Dr. Antonio Moreno, and Dr. Joan Borràs. I can't express how much I appreciate their guidance and advice. They made the thesis work progress seamlessly with their effort in correcting drafts, offering great ideas and feedback, and ensuring the ideas in this work are sound and logical.

I would like to thank Prof. Dr. Salvador Anton Clavé, Dr. Assumpció Huertas, Dr. Aïda Valls, and Dr. Ludovico Boratto for their advice and contributions in several aspects of this work.

I am also grateful to my past and present colleagues at Eurecat, Vila-Seca, especially Giovanni Natale for their friendship and help throughout the course of this work.

Finally, I would like to appreciate my mother and siblings, and I want to give all glory to the Almighty GOD for the life, knowledge, and wisdom He gave to me to enable me through the period of this work.

Abstract

In recent years, recommender systems have become ubiquitous on the web. Lots of web services including movie streaming, web search, e-commerce, etc., use recommender systems to aid human decision-making. Tourism is one industry that is highly represented on the web. There are several web services (like TripAdvisor, Yelp, Booking.com, etc.) that benefit from integrating recommender systems to aid tourists in exploring various tourism destinations. This has brought about an increase in research focused on improving tourism recommender systems and solving the main issues they face. This thesis proposes new algorithms for web-dependent tourism recommender systems that learn tourist preferences from their social media data to suggest a sequence of touristic activities that align with various contexts and include affine activities. To accomplish this, we propose a method for identifying tourist posts from their Twitter data by identifying their frequent post locations and counting the number of days they post. Afterward, we identify the points of interest (activities) being experienced in the tourist's tweets using the location of the tweets and a priority ranking for different categories of activities. These steps serve as input into a user profiling phase where similar tourists are clustered using features extracted from their social media data that embed their interests, contextual information, and activity periods. To complement user profiling, we propose combining it with an association rule mining algorithm for capturing implicit relationships between points of interest. The rules mined are run through a rule ranking and point-of-interest selection process that maintains the causality of the rules and produces a set of recommendable activities. We evaluate the accuracy of the recommendations and the importance of the user profiling phase. We also propose a multi-objective algorithm for ordering the set of activities while balancing certain criteria that are essential for a rich tourist experience. We further show the application of Artificial Intelligence techniques for extracting and analysing the mobility of tourists from user profiles that characterize them. This is useful to destination management organisations seeking to understand different kinds of tourist flows to better organize destinations and improve the tourist experience. Overall, the methods and algorithms proposed in this thesis are shown to be useful in various aspects of tourism recommender systems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Motivation

The tourism industry is highly lucrative. At its peak, it accounted for 10.4% of the global gross domestic product (GDP), about 9.2 trillion U.S. dollars (WTTC, 2021a) which comes from international tourists arriving and spending at destinations around the globe. There was a 5% average yearly increase in international tourist arrivals between 2009-2019, marking 10 consecutive years of growth in arrivals and recording a 117% overall growth since 2000 (UNWTO, 2019, 2020). This growth can be attributed to the attractiveness and marketability of destinations, and the needs of the 'experience hungry' tourist. Despite the adverse effects of the Covid-19 pandemic on tourism, it still accounted for approximately 4.7 trillion U.S. dollars of the global GDP in 2020 (WTTC, 2021b). As a result, countries invested in tourism are incentivized to increase the attractiveness of their destinations by converting natural and cultural sites into tourist attractions, or by going a step further to create entirely new attractions. Table 1.1 shows the top 20 countries ordered by their contribution to the travel & tourism global GDP in 2019 (WTTC, 2021b) and their respective number of attractions gotten from things to do suggested by TripAdvisor. There is a positive correlation (Figure 1.1) between GDP contributed and the number of attractions, which leads us to believe that countries with more tourist attractions invite more spending from tourists. Also, the number and variety of attractions at a destination may account for tourists revisiting and staying for longer periods. Although a destination's number of attractions has not been explored as a possible determinant of tourists' length of stay, it is considered to be a factor in their trip planning (Almeida et al., 2021). So, if tourists spend more time and money at destinations with lots of attractions, then more destinations would increase their number of attractions to keep tourists interested.

Consequently, the more attractions created, the more difficult it is for tourists to plan a trip. When faced with numerous choices of things to do, they become overwhelmed and require tools to aid their decisions on attractions to visit. Online travel guide services such as TripAdvisor and Yelp try to filter the choices but do not suggest attractions that match a tourist's preferences. These websites

Table 1.1

Top 20 contributers to global tourism GDP, including their contribution in billions, and their number of attractions from Tripadvisor.

Country	Total GDP contributed (USD billion)	Number of attractions
United States	1869.7	256857.0
China	1665.6	41680.0
Germany	393.1	42418.0
Japan	373.0	113151.0
United Kingdom	305.0	83192.0
Italy	269.8	129615.0
France	240.5	78254.0
Spain	202.1	56804.0
India	191.3	34455.0
Mexico	175.6	19507.0
Australia	149.1	38889.0
Brazil	115.7	33589.0
Canada	111.6	38926.0
Thailand	106.5	16218.0
Netherlands	101.6	14688.0
Philippines	90.0	7303.0
Saudi Arabia	79.2	807.0
Turkey	77.6	14862.0
Russia	75.5	45090.0
South Korea	73.2	10202.0

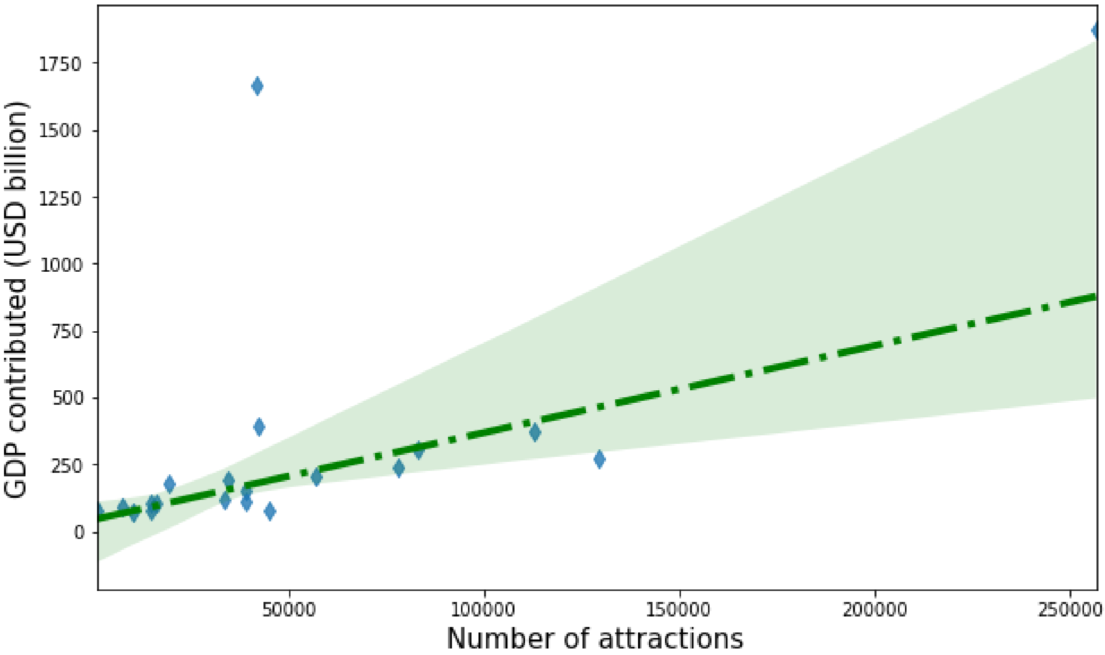


Figure 1.1: Regression plot of countries' contributed GDP and their number of attractions.

usually suggest attractions by the number of good reviews and ratings posted by other tourists. This method, while easy to implement, would only recommend very popular attractions to tourists which may exclude lesser popular choices that suit the tourist's preferences. Also, the most popular attractions become overcrowded leading to overtourism and pollution concerns (Dodds and Butler, 2019). *Recommender systems* (RS) are a viable solution to this problem. Unlike online travel guide services, they can suggest suitable attractions tailored to a user's preferences without human intervention. They also function as filters but with the added constraint that suggestions are relevant to a specific user or group of users in terms of their preferences, behaviour, and environmental or personal contexts. RS have seen successes in various fields of commerce and have gained traction in the field of tourism in recent years (Borràs et al., 2014). To make precise recommendations, RS need to study and analyse tourists' behaviour and patterns to learn their preferences and any other useful information.

In traditional tourism research, data needed to analyse tourists' behaviour are sourced directly from tourists through surveys conducted by international tourism statistics and regulation bodies, such as the United Nations World Tourism Organization (UNWTO), and World Travel and Tourism Council (WTTC), or conducted on a smaller scale by researchers and research institutes. In most cases, these surveys are undertaken using questionnaires, and they may include GPS devices handed to tourists to analyse their mobility and trajectory patterns (Massimo and Ricci, 2018). Since surveys are sourced directly from tourists, they are crafted to get specific information for a particular use case. The downside is that surveys require a lot of work, setup time, and willing and truthful participants, and also the data is not readily available for analysis. On the other hand, social media platforms have risen in popularity among researchers as a data source for tourism analysis because they are readily available and widely used by tourists looking to share their experiences online.

The digital footprints of tourists are seemingly endless. If they are properly identified, data can be gathered from as far back as required and actively tracked to obtain real-time entries. It is possible to get these data in various forms, such as texts, images, and GPS coordinates, and also glean contextual data from them without having to specially craft questionnaires for this purpose. However, social media data (SMD) is not without problems. Its unstructured and noisy nature makes it necessary to extensively preprocess the data before commencing any analysis. Preprocessing usually involves identifying tourists' data, discarding data from automated software (bots), and formatting the data in an appropriate manner. The methods applied to preprocess the data vary depending on the research focus and the social media platform, and they are specific to the dataset. For instance, in research focused on text analysis, preprocessing steps would include discarding data entries without text, discarding data entries with insufficient text length, stop word removal, stemming, and other popular text analysis preprocessing steps. Another problem of SMD is its reliability. It is mainly criticized for its misrepresentation of the population to be analysed, as the results could be skewed in favour of the young and technology reliant, while excluding rural populations with limited access to social media (Hargittai, 2020; Malik et al., 2021; Johnson et al., 2017). Despite this misrepresentation, some works have shown SMD to be correlated with data from surveys and other reliable sources (Ma and Kirilenko, 2021; Liao et al., 2022). In addition, SMD being big data provides a broader view of the population than the most diversely sampled survey datasets. Finally, SMD has a problem with consistency. Take for example, a user visits some attractions on a particular day, taking and posting geotagged photos on a social media platform. A researcher could build and study that user's timeline with those photos, but since users are not obligated to keep posting, often times gaps exist in the timeline. This problem is termed 'data sparsity' and is a big challenge to research focused on sequence analysis (Quadrana et al., 2019). To soften the effects of this problem, researchers use cluster models that represent the interests of similar users which fills in the gaps of individual interests with the interests of the group (Quadrana et al., 2019). All these problems do not take away from the value and utility of SMD.

SMD can be exploited for various use cases in tourism research, among which mobility, spatial and sentiment analysis are most prevalent. In mobility analysis, the focus is the movement of tourists within a destination or between destinations to discover movement patterns and popular travel routes, which may help in city planning and destination management. Ideally, data needed for mobility analysis must be geolocated data that provides a certain level of granularity to allow pattern mining, sourced from social media platforms that allow geotagging (Twitter, Instagram, etc.). Like mobility analysis, spatial analysis requires geolocated data but focuses on the geospatial distribution of the data, highlighting concentrations at different time periods to discover hotspots or popular locations. For the case of sentiment analysis, texts (preferably feedback and review texts) from posts are used to understand individual and group sentiments about a particular topic, product, destination, etc. Understanding tourists' feelings is useful to managers of destinations or attractions, seeking to improve the tourist experience and boost consumption. Apart from these use cases, SMD can also be used in collecting tourism statistics (alternative to UNWTO statistics) and in recommender systems. Its use in recommender systems is unique. Unlike other cases which are purely analytical, SMD serves as a source for learning user interests (whether they be interests in certain activities, popular or unpopular attractions, tour times, etc.) to make relevant recommendations.

Recommender systems are designed to suggest one or more items at a time. It is currently more popular to suggest multiple items at a time, either to give the user a choice (which provides preferential feedback for system fine tuning), or to provide the user with an ordered or unordered set of items to consume. A prominent research area in recommender systems deals with the benefits of suggesting a sequence of items for consumption (Quadrana et al., 2019). In tourism recommenders, suggesting a sequence of activities is more desirable than a set of individual activities. A sequence of activities offers a more complete recommendation to a tourist. For instance, a purchasable tour containing the popular attractions in a city arranged in a meaningful order is more appealing to a tourist than the list of all popular attractions in the city. In addition, a sequence of activities provides the opportunity for the recommender system to balance certain criteria that could positively or negatively affect the experience of the tourist. The distance between attractions, optimal visitation route, popularity of attractions, or the diversity of attractions are some of these criteria that could be efficiently balanced. It is difficult to take into account all these criteria without trade-offs, which is why they need to be efficiently balanced.

1.2. Objectives

The primary objective of this thesis is to develop and implement new recommendation algorithms that work by exploiting social media content to analyse user preferences, detecting clusters of users with affine interests/time preferences/popularity preferences, using association rules to analyse and incorporate Point of Interest (POI) affinity, and then optimising routes according to diverse criteria with the purpose of recommending a set of ordered POIs that maximize the user satisfaction.

The primary objective will be accomplished by completing the following sub-objectives:

1. Study of the state of art on social media analysis and recommender systems, specifically focused on their application in the tourism domain.

2. Choose and prepare a data source to analyse tourist behaviour, collect data, and study methods for learning user preferences, detecting clusters of users with similar travel behaviour, and incorporating POI affinity.

3. Design and develop novel recommender mechanisms based on chosen Artificial Intelligence analytical and modelling methods. Extend recommender system to optimize visit routes according to several criteria.

4. Evaluate and test the recommender system and the optimized routes.

1.3. Contributions

In this section, we summarize the scientific contributions of this thesis work in the form of the publications and conference presentations.

1. We have developed a context-aware tourism recommender system called ReCLARM which uses a cluster-based user profiling technique combined with association rule mining to recommend points of interest to tourists based on data downloaded from Twitter. We describe this system in Chapter 3. The results of this work have been published in the following paper:

   - Orama, J. A., Borràs, J., and Moreno, A. (2021). Combining cluster-based profiling based on social media features and association rule mining for personalised recommendations of touristic activities. *Applied Sciences*, 11(14):6512.

2. We have developed an algorithm for identifying tweets from tourists considering their duration of stay and frequency of posts. This method was used to

study the effect of Covid-19 on tourism-oriented theme parks, a case study on PortAventura. The study was published in the following book:

- Anton Clavé, S., Borràs Nogués, J., Orama, J. A., and Soto, M. T. R. (2021). The changing role of tourism-oriented theme parks as everyday entertainment venues during COVID-19. In Condevaux, A., Gravari-Barbas, M., & Guinand, S. (Eds.) *Tourism Dynamics in Everyday Places* (pp. 245-262). Routledge, London.

3. We have applied Artificial Intelligence techniques in the design and development of a method for analysing social media data that creates visitor profiles according to their travel preferences and mobility patterns which is useful to destination management organisations. We describe this study in Chapter 4. The results of this study were published in the following article:

   - Orama, J. A., Huertas, A., Borràs, J., Moreno, A., and Clavé, S. A. (2022). Identification of mobility patterns of clusters of city visitors: an application of Artificial Intelligence techniques to social media data. *Applied Sciences*, 12(12), 5834.

4. We have presented our method for learning and modelling the preferences of tourist from their Twitter data detailed in Chapter 3 at the 32nd European Conference on Operational Research in Aalto University, Espoo, Finland https://euro2022espoo.com/. The abstract of this presentation was published by EURO in:

   - Orama, J. A., Moreno, A., and Borràs, J., (2022). Learning the preferences of tourists through the analysis of social media data. [Abstract]. In *32nd European Conference on Operational Research*, Espoo, Finland, 34. https://www.euro-online.org/conf/admin/tmp/program-euro32.pdf

5. Finally, we have developed and presented our multi-objective genetic algorithm for optimizing route selection in recommendation of touristic activities which is an extension of our tourism recommender system detailed in Chapter 3 at the Conference of the Catalan Association for Artificial Intelligence (CCIA 2022). The results of this work were published in the following article:

   - Orama, J. A., Moreno, A., and Borràs, J. (2022). Multi Objective Genetic Algorithm for Optimal Route Selection from a Set of Recommended Touristic Activities. In Cortés, A., Grimaldo, F., & Flaminio, T. (Eds.) *Frontiers in Artificial Intelligence and Applications* (pp. 9-12). IOS Press.

1.4. Structure of the Thesis Document

The rest of this thesis document is structured as follows:

- Chapter 2 details the state of the art of the main aspects of this dissertation. Tourism focused social media analysis, contextual recommendations in tourism, and social media tourist flow analysis.

- Chapter 3 presents our proposed tourism recommender system based on the aspects detailed in chapter 2. Detailing its algorithms for tourist identification, activity identification, user profiling, and ordering touristic activities.

- Chapter 4 presents a tourism focused study on extracting mobility patterns of clusters of tourist using AI techniques to aid destination management organisations in improving the tourist experience.

- Chapter 5 provides the final conclusions and presents some possible future research from this work.

Chapter 2

State of the Art

2.1. Introduction

In this chapter, we review the state of the art of the important aspects of this thesis, including social media analysis in tourism, contextual recommendation, and flow/mobility analysis based on social media data. In Section 2.2, we review the current state of social media analysis in tourism including data sources, analytical methods, tourist identification, point of interest identification, and user profiling. In Section 2.3, we briefly introduce contextual recommendations in tourism, also touching on traditional recommenders, contexts employed in tourism analysis, and the AI techniques applied by tourism recommender systems. Finally, in Section 2.4, we review the analysis of social media data for studying tourist flows.

2.2. Social Media Analysis in Tourism

All research includes some form of analysis necessary to make an empirical conclusion about a particular subject. Initially, data surrounding the research subject are collected for analysis. The source of these data is highly dependent on the goals of the research and the topic area. In tourism research, data are typically sourced through surveys or social media, each having its advantages and disadvantages as described in chapter 1. The analysis set around social media data is called social media analysis or analytics (SMA).

SMA primarily involves developing methods for tracking, collecting, and studying social media data to discover unique trends and offer explanations for certain phenomena (Stieglitz et al., 2014). Over the years, it has become very popular despite concerns about its reliability (Ma and Kirilenko, 2021; Xiang et al., 2018). It is applicable in a wide range of areas such as agriculture, banking, business intelligence, communication, disaster management, disruptive technology, education, ethics, government, health care & public health, hospitality & tourism, journalism, management, marketing, terrorism, etc. (Zachlod et al., 2022; Mirzaalian and Halpenny, 2019). This applicability stems from three main facts. Firstly, social media data are readily available and easily downloadable for use within privacy policies. Secondly, users are ready to share their opinions about products and topics on social media platforms. Lastly, social media platforms are mostly not restricted by location; hence the population of users is diverse, consisting of people from various races, ethnicities, and walks of life.

SMA has many essential aspects. In this section, we describe the aspects relevant to this thesis and review some works focused on them.

2.2.1 Data sources and Data retrieval

The choice of a data source is a very crucial step in SMA. Social media platforms differ in various ways but they can be broadly classified into the following categories as proposed in Mirzaalian and Halpenny (2019): (1) social networking sites, (2) media & content communities, (3) discussion forums, and (4) customer review sites. These categories correspond with the services and activities encouraged on these platforms: networking, content sharing, conversations, and opinion sharing. In the next sub-sections, we review these categories and highlight relevant articles that use them as a data source.

Social Networking Sites as a Data Source

Social networking sites (SNSs) are web-based platforms accessible on mobile or computer devices that allow users to build social networks with other users who share similar interests or already belong to their personal or career networks. These social networks are used to share information in the form of texts, images, videos, links (URLs), geotagged locations, etc. Usually, SNSs have a central purpose; for example, Twitter focuses on trending topics, while Facebook focuses on reconnecting with old friends and schoolmates. Despite this, their utility has grown in the hands of the public, transforming them into rich banks of information usable for SMA. Twitter in particular has been used to petition governing bodies, effect large-scale adoption of products or policies, increase the visibility of locations, broadcast disaster warnings, etc.

Twitter and its Chinese counterpart Sina Weibo are the most used SNSs in tourism research. This adoption is partly due to their easy-to-use application programming interface (API), which provides access to posts published on their platforms, and their lenient policies that allow researchers to use these posts in their work. Mirzaalian and Halpenny (2019), in their review of SMA in tourism and hospitality, documented Twitter and Sina Weibo among the top five sources, as 78% of articles with data sourced from SNSs got data from them. This trend followed through in our study of SMA in tourism within 2019-2022. Of 13 articles with data sourced from SNSs, 6 were from Twitter (Obembe et al., 2021; Yan et al., 2020; Morgan et al., 2021; Park et al., 2020; Liao et al., 2022; Petutschnig et al., 2021), 5 from Sina Weibo (Chen et al., 2022; Sun et al., 2020; Zhang et al., 2020; Ebrahimpour et al., 2020; Jiang et al., 2021), and 2 from Facebook (Yang et al., 2021; Önder et al., 2020). In combination, Twitter and Sina Weibo had an 84% representation in the articles sourced from SNSs.

In addition, Twitter and Sina Weibo are preferred by researchers because of their popularity and the nature of their data. Twitter had 436 million active users in January 2022, while Sina Weibo had 573 million (Statista, 2022). Although they both fall behind Facebook in popularity, their data is easily accessible and works well for the analytical methods used by tourism researchers. For instance, sentiment analysis works well with Twitter data because of the limited length of text in posts (up to 280 characters). The limited text length is desirable to researchers for three main reasons. First, processing the text requires less time and computing power, allowing analysis of large datasets without issues. Second, it ensures that users are concise with their posts, selecting certain keywords and short phrases that match their thoughts which makes thematic analysis and topic modelling effective. Lastly, it makes it easier for users to post frequently, which results in highly granular data that is best for spatial, temporal, and mobility analysis.

In contrast, Facebook doesn't do so well in terms of text length (up to 63,206 characters) and data availability, making it less desirable to researchers. Some works have used the metadata of Facebook posts instead of the text. Yang et al. (2021) turned Facebook metadata (post likes, number of shares, number of comments, etc.) into variables in quantitative analysis to understand the effect of Facebook posts on the demand for peer-to-peer accommodation (Airbnb) in an epidemic. Önder et al. (2020), used not only likes and comments but also the different reactions provided by Facebook (*angry*, *haha*, *like*, *love*, *sad*, and *wow*) to predict the demand for destinations based on reactions to their Facebook posts. They also expressed how Facebook's data policies limit the research that is possible (Önder et al., 2020, p. 198). Lee et al. (2021), used the number of followers, comments, and likes accrued by destination management organisations' Facebook pages to evaluate their richness and how they engage tourists.

Media & Content Communities as a Data Source

Like SNSs, *Media & content communities* (MCCs) are also online social sites where users connect with others, but solely for sharing pictures and videos. The concepts of likes and comments remain the same, but posts are centered around pictures or videos and not text. Although the content (photos & videos) shared is the focus of MCCs, researchers are more interested in the ability to geotag photos and videos before sharing. Geotagged photos are generally better than geotagged texts because users tend to post photos from the sight they were taken, while text about a sight might contain thoughts crafted after visiting the sight.

MCCs data are usable in a wide variety of social sciences research, such as tourism, politics, communication studies, sociology studies, etc (Chen et al., 2021). Flickr, Instagram, and Panoramio are common MCCs data sources for tourism research. The differences are Flickr and Panoramio have quite a few popular pre-compiled datasets for research, which is not the same for Instagram. Even though Panoramio was shut down in 2016, these datasets keep it relevant in tourism research. Zhang et al. (2021), used Panoramio and Flickr datasets collected between 2006-2014 before it was shut down, to study the spatial distribution of visits to national parks in Utah, USA.

Flickr is the most used MCC as a data source for tourism research because photos published on its platform are publicly available and free to access. On the other hand, Instagram is popular but not well used because of its privacy policy. It is currently possible to get data using their API but only from business accounts. Another problem with Instagram data is user anonymity. It is possible to get data using a hashtag and post ID, but posts cannot be linked to users in any way. This means that while Instagram data is usable in spatial analysis and trend analysis it is not useful for mobility analysis. Other available MCCs that are scarcely used include Tumblr, Pinterest, etc.

As previously mentioned, data are sourced from MCCs mainly for the geolocation information which is used to visualize the spatial and temporal distribution of users (Ghermandi et al., 2020; Zhang et al., 2021; Ciesielski and Stereńczak, 2021; Domènech et al., 2020b). Some works have also identified the content of the images. Giglio et al. (2019), identified and classified the contents of Flickr images to evaluate the attractiveness of some Italian cities. Similarly, Arefieva et al. (2021), labeled photos from Instagram based on their content (using Google's Cloud Vision API) to understand tourists' experiences and impressions of destinations. Han et al. (2021), embedded the content of Flickr photos into a recommender model to account for the visual interests of tourists in the recommendation process. Other works have studied the text accompanying photos. Gon (2021), utilised hashtags and text in Instagram photos to discover themes and relate them to the local experience of a destination. Similarly, Ghermandi et al. (2020), used titles, tags, and text of Flickr photos to classify them under specific cultural services.

Discussion Forums and Customer Review Sites as a Data Source

Discussion forums and *customer review sites* are online platforms for users to create topics and share their opinions about different products and services. They cover a wide range of topics like politics, electronics, commercial products, tourism destinations, etc. Both discussion forums and customer review sites serve the same purpose, to allow users the opportunity to make complaints, suggestions, and rate services. Most times they come bundled together, for example, TripAdvisor.com and TripAdvisor Travel Forum. In this way users can find suggestions on destinations to visit on TripAdvisor.com and then comment about their experiences on the forums which can then be used to improve the services on TripAdvisor.com. This feedback mechanism is also currently applied in servicing companies, such as hotels, airline companies, restaurants, etc.

Customer review sites and discussion forums are good data sources in tourism research. Specifically, the forums focused on travel, hospitality, and tourism, such as TripAdvisor, Yelp, Booking.com, etc. The data sourced from these platforms are textual, comprising of the reviews and opinions of tourists about a destination, attraction, or purchasable guided tour. These texts are valuable to tourism researchers and other interested parties for qualitative analysis aimed at understanding and improving the experience of tourists. TripAdvisor, Yelp, and Baidu are common data sources in this category. TripAdvisor is the main data source in tourism because of its international popularity. The TripAdvisor site offers suggestions on destinations, attractions, tours, restaurants, hotels, etc. These suggestions include ratings from other users and their textual reviews. Yelp does the same things as TripAdvisor but is only popular in the USA. Others like Baidu Travel and Mafengwo are focused on Chinese nationals.

The data sourced from any customer review site or discussion forum are the same, therefore the difference lies in the focus of the research. Most tourism research using review data is concerned with the tourist's experience, perception, and the effect of electronic word-of-mouth on a destination or attraction (Cassar et al., 2020; Zhang et al., 2022; Wang and Kirilenko, 2021; Yu et al., 2021; Nakayama and Wan, 2019; Luo and Xu, 2019). Others attempt to learn tourist preferences (Nilashi et al., 2021), evaluate tourist loyalty to a destination (Mirzaalian and Halpenny, 2021), or evaluate a destination's image (Cao et al., 2020).

Data Retrieval

Social media data qualifies as big data; therefore, it is important to decide on parameters that guide data retrieval to ensure the dataset is streamlined and relevant to the research goals. The following are the most common retrieval parameters or combinations of parameters in tourism research.

- *Date-focused retrieval*: As the name suggests, the dataset is comprised of data retrieved within a particular date period or a specific date. This means that data published on the chosen social media platform within the date are retrieved (Obembe et al., 2021; Zhang et al., 2022).

- *Location-focused retrieval*: In this case, the dataset is comprised of posts published at a particular location, such as a city, country, attraction, or geographical bounding box (Sun et al., 2020; Yu et al., 2021; Ghermandi et al., 2020; Önder et al., 2020; Cassar et al., 2020; Nilashi et al., 2021; Nakayama and Wan, 2019; Han et al., 2021).

- *Location- and date-focused retrieval*: This is the most preferred parameter combination for tourism research. It combines a specific date or date period with a location to retrieve data (Chen et al., 2022; Zhang et al., 2020; Mirzaalian and Halpenny, 2021; Yang et al., 2021; Liao et al., 2022; Ebrahimpour et al., 2020; Jiang et al., 2021; Petutschnig et al., 2021; Wang and Kirilenko, 2021; Luo and Xu, 2019; Ciesielski and Stereńczak, 2021; Domènech et al., 2020b; Lee et al., 2021).

- *Content- and date-focused retrieval*: In this case, the dataset is not only comprised of data retrieved within a specific period but also the data must contain a specific textual identifier. The most common identifiers are hashtags, which are mostly used in Twitter and Instagram posts. Adding specific hashtags can streamline the data around a certain event, topic, location, or attraction.

This combination is sometimes used instead of *'Location- and date-focused retrieval'* because hashtags can be used to identify locations (Yan et al., 2020; Morgan et al., 2021; Arefieva et al., 2021; Gon, 2021).

2.2.2 Analytical Methods and Use Cases

Analytical methods are a core part of SMA. As previously mentioned, some form of analysis is necessary to make conclusions in research. This analysis process consists of analytical methods that derive meaning from the collected data. In tourism research, analysis is centered around the tourists (their behaviour, perception, preferences, and experiences), the tourist attractions or destinations, the factors that affect them, and how they relate to each other. In this section, we review some common analytical methods in tourism and relevant articles that employ them.

Text Analysis

Text analysis derives meaningful conclusions from text data by automatically extracting keywords, themes, or concepts from the text. A large percentage of data published on social media platforms are textual, which makes text analysis very valuable in social media analysis. In addition, text analysis helps tourism researchers gain insight into the tourist perspective. For instance, tourists' reviews about a certain attraction could include frequently used words like *'great scenery'* or *'gorgeous view'* which highlights what aspects of the attraction they found interesting. Usually, the text analysis process requires some preprocessing steps. These steps may include the following:

- Removal of punctuations, signs, and symbols. They are irrelevant in text analysis because they do not contain any meaningful insights.

- Removal of stop words. Stop words are parts of a language that mainly serve to make a text more coherent and do not add much meaningful insights, such as determinants (e.g., 'a', 'the', 'another', etc.), conjunctions (e.g., 'an', 'for', 'but', etc.), prepositions (e.g., 'in', 'under', 'above', etc.), etc. Some tools and articles attempt to provide a database of these words (Bird et al., 2009; Sarica and Luo, 2021).

- Lemmatisation. This step involves reducing words to their base form found in the language dictionary. For example, *'walking'* is reduced to *'walk'*. It helps to ensure that words which are essentially the same are grouped in word frequency calculation.

After preprocessing, keywords or concepts are usually identified by word frequency calculation and then used to classify the text under meaningful themes extracted from the dataset. For instance, Yu et al. (2021) extracted themes and concepts that represent tourists' memorable experiences from their reviews of attractions in London, UK. They then used those themes to classify the attractions under four categories (nature-marker, human-marker, nature-sight, and human-sight) that embody tourist perception. Along the same line, Cassar et al. (2020)

extracted themes and concepts from reviews of 1000 restaurants to understand the role wine plays in customer satisfaction. They found that concepts like *taste*, *wine*, *chocolate*, etc., were closely related to positive reviews, while *waiter*, *bill*, etc., were related to negative reviews.

Text clustering is another way to extract meaning from textual data. Clustering is an unsupervised machine learning technique that identifies distinct segments of similar data points in a dataset. In text clustering, a document is represented by frequencies of keywords from a corpus extracted from the dataset. This means that clusters are made up of documents with similar keyword frequencies. Arefieva et al. (2021), explored text clustering based on keyword frequency to build models that describe the image of a destination based on Instagram posts. The clustering process provided clear classifications of the interesting concepts associated with the image of a destination. Similarly, Mirzaalian and Halpenny (2021) utilised a keyword clustering approach to identify four loyalty-focused categories in tourist online reviews.

Similar to text clustering, topic modelling groups documents under topics by their keywords but considering their similarity in a vector embedding. *Latent Dirichlet Allocation* (LDA) is the most common topic modelling algorithm used in tourism research for extracting important aspects in documents (topics) that are linked to keywords that embody them (Mirzaalian and Halpenny, 2021).

A popular toolkit among tourism researchers for text analysis is Leximancer (Smith and Humphreys, 2006). It combines tools for both conceptual and relational text analysis to extract themes from text without using a dictionary (Cassar et al., 2020; Yu et al., 2021; Gon, 2021). Leximancer's conceptual map is valuable in understanding the semantic relationship between concepts and themes extracted from the text. Figure 2.1 shows a sample concept map (from (Cassar et al., 2020)), which illustrates concepts that affect customer satisfaction at 1000 restaurants. Concepts in the same circles belong to the same theme, and the closer concepts are on the map, the higher their co-occurrence in the dataset.

Sentiment Analysis

Sentiment analysis or *opinion mining* is a type of text analysis that is concerned with identifying the feelings or emotions portrayed in textual data. This analytical method does not extract keywords that characterize the text but classifies it by the feeling it portrays, usually positive, negative, or neutral. There are two basic approaches to sentiment analysis: lexicon-based sentiment analysis and machine learning sentiment analysis. Lexicon-based sentiment analysis uses a lexicon (a dictionary of words that represent the different sentiment categories) to determine the inclination of the words in a text. On the other hand, machine learning sentiment analysis trains a classifier on texts already labeled according to their sentimental inclination and uses the trained model to classify unlabeled text. Both approaches perform reasonably in sentiment analysis, but the lexicon-based based approach is highly domain dependent compared to the machine learning approach (Mirzaalian

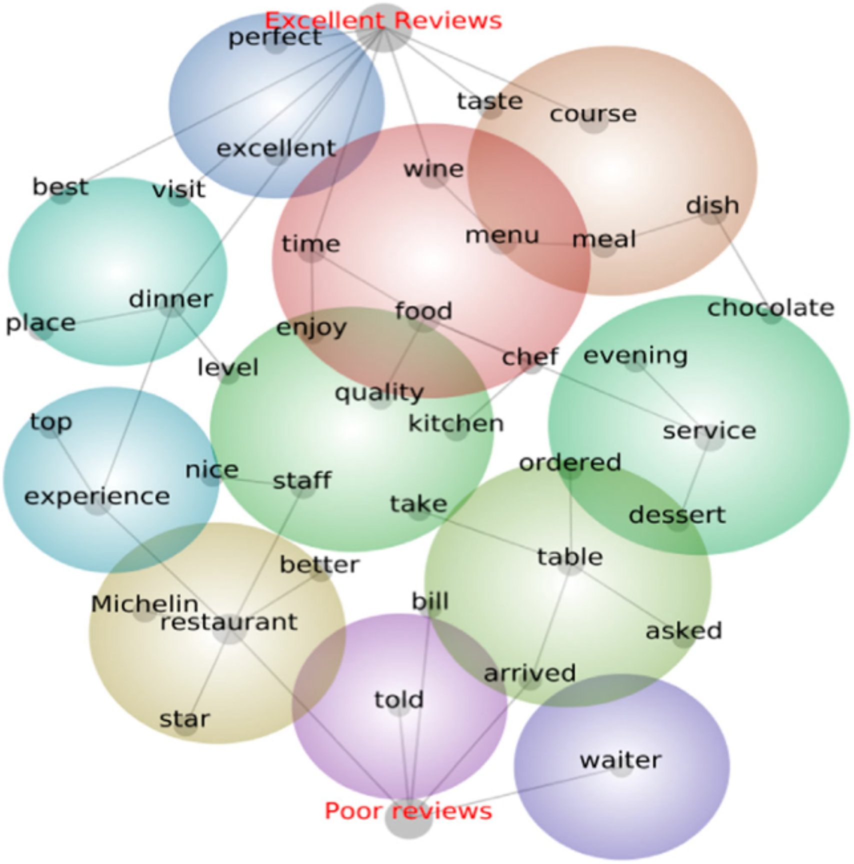


Figure 2.1: Sample concept map created using leximancer (from (Cassar et al., 2020)).

and Halpenny, 2019).

The notion of sentiments and opinions is very valuable in tourism research. It makes it possible to understand how a destination is perceived and how it may be evaluated. For instance, analysing the sentiment of tourists portrayed in destination reviews can be used to build an emotional and cognitive image of the destination (Cao et al., 2020). Also, the degree of positive to negative sentiments can be used to score destinations to evaluate their performance (Chen et al., 2022; Mirzaalian and Halpenny, 2021; Zhang et al., 2022). In addition, the effects of crisis communication (Obembe et al., 2021), crisis management (Morgan et al., 2021), air pollution (Zhang et al., 2020), natural disasters (Yan et al., 2020), cultural differences (Wang and Kirilenko, 2021; Nakayama and Wan, 2019), and other factors on the sentiments of tourists could be studied. For instance, Jiang et al. (2021) studied the effects of distance, historical background of attractions, and traffic conditions on the sentiments of tourists as they navigate through a destination.

Sentiment analysis and topic modelling are commonly combined to understand tourist sentiments concerning certain topics. For instance, Luo and Xu

(2019), grouped tourists' reviews under topics (taste, experience, value, etc.) about a destination and then extracted sentiments from grouped reviews to develop the consensual sentiments tied to the topics. This allows destination managers to focus on specific aspects with poor sentiments to improve tourists' experience. Also, combining both analytical methods allowed researchers to reduce large amounts of text to major topics that are representative of the analysed documents (Yan et al., 2020).

Social Network Analysis

Social network analysis (SNA) explores entities on a social network, their relationships, and the overall network structure. In SNA, social connections are represented by a graph with nodes as entities and links as their relationships. SNA is not common in tourism research but can be used to understand the relationships between attractions or destinations from the tourist's perspective. For instance, Sun et al. (2020) studied the social network of villages and attractions in a peri-urban area. They constructed a co-visitation map using tourists' social media data, where nodes are villages and edges represent their connections to attractions and the co-visitation of tourists. The authors studied this network to understand the roles these villages play in the network of the peri-urban area.

Spatial Analysis

In *spatial analysis*, the geographical distribution of entities on a map is studied to reveal spatial relationships and the physical activity of entities at locations. Spatial analysis is sometimes combined with temporal analysis to include a factor of time, called spatio-temporal analysis. In tourism research, spatial analysis could be used to study the demand for certain attractions, study the active periods of attractions, and highlight the presence of overtourism in these attractions. For instance, Ghermandi et al. (2020) compared the spatial distribution of different tourist groups among attractions in a destination. The groups included international, locals, and other domestic tourists. Their analysis showed attractions favoured by individual groups and areas in the destination that are not frequented by any group. Similarly, Giglio et al. (2019) studied the distribution of geotagged photos published by tourists to locate hotspots in a destination. They then utilised unsupervised machine learning to group distributed points in an attempt to automatically detect points of interest.

Spatial analysis can also be used in conjunction with the other analytical methods to provide new insights. For instance, Park et al. (2020) combined spatial analysis and sentiment analysis to understand the rides in a theme park that were associated with different emotions. Similarly, Zhang et al. (2020) employed sentiment analysis and spatial analysis to investigate the effects of air pollution on the tourist visit pattern and found that areas with higher pollution had fewer positive sentiments and fewer visits. Zhang et al. (2021) used statistical methods to validate the use of social media data for spatial distribution analysis. Along the same line, Ciesielski and Stereńczak (2021) used statistical models to determine factors that significantly impact the spatial and temporal distribution of tourist activities in forest areas.

Trend and Predictive Analysis

Trend and predictive analysis are statistical analysis methods. Trend analysis studies past trends of a phenomenon in a time series to forecast future trends, while predictive analysis attempts to estimate or predict a future event. Both methods are quite similar and apply regression techniques. In tourism research, they are mostly used in demand and supply estimation, and for predicting tourist visits or preferences. Yang et al. (2021) showed the impact of social media posts from the Centers for Disease Control and Prevention (CDC) on the supply and demand for Airbnb rentals during a virus epidemic. Liao et al. (2022) explored the feasibility of using social media data for travel demand estimation. Similarly, Önder et al. (2020) explored the use of Facebook posts statistics of Destination Management Organisations and tourist arrival statistics to predict the tourism demand of destinations. In the case of predictive analysis, Nilashi et al. (2021) used classification and regression trees to analyse tourist online reviews for preference prediction.

2.2.3 Tourist Identification in Social Media Analysis

The tourist is the subject of tourism research. Thus, for tourism-focused social media analysis, it is necessary to identify data from tourists for analysis. This is a difficult task due to the unstructured nature of social media data, and it is even more difficult when the social media platform is not built specifically for touristic purposes. Several research works have approached the problem of tourist identification in social media analysis with a wide range of techniques, which can be categorised into simple/straightforward methods and complex/advanced methods. In this section, we review these methods for tourist identification and relevant articles that employ them.

Simple/Straightforward Methods

The simple/straightforward approaches are easily applicable and they don't have many steps. They include:

- *Home location*: The user's home location could be compared with the tourist destination to be analysed. A user's home location is usually extracted from his/her social media profile (Giglio et al., 2019; Yan et al., 2020) or inferred from locations he/she frequently posts from. Ghermandi et al. (2020), determined a user's home location as the location with the highest number of active days (i.e., days with posts published on their social media timeline). Finally, a tourist is identified as a user with a home location different from the tourist destination in focus. This method fails when it is not possible to determine the user's home location. However, the home locations extracted from user profiles may be wrong, because users are allowed to input whatever they like.

- *Activity period*: A user's period of activity at the tourism destination could be used to determine if he/she is a tourist. Domènech et al. (2020b), identified tourists as users with less than or equal to one month of activity per year at a destination. The authors also defined repeat tourists/visitors as users with several active years with only one month of activity per year.

- *Tourist activities*: The kind of activities experienced by the user may determine if he/she is a tourist. If the goal of tourist identification is to capture all tourists both local and international, the activities they experience could help differentiate users touring a destination and users engaged in other activities. Xue and Zhang (2020) identified a tourist as a user with at least one visit to a tourist attraction from a pre-defined attraction list. This method requires tourist attractions to be explicitly defined.

- *Manual scan*: This method involves a human manually reviewing users' posts to determine if they are tourists (Sun et al., 2020). This method is not feasible in big data analysis and only works with small datasets. It is also taxing and prone to human error but it will provide a set of tourists that perfectly matches the researcher's idea of what a tourist is.

Complex/Advanced Methods

The complex/advanced approaches usually utilise multiple steps and machine learning techniques. They include:

- *Machine learning*: It is possible to use machine learning techniques on features extracted from a user's timeline to determine if he/she is a tourist. The features extracted from the dataset should be engineered to provide a distinction between tourists and locals. Bustamante et al. (2019), extracted features that describe the user's behaviour such as posting period, number of posts at attractions, user's time zone, the total number of posts, etc. Then they applied unsupervised machine learning to cluster users into two groups that represent tourists and locals. This method, while pretty accurate, is very case-dependent, and the clustering process must be performed every time a new user is added to the dataset. Also, the feature engineering step is very crucial to get usable results.

- *Multiple steps with conditions*: These cases usually start out employing one or a combination of the methods in the simple/straightforward category and then they add conditional steps for users that can't be identified in the first steps. Jiang et al. (2021), used a combination of home location (the location the user registered to the social media platform) and tourist activities visited by the user to determine if he/she is a tourist. They added a further condition, using a probability index based on the average amount of days spent at the tourism destination, to classify users who cannot be identified using the initial steps.

- *Multiple-step elimination*: This method focuses on eliminating users that cannot be classified as tourists rather than identifying tourists. This process usually involves multiple steps. Arefieva et al. (2021), eliminated posts that were not tourism-related using hashtags, then they eliminated spam accounts by dropping users who follow an unreasonably large number of users, and then they finally eliminated business accounts using machine learning to find accounts with names that are similar to businesses like hotels, etc. Along the same line, Park et al. (2020), eliminated staff at a theme park by deleting users with posts outside the park's operating hours, and then they manually checked the remaining user's timeline to remove business accounts.

- *Combination of simple methods*: This method combines methods from the simple/straightforward category to get better results in identifying tourists. The user's home location is usually combined with another method to make identification more accurate. Ebrahimpour et al. (2020) combined home location (defined as the place where most of the user's messages are posted during non-working hours) with activity period, where users who published only within a 10-day period are considered tourists. Similarly, Liao et al. (2022) combined home location with activity period, where a user's home location is the most visited location on the weekends, and on weekdays between 7 pm and 8 am. Also, their most visited non-home location is the most visited location on weekdays between 8 am and 8 pm. Finally, Dietz et al. (2020) combined home location and the distribution of a user's posts to determine if he/she is a tourist. The authors defined a home location as the location with the highest number of posts and only considered a user as a tourist if the number of posts at their home location consisted of 50% or more of all posts on their timeline.

2.2.4 POI Identification in Social Media Analysis

The POI is an important part of tourism analysis. It is a term used to define a place or attraction a tourist is expected to see when visiting a destination, also known as a 'sight' (Neff, 1938; Enzensberger, 1996). POI identification works in tandem with tourist identification. After identifying tourists within the dataset, the POIs they visit must also be identified to ascertain their interests. This is only necessary when dealing with a singular dataset containing both tourist profiles and check-in data. In most cases, it is not necessary to identify POIs because the POI dataset is downloaded separately from tourists' data (Ebrahimpour et al., 2020), or the POIs are decided before downloading the data (Zhang et al., 2020). In this section, we discuss methods for automatically identifying POIs within a singular dataset.

Spatial Clustering

Spatial clustering is a popular method for POI identification. It involves studying the spatial distribution of geolocated data points to determine hotspots or popular points on a map that could be identified as POIs. An easy way to accomplish this is through clustering, which is an unsupervised machine learning technique for grouping close data points and separating those far apart. Although clustering helps to identify actively visited parts on a map, knowledge of the actual POIs at those points must be acquired to properly identify tourist interests.

Giglio et al. (2019), identified POIs and landmarks by clustering geolocated Flickr photos. They used the *FindCluster* module from the Wolfram Mathematica toolkit to find hotspots that users prefer photographing by clustering their photos. The clusters generated are plotted on city maps and the active areas are recognised and compared against popular attractions in the area to discover the city's identity from the tourists' perspective (Fig. 2.2).

Han et al. (2020, 2021), defined a new spatial clustering process for geolocated Flickr posts that considers not only the spatial distance between geolocated photos but also the users who post them and the semantic similarity of their tags. Their clustering process is illustrated in Figure 2.3. An initial clustering is performed using a modified density-joinable cluster method on photos with similar tags to generate clusters. The clusters are then evaluated to check if they meet a minimum time threshold and a minimum number of users, if not they are classified as noise. The final clusters were found to represent fine-grained tourist attractions when applied to a dense area of geolocated Flickr photos.

Textual-based Identification

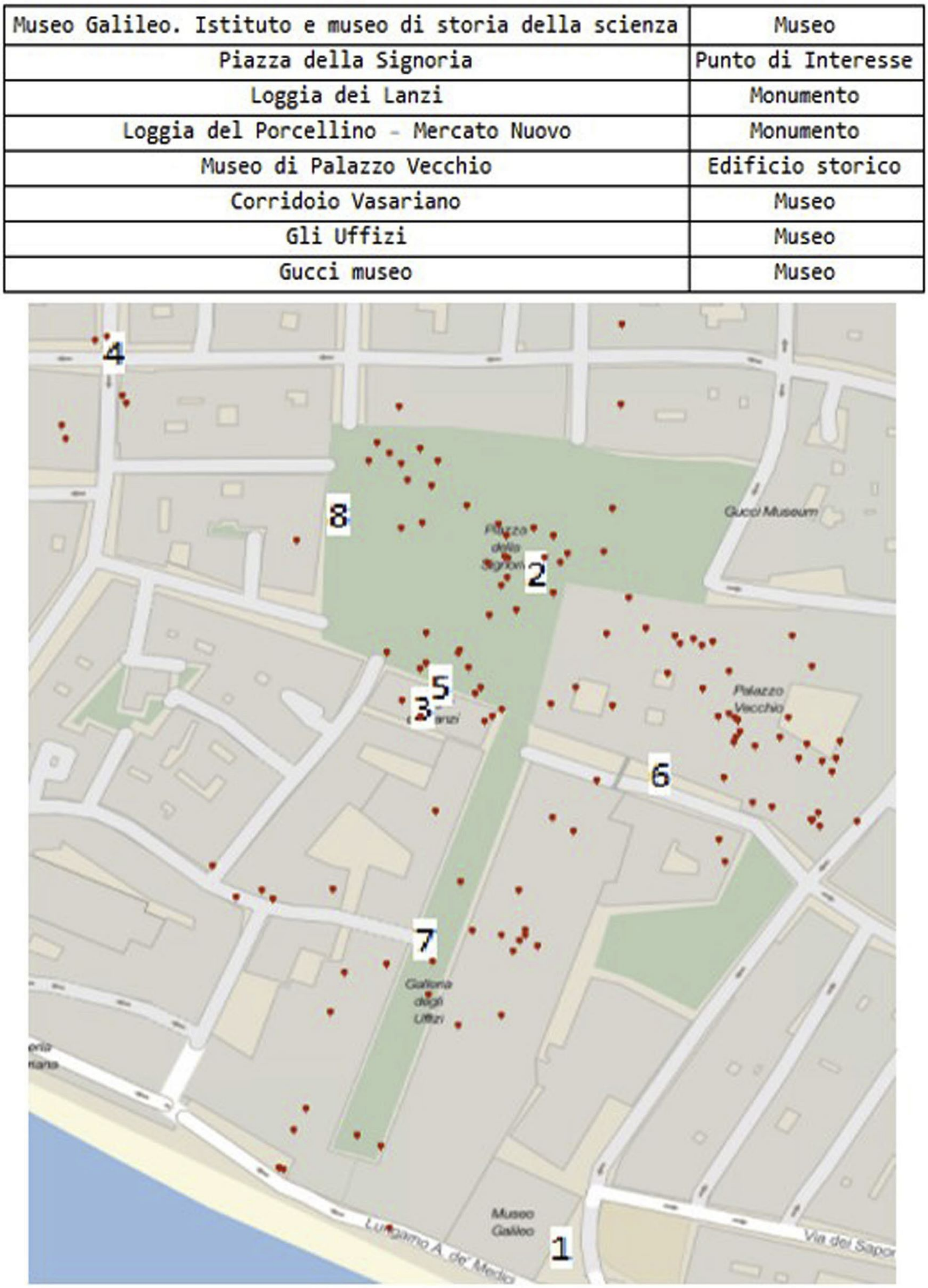
In this method, the POIs are extracted from the textual data present in tourists' posts. It usually involves extracting keywords from the text which are associated with certain POIs or extracting text that matches POI names. It is a simple and easily implemented way of identifying POIs.

Ghermandi et al. (2020), identified the cultural activities associated with Flickr photos by extracting keywords from the title and tags that accompany the photos. The title and tags are pre-processed, removing stop words and duplicates, and then keywords related to cultural activities are extracted. The keywords are then classified under the respective cultural activity which forms the basis of the authors' analysis. To confirm the classification process, the photos are manually checked to ensure they align with the cultural activity selected based on their keywords.

Chen et al. (2022), focused on identifying cities and regions as destinations instead of specific POIs. They identify these destinations by extracting postcodes from the textual data in the posts and grouping them using the first two digits which represent the regions in the Australian postcode system. The grouped postcodes are then geocoded using the *Geopy* python library to identify the nearest urban centre.

Inferred by Association

In this method, it is assumed that places visited by tourists are POIs. This means that if a user is identified as a tourist or identified to be engaging in touristic



Museo Galileo. Istituto e museo di storia della scienza	Museo
Piazza della Signoria	Punto di Interesse
Loggia dei Lanzi	Monumento
Loggia del Porcellino - Mercato Nuovo	Monumento
Museo di Palazzo Vecchio	Edificio storico
Corridoio Vasariano	Museo
Gli Uffizi	Museo
Gucci museo	Museo

Figure 2.2: List of cultural POIs, cluster & POIs in an area of Florence. (from (Giglio et al., 2019)).

activities, then the locations he/she visits during a specified time range are assumed to be POIs. For this method to work, the check-in data of every place visited by the user must be logged.

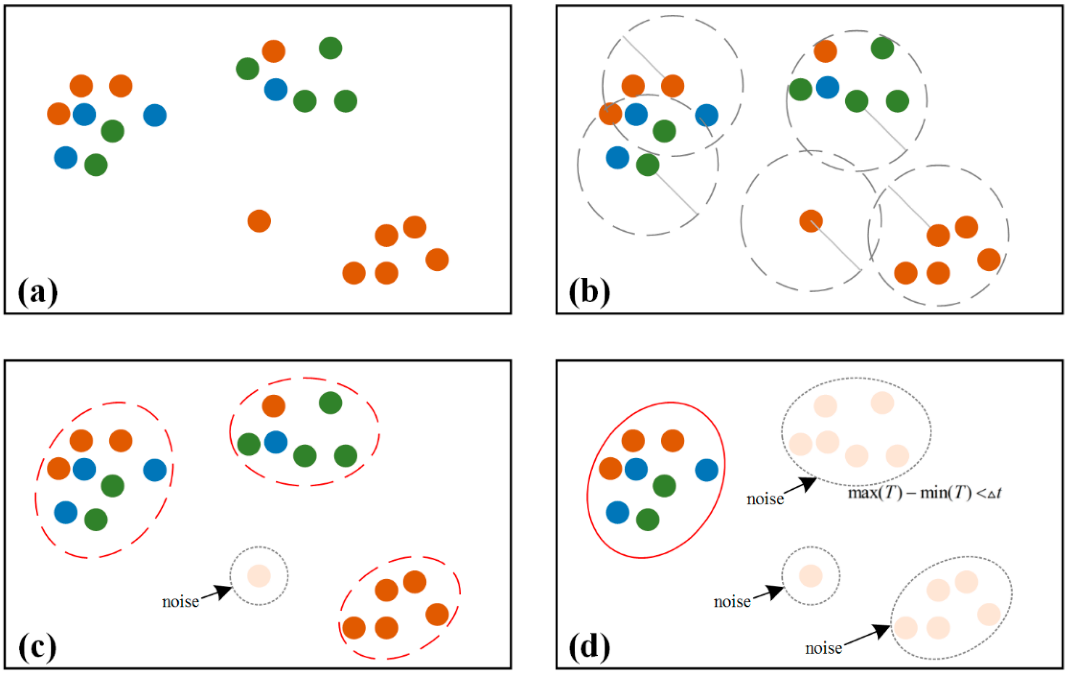


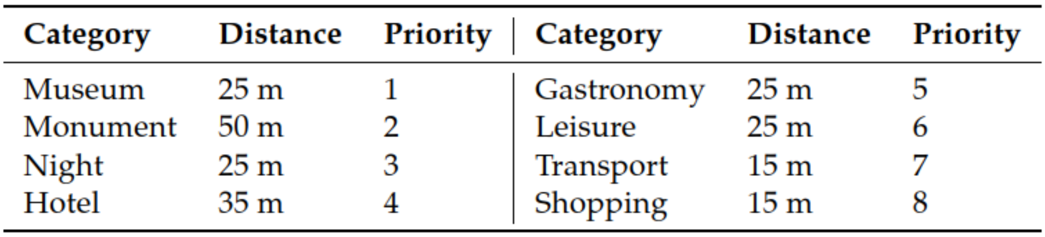
Figure 2.3: Illustration of TU-DJ-Cluster: (a) the dataset; (b) the process of calculating the neighbourhood; (c) initial cluster results; (d) final cluster results and noise points. (from (Han et al., 2020)).

Xue and Zhang (2020) identified every check-in made by a user who visited POIs from a specified category as a POI. Sina Weibo, the microblogging platform used in this analysis, allows users to select specified categories for the places they visit. The authors then identify any place visited by a user as POI if it belongs to the *'Park and Outdoor'* category, or if it was visited on the same day as a POI from the *'Park and Outdoor'* category. The *'Park and Outdoor'* category was a main part of the analysis and was also used to ascertain if a user was a tourist or not.

Proximity-based Identification

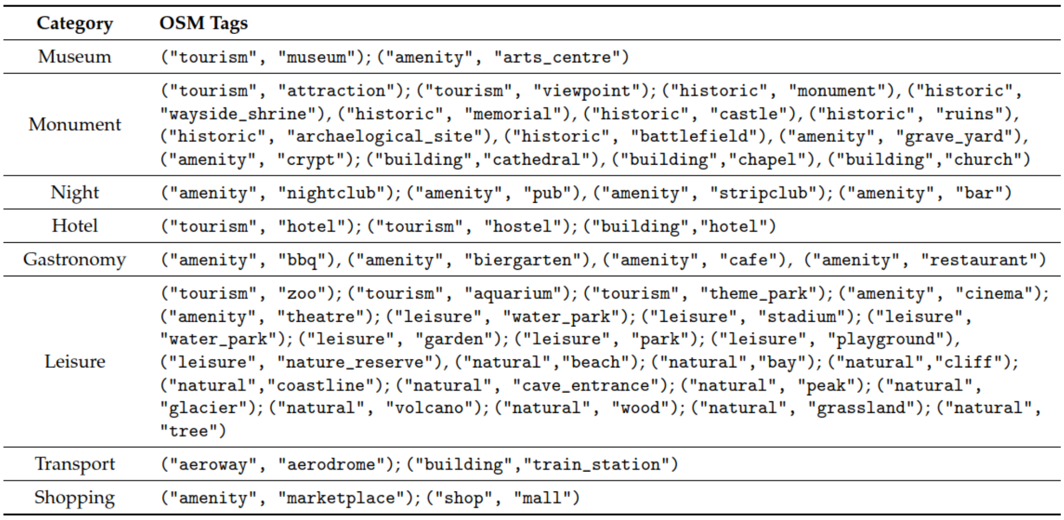
This method aims to select a POI from a list of POIs in the proximity of the user's post location. It works on the simple premise that a tourist will post from a location while experiencing a tourist attraction. While not always correct, this premise is good enough to provide correct identifications. The simplest form of this method is to select the POI closest to the user's post location. While simple, this solution can be erroneous because a tourist might be experiencing multiple POIs simultaneously or just passing by. To remedy this, other filters may be added to remove unlikely POIs.

Bustamante et al. (2019) considered not only proximity to the post location, but also the POI category based on its characteristics. They accomplish this using a priority table (Fig. 2.4), where POI categories are given higher priority based on their association with tourist activities. POIs are queried from the Open Street Maps (OSM) server engine and their descriptive tags are used to categorise them



Category	Distance	Priority	Category	Distance	Priority
Museum	25 m	1	Gastronomy	25 m	5
Monument	50 m	2	Leisure	25 m	6
Night	25 m	3	Transport	15 m	7
Hotel	35 m	4	Shopping	15 m	8

Figure 2.4: Priority table. (from (Bustamante et al., 2019)).



Category	OSM Tags
Museum	("tourism", "museum"); ("amenity", "arts_centre")
Monument	("tourism", "attraction"); ("tourism", "viewpoint"); ("historic", "monument"), ("historic", "wayside_shrine"), ("historic", "memorial"), ("historic", "castle"), ("historic", "ruins"), ("historic", "archaelogical_site"), ("historic", "battlefield"), ("amenity", "grave_yard"), ("amenity", "crypt"); ("building","cathedral"), ("building","chapel"), ("building","church")
Night	("amenity", "nightclub"); ("amenity", "pub"), ("amenity", "stripclub"); ("amenity", "bar")
Hotel	("tourism", "hotel"); ("tourism", "hostel"); ("building","hotel")
Gastronomy	("amenity", "bbq"), ("amenity", "biergarten"), ("amenity", "cafe"), ("amenity", "restaurant")
Leisure	("tourism", "zoo"); ("tourism", "aquarium"); ("tourism", "theme_park"); ("amenity", "cinema"); ("amenity", "theatre"); ("leisure", "water_park"); ("leisure", "stadium"); ("leisure", "water_park"); ("leisure", "garden"); ("leisure", "park"); ("leisure", "playground"), ("leisure", "nature_reserve"), ("natural","beach"); ("natural","bay"); ("natural","cliff"); ("natural","coastline"); ("natural", "cave_entrance"); ("natural", "peak"); ("natural", "glacier"); ("natural", "volcano"); ("natural", "wood"); ("natural", "grassland"); ("natural", "tree")
Transport	("aeroway", "aerodrome"); ("building","train_station")
Shopping	("amenity", "marketplace"); ("shop", "mall")

Figure 2.5: Association between OPENSTREETMAP tags and categories. (from (Bustamante et al., 2019)).

(Fig. 2.5). The priority table is then used to assign a POI to a post if that POI belongs to a category with the best priority and is within a stipulated distance from the post's location. The distance in the priority table ensures that POIs that could be experienced from larger distances are not excluded. Similarly, Mariescu-Istodor et al. (2019) uses the location of users and their GPS trajectory to determine the next POI they are likely to visit. This is done by determining their mode of transport, then calculating their expected travelled distance based on their mode of transport and historical data. Finally, POIs determined by their expected travelled distance are evaluated as possible POIs.

2.2.5 User Profiling in Social Media Analysis

Another important aspect of social media analysis in tourism is user profiling. It is the process of representing or modelling user preferences, behaviour, and characteristics, such that it is possible to split users into distinct groups. This allows researchers to study the behaviour and perspective of different tourist groups, which is beneficial to destination managers when planning activities to target certain groups of tourists. It also makes it easier to recommend activities to members of groups as similar tourists are interested in the same activities (Esmaeili et al., 2020). This section gives an overview of user profiling methods and some relevant articles.

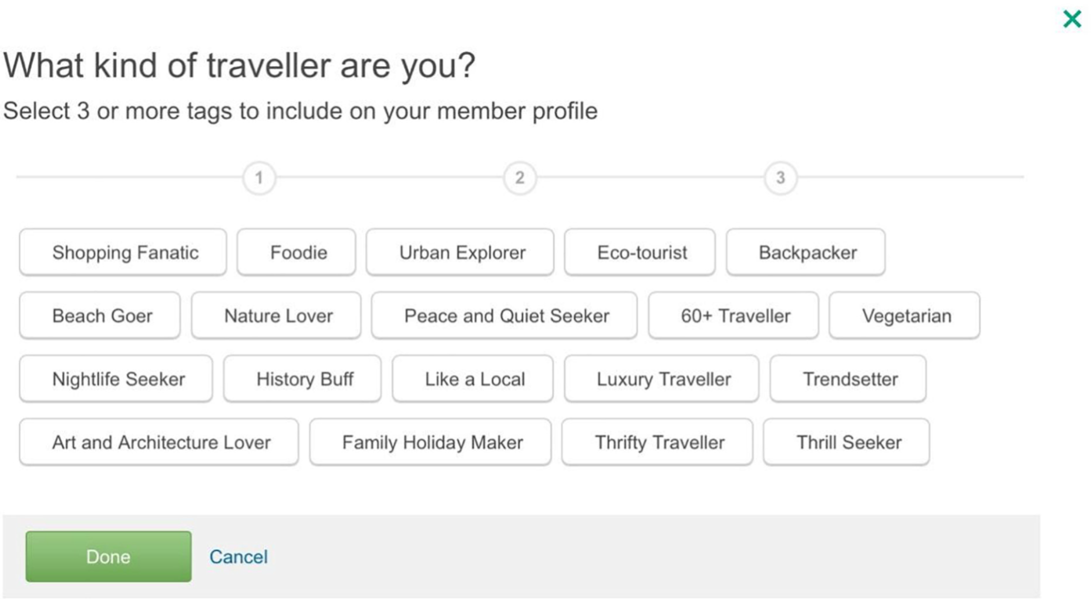
Vector Representation

Vectors are a finite set of n values that determine the position of an entity in a data space. They make it possible to compute the distance relation of multiple entities in an n dimensional space. An entity (user) can be represented as a vector in a data space when it is possible to compute a fixed number of predetermined values from its characteristics or properties. In tourism research, it is possible to represent a tourist as a vector by extracting values that define them, like, their age, the number of attractions they visited, the distance they travelled, etc. These must be chosen carefully to properly represent the main characteristics of the tourist that the research wishes to capture. Representing a tourist in this way makes it possible to compute the distances between them so that smaller distances mean more similarity and larger distances mean more differences.

Bustamante et al. (2019), extracted values from users' Twitter data that included their posting period (emphasizing their activity at the tourist destination), their percentage of posts at tourist attractions in the destination (emphasizing their interest in tourist activities), and finally, their check-ins at different tourist activity categories (like hotels, gastronomy, monuments, etc.). In this case, the goal was to create a clear distinction between locals and tourists, and these attributes worked well to that effect. This allowed the authors to study visitor activities.

Nilashi et al. (2021), represented tourists by a vector of their TripAdvisor rating of attractions. The tourists were then clustered to form groups of tourists with similar interests in attractions. High Order Singular Value Decomposition was further used for dimensionality reduction to help with estimating the nearest neighbours of tourists in the data space. Finally, regression trees were used to learn the nearest neighbors' preferences to predict tourists' overall ratings on specific spa hotels at a destination. Similarly, Pantano et al. (2019), represented tourists by a vector extracted from their TripAdvisor ratings and reviews of attractions linked to certain topics. 19 topics (Fig. 2.6) provided by TripAdvisor were considered. They are linked with attractions that embody the characteristics of the topics, e.g., the topic *Shopping Fanatic* is linked to shopping malls, the topic *Foodie* is linked to restaurants, etc. A support vector machine (SVM) model was used to predict the users' interests in these topics based on their review data and their initial selection of 3 topics required when registering to TripAdvisor. The final output of the SVM is an 19-digit binary vector in which 0 represents 'no interest' and 1 represents 'interest'.

Sometimes, the vector representation can be a mix of numeric and categorical variables. Categorical variables, unlike numerical attributes, have a fixed number of possible values, like the gender of a tourist. Torres-Ruiz et al. (2018), extracted vectors from different data sources including online reviews of museums and physical sensors at museums to represent the interests of users in museums. The



What kind of traveller are you?

Select 3 or more tags to include on your member profile

Figure 2.6: The 19 possible characteristics defining a tourist to be chosen to create a profile on TripAdvisor. (from (Pantano et al., 2019)).

vector included the number of visits to the museum (to assess its popularity), the museum type, and the polarity of users' opinions about the museum (positive or negative). Users were classified based on the types of museums they visit. Similarly, Logesh et al. (2019), constructed two user profiles, one using mixed variables representing the demographic profile of a user (gender, employment, age), and the other based on numerical variables representing the preference profile of a user. Their proposed system *ABiPRS* then used the fuzzy C-means algorithm to group similar users considering both profiles.

Fararni et al. (2021), constructed four user profiles based on the type of data mined from the user. A *content-based* user profile contained keyword vectors representing interests in activities, which are explicitly provided by the user or extracted from past visits to attractions. A *collaborative* user profile contained user ratings of activities presented to them. A *social* user profile contained the user's relationships with other individuals. Finally, a *demographic* user profile contained their demographic data like age, profession, geographical location, etc. Depending on the mode of accessing the system, the appropriate user profile was constructed and used to rank a list of attractions to be recommended.

Although it is not very common, a user may be represented by a vector of words. These words could then be modelled with a numerical vector using *Word2Vec*(Mikolov et al., 2013) to allow computing similarity between users. Abbasi-Moud et al. (2021), represented tourists by their frequently used nouns, which are semantically related, and which portray their sentiments. This was accomplished by extracting nouns from tourists' TripAdvisor reviews after pre-processing them. Then these nouns were clustered by their semantic relationships, and finally the clusters of nouns that best portray a user's sentiment about an attraction were selected. The final user representation was a vector of nouns that were matched with features extracted from attractions to evaluate their appeal to the tourist.

Modelling

In modelling, users are represented by mathematical models based on their behaviour or in textual content extracted from their social media data. The benefit of mathematical models is that they make it much easier to compute the similarity between users, which may be used to group users or to locate nearest neighbours for a recommender system.

LOOKER (Missaoui et al., 2019), utilised statistical language modelling to determine users' interests in tourism-related activity categories from their online reviews. Users' positive reviews posted on TripAdvisor were selected using a positive filtering strategy that considered the *ratings* associated with the reviews or the sentiment portrayed in the text. These reviews were then classified under the tourism-related activity categories including food, shopping, health, and attractions. Finally, a multi-layered language model was constructed to represent the user, where each layer was associated with one of the tourism-related activity categories. The user profile was then combined with a content-based filtering algorithm to recommend attractions to tourists.

Massimo and Ricci (2018), defined a method for constructing user behaviour models based on trajectory data. To accomplish this, a user's trajectory was modelled as a Markov Decision Process (MDP), and the Maximum Log-likelihood Inverse Reinforcement Learning (MLIRL) algorithm was used to compute the optimal reward function of MDP. In addition, they constructed a generalized user behaviour model to solve the *cold start* problem, where user trajectories were clustered using Non-Negative Matrix Factorization (NMF) which represents trajectories as documents with features as terms. After clustering, the specific reward functions of clusters were computed using MLIRL. In recommendation, a new user's current trajectory modelled using MDP was combined with the generalized user behaviour model to determine the next POI in the sequence to visit.

Although they are not entirely mathematical models, ontologies are knowledge-based graphical or hierarchical representations of a domain. It is possible to profile users using ontologies built on a particular knowledge domain. Moreno et al. (2013), built an ontology in the tourism domain with five levels of hierarchy, and concepts representing different abstractions of touristic activities. They then profiled users with a static vector containing demographic and contextual information, and a part of the tourism ontology, with assigned degrees of interest to each concept that interests individual users.

Neural Networks (NNs)

A neural network is an advanced supervised machine learning technique that models the neurons and synapses in the human brain. It learns by adjusting certain weights that determine if a set of interconnected artificial neurons are activated or deactivated to provide a desired vector output. Although it is not a popular technique in tourism research, it is possible to capture the preferences of users from features extracted from their social media data using NNs.

Sertkan et al. (2020), used a type of NN called Convolutional Neural Network (CNN) which is specialized in image decomposition, to learn the preferences of users based on a set of pictures selected from a larger set in a specific order that matches their preferences and interests in touristic activities. CNNs were trained to score the user's selections according to seven factors (*Sun & Chill-Out*, *Knowledge & Travel*, *Independence & History*, *Culture & Indulgence*, *Social & Sports*, *Action & Fun*, *Nature & Recreation*) that represented his/her character in terms of touristic activities. The final user representation was an aggregate of his/her score in these factors.

Relationship Representation

Although not as technical as other methods, it is possible to profile users solely by their relationships. It could be their relationship with other users in a social context or common relationships with a geographic location or concept. Esmaeili et al. (2020), based user similarity on the homophily principle, meaning that users in a community were seen as similar. They used a community detection technique to determine communities from social media data. If a user did not belong to any community, then demographic data were employed to find similar users. The user profiles were then used to predict the users' ratings based on the most similar users.

2.3. Contextual Recommendation in Tourism

2.3.1 Traditional Recommender Systems

Recommender systems have become staple software tools in this information age. They aid people in making decisions, ranging from products to buy, movies to watch, songs to listen to, places to visit, and many more. Recommender systems excel in cases where there are numerous options to choose from that appeal differently to the decision maker and are also affected by his/her context and environment. For this reason, they have been successfully applied in e-commerce, tourism, video, and music streaming services.

Primarily, recommender systems aim to personalise and streamline a user's options in order to simplify the decision-making process. Traditional recommender systems accomplish this by modelling the relationships between items and users as a user-item matrix and operating as a 2-dimensional system (Adomavicius and Tuzhilin, 2001). The relationship modelled in the user-item matrix is usually the *rating* a user gives to the items considered in the recommendation process, but it may include other content. The accepted traditional techniques used in recommender systems are briefly described below as proposed in Burke (2002):

- *Collaborative filtering recommender systems* (Schafer et al., 2007) work on the premise that users that liked similar things in the past will continue to like the same things. Items are filtered by user ratings of similar items and the collaborative ratings of similar users. Usually, a user-item matrix is built with cells containing each user's rating for a particular item, which are left empty when a rating is missing (i.e., an item has not been evaluated by a user). Then the RS predicts the missing ratings based on the ratings of similar users.

- *Demographic-based recommender systems* (Al-Shamri, 2016) suggest items based on the demographic information of the user. They work on the premise that people belonging to the same demographic prefer the same things due to their influence on one another. DRSs use age, gender, marriage status, country of origin, hobbies, etc. For instance, in e-commerce, a woman will be suggested gender-specific items different from those suggested to men. Demographic data of users are a crucial aspect to consider in recommending items; thus, if present, they should be incorporated into every recommender system.

- *Content-based recommender systems* (Pazzani and Billsus, 2007) suggest items based on textual data that describe or differentiate them from other items. Items are suggested to a user when their descriptions match the descriptions of previously liked items. This RS is applicable in e-commerce where products have descriptions that highlight their use, quality, and features. The accuracy of content-based RSs hinges on adequate representations of items by their descriptions and adequate user profiles created from previously liked items.

- *Utility-based recommender systems* (Burke, 2002) suggest items based on their utility to the user. This is done by defining a user-specific utility function that evaluates the utility of each item to a user's needs. The utility function is usually a mathematical model that can incorporate not only item-related attributes but also attributes about the user and the service provider (or vendor).

- *Hybrid recommender systems* (Burke, 2002) combine multiple techniques to overcome the shortcomings of individual techniques for better recommendations. Collaborative filtering is the most combined technique, because of its popularity and wide acceptance. Hybrid recommender systems combine techniques in various ways, including but not restricted to, 1) weighted aggregation of item ranking provided by the combined techniques, 2) switching between techniques when appropriate, 3) refining recommendations of one technique with another, and 4) suggesting items from combined techniques simultaneously.

While suitable in most cases, traditional recommenders do not consider other factors outside of the user-item system, that are important in making relevant suggestions. For instance, in suggesting travel locations to tourists, it is necessary to consider weather conditions, seasons, availability of location, etc. These factors are not in view when considering just the relation of a tourist to a location in terms of utility, rating, and preferences. Adomavicius and Tuzhilin (2001) proposed a shift from the 2-dimensional user-item system to a multiple-dimensional system based on the context surrounding the user, item, and environment, specific to the domain. The embodiment of this proposal has been termed *Context-Aware Recommender Systems* (CARS).

2.3.2 Context-Aware Recommender Systems (CARS)

The Context in CARS is too broad to put into one definition, but in simple terms it can be defined as any circumstance that might play a role in whether a suggested item is appealing to a specific user during consumption. Context could be *static* (not changing over time, e.g., intent) or *dynamic* (variable with time, e.g., season), related to the user (e.g., mood, budget) or product (e.g., discounted price), or non-specific (e.g., weather conditions). In theory, as reality is riddled with multiple ever-changing variables that affect our daily decisions, it is the same in contextual recommendations. CARS extract context and user preferences from user histories and combine them into a working recommender system model that would not only match users to items but also ensure the context is ideal.

As it is difficult to account for all possible variables around a decision, some specialised recommender system categories that adapt to specific contexts have been established and successfully applied in several domains. They include but are not limited to:

- *Time-aware recommender systems* (TARS), (Campos et al., 2014) adapt the temporal context into the recommendation process. TARS are popular due to the role *time* plays in many situations that require a recommender system. As a result, this category has become integrated into the general CARS category, and it has lost its specialty. This RS is useful in any case where time in any of its forms could affect user decisions. For instance, a Christmas song recommendation from a music RS could be preferable in the Christmas season.

- *Session-based recommender systems* (Wang and Kirilenko, 2021) adapt the user-based context of *session*. They consider only the short-term history of user transactions when filtering items for recommendation. Particularly, only transactions logged within the same session are considered in the recommendation. This RS is useful in cases where past sessions of user interactions with the system are not very relevant to current sessions. For instance, a web search engine should only recommend pages based on the current search of the user and not on previous searches.

- *Sequence-aware recommender systems* (Quadrana et al., 2019) adapt the context of the *sequence* in which users interact with items into the recommendation process. Item sequence is very essential in cases where the order of user interactions matters and can be used to enhance recommendations. For instance, the previous song consumed in song recommendation is relevant in recommending the next song. This idea is also applicable to tourism recommender systems, where tourist activities visited in the right order enhance the overall experience. Sequence-aware recommender systems are different from recommenders that suggest a set of items without a specific order of interaction. They either suggest a set of ordered items or suggest the next logical item considering the previous item. For instance, in tourism, there is a group of recommender systems called Next-POI recommenders that suggest the next logical POI to visit from the current visited POI (Massimo and Ricci, 2021).

- *Trust-aware recommender systems* (Massa and Avesani, 2007) adapt the *trust* of users into the recommendation process. They are based on the premise that users in a social network are likely to prefer items suggested by people they trust (e.g., family members, friends, influencers, and mentors). If this premise is considered true, and trust can be quantified, it means that a user would prefer items that are highly rated by their most trusted connections. Trust-aware RSs are applicable in just about any situation, including movie, song, place, commodity recommendations, etc.

- *Group recommender systems* (Felfernig et al., 2018) adapt the context of *companionship* or *social relationships*. They focus on suggesting items that represent a group's combined interests. They are applicable in situations where people socialise, like tourism (Garcia et al., 2011), movie streaming, etc. The preferences of users in a group are aggregated into a single profile that represents the combined interests of the group. Then a regular RS suggests items using the aggregated profile. Group RSs are invaluable in tourism, where people visit attractions in groups (like families, co-workers, friends, etc.) (Garcia et al., 2009).

Apart from these categories described above, other specialisations exist that, despite not being as popular, are relevant in some cases (e.g., mood-aware recommender systems). In fact, it is possible to introduce any context as a recommender system specialisation. The latest works on tourism recommender systems are in some way contextual. In the next subsection, we discuss some contexts applicable in the tourism domain.

2.3.3 Context Awareness in the Tourism Domain

The tourism domain is overwhelmed by many possible circumstances that could affect the decisions of tourists. These circumstances can be modelled by recommender systems as contexts. As a result, most contexts applicable in other domains exist in tourism recommender systems. We can outline the following relevant contexts in the tourism domain, some of which can be found in Haruna et al. (2017).

- *Temporal context.* As previously mentioned, time is a very important context in many recommender system domains. The particular moment an item was consumed could have played a major role in why it was chosen. Time in recommender systems can be modelled as a continuous contextual variable or as a categorical contextual variable (Campos et al., 2014). When modelled as a continuous variable, the specific time items are consumed is considered, while as a categorical variable, the season (summer, autumn, etc.), time of day (morning, evening, etc.), day of the week (Monday, Tuesday, etc), or month of the year (January, February, etc.) are considered. In tourism, it makes sense to model time because certain attractions can only be fully experienced in certain seasons or times of the day. For instance, the *Northern lights* in Norway can only be seen at certain times of the year. The temporal context is also related to the *sequence* and *session* contexts because sequences are progressions over time while sessions are time restrictions (Kolahkaj et al., 2020; Fogli and Sansonetti, 2019; Pan and Zhang, 2019; De Pessemier et al., 2013; Bahramian et al., 2017; Biancalana et al., 2013; Yuan et al., 2015; Korakakis et al., 2017; Afsahhosseini and Al-Mulla, 2021).

- *Weather as context.* The weather is another important context, especially important when suggesting outdoor activities. This makes it very applicable in the tourism domain. Weather in tourism recommender systems is usually modelled by considering different weather conditions like rainy, sunny, windy, snowy, etc. Due to decent weather recording and forecasting bodies, it is possible to match past user records to different weather conditions and suggest items based on forecasted weather conditions. There are also cases where it is not necessary to check previous records as some tourist attractions cannot function in certain weather conditions. For instance, picnics in the park are not feasible in a snowstorm, but skiing trips require snow (Fogli and Sansonetti, 2019; De Pessemier et al., 2013; Kashevnik et al., 2017; Biancalana et al., 2013).

- *Location as a context.* The geographical location of a user is another useful context in the tourism domain. Tourism recommender systems usually suggest places or attractions at different geographical locations within a destination or city. To experience these suggested attractions, tourists must move from their current locations which requires time, effort, and the means to do so. This means that the location of users and items must be modelled into the recommender system. Location can be modelled such that only places within a certain distance from the tourist are suggested to them, or an optimal route to visit a group of POIs can be created after matching them with user preferences (Moreno et al., 2013; Kolahkaj et al., 2020; Fogli and Sansonetti, 2019; Pan and Zhang, 2019; Bagci and Karagoz, 2015; De Pessemier et al., 2013; Benouaret and Lenne, 2015; Bahramian et al., 2017; Kashevnik et al., 2017; Levi et al., 2012; Biancalana et al., 2013; Yuan et al., 2015; Korakakis et al., 2017).

- *Budget as context*. The tourism domain, like most domains, is commercial, so it is important to consider the financial capabilities of tourists when suggesting items. Taking into account the budget of tourists when suggesting attractions will ensure they are not rejected for being too expensive. Budget is modelled as a constraint in tourism recommenders (Afsahhosseini and Al-Mulla, 2021).

- *Companionship as context*. As described in the previous section, the context of companionship has been formulated into a specialization of CARS called Group RS. This context is also very important in the tourism domain because it can play a major role in selecting places to visit. For instance, a new married couple on their honeymoon trip would visit romantic sites and engage in romantic activities while a married couple with young kids will visit amusement parks and leisure sites. This context is also modelled as a means of grouping tourists by their companionship (Moreno et al., 2013; De Pessemier et al., 2013; Levi et al., 2012).

- *Intent as context*. The knowledge about the intent of a user is very valuable in the tourism domain. In essence, if we know what a tourist wants from his/her trip it is very easy to suggest places and activities to them. For instance, a tourist that came to visit Barcelona to see the Sagrada Familia can easily be suggested other buildings designed by Antoni Gaudí, the architect of Sagrada Familia. They could also be suggested attractions within the proximity of Sagrada Familia because of the convenience of location. Intent can be modelled as the theme of the trip and all other attractions suggested are made to fit that theme (Levi et al., 2012; Biancalana et al., 2013).

These contexts are just a few possibilities and could be extended as far as the situation requires. It is also important to note that some can be inferred from previous user records, others can be gathered from physical sensors and devices, and others must be provided explicitly by the tourist. In any case, it is tedious to incorporate all possible contexts into one recommender system but incorporating a few can enhance recommendation accuracy significantly.

2.3.4 Artificial Intelligence Techniques in Tourism Recommender Systems

Artificial Intelligence techniques are commonly used in recommender systems in different aspects of the recommendation process, including preference learning, similar user identification, item sequence optimization, item relationship identification, item/user knowledge representation, etc. In this subsection, we briefly outline the common AI techniques used in tourism recommender systems, in line with Borràs et al. (2014) and Renjith et al. (2020).

Multi-agent systems

Agent-based systems utilise independent intelligent computer programs called agents, to perform specific tasks in the interest of a user. In the case of recommender systems, the agent works to fulfil a certain goal necessary in the recommendation process. This process contains several different phases and tasks to accomplish which is more suited for a multi-agent system. Multi-agent systems consist of more than two agents performing different tasks while in communication with each other to accomplish the main goal efficiently.

MARST (Bedi et al., 2014) is a multi-agent system designed to make hotel, restaurant, and place recommendations to a tourist. The system consists of agents for collecting data from web services, finding similar users, evaluating the reputation of items to be recommended, collecting information about a user, and for making different recommendations. The recommendation stage includes three agents for hotel, restaurant, and place recommendations. The agents are split into those that work offline and online. Offline agents collect the data, find similar users, and evaluate and store the reputations of items, while online agents collect user preferences and use the data store offline to make recommendations.

Turist@ (Batet et al., 2012) is an agent-based tourist recommender system designed to suggest activities to tourists that have already arrived at a tourist destination. The system consists of several agents performing different tasks. A user agent requests recommendations for a user, a broker agent acts as a go-between for the user agent and activity agents, activity agents that manage the information of specific attractions or categories of attractions, and finally a recommender agent that returns recommendations to the user based on their preferences and profiles. The system is built to be completely modular and easy to expand.

Similarly, Sebastia et al. (2010) introduces a multi-agent tourism recommender with a user agent that not only collects a user's demographic and preference data, but also updates his/her preferences based on selected activities. The recommender agent (Generalist Recommender System Kernel agent in Fig 2.7) manages the recommendation process by invoking different RS agents (content, demographic, etc.) to filter activities for recommendations. Finally, a planner agent creates a visitation plan from the activities selected by the user, considering their availability, time, and location. The architecture of the system is shown in Figure 2.8.

A big advantage of multi-agent RS is the ability for agents to run simultaneously, making it possible to adopt the full potential of distributed computing in recommendations (Borràs et al., 2014). In addition, any task in the recommendation process can be designed as an agent which makes the system to be highly modular and easily scalable (meaning existing agents could be easily replaced or new agents could be added) (Sebastia et al., 2010).

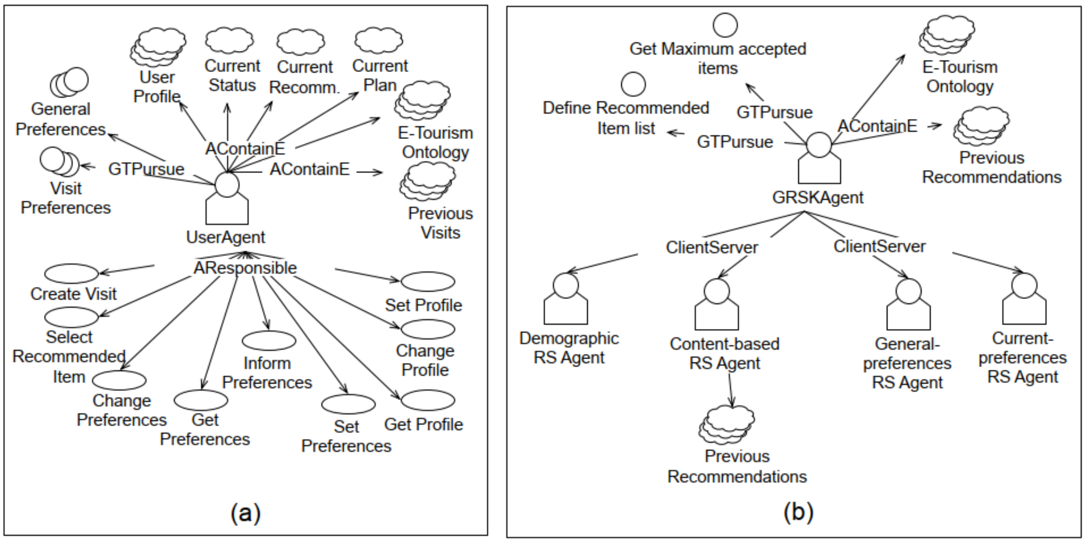


Figure 2.7: (a) Agent diagram of the UserAgent and (b) Agents of the GRSK organization. (from (Sebastia et al., 2010)).

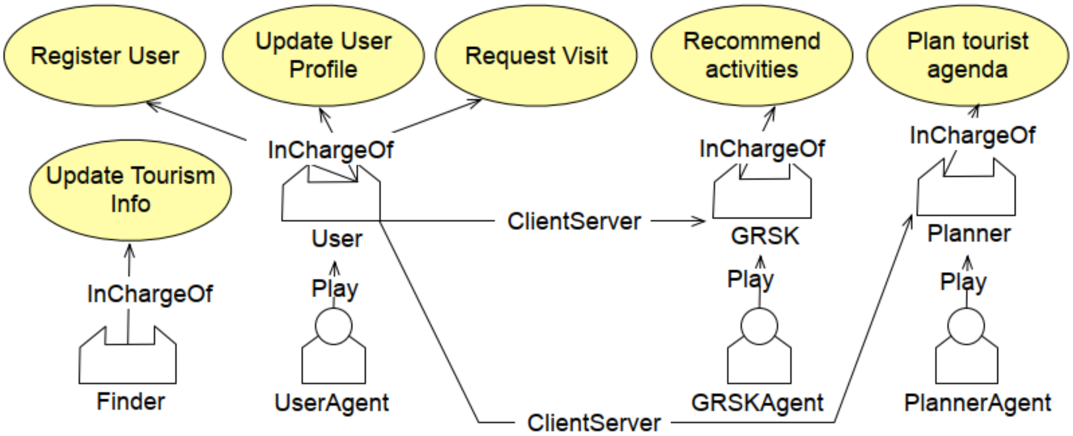


Figure 2.8: Organization architecture and use cases of the e-Tourism system. (from (Sebastia et al., 2010)).

Automatic Clustering

A big part of recommender systems is the premise that similar users are interested in the same things. Most tourism recommender systems include a user profiling stage as detailed in Section 2.2.2, in which users are defined by their demographic data or preferences, and similar users are grouped. This grouping process is the essence of clustering in the recommendation process. Automatic clustering techniques can be used to create distinct groups of tourists to recommend them the same activities.

Most automatic clustering techniques are applicable in tourism recommender systems. Nilashi et al. (2017, 2018) designed a clustering ensemble that combined Self-Organsing Maps (SOM) and Expectation Maximization (EM) clustering methods to group tourists by their ratings of activities for recommendation. This ensemble performed better than recommendations based on a single clustering method. Similarly, Arvianti et al. (2019) grouped tourists by their ratings using the k-means algorithm. They also computed the Pearson cosine similarity of items so that the ratings of users may be predicted for similar items in the recommendation process.

State of the art classification techiques can also be used for clustering in tourism recommender systems. This is because classification techniques assign users to classes which are considered as groups. Riswanto et al. (2019) used the k-nearest neighbour (KNN) algorithm to group tourists by their ratings of food, places, environments, and services in a collaborative filtering based tourism recommender system. Wang et al. (2012) explored Naive Bayes classifier, Bayesian Networks, and Support Vector Machine (SVM) for grouping tourists by their demographic data, preferences, and ratings. The tourist groups are used to predict the ratings of new users and recommend items to them.

Some works have also used clustering techniques to group POIs together by their proximity. The k-means algorithm was used to build daily tourist itineraries where the number of clusters are the number of available days the tourist has to visit places and the geographical location determines the POIs in each cluster (Tlili and Krichen, 2021; Tenemaza et al., 2020).

Knowledge Representation

Knowledge representation techniques are very useful in recommender systems. They provide a means to construct a structured knowlegde base for the recommender domain knowledge. In the case of tourism recommender systems, the domain knowledge includes activities, tourist attractions, tourism areas, etc. It could also include the relationships between these items and their categories. An efficient knowledge representation will help the recommender system to understand rules of the domain.

Ontologies are the foremost methods for knowledge representation. They include classes, sub-classes, and instances in a hierarchical structure that makes it easy to navigate the domain knowledge. Moreno et al. (2013) developed an ontology for the tourism domain. Their system *SigTur* used the ontology to define and categorise touristic activities and store the preferences of users, which are then used to make recommendations to users. *SigTur* also dynamically manages users preferences by propagating them down the hierarchy such that the preferences are distributed from parents to children. Similarly, *e-Tourism* (Sebastia et al., 2010) an ontology-based multi-agent tourism recommender system, also constructs an ontology on the tourism domain to describe the features of touristic activites. Their ontology includes a taxonomy of features commonly managed in the tourism domain, and specific activities are described using these features. Abbasi-Moud et al. (2022) took the ontology construction further by introducing fuzzy weights to relationships between parent classes and child classes. Their ontology represents in a more real way the relationships of touristic activities and categories, and also provides a more generalised representation of the tourism domain.

Uncertainty Management

Uncertainty is an important part of real life decision making. Many aspects of choosing an item for consumption are not clear cut but rather apparent in degrees of dependence on several factors. In the same sense, recommendation of touristic activities needs to incorporate the uncertainty of relationships of activities and tourists. One common way to represent uncertainty is fuzzy logic, which models uncertainty with fuzzy variables that express the degree of membership of domain aspects as fractional values between 0 and 1. Many parts of the recommendation process can be represented in fuzzy logic. For instance, the clustering of tourists can be modelled with fuzzy logic such that a tourist doesn't belong to one cluster but is a member of multiple clusters to certain degrees (fuzzy C-means clustering (Selvi and Sivasankar, 2017)). Fuzzy logic could also be applied in knowledge representation (Abbasi-Moud et al., 2022) and in the definition of contextual information (Tiwari and Kaushik, 2015).

Optimization Techniques

In tourism recommender systems, optimization techniques are used primarily for planning and optimizing travel routes. This is an important task in activity recommendation because it can improve the tourist overall experience when visiting new places. Several factors must be considered when optimizing routes like time, location, distance, weather, traffic, etc., which complicates the optimization process.

Researchers have applied different types of optimization techniques that might not provide the optimal solution but solutions close to the optimal. One of such algorithms is simulated annealing, which is a probabilistic metaheuristic for approximate global optimization of a large discreet optimization problem. It is useful in finding an approximate solution to an optimization problem that is close to the optimal solution. Tlili and Krichen (2021) proposed *StayPlan* a tourism recommender system that combines k-means and simulated annealing to build optimal travel routes for tourists. In contrast, Forouzandeh et al. (2022) explored an *artificial bee colony* algorithm that simulates the foraging abilities of bees to suggest preferred touristic places that are optimised to the location of the tourist for efficient touring. Tenemaza et al. (2020) used a *genetic algorithm* which simulates the survival of the fittest property of Charles Darwin's theory of evolution, to build daily tour itineraries for tourists considering the distance and opening times of attractions. Similarly, Hang et al. (2018) proposed a genetic algorithm for route optimization in a city. The optimised routes incorporated the time, date, season, and frequent item sets of POIs extracted from past tourist visits. Genetic algorithms in general are quite popular for the route optimization task and they perform well with decent run time. In Chapter 3, we utilise a multi-objective genetic algorithm to optimise routes.

Prediction Techniques

Prediction techniques attempt to predict a dependent variable based on one or more independent variables. They are useful in studying the relationships between variables in a dataset. In tourism recommender systems, they are used to predict tourists ratings of touristic activities by analysing previous ratings of similar activities (in collaborative filtering based recommender systems). This is useful because it provides a method to learn tourist preferences and recommend ideal touristic activities. Most classical classification techniques like KNN, Naive Bayes, SVM, Classification and Regression Tree (CART), etc., can be used as prediction techniques by switching their output from binary (0 or 1) to fractional [0-1]. For instance, Nilashi et al. (2018) used CART to predict ratings of clusters of tourists from their preferences extracted from TripAdvisor. The predicted ratings were a part of their collaborative filtering based tourist recommender system.

Association Rules Techniques

Association Rule Mining or learning (ARM) is an AI technique used to mine frequent itemsets i.e. items frequently consumed together from a dataset of transactions. ARM can be applied in tourism recommender systems to mine association rules containing POIs with high affinity from a dataset of tourists previous visits. ARM utilises different algorithms for mining frequent itemsets, of which the most popular are Apriori and FP-Growth. Most ARM algorithms provide the same results with similar run time so any one algorithm can be used in implementation.

Association rules mined from tourist previous visits can be used to recommend places to new tourists. Viktoratos et al. (2018) mined association rules from tourist check-in data in New York City. Their goal was to alleviate the cold-start problem detailed in chapter 1 by mining rules from previous tourists to augment those mined from a new user. They scored rules by their relevance to the new tourist and recommended the top N-rules. Similarly, Lou (2022) proposed an association rule based tourism recommender system, that recommends rules mined with a modified ARM algorithm that introduces weights for interesting tourist characteristics. In the same line, Hang et al. (2018) used ARM to mine frequent routes incorporating contextual information about the tourist in the process. The mined rules are then ordered using a genetic algorithm.

Association rules can also be used for classification in a tourism recommender system. The idea is that similar tourists would have similar association rules minied from their previously visited attractions. This means that classes of tourists could be built based on this similarity, and new tourists could be assigned to a class during the recomendation process (Lucas et al., 2013).

In Chapter 3, we also make use of the ARM technique to mine POIs with high affinity for our recommendation process.

2.4. Social Media-Based Tourist Flow Analysis

Tourist flow analysis studies the directional spatial distribution of tourists traversing a destination, to detect unique mobility patterns within the destination. Tourists are usually on the move, whether they are arriving at a new destination or visiting attractions within the same destination. This can be difficult with a lack of proper information and adequate tourism management. Destination Management Organizations (DMOs) are charged with enhancing tourists' experiences and improving their management. Progress in this area requires DMOs to know who their tourists are, what needs and travel preferences they have, what they visit the most, and which are their mobility patterns and flows around the destination (Xiang and Fesenmaier, 2017). Nowadays, it is possible to extract this information from Geotagged Social Media Data (GSMD) and analyse them to know the movements or flows of tourists (Li and Law, 2020).

In the following subsections, we briefly review the evolution of tourist flow analysis in relation to social media analytics and the identification of mobility patterns in tourist clusters presented in the latest articles on these topics.

2.4.1 Tourist Flow Analysis and Social Media Analytics

Prior to recent technological developments, studies on tourist mobility were based on tourist surveys (Salas-Olmedo et al., 2018), and they were rather limited. Tourist flow analysis grew enormously with the development of tracking mechanisms like GPS, cell-tower identification, or Wi-Fi positioning (Jin et al., 2018), which has made it possible to obtain big data from the movement of tourists at destinations. Several studies have used GPS (Orellana et al., 2012; Edwards and Griffin, 2013; Shoval et al., 2011) to find out which were the most visited places in a particular destination and when they were visited. In destinations, the proliferation of sensor networks and portable devices like smartphones has also made it possible to obtain big data from tourists and to know their movements or flows (Chua et al., 2016). In general, the increasing effectiveness and reliability of GPS data and mobile positioning data have increased the possibilities of analysing spatial-temporal behaviours, widening the research objectives beyond the initial aim to know where and when visitors went. In this vein, they have been used to identify seasonal demand patterns by Ahas et al. (2007) or to improve the management of destination marketing by Kuusik et al. (2011). Other authors have considered data obtained from mobility services, such as the subway smart card (Roth et al., 2011) or bike sharing systems (Beecham et al., 2014).

Social media also has very useful platforms for knowing the movements of tourists. Research in social media analytics (SMA) has advanced heavily since 2014, but it is still in an early stage of development (Mirzaalian and Halpenny, 2019). The potential of social media as sources for big data research in the field of tourism has increased in the last decade (Xiang and Fesenmaier, 2017; Mariani et al., 2018), and studies on big data, social media, UGC (User-Generated Content), and online reviews have proliferated in hospitality and tourism (Mariani et al., 2018; Mirzaalian and Halpenny, 2019; Li and Law, 2020; Li et al., 2018b; Lu and Stepchenkova, 2015). Moreover, the innumerable footprints that millions of tourists leave online using technological platforms constitute an interesting source for knowing the tourists' movements and flows (Lu and Stepchenkova, 2015; Provenzano et al., 2018), although big data-based theoretical studies still remain limited (Li and Law, 2020).

As described in Section 2.2.2, there are several analytical methods applied to social media data in the tourism domain. For example, text analytics and data mining studies try to find out tourists' interests and predict their decisions and behaviours (Sohrabi et al., 2020). Also, trend analysis studies also try to predict tourists' behaviours (Vecchio et al., 2018) or future trends in tourist behaviour at destinations (Pantano et al., 2017). Nevertheless, one of the best ways to know the behaviour of tourists during the trip is through spatial data analysis.

In tourism research, spatial data analysis studies the distribution of tourists at a destination (Huang et al., 2017), but to understand tourists' flows, their directionality and displacement on a map are studied (Chua et al., 2016; Önder, 2017; Jin et al., 2018). Flow is the collective movement of people (Chua et al., 2016), and flow analysis shows the movement of tourists in a location (Miah et al., 2019). GSMD are key sources of information to analyze tourists' flows (Hawelka et al., 2014) in order to uncover their travel preferences and behaviours (Chua et al., 2016) or the tourists' experiences through a directional spatial analysis (Zhang et al., 2020). Provenzano et al. (2018) compared GSMD results with the UNWTO record-based network demonstrating the usefulness of GSMD to discover tourist flows.

GSMD analysis allows for knowing the dispersion of tourists and the routes and activities they carry out in the destination (Li et al., 2016; Önder et al., 2016; Orsi and Geneletti, 2013; Vu et al., 2018; Wood et al., 2013; Zhou et al., 2015), their density of movements (García-Palomares et al., 2015), their flows (Chua et al., 2016; Miah et al., 2019; Cheng and Edwards, 2015; Miah et al., 2017), and the most popular resorts, attractions, or points of interest in the destinations (Hu et al., 2019; Zhou et al., 2015; Chen et al., 2011; Zanker et al., 2009). Like other analytical methods, Twitter and Tripadvisor are the two most used GSMD platforms for tourist flow analysis, because they allow for the analysis of multimodal data such as User Generated Content (including text, images, and even videos), and geotagged information (Mirzaalian and Halpenny, 2019; Jurdak et al., 2015). However, studies based on geotagged photos and other social media like Flickr (Miah et al., 2019; Barchiesi et al., 2015), Foursquare (Vu et al., 2018), or Instagram (Ma et al., 2020) have also proliferated showing tourists' flows, movements, and behaviors in destinations.

It has also been shown that analysing different social media sources or platforms is useful because they provide complementary information and enrich the knowledge of different tourists' movements (Salas-Olmedo et al., 2018). In this line, Dietz et al. (2020) analysed tourists' movements at destinations through three social media (Twitter, Foursquare, and Flickr) and identified different types of trips according to the origin of the tourists. A study by Sugimoto et al. (2019) combined tracking technologies and surveys to study the relationship between visitor mobility and urban spatial structures. Salas-Olmedo et al. (2018) analysed the digital footprint of urban tourists through photos, check-ins, and tweets from three social media (Panoramio, Foursquare, and Twitter). In addition, they used a clustering methodology to identify certain areas of the destinations according to the tourist activities that visitors carried out in them. However, they did not cluster or segment tourists to know their different preferences, movements, and behaviours.

Many studies on GSMD have focused on tourists' mobility patterns (Orsi and Geneletti, 2013; Gabrielli et al., 2015; Li et al., 2018a; Wu et al., 2018). However, the difference in mobility patterns between sub-groups or clusters of tourists has not been fully researched (Liu et al., 2018).

2.4.2 Uncovering the Mobility Patterns of Clusters of Tourists

Some studies have shown that different types, sub-groups, or clusters of tourists may present different travel behaviours (Batra, 2009; Vu et al., 2015; Ahn and McKercher, 2015; Phillips and Jang, 2010). Domènech et al. (2020a), for instance, identified that cruise passengers with different expenditure levels have different mobility patterns in port destination cities. However, these kinds of studies usually apply an ad-hoc combination of analytic techniques that is not easy to generalize.

From a complementary perspective, many researchers have followed the digital footprint of tourists (Salas-Olmedo et al., 2018) to know their mobility in destinations, but the current research shows that it is difficult to analyse all these data by segmenting tourists. In fact, many studies have analysed tourists' mobility patterns (Gabrielli et al., 2015; Li et al., 2018a) without taking into account the diversity of tourists (Liu et al., 2018) because of the difficulty of obtaining their socio-demographic data. This aspect can be considered only in those cases in which user information is available, such as the one described by Massimo and Ricci (2019). In that case, they use information on users' past POI visits to segment visitors with similar visit trajectories, to make POI recommendations.

GSMD-based studies have focused on the clustering of tourists according to diverse factors. Following Liu et al. (2018), studies that segment visitors according to their mobility patterns can be based on non-spatial factors (socio-economic status, gender, age, income, education, race) or on spatial factors. For instance, Manca et al. (2017) focused on spatial data, and Jin et al. (2018) on spatial and temporal data. Nevertheless, very few studies have focused on the analysis of mobility patterns according to the socio-demographic data in order to segment visitors in a destination because, with the currently applied methods of analysis, very little socio-demographic data can be obtained from users. In this vein, several SMA studies based on GSMD analysis have claimed to obtain demographic data from tourists to better understand who they are and to be able to classify them (Chua et al., 2016; Barchiesi et al., 2015; Fuchs et al., 2014). However, the available information is still very limited (Vu et al., 2018) and, in some cases, it is even reduced to the country of origin (Hawelka et al., 2014).

To name some of those studies, Chua et al. (2016) focused on spatial, temporal, and also demographic data from Twitter to discover the tourist flows in a destination, creating tourist profiles and segmenting them by country of origin. Similarly, Vu et al. (2015) analysed the different mobility patterns, popular locations, and routes of Western and Asian tourists in Hong Kong. In the same line, Paldino et al. (2015) analysed geo-tagged picture data from Flickr, segmenting domestic and foreign tourists, and Ma et al. (2018) also analysed the mobility of tourists in destinations and their most visited attractions by classifying tourists into foreign tourists and domestic tourists. Van der Zee and Bertocchi (2018) analysed the spatial behaviour of visitors at a destination through a relational approach and Trip Advisor data, classifying visitors as local, national, European, and non-European. Vu et al. (2020) analysed the activities carried out in a destination by different groups of tourists also segmenting them by their country of origin. Xu et al. (2021) analysed mobility patterns of tourists in South Korea, and its diverse destinations, according to their nationality or country of origin. Liu et al. (2018) analysed mobility via GSMD from Twitter by considering homogeneous segments of users (state visitors, national visitors, and international visitors), created according to their past visits.

However, despite the difficulty of obtaining other socio-demographic data than the origin of users from the GSMD, some mobility studies have tried to take a step further in segmenting visitors. Huang and Wong (2016), for example, analysed their mobility segmenting them by their socio-economic status through Twitter's GSMD. They identified this status from the home and work location of the users. In addition, they showed that socio-economic status and urban spatial structure are the factors that have a stronger influence on the mobility of visitors. On the other hand, Han et al. (2018) analysed the mobility patterns of visitors from the analysis of social media check-in. They used a deep learning method to try to classify tourists by the purpose of their travel.

Studies of mobility patterns that employ Artificial Intelligence techniques are still emerging, and very few of them try to identify meaningful clusters of tourists. Liao (2020) obtained trajectory data from different location-based services and tourism applications, and then they applied cluster analysis to identify the most popular tourist attractions. DBSCAN clustering was used to identify spatial clusters of trajectories at the points of greatest interest, but not to identify clusters of tourists. Xu et al. (2021) analyzed a mobile positioning data set in order to know the nationality and movement patterns of foreign tourists in South Korea. They used network analysis to identify the structure of tourism destinations based on patterns of travel flow, and clustering analysis to identify similar patterns. They identified areas of destinations with different visit patterns of tourists according to their nationality, but not clusters of tourists. Instead, Giglio et al. (2020) used cluster analysis to automatically identify clusters of tourists around points of interest at destinations. They studied the relationship between human mobility and tourist attractions through geo-located images of Italian destinations provided by Flickr users. The results showed that social media data are a valuable source to understand the behavior of tourists in a destination. However, the study did not define or specify the different clusters of tourists and their different mobility flows between the most popular attractions.

Considering that, despite the difficulties and limitations, GSMD data can be analysed in order to make segmentations more precise than the ones based on the country of origin, and also considering that mobility patterns are a key issue for DMOs, this thesis aims to make a contribution to the current challenges of analysing tourism flows in destinations. Chapter 4 is focused on identifying unique mobility patterns from tourists' clusters, targeted at DMOs to improve their management of critical flows in certain points of interest of tourism destinations, helping to minimize social stresses, environment management difficulties, transportation issues, and frictions between tourists and local population in situations of congestion and overcrowding (Anton Clavé, 2019).

2.5. Summary

This chapter detailed the relevant aspects of this thesis. We introduced social media analysis in the tourism domain, touching on the data sources, analytical methods, identification, and profiling methods applied by researchers in this domain. We further introduced contextual recommender systems, their applications in tourism, their benefits over traditional recommender systems, and the AI techniques they commonly apply in recommendations. Finally, we reviewed the current state of social media-based tourist flow analysis.

In the next chapter, we detail our proposed recommender system with its particular methods for data retrieval, data processing, tourist identification, POI identification, user profiling, route planning, and flow analysis.

Chapter 3

Contextual Touristic Recommendations Based on Twitter Data Analysis

3.1. Introduction

In chapters 1 and 2, we introduced social media analytics (SMA) as methods developed for studying data collected from social media platforms to extract relevant information for a particular task. Although this information is generally used to understand and explain a certain phenomenon, it could also be incorporated into a recommender system to aid a decision-maker. This chapter is based on the premise that recommender systems can also profit from the data shared on social media platforms to gain insights into a user's travel preferences.

A tourism recommender system needs to learn from tourists' behaviour and understand their preferences and context to suggest touristic activities to them (Borràs et al., 2014). It is possible to accomplish this task by applying analytical methods to social media data in an information-gathering phase where social media users are identified as tourists, their interests extracted, and contexts established. Then there is a modelling phase in which the information gathered is incorporated into a recommender model that filters options relevant to a tourist. In a final recommendation phase, items are suggested and evaluated according to the preferences of the tourists.

Following this process, in this chapter, we propose a tourism recommender system (**ReCLARM**, which stands for **Re**commender built on **CL**ustering and **A**ssociation **R**ule **M**ining) that utilises features extracted from tweets of tourists in order to create user profiles which are employed to create personalised recommendations of touristic activities. The features, extracted from the activity of the users on Twitter, represent not only the users' cultural preferences but also the context, their travel habits, and the popularity of the visited points of interest (POIs). ReCLARM combines in an original way clustering based on social media features with association rule mining to find the preferred combinations of POIs from those visited by similar users. These combinations capture in a novel way the relatedness of certain activities, which may be due to their similarity, their physical proximity, the ease of travel from one to the other, and even their popularity. Additionally, ReCLARM ranks the mined association rules using methods that adapt to the characteristics of the user to ensure that the recommended POIs fit the user's unique interests and his/her attraction toward popular or unpopular POIs. The performance of ReCLARM has been evaluated with the initial clustering step and without it, to study the influence of social media-based clustering in the recommendation process. The obtained results confirm the usefulness of the clustering phase in improving ReCLARM's performance in several metrics.

The rest of the chapter is structured as follows. Section 3.2 presents some related works that employ methods similar to those of the proposed system. In section 3.3, we describe the proposed system in detail, including the steps, mathematical formulations, and algorithms. The experiments and results are presented and discussed in Section 3.4. In Section 3.5, an extension of ReCLARM is detailed, that suggests an ordered set of POIs to improve the tourist experience. Finally, the chapter is concluded in the last section.

3.2. Related Works

SMA in tourism recommender systems is normally used to profile users by their behaviour and preferences in order to provide them with appropriate personalised suggestions. It is indeed feasible to take into account not only the textual content of the messages they send but also additional information, such as the moment in which they are sent (day of the week and time of the day), the language in which messages are written, the exact geospatial location from which each message is sent, etc. Thus, different features may be extracted from the social media data in order to build user profiles. Having a personalised profile for each user makes it possible to apply clustering procedures to detect different types of users. For example, previous works have suggested that the text in tweets could be used to capture emotions and cluster users by their personalities so that it would be possible to make travel recommendations to similar users (Ishanka and Yukawa, 2018).

Features extracted from social media data can incorporate information about the context of the users to improve recommendations. For example, geolocation extracted from social media posts can be used to distinguish local citizens from tourists in a particular destination (Manca et al., 2017; Jabreel et al., 2017, 2018). It is also possible to infer the purpose of a trip by utilising check-ins, timestamps, and point-of-interest labels from Gowalla and Foursquare (Huang et al., 2021). Check-in information has also been used to study the trajectory of users within a city, allowing next-POI recommendations to be made to users based on historical data (Huang et al., 2021; Massimo and Ricci, 2021; Baral et al., 2018; He et al., 2017). Some works also included temporal and social factors extracted from check-in information [24]. When dealing with check-ins, it is often necessary to link a geolocation to a specific POI name and category, especially in travel planning. Some location based social networks (LBSNs) such as TripAdvisor or Gowalla, provide these labels but, in other cases, they must be inferred.

There are also works that have combined different data modalities or data from different social networks when creating user profiles. For example, Farnadi et al. (2018) proposed a hybrid deep neural network that aggregates textual, visual, and social-relational data extracted from Facebook profiles into a user profile which is evaluated by predicting the user's age, gender, and personality. A platform-independent system that automatically extracts textual features, including comments, check-in labels, and links from multiple social networks to build user profiles was proposed by Orlandi et al. (2012). These profiles, which represent the interests of users, were enriched with information from DBpedia[1].

The application of Artificial Intelligence techniques to analyse the features extracted from social media data could further enhance recommendations by solving some problems like *cold start* and *data sparsity* (described in chapter 1). Clustering has proven useful in this case. For instance, Liji et al. (2018) proposed an evolutionary algorithm to cluster user attributes before building a user-item matrix for collaborative filtering. Ma et al. (2016) went a step further by combining three clustering processes (based on user trust, user similarity, and item similarity) to form the user-item matrix and then using the matrix factorisation model to make predictions. Following the same trend, Nguyen et al. (2020b,a) proposed grouping users by their cognitive similarity, determined by their interest in similar items, to handle the cold start and data sparsity problems. Along the same lines of user similarity, Fränti et al. (2015) compared the similarity of users via different factors (location visit frequencies, opinions, and liked pages) to improve the recommendations. The same authors also considered the sparse location histories of mobile users to find similar users even if a user's trajectory was incomplete (Fränti et al., 2018).

Another Artificial Intelligence technique that may be useful in recommender systems is *association rule mining* (ARM). Rules provide common relationships between items (objects frequently bought together, or POIs frequently visited together), which may be helpful in the recommendation process. For example, the use of association rules has been suggested to recommend songs (Najafabadi et al., 2017). More concretely, song clusters are used to build a user's profile, and then rules are mined from the user's listening history to make recommendations that fit his/her preferences.

The use of ARM in tandem with clustering is a popular concept that has been successfully applied in several fields. Pandya et al. (2016) proposed this combination to battle data sparsity and cold start problems in e-commerce, using k-means for user-item matrix clustering and mining association rules from Boolean data extracted from user clusters. Similarly, Jalalimanesh et al. (2012) proposed a

[1] https://wiki.dbpedia.org/ (last accessed October, 2022)

recommender system in the inter-library domain, to recommend books by assigning users to clusters (built on users' categorical features extracted from library logs) that are attached to association rules mined using decision trees. Despite the popularity of this concept, it has been scarcely applied in the field of tourism. Fenza et al. (2011) proposed a tourism recommender system based on separate fuzzy clustering of users and POIs, allowing new users and POIs to be assigned to clusters. Users are then matched with previously mined association rules that relate users and their context to POI clusters. However, this system was specifically tailored to the data collected by their application and cannot be used out of the bounds of their project. We focus on the combination of clustering and ARM in our proposed system, which is detailed in the next section.

3.3. System Description

This section presents the main aspects of our proposed recommender system (**ReCLARM**) for tourist activities, which combines in a novel way the use of clustering to aggregate users that have similar preferences when visiting a city (according to the data they provide on social media) and the use of association rules, which indicate the preferred combinations of items for classes of users (also obtained from the analysis of the sequences of POIs visited by users, registered in their social media posts).

User profiles are not built with a specific generalisation in mind (for instance, personality (Ishanka and Yukawa, 2018)). They are solely based on the features extracted from Twitter that embed users' interests, travel habits, and degrees of interest in popular items.

Figure 3.1 presents the architecture of ReCLARM, which has three basic stages. First, data from a collection of users that have visited a destination is collected from Twitter, pre-processed, and stored in a PostgreSQL database. Some pertinent information about geolocations is extracted from Open Street Map (OSM). Features are then extracted after filtering unwanted data. In the second stage, a clustering process is performed to uncover user profiles (groups of visitors with similar characteristics, preferences, and patterns of travel). In the final stage, a set of association rules is mined for each cluster. These rules describe POIs frequently visited together by the members of that cluster. Finally, these rules are employed to provide personalised recommendations to a particular user. The following subsections describe these stages in detail.

3.3.1 Data Collection and Pre-Processing

This is the main stage in the information-gathering phase. The goal is to gather and pre-process all the necessary information needed in understanding and modelling tourist preferences and behaviour.

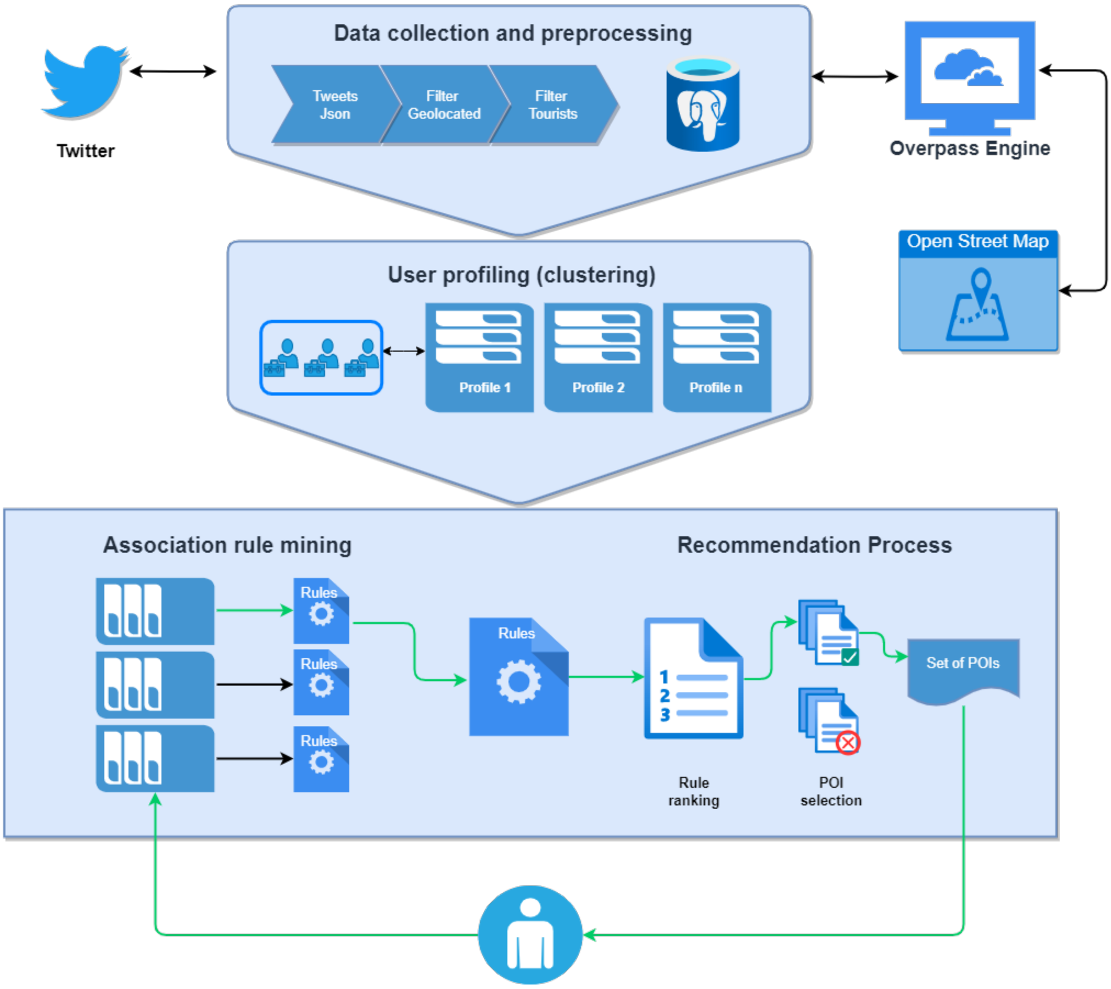


Figure 3.1: Architecture of ReCLARM.

Data Source

Twitter is an LBSN and a microblogging platform on which users post short pieces of text called tweets, which may contain URLs and references to other users. It was created by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams in March 2006 and launched in July 2006[2]. Twitter boasts of approximately 450 million active users per month who send approximately 500 million tweets per day[3], making it the fourth largest website globally by user traffic as of September 2022[4]. It is a popular and widely available medium for users to share their tourism experiences.

Twitter was chosen as our data source for the same reasons it is popular among researchers as described in chapter 2. Its policies allow access to data for research purposes, also its Application Programming Interface (API) is easy to use and free to access. Twitter has been widely used in Tourism fields, such as destination brand communication analysis (Lalicic et al., 2019, 2020), sentiment analysis (Jabreel

[2] https://en.wikipedia.org/wiki/Twitter (last accessed September, 2022)
[3] https://www.businessofapps.com/data/twitter-statistics/ (last accessed October, 2022)
[4] https://www.similarweb.com/website/twitter.com/ (last accessed October, 2022)

et al., 2017, 2018) and detection of trajectories. It also includes shared links from posts on other LBSNs, such as Instagram, which might be relevant in more complex multi-platform and multi-modal (text, images, videos) analysis.

We use their freely available API[5] to download tweets posted in Catalonia. Tweets are returned as JavaScript Object Notation (JSON) files with different kinds of information. Table 3.1 shows the attributes of a tweet relevant to this thesis work. A full list of tweet attributes can be found at Twitter's developer manual[6].

Table 3.1

Tweet attributes in Twitter data.

Attribute	Data Type	Short Description
Created at	String	UTC time when the tweet was created.
Id str	String	Unique identifier of the tweet.
Text	String	Actual UTF-8 text of the tweet.
User	User object	Data dictionary containing information about a user including id, screen_name, geo_enable, etc.
Coordinates	Coordinates	Geographical location of the tweet, if shared by the user.
Place	Place object	Geographical data dictionary that indicates the place from which the tweet was sent. It can be a country, a region or even a POI.
Lang	String	BCP 47 language identifier of the machine-detected language of the text of the tweet.

Identification of Tweets from Tourists

Tourist identification is an important aspect of SMA in tourism (as described in chapter 2). This is also true in tourism recommender systems that seek to build tourist profiles from social media data. In ReCLARM, it was important to distinguish the tweets sent by visitors from the tweets sent by local citizens. Twitter does not provide any specific information in this regard.

To identify the tweets of the tourists, we extended the solution proposed by Manca et al. (2017) to consider *locations frequently visited* by users. A tweet was touristic if it was sent from a location other than the user's frequently visited locations or his/her profile location (which the user may indicate by free text in his/her public profile). The frequently visited locations are defined as those cities from which the user posted tweets for at least 20 days (at least one tweet per day).

Algorithm 1 details the steps performed to determine if a tweet was sent by a tourist. The frequent locations of all users are first determined by counting the number of days a user has tweeted from each place. After that, the coordinates of

[5] https://developer.twitter.com/en/docs/twitter-api (last accessed April, 2021)

[6] https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet (last accessed October, 2022)

each geolocated tweet are checked against the set of home locations. If they do not fall within any home location, then the tweets are considered *touristic*. We classify the tweets rather than the users, because a user might be posting as a tourist in one tweet and as a local citizen in another.

Algorithm 1 Identification of tourists' tweets (pseudocode).

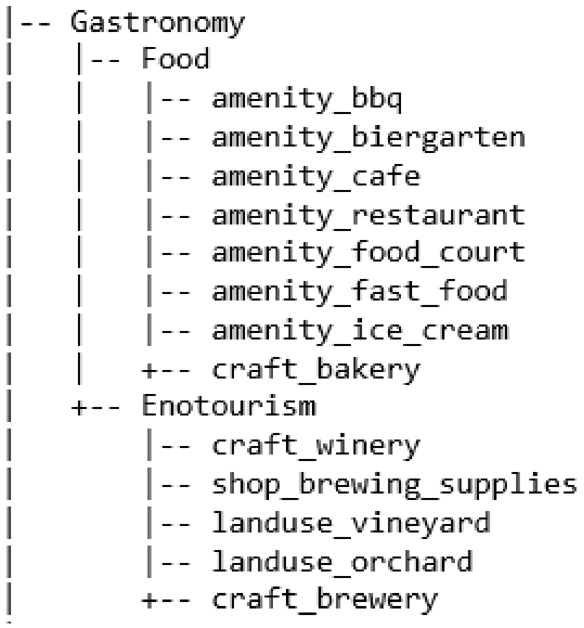
```
Let $T$ be the set of all geolocated tweets;
Let $U : L$ be the set of all users and their profile locations $U : L = \{u_1 : l_1, ..., u_i : l_i\}$;
Let $\alpha = 20$ be the threshold of days tweeting from a location;
**for** users in $U$ **do**
    Initialize an empty set of home locations for user $u_i(H_{u_i})$
    Add user's profile location $l_i$ to $H_{u_i}$, if available
    Get tweets $t_{u_i}$ of user $u_i$ from $T$
    Group tweets $t_{u_i}$ by their tweet location
    **if** the count of days with tweets in a tweet location exceeds $\alpha$ **then**
        Add tweet location to $H_{u_i}$
    **end if**
    **for** tweet in $t_{u_i}$ **do**
        **if** tweet location in $H_{u_i}$ **then**
            Set tweet as local
        **else**
            Set tweet as touristic
        **end if**
    **end for**
**end for**
```

Activity Identification

In order to capture the preferences of users, it was necessary to analyse their tweets to determine the POIs and types of activities that they have taken part in. The term *point of interest* (POI), as described in chapter 2, has appeared in many tourism-related articles, and was used as early as the 1930s. It usually refers to a place a tourist is expected to see or visit, also known as a "sight" (Enzensberger, 1996; Neff, 1938). These POIs may be defined by verified administrative bodies or predicted from social media data based on other factors detailed in chapter 2. Twitter does not provide information on the POIs visited by users, so it has to be inferred from the available data. In ReCLARM, the POIs and types of activities were determined from Open Street Map (OSM) and represented in a hierarchical structure. This approach is an extension of the method proposed by Bustamante et al. (2019), but we have considered more activity categories and we have added text analysis to the identification process.

The following services have been used in the identification process:

- *Open Street Map (OSM)* [7]: It is an open-source map server which includes cartographic documentation of roads, streets, water bodies, buildings, etc. It also provides geocoding and geoparsing services. As it is open source, it has become the first choice for academic research. In the OSM database, physical features (buildings, roads, etc.) are represented by tags, which describe the geographic attributes of those features[8]. These tags provide information about an element, such as its "name" and "purpose". For example, the tag `nature:beach` is used to identify a beach. We have used these tags to create a tree structure to categorise activities experienced by tourists. A fully comprehensive list of tags and their descriptions can be found at OSM taginfo[9]. It is important to note that tags may change overtime or be discontinued and replaced by OSM. The activity tree has a root node named "Activities". Its children are the main categories, which were inspired by an ontology developed by Moreno et al. (2013): Routes, Sports, Gastronomy, Leisure, Accommodation, Transportation, Nature and Culture. The tree also includes numerous subcategories that are descendants of the main categories. The leaves of the tree correspond to the OSM tags. Figure 3.2 shows a sample section of the activity tree. The complete tree shown in appendix A.1 consists of 32 subcategories and a total of 175 OSM tags in the leaves.



```
|-- Gastronomy
|    |-- Food
|    |    |-- amenity_bbq
|    |    |-- amenity_biergarten
|    |    |-- amenity_cafe
|    |    |-- amenity_restaurant
|    |    |-- amenity_food_court
|    |    |-- amenity_fast_food
|    |    |-- amenity_ice_cream
|    |    +-- craft_bakery
|    +-- Enotourism
|         |-- craft_winery
|         |-- shop_brewing_supplies
|         |-- landuse_vineyard
|         |-- landuse_orchard
|         +-- craft_brewery
```

Figure 3.2: Sample chunk of the Activity tree, showing the main category *"Gastronomy"*, its subcategories *"Food"* and *"Enotourism"* and the OSM tags associated with them.

- *Overpass turbo* [10]: It is a query server for requesting specific features in the

[7] https://www.openstreetmap.org/ (last accessed October, 2022)

[8] https://wiki.openstreetmap.org/wiki/Map_features (last accessed October, 2022)

[9] https://taginfo.openstreetmap.org/tags (last accessed October, 2022)

[10] https://overpass-turbo.eu/ (last accessed October, 2022)

OSM database. Overpass provides a query language[11], similar to Structured Query Language (SQL), to help users gain access to specific information in the OSM database. For example, physical features within a certain radius from a `[Latitude, Longitude]` coordinate pair can be requested and filtered by their OSM tags. We used the Overpass query language to request all POIs within a certain range around the coordinates of a touristic tweet. These POIs were OSM map features categorised as *Nodes*, *Ways*, *Relations* or *Areas*. *Nodes* are single structures, such as office buildings, which include coordinates to represent their locations. *Ways* consist of several nodes with individual coordinates, which represent structures such as roads, highways, streets, pathways, plazas, fountains, parks and steps. *Relations* are compound structures which include several nodes and ways. For example, complex attractions comprising multiple buildings such as Sagrada Familia are relations. Finally, *Areas* are large physical features that are represented by bounding boxes. Areas contain several nodes, ways and relations. For example, the Port Aventura theme park in Spain is categorised as an area, because it contains several attractions over a large area. This Overpass query requests all named nodes, ways and relations within a 50 m radius from the coordinates of a tweet, also including the areas if they are not cities, countries, towns or time-zone boundaries. Figure 3.3 shows the code snippet written in Overpass QL. Line 1 `[out:json]` sets the query output as JSON and `[timeout:1000]` sets the waiting time in seconds before the query is terminated. Line 2 `'nwr'` requests all nodes, ways and relations `'around'` <*Lat*>,<*Lon*> within the radius of <*displacement*> meters, and `[~"^name(:.*)?$"~"."]` filters out unnamed map features. Line 3 requests all areas bordering the <*Lat*>,<*Lon*>. Lines 4 and 5 filter out all areas that are cities, towns, countries and time-zone boundaries. Finally, line 6 formats the results, `'out geom'` gets the full geometry of results, `'tags'` includes IDs and tags of the results and `'qt'` sorts the results by their geometry.

```
1: [out:json][timeout:1000];
2: (nwr(around:<displacement>,<Lat>,<Lon>)[~"^name(:.*)?$"~"."];
3: is_in(%s,%s);)->.a;
4: (((nwr.a; - nwr.a[type=boundary];); - area.a[place=town];);
5: ((area.a; - area.a[type=boundary];); - area.a[place=town];););
6: out geom tags (<bounding-box>) qt;
```

Figure 3.3: Code snippet of an Overpass query.

- *Natural Language Toolkit (NLTK)* (Bird et al., 2009): NLTK is an open source toolkit for natural language processing written in Python. It is widely used because it includes a large number of tools for text analysis and it is

[11] https://wiki.openstreetmap.org/wiki/Overpass_API/Overpass_QL (last accessed October, 2022)

very well documented. The NLTK library is used to pre-process the text of each tweet. The text is first stripped of stop words based on its language, and then separated into tokens with the NLTK tweet tokenizer. URLs and links of any form are removed, and hashtags composed of several capitalised words are split (e.g., *#SagradaFamilia*). Finally, numbers, icons, accents, punctuation, user mentions (i.e., user tags beginning with ”@”) and excess letters in words, such as ”funnnn”, are also eliminated. Once tweets have been processed, they can be compared with the names of the POIs returned from Overpass to find matches. The NLTK evergram tool is used to make n-grams of the POI names returned from Overpass. In this way it is possible to detect hashtags that contain POI names which could not be split in the pre-processing step.

To associate an activity with each geolocated tweet in the dataset, POIs are requested from Overpass as described. After the resulting POIs are returned, we adopt a priority-based method, as suggested by Bustamante et al. (2019), to determine the categories to be assigned to the tweet.

Table 3.2 is used to assign a category when there are conflicting OSM tags in the proximity of the tweet. We established the priority according to the importance of a category to a tourist, 1 being the highest priority and 8 the least. Additionally, the distance shown in the table is the maximum range in meters at which the priority is relevant. Thus, if a tweet matches two or more POIs with OSM tags from different categories, the category with the highest priority is chosen if the tweet is within the stipulated distance of that POI.

Table 3.2

Category priority table.

Category	Distance	Priority
Culture	50 m	1
Leisure	25 m	2
Accommodation	35 m	3
Gastronomy	25 m	4
Nature	15 m	5
Routes	15 m	6
Sports	15 m	7
Transportation	15 m	8

The activity identification steps are detailed in Algorithm 2:

- **Step 1:** The Overpass server is queried to return POIs within a 50 m radius of each tweet in the dataset.
- **Step 2:** Names of POIs returned from Overpass are analysed to find matches with the tweet text or the place name provided by Twitter if it is a POI. The tokens from the tweet are compared with the POI names returned by

Overpass, and also with the n-grams made from the POI names that are between 2 and 5 words.

Algorithm 2 Activity identification (pseudocode).

```
Input: Geolocated tweets with exact coordinates
Output: category of each tweet
for all tweets do
    Get set of POIs within 50m radius of tweet $(P)$
    Preprocess text in tweet or in POI name when provided by Twitter $(t\_tokens)$
    if $P$ is empty then
        Return $CategoryNull$
    end if
    for POIs in $P$ do
        Tokenize and remove stop words from POI name $(p\_tokens)$
        Find intersection between $t\_tokens$ and $p\_tokens$
        if size of intersection greater than two then
            Add POI to matched list. $(M)$
        end if
        if size of matching list $M$ is equal to one then $/{*}Case\ 1 : One\ match{*}/$
            Find category/ies of POI tags using the activity tree
            if multiple categories found then
                Return category with highest priority
            end if
        end if
        if size of matching list $M$ is greater than one then $/{*}Case\ 2 : Multiple\ matches{*}/$
            for all matching POIs do
                Find category/ies of POI tags using the activity tree
                if multiple categories found then
                    Add category with highest priority to list $C$
                end if
            end for
            if categories in $C$ conflict then
                Return category with highest priority
            end if
        end if
        if matching list $M$ is empty then $/{*}Case\ 3 : No\ match{*}/$
            Repeat steps in $Case2$ with all POIs returned from Overpass.
        end if
    end for
end for
```

The following cases occur as a result of the text analysis:

1. *One Match*: When only one POI matches the text, the tags of that POI are checked against the activity tree, and the best suited category based on Table 3.2 is assigned to that tweet.
2. *Multiple Matches*: When more than one POI matches the text, the tags of all matching POIs are checked against the activity tree, and the best suited category based on Table 3.2 is assigned to that tweet.
3. *No Match*: When no match is found in the text, the tags of the returned POIs are checked against the activity tree, and the category with the highest priority rank based on Table 3.2 is assigned to the tweet.
4. *No POI*: If the text analysis did not return any POI, the tweet is not assigned to any category.

Data Summary

The data used in this thesis work was streamed from Twitter and comprised of over 4 million tweets posted in Catalonia, Spain in 2019. Since we are concerned on providing a city based recommender system, we selected posts published in the city of Barcelona in 2019, totalling 1,523,801 tweets from 108,515 users. Before the analysis and experimentation, we removed the following information:

- Tweets from Barcelonian citizens, not from tourists.
- Tweets that could not be assigned to any category in the activity identification process.
- Users with less than three tweets and their tweets.

Table 3.3 shows the summary of the data set before and after this filtering process.

Table 3.3

Data set summary.

Statistics	**Value**
Total number of tweets in Barcelona	1,523,801
Total number of users in Barcelona	108,515
Statistics after filtering	**Value**
Total number of tweets in Barcelona	37,302
Total number of users in Barcelona	6,066

3.3.2 Cluster analysis for user profiling

The next stage after data collection and pre-processing was cluster analysis, which is a part of the modelling phase. The goal of this stage was to identify groups of users that had a similar travel behaviour (they enjoyed the same kinds of activities, they had similar mobility patterns, they visited POIs at the same times of the day, they enjoyed (or not) visiting popular places, etc.). This knowledge permitted ReCLARM to recommend to a user POIs that were visited by similar users. A clustering procedure identifies the users that have a high similarity in the values of a set of features. First, we describe the features that were chosen to represent different aspects of the travel behaviour, context, and preferences of tourists; after that, the clustering process is described.

Clustering Features

The clustering features are a vector representation of a user. Each user is described by four kinds of features that represent the preferences of the user with regard to cultural and leisure activities, travel characteristics (length of stay and degree of mobility within the city), degrees of interest in popular POIs and period of the day with more touristic activity.

- *Activity interest features*. These features embed the users' interests in different kinds of touristic activities. They represent different levels of abstraction in the activity tree, as the analysis would be too general if we only considered the eight main categories of the first level. The activities associated with higher percentages of users in Barcelona were selected for the clustering process. All these features were scored as the percentages of tweets by users that were related to the particular types of activity. The selected features were the following:

  1. **Top-tier features**. These features represent some of the main categories in the activity tree. They are %Routes, %Sports, %Accommodation, %Transportation and %Nature.

  2. **Middle-tier features**. These are activity features selected from the subcategories of the activity tree. They are %Food, %Enotourism, %AmusementParks, %RecreationFacilities, %Beach, %Health&Care, %NightLife, %Shopping, %Viewpoint, %CulturalAmenities, %Historic and %Religious.

  3. **Bottom-tier features**. These activity features are OSM tags represented as leaves of the activity tree. They are %tourism_museum, %amenity_arts_centre and %tourism_gallery.

  4. **Other features related to the activity tree**. In the analysis it was found out that the OSM tag `{tourism, artwork}` was quite popular in our data set, but we did not know what type of artwork was being experienced. Thus, it was decided to break down this tag into several features that represent the type of artwork, using other tags associated with the POIs. These features are %artwork_type_sculpture, %artwork_type_architecture, %artwork_type_statue and %other_artwork. The last one represents cases of undetermined type or works of art that do not belong

to the other three types. Figure 3.4 shows the 24 activity features and the percentages of users that visited them in Barcelona (according to the content and the location of their tweets). It may be seen that the top categories were RecreationFacilities, Religious, Historic and Food, followed by Museums and Accommodation.

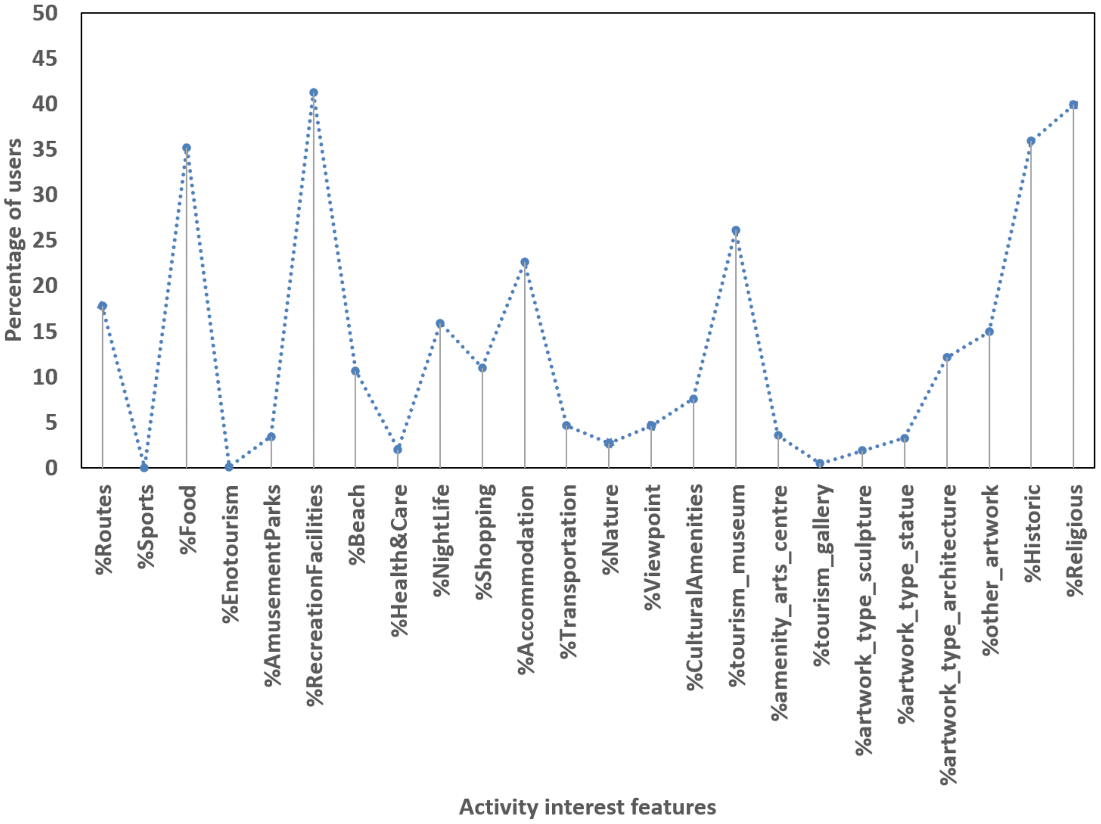


Figure 3.4: Activity features by percentages of users.

- *Travel features*. These features are related to the travel habits of the user and model the user-based mobility context. Concretely, they contain information on the durations of trips to Barcelona and the degrees of mobility within the city. They are the following:

    1. **Length of stay**. It is the maximum number of consecutive days in which the user posted tweets from Barcelona.

    2. **Tweet distance maximum and average features**. Twtdistance_max and Twtdistance_avg are the maximum and average distances between the locations of the tweets of the user in Barcelona. They constitute contextual information on the user's ability to explore the city.

- *Popularity features*: These features represent the interest of the user in visiting the most well-known and popular POIs. In order to obtain a popularity order, the POIs were ranked according to the numbers of users in the database that had visited them. Popularity is split into five features. The first four (%top10_tweets, %top10–20_tweets, %top20–50_tweets and %top50–100_tweets) are the percentages of tweets of the user from POIs in positions 1–10, 10–20, 20–50 and 50–100 of the ranking. The feature %top100_tweets codifies the percentage of tweets that were not sent from any of the top 100 POIs in the city. For example, Sagrada Familia and a few other POIs are the top 10 POIs in the dataset, the percentage of a user's tweets in these top 10 POIs is the feature %top10_tweets.

- *Temporal features*: These features embed the time of the day favoured by the user in his/her trips. They incorporate the temporal context. There are 4 features representing the percentages of tweets that the user posted by period of the day. The features are: %Dawn_tweets (00:00–07:00), %Morning_tweets (07:00–12:00), %Afternoon_tweets (12:00–20:00) and %Night_tweets (20:00–00:00).

In summary, the preferences and travel habits of each user are represented by a vector of 35 numerical features (24 for the interest in different kinds of activities, 2 for the travel features, 5 for the interest in popular POIs and 4 to codify the time the tweets were sent out). All of them are percentages (values between 0 and 1), except the two travel features.

Clustering Parameters

The clustering process was done using the scikit-learn (Pedregosa et al., 2013) Python library. The choice of parameters used in the cluster analysis was the following:

- *Algorithm*. The k-means algorithm was selected because of its speed and ability to adapt to new data samples.

- *Feature scaling*. In cluster analysis it is necessary to ensure the data are scaled appropriately, as features having different scales would affect the clustering process negatively. The clustering features were standardised using the *Z-Score*.

- *Number of clusters (k)*. The k-means algorithm requires the number of clusters as an input parameter. Clustering is by nature an unsupervised analysis process, and therefore, the optimal number of clusters is case-dependent. We were not concerned about having equally sized clusters with clear dividing boundaries, but rather clusters that represent different combinations of the clustering features in order to create user profiles with different interests and contexts. After some experimentation with the k-elbow method and some business decisions based on our data, we found $k = 25$ clusters was a suitable number.

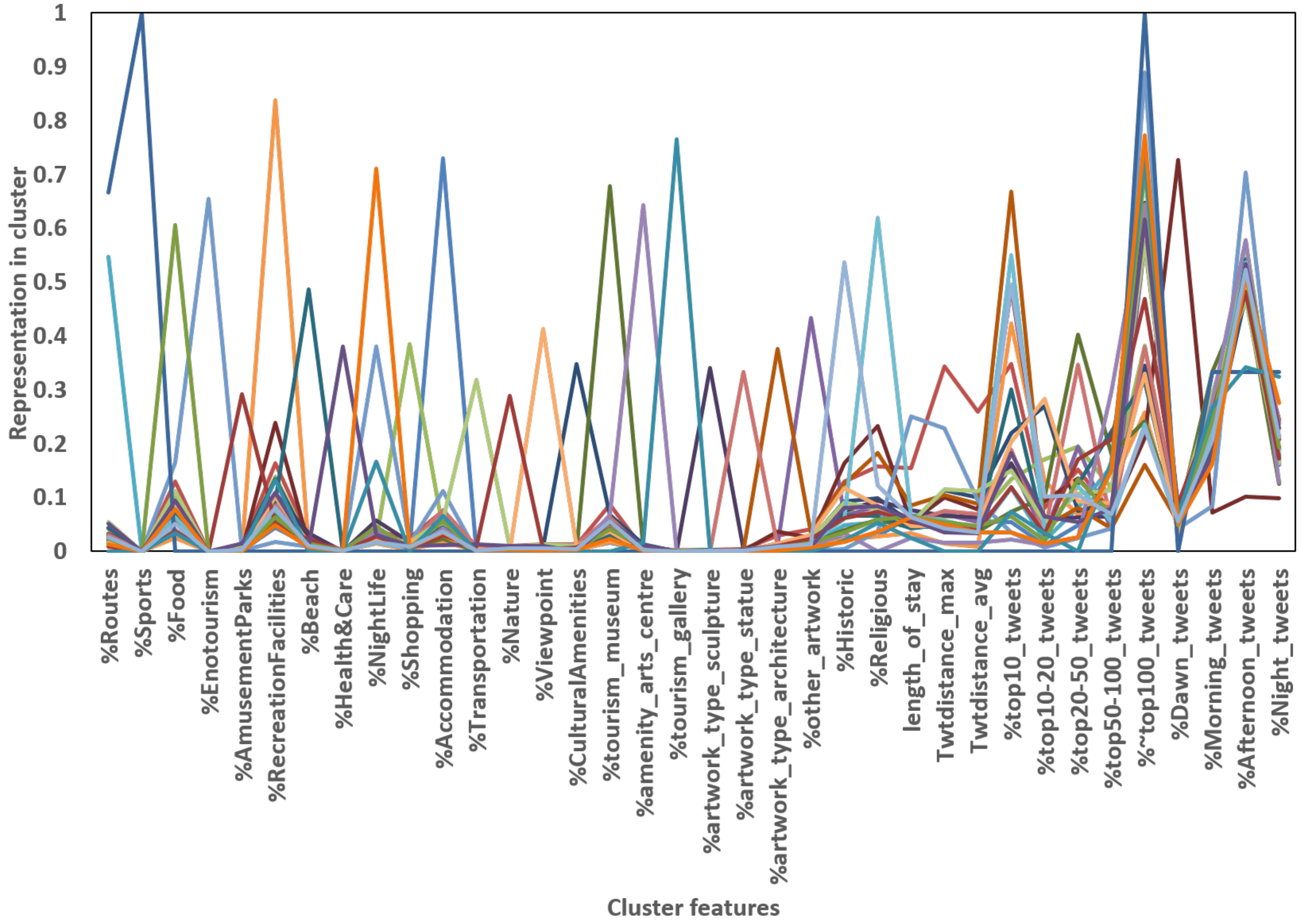


Figure 3.5: Cluster model showing distribution of features in the clusters.

Figure 3.5 shows a graphical representation of each cluster with a different colour; features are on the x-axis and the mean value of each is on the y-axis. The clusters display different combinations of peaks across the clustering features, showing high heterogeneity (especially in the activity interest features). On the contrary, temporal features do not exhibit great differences among the clusters. The high heterogeneity in activity features allows the recommendation of different POIs across clusters.

3.3.3 Association Rule Mining

After clustering, the next stage is *association rule mining* (ARM), the second part of the modelling phase. ARM is a popular technique that enables the identification of items that co-occur frequently within a specific data set. It has been successfully applied in marketing to study purchases in retail markets, under the name *market basket analysis* (MBA). The idea is that the owner of a supermarket could potentially increase the sales of two products with high affinity by shelving them together. Similarly, the idea is to recommend together POIs with high affinity, which will have been detected in this step. Two POIs will have been visited together on the same day by many tourists if they are close (or well connected by public transport) and if they match the same set of preferences. The ARM process was implemented with the help of the mlxtend (Raschka, 2013) Python library.

ARM Parameters

- *Pre-processing*: In MBA, the analysis is usually performed in shopping sessions; i.e., one user may have multiple baskets from different shopping sessions. In our case, it was decided to split the users into POIs experienced in the same day. This was beneficial because ReCLARM aims to recommend POIs for daily trip planning. However, this decision also led to some loss of information, because we dropped the days in which less than two POIs were experienced.

- *Frequent itemset mining algorithm*: Frequent itemset mining (detecting sets of items that appear frequently together) is the main step in ARM. There are a variety of similar algorithms with which to perform this step. The *Apriori* algorithm (Agrawal and Srikant, 1994) was chosen because of its popularity and widespread acceptance. This algorithm requires *minimum support* to be provided, which is the minimum amount of times an itemset has to occur for it to be considered as frequent. This parameter was given a low value because the data set is sparse and it needed some leeway to function. The algorithm also requires the *maximum length* of the itemsets (maximum size of the sets of items appearing together frequently). The values chosen for these parameters are shown in Table 3.4.

- *Association rule parameters*: In ARM, multiple metrics are computed for each mined rule to evaluate its performance (they are detailed in the following subsection). In order to provide useful rules, these metrics may be used as filters, so that rules that do not reach a given threshold are discarded. The usual choices are *confidence*, *support* or *lift*, but in our case this filtering step was not relevant because the posterior selection algorithm performed a ranking of the rules, as will be shown later. Thus, the filtering parameters were set as shown in Table 3.4.

Table 3.4

ARM algorithm parameters.

Apriori Params		**Association Rule Params**	
Min support	Max length	Metric	Minimum value
0.001	3	Lift	0

ARM Metrics

Several metrics can be used to evaluate the usefulness of mined rules. The three most popular ones are support, confidence and lift. They were used to select the best rules for the recommendation process, as will be described in the next section.

- **Support**. It indicates how frequently an itemset occurs in a data set. It is the fraction of times an itemset appears among all the transactions being analysed. It can be denoted as the probability of occurrence of the itemset $P\left(itemset\right)$. The support of a rule is the percentage of times that the antecedent and the consequent of the rule appear together.
- **Confidence**. It indicates how often a rule is found to be true. It is the proportion of times the consequent is found in the same transaction as the antecedent. It can be denoted as the conditional probability of the consequent appearing in the same transaction after the antecedent is found to be true $P\left(consequent|antecedent\right)$.
- **Lift**. It was established to solve the problem of the confidence being dependent only on the support of the antecedent. The order of the consequent and the antecedent in the rule does not matter, which makes the confidence metric a bit skewed because it considers the consequent to be dependent on the antecedent. The lift metric modifies this fact by considering the support of both the antecedent and the consequent.

The mathematical formulation of these measures is the following:

For a rule: $X \to Y$, *where X is the antecedent and Y is the consequent*

Support:

$$supp\left(X\right) = P\left(X\right);\ supp\left(Y\right) = P\left(Y\right);\ supp\left(X \to Y\right) = P\left(X \cup Y\right) \tag{3.1}$$

Confidence:

$$conf\left(X \rightarrow Y\right) = P\left(Y|X\right) = \frac{supp\left(X \rightarrow Y\right)}{supp\left(X\right)} \tag{3.2}$$

Lift:

$$lift\left(X \rightarrow Y\right) = \frac{supp\left(X \rightarrow Y\right)}{supp\left(X\right) * supp\left(Y\right)} \tag{3.3}$$

The three metrics were combined to rank the mined rules in the selection stage, as will be detailed shortly. In summary, the following ARM steps were executed for each individual cluster to mine useful rules:

1. Create separate baskets with the POIs visited in the same day by each user.
2. Mine frequent itemsets of visited POIs with a maximum length of 3 and a minimum support of 0.001.
3. Build association rules from the itemsets uncovered in the previous step.
4. Compute the previous metrics for each rule.

3.3.4 Personalised Recommendations of Touristic Activities

This is the final stage of ReCLARM and the main part of the recommendation phase. First, ReCLARM ranks the association rules obtained in the previous stage, which are used to decide the POIs to be recommended to a user. The set of recommended POIs should ideally fulfil these conditions:

- It should contain only POIs that have been visited by other members of the same cluster.
- It should fit the user's interests regarding the preferred types of activities and the attraction towards popular items.
- It should reflect the causality of the association rules of the cluster, in order to recommend POIs with high affinity.

To achieve these conditions, ReCLARM's recommendation process was formulated as a ranking problem. The association rules of the user's cluster were ranked based on their performance according to certain evaluation metrics, and POIs were selected while taking into account the associations expressed in those rules. The employed evaluation metrics were antecedent support, consequent support, support, confidence, lift and two new metrics that evaluate the rules in terms of the preferences of the user towards certain kinds of touristic activities and towards popular spots. Those two new metrics are the following:

- *Preference ratio*: This metric evaluates if the POIs that appear in a rule belong to any of the activity categories preferred by the user. The activity categories coincide with the activity features used in the clustering process. Let CAT_{pref} be a user's preferred categories, i.e., the activity categories for which the user has at least one tweet. $1_{CAT_{pref}}\left(\psi\left(POI\right)\right)$ is the indicator function that has value 1 if $\psi\left(POI\right) \in CAT_{pref}$ or 0 otherwise. $\psi\left(POI\right)$ is a function that gets the category to which a POI belongs by looking up the activity tree. Furthermore, let $D_{CAT}\left(POI\right)$ be a function that signals the user's degree of interest in a preferred category, where the degree of interest is the percentage of the user's visited POIs belonging to the preferred category. If P_{rule} is the set of POIs in a rule, the preference ratio is calculated as follows:

$$\textit{preference ratio} = \frac{1}{|P_{rule}|}\sum_{p\in P_{rule}}\left(1_{CAT_{pref}}\left(\psi\left(p\right)\right) * D_{CAT}\left(p\right)\right) \tag{3.4}$$

- *Popularity ratio*: This is the percentage of popular POIs in a rule. Let P_{top10} be the top 10 POIs extracted from the data set, and $1_{P_{top10}}(POI)$ is the indicator function with value 1 if $POI \in P_{top10}$ and 0 otherwise. The popularity ratio is computed with the following formula:

$$\textit{popularity ratio} = \frac{1}{|P_{rule}|}\sum_{p\in P_{rule}} 1_{P_{top10}}(p) \tag{3.5}$$

After computing all the metrics as shown in Equations (3.1) to (3.5), they were combined to get an overall score for each association rule, using a weighted arithmetic mean (also known as the *weighted average* (WA) aggregation operator in decision support systems). Let A be the set of metrics to be aggregated $A = (a_1, ..., a_n)$ and $W = (w_1, ..., w_n)$ be the set of weights for each metric. Then,

$$WA\left(A\right) = \sum_{i=1}^{n} w_i a_i \tag{3.6}$$

The WA operator allows one to determine the relevance of each metric using weights. In ReCLARM the weight of each metric has been manually set, but it depended on the user's degree of interest in popular POIs. Two sets of weights were defined, one for users interested in popular POIs and the other for users interested in off the beaten track POIs. The weights were linearly combined as follows:

$$W_{comb} = \left(1-\beta\right) W_{unpop} + \beta W_{pop} \tag{3.7}$$

In this expression, W_{comb} is the set of adapted weights, and W_{unpop} and W_{pop} are the predefined sets of weights for the users interested in unpopular and popular POIs, respectively (shown in Table 3.5). The parameter β expresses the degree to which the user is interested in popular POIs. It is computed as the fraction of the user's visited POIs that are in the top 10. Thus, when all POIs visited by the user are in the top 10, β is 1 and the W_{pop} weights are applied. Inversely, when none of the POIs visited by the user appear in the top 10, β is 0 and W_{comb} coincides with W_{unpop}.

Table 3.5

Weights for different cases of interest.

Case	Preference Ratio	Lift	Confidence	Support	Antecedent Support	Consequent Support	Popularity Ratio
W_{unpop}	0.5	0.15	0.15	0.1	0.05	0.05	0.0
W_{pop}	0.3	0.05	0.05	0.1	0.1	0.1	0.3

In the W_{pop} case, we wanted to give high priority to the rules containing POIs in the top 10 (while still considering the user's preferences), so we gave higher importance to *popularity_ratio* and the three support metrics because they reflect popularity in the rule and in the cluster respectively. In the case of W_{unpop}, we zeroed out the *popularity_ratio* and relied on the other metrics, especially on the *preference_ratio*.

Given the definition of W_{pop} and W_{unpop} in Table 3.5, W_{comb} was adapted for each user using Equation (3.7). These weights were then used in Equation (3.6) to combine the metrics of all the association rules of the user's cluster. The steps of ReCLARM's final recommendation process are as follows:

1. The user that desires a recommendation is assigned to a cluster. To make this assignment, first the values of the clustering features are extracted from the analysis of the Twitter history of the user (in the future, a survey could be used to gauge the user's preferences). Then the user is assigned to the closest cluster comparing the Euclidean distance between the user's features and the mean of the members in each cluster. The Euclidean distance was used because it was the distance metric employed in the previous k-means clustering process.

2. Then ReCLARM takes the association rules of the user's cluster and their metrics.

3. The popularity and preference metrics are computed for each rule, based on the user's data.

4. The user's personalised weights are then computed as described in Equation (3.7).

5. The metrics in Table 3.5 are combined using the WA operator to give an overall score for each rule.

6. A final selection procedure (see Algorithm 3) is used to select the set of items

to be recommended to the user.

Algorithm 3 Selection pseudocode.

```
Input: ARM rules
Output: Set of recommended items $R$
Let $N$ be the number of POIs to recommend
Let $R$ be the empty set of POIs to recommend
Sort ARM rules by their score
**while** ARM rules have not been exhausted **AND** (size of $R < N$) **do**
    Select the highest ranked rule not considered in the previous iteration
    Add POIs in the rule to $R$ (if not already in $R$)
    **for** $POI$ in $R$ **do**
        Select the highest ranked rule where $POI$ appears in the antecedent
        Add POI in the rule to $R$ (if not already in $R$)
    **end for**
    Repeat the for loop for newly added POIs
**end while**
return $R$
```

Algorithm 3 starts by ranking the association rules by their overall scores, and then the POIs in the highest ranked rule are added to the set of POIs to be recommended R. It then loops through R, adding the POIs in the highest ranked rule for which any POI in R appears in the antecedent. This process is repeated until we have the requested amount of POIs for recommendation or the association rules of the cluster have been exhausted.

3.4. Experiments and Results

3.4.1 Experimentation Details

The primary focus of the experiments was to evaluate the effect of the social media-based clustering process on the quality of the recommendations. ReCLARM was not compared directly to other tourism recommenders, as there are not any similar approaches combining clustering and association rules. We used the final data set of tweets posted in the city of Barcelona, obtained after the pre-processing and filtering steps detailed in Section 3.3.1. The dataset consisted of 37,302 tweets and 6,066 users. It was then partitioned by users—80% users in the training set and 20% in the test set—which is a common method used in machine learning studies for testing an algorithm's accuracy on data it has never seen.

The training set was run through ReCLARM's proposed stages. Firstly, the clustering features were extracted from the tweets of the users of the training set and then they were clustered using the parameters described in Section 3.3.2. Secondly, the association rules for POIs were mined with respect to each cluster, as described in Section 3.3.3. The clusters for which no rule was discovered in this stage were removed, so they were not considered when assigning users to clusters. Finally, the recommendation process was carried out as described in Section 3.3.4. The popularity ratio and preference ratio for each association rule with respect to every individual user in the test set were computed. The weighting vector was adapted for each user based on his/her interest in popular POIs. Then, all the rule evaluation metrics were combined using weighted averaging and the final selection (Algorithm 3) was carried out. It was decided to recommend 10 items to each user in the test set. POIs related to Gastronomy, Accommodation and Transportation were not considered, as we decided to focus the recommendations purely on touristic activities (a total of 1,363 in the considered data set). These 1,363 POIs were only considered for recommendations when clustering was not applied. After the clustering stage, the pool of POIs to select from was, on average, approximately 80 POIs per cluster, with very little overlap between pools. The average pairwise Jaccard similarity coefficient of all cluster pools was 0.129. The next subsection describes the metrics used to evaluate the quality of the recommendations.

3.4.2 Evaluation Metrics

The metrics were modelled as suggested by Massimo and Ricci (2021). Let U be the set of users in the test set, R the set of POIs recommended to the users in U, R_u the set of POIs recommended to a particular user u, V the set of POIs visited by users in the test set, V_u the set of POIs visited by a specific user u, and P the set of all possible recommendable POIs.

- **Average Precision (AP)**: It is the ratio of *correct* POI recommendations made to the users in the test set. A correct recommendation was determined by the user's degree of preference in the category of the POI. AP is formulated as follows:

$$AP = \frac{1}{|U||R_u|} \sum_{u \in U} \sum_{r \in R_u} 1_{CAT_{pref}} \left(\psi \left(r \right) \right) \tag{3.8}$$

  The function $\psi(.)$ gets the category of a POI. $1_{CAT_{pref}}$ is an indicator function that has value 1 if $\psi(.)$ returns a category that is preferred by the user and 0 otherwise. A category is preferred by a user if at least one of the user's tweets has been associated with it.

- **Average Category Recall (ACR)**: It is the ratio of *preferred* recommendable POIs that are actually recommended. ACR is formulated as:

$$ACR = \frac{1}{|U|} \sum_{u \in U} \frac{|R_u \cap PREF_u|}{|PREF_u|} \tag{3.9}$$

  where $PREF_u$ is the set of POIs preferred by a concrete user u in the test set. Preference is determined by the user's interest in the activity category to which the POI belongs. A category is preferred if at least one of the user's tweets has been associated to it.

- **Average Item Recall (AIR)**: It is the ratio of visited POIs that are actually recommended. AIR is formulated as:

$$AIR = \frac{1}{|U|} \sum_{u \in U} \frac{|R_u \cap V_u|}{|V_u|} \tag{3.10}$$

- **Unified Item Recall (UIR)**: It is the unified fraction of times in which the POIs recommended to a user were among the POIs actually visited by the user. The UIR is formulated as:

$$UIR = \frac{|\bigcup_{u \in U} R_u \cap V_u|}{|\bigcup_{u \in U} V_u|} \tag{3.11}$$

- **Similarity**: It measures how similar the POIs recommended to a user and the POIs visited by the user are in terms of their paths in the activity tree. It is formulated as:

$$similarity = \frac{1}{|U||R||V|} \sum_{u \in U} \sum_{r \in R_u} \sum_{v \in V_u} sim\left(\rho\left(r\right), \rho\left(v\right)\right). \tag{3.12}$$

  $sim(.,.)$ is the Jaccard similarity coefficient that estimates the similarity of two sample sets, and $\rho(.)$ produces a set representing the path of the POI in the activity tree.

- **Coverage**: It measures the width of the recommendations. It is the percentage of items recommended to the users in the test set with respect to the total set of recommendable items. It is formulated as:

$$coverage = \frac{|\bigcup_{u \in U} R_u|}{|P|} \tag{3.13}$$

- **Popularity**: It measures the degree to which the popularity of recommended POIs matches the user's popularity preference. *VPOP*$_u$ is the fraction of top 10 POIs visited by the user u, and *RPOP*$_u$ is the fraction of top 10 POIs recommended to the user.

$$VPOP_u = \frac{|P_{top10} \cap V_u|}{|V_u|} \tag{3.14}$$

$$RPOP_u = \frac{|P_{top10} \cap R_u|}{|R_u|} \tag{3.15}$$

Then, the popularity measure is computed as follows:

$$popularity = \frac{1}{|U|} \sum_{u \in U} 1 - |VPOP_u - RPOP_u| \tag{3.16}$$

- **Personalisation**: It measures the mean dissimilarity of users in the test set, based on their recommended POIs. It is formulated as:

$$personalisation = 1 - \left(\frac{1}{|C_2^{|U|}|} \sum_{u,w \in U} cossim\left(\gamma\left(u\right), \gamma\left(w\right)\right) \right) \tag{3.17}$$

$C_2^{|U|}$ is the set of all pairs of different users without repetitions, $cossim(.,.)$ is the cosine similarity of two vectors and $\gamma(.)$ generates one hot vector representing the POIs recommended to a user from all recommendable POIs.

- **Diversity**: It is the pairwise dissimilarity of POIs recommended to users in the test set, modelled after the diversity measure used in (Borràs et al., 2017). They replaced the arithmetic mean for an *ordered weighted average* (OWA) aggregation operator to avoid situations in which high values compensate for low values (e.g., aggregating (0,0,0,1,1,1) and (0.5,0.5,0.5,0.5,0.5,0.5) with arithmetic mean will have the same result, 0.5, but they are very different situations). The OWA weights are defined using a regular increasing monotone linguistic quantifier, which invokes a disjunctive policy where the lowest similarity values have higher weights. If $A = (a_1, a_2, a_3, ..., a_n)$ is the set of values to be aggregated (in decreasing order) and $W = (w_1, w_2, w_3, ..., w_n)$ is the weighting vector, where $\sum_{j=1}^{n} w_j = 1$, then OWA is defined as:

$$OWA_w(A) = \sum_{j=1}^{n} w_j a_j \tag{3.18}$$

where W is a regular increasing monotone quantifier defined as:

$$w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right) \text{ , } and\ Q(x) = x^2 \tag{3.19}$$

Diversity is then formulated as follows:

$$diversity = \frac{1}{|U|} \sum_{u \in U} \left(1 - OWA_w \left(\bigcup_{r,t \in R_u} sim\left(\rho\left(r\right), \rho\left(t\right)\right) \right) \right) \tag{3.20}$$

In that expression, $sim(.,.)$ and $\rho(.)$ are the same as in the similarity measure.

3.4.3 Results and Discussions

We performed three identical experiments to compare the results of ReCLARM with and without the clustering stage (when the clustering stage was not used, all users were assumed to belong to a single unique cluster). In each experiment, the data set was randomly split into a training set (80% of 6,066 users: 4,853) used in the clustering and association rule mining stages, and a test set (20% of 6,066 users: 1,213) used to make recommendations which were evaluated with the metrics in Section 3.4.2. Figure 3.6 shows ReCLARM's performances in the two cases. In almost all metrics, the use of clustering improved ReCLARM's performance (noticeably better in some metrics, such as coverage and personalisation). We determined the noticeable performances by the differences between the scores in the two cases. If $diff \geq 0.05$, it was noticeably better; if $0.02 \leq diff < 0.05$ it was slightly better; and $diff < 0.02$ shows a negligible difference.

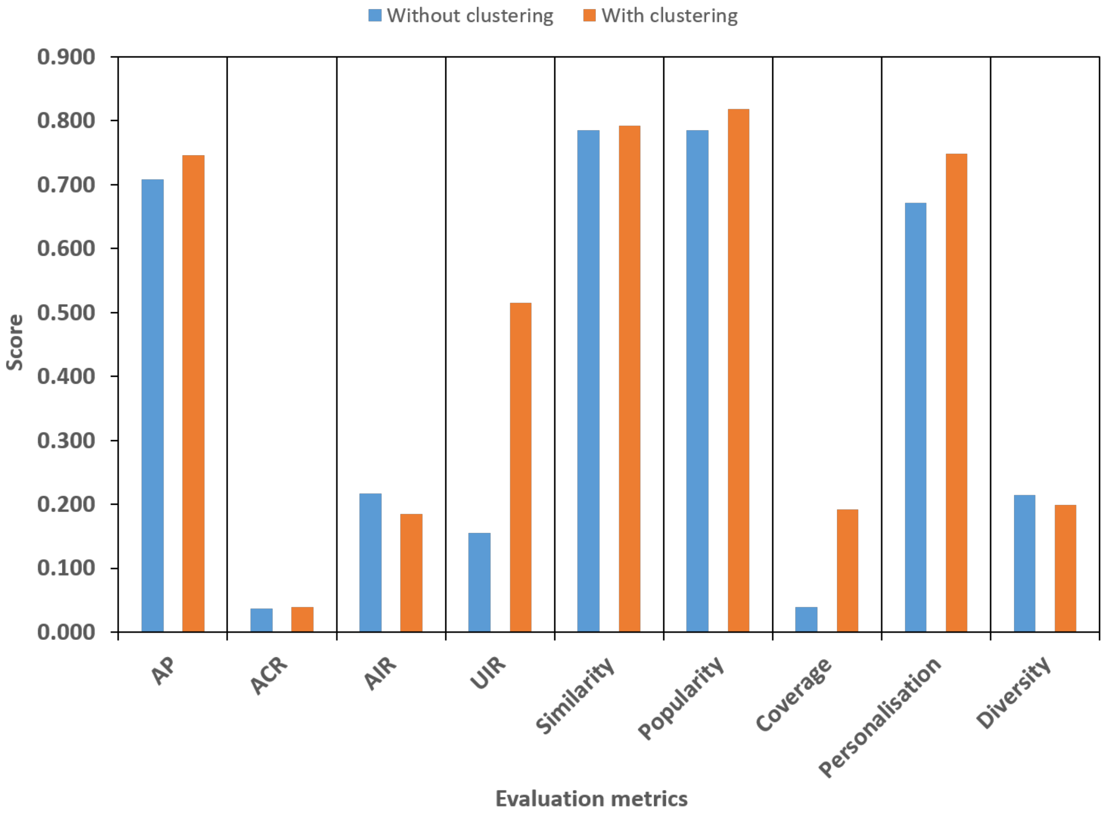


Figure 3.6: A bar plot of the average performance scores of ReCLARM across the three experiments. Cases with and without clustering are shown in orange and blue, respectively.

The actual performance scores in the experiments are detailed in Table 3.6. Noteworthy *difference* values are in bold and negligible differences are underlined. The red values show the cases in which the performance decreases when using clustering. In *UIR*, *coverage* and *personalisation*, the use of clustering improves the results noticeably.

Table 3.6

Results of evaluation with and without clustering.

Case	AP	ACR	AIR	UIR	Similarity	Popularity	Coverage	Personalisation	Diversity
Experiment 1									
Without clustering	0.707	0.036	0.223	0.161	0.785	0.779	0.039	0.654	0.220
With clustering	0.735	0.039	0.197	0.484	0.791	0.826	0.172	0.733	0.208
Experiment 2									
Without clustering	0.712	0.038	0.223	0.164	0.784	0.780	0.041	0.676	0.210
With clustering	0.762	0.040	0.181	0.543	0.794	0.806	0.206	0.753	0.191
Experiment 3									
Without clustering	0.705	0.035	0.206	0.142	0.788	0.794	0.038	0.684	0.213
With clustering	0.742	0.039	0.178	0.519	0.792	0.822	0.199	0.758	0.199
Experiment average									
Without clustering	0.708	0.036	0.217	0.156	0.786	0.785	0.039	0.671	0.214
With clustering	0.746	0.039	0.185	0.516	0.792	0.818	0.192	0.748	0.199
Difference (diff)	0.038	0.003	−0.032	**0.360**	0.006	0.034	**0.153**	**0.077**	−0.015

These metrics benefit from the clustering process because they depend on the POI pool considered for recommendation. *Unified item recall* and *coverage* increased because users were placed in clusters with pools that match their preferences, so the recommended POIs were likely to have been visited. *Personalisation* increased because the pools of different clusters were quite different, as explained in Section 3.4.1, so users across clusters received different recommendations, causing more uniqueness. The use of clustering provoked a slight decrease in *diversity*, due to data sparsity. *Diversity* scored the set of recommended items based on their dissimilarity, which was affected by the *preference ratio* in Section 3.3.4. In the experiments we used the categories visited by the test users as their preferred activity categories, and therefore, diversity dropped when there were clusters and the set of visited activity categories was small. This could be solved by incorporating mechanisms to enhance diversity in ReCLARM's algorithms (Borràs et al., 2017). A similar drop was also seen in *average item recall*, but in this case it was due to the streamlining of the POI pools for recommendation during the ARM process.

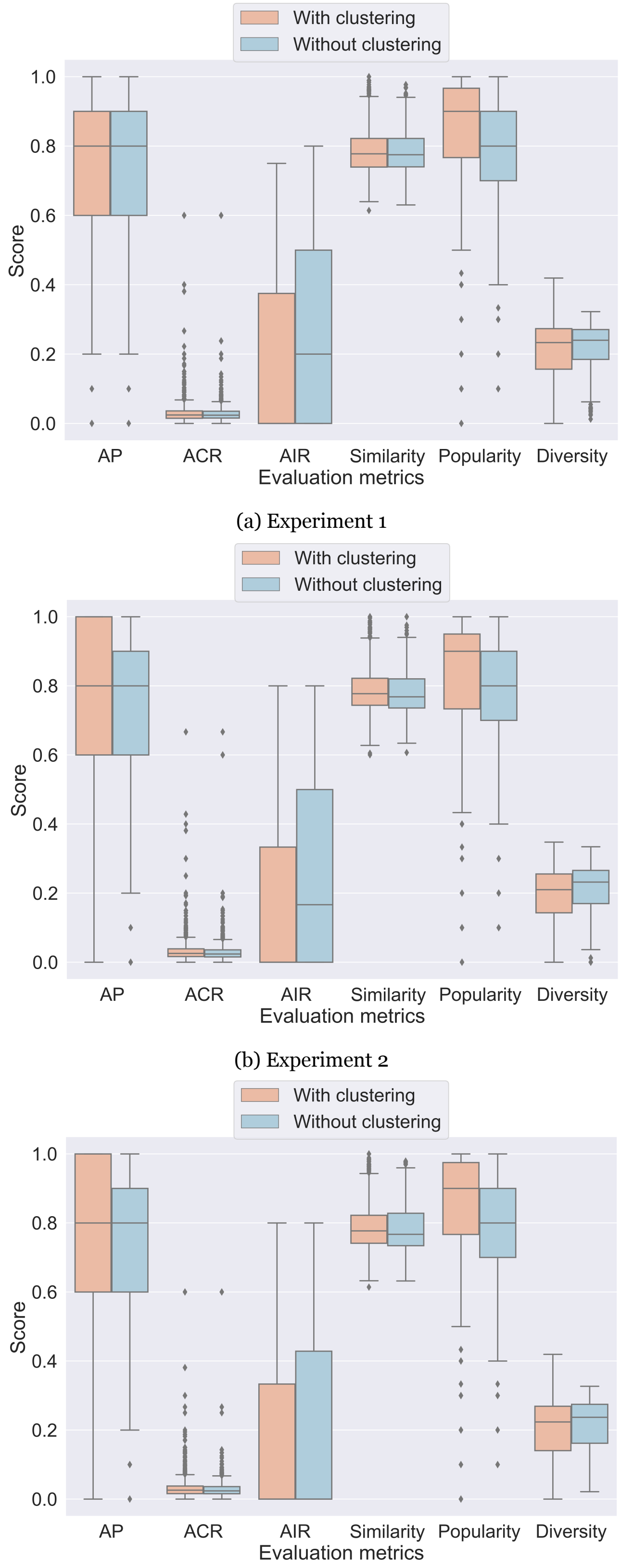
Average category recall and *similarity* were not very affected by the use of clustering; they only improved slightly. These metrics are largely dependent on data sparsity and the number of recommendable POIs, as detailed in Section 3.4.1; thus, unlike other metrics, they are indifferent to clustering. Finally, *precision* and *popularity* saw just noticeable increases when clustering was applied.

A deeper analysis of the results is presented in Figure 3.7, which shows box plots of *precision*, *category recall*, *item recall*, *similarity*, *popularity* and *diversity* scores for each user in the test set. In experiments 2 and 3, the use of clustering was able to raise the upper quartile of the precision plot to 1, so 25% of the users in the test set were given recommendations with near 100% accuracy.

However, the most important differences in the plots are present in the popularity metric. In the case with clustering, although the increase in popularity in Table 3.6 is only slightly noticeable, the box plots show that in all the experiments the majority of the test set scored better than without clustering. The outliers shown in the box plots were responsible for the lower overall difference value.

Concerning category recall, the box plots show that, although the concentration of the test set did not result in any differences in both cases, there were more outliers with good scores when clustering was used. In the case of item recall, the performance increased without clustering because a larger pool of POIs was considered, thereby creating the chance to have more POIs recommended that could have been visited by users. The diversity measure was the only one with better scores for all quartiles in all experiments when clustering was not used.

To summarise, we evaluated the ReCLARM's performance to quantify the benefit of a social media-based clustering step in the recommendation process. When clustering was applied, the performance scores increased across all metrics except diversity and item recall. The clustering process helped to refine the pools of POIs considered for recommendations.



(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

Figure 3.7: Box plots of *precision*, *recall*, *similarity*, *popularity* and *diversity* metrics for the test set in each experiment.

3.5. ReCLARM-GA: An Extension of ReCLARM

This section presents ReCLARM-GA, an extension of ReCLARM with tourist experience and convenience in focus. The goal of ReCLARM-GA is to select and order a subset of POIs recommended by ReCLARM as described in Section 3.3, such that the POIs are easy to traverse while maintaining the tourist's interests.

It is known that the order in which POIs are visited enhances tourist satisfaction. Some works have evaluated this fact by studying the impact of guided tours on tourist satisfaction (Çetinkaya and Öter, 2016). Purchasable guided tours started as a means for destination marketers to provide a way for tourist to see the essence of the destination in an effective stress-free way, while benefiting from the opportunity to increase the visibility of certain attractions and make a good profit on sales. Although performing great to this effect, guided tours are expensive, do not allow room for personal exploration and are built as a one-size-fits-all which does not consider the individual preferences of the tourist.

Over the years, quite a bit of research has gone into developing algorithms for building and recommending personalised tours (Lim et al., 2019). One of them is the Genetic Algorithm (GA), which is a heuristic search algorithm that is relatively fast at finding a good solution for a combinatorial optimization problem. A variant called Multi-Objective Genetic Algorithm (MOGA) is perfect for building personalised tours because solutions are optimized to meet multiple objectives. Notable works that use MOGA for tour recommendation start by defining constraints relevant to tourists (e.g. time, distance, budget), and then they evaluate possible solutions from a list of Points of Interest (POIs) based on these constraints while ensuring solutions match tourist preferences (Yuan and Uehara, 2019; Yochum et al., 2020; Zheng et al., 2021).

We propose an MOGA algorithm in combination with ReCLARM that addresses the problem of selecting and ordering a subset of recommendable POIs to minimize the distance between them and maximize their diversity while maintaining the ratio of popular POIs and the types of POIs preferred by the user. These four criteria are formulated into objective functions which are balanced using a weighting vector.

3.5.1 Multi-Objective Route Selection Algorithm

Problem Definition

The problem is formulated as a combinatorial optimization with the objective of finding an optimal combination of objects from a finite set of objects. In this case, we are looking for a combination of five POIs from a set of ten recommended POIs. Let $R = \{r_1, ..., r_n\}$ be the set of recommended POIs, and $S = \{s_1, .., s_k\} \equiv \{(r_{n1}, .., r_{n5}), ..., (r_{m1}, .., r_{m5})\}$ be the search space of possible solutions. The goal of the MOGA is to pick a solution s_i from S that best balances the objectives.

Our proposed algorithm comprises four objective functions that are combined using weighted averaging and a predefined weighting vector to form the fitness function. The objectives are defined as follows:

Given a possible solution $s_k = (r_{k1}, .., r_{k5})$

- **Proximity**: This objective ensures that the mean distance between adjacent POIs is minimized. It is formulated as follows:

$$Proximity_{(GA)} = \frac{1}{|s_k|} \sum_{i,j=1}^{n} Haversine\_Distance\left(\phi\left(r_{ki}\right), \phi\left(r_{kj}\right)\right) \quad (3.21)$$

where r_{ki}, r_{kj} are recommended POIs in the possible solution s_k that are adjacent in its ordering; $\phi(.)$ is a function that gets the longitude and latitude of the recommended POI; and $Haversine\_Distance(.,.)$ finds the distance in kilometers of the two recommended POIs. This objective is a minimizing function, hence solutions with lesser mean distance are better.

- **Diversity**: This objective ensures that the solution contains a diverse set of POIs. It is similar to the diversity measure in Equation (3.20). POIs are diverse when they share less categories and subcategories in their path. It is formulated as:

With *OWA*$_w$ defined as in Equation (3.18)

$$Diversity_{(GA)} = 1 - \mathit{OWA}_w \left(\bigcup_{r_{ki}, r_{kj} \in s_k} sim\left(\rho\left(r_{ki}\right), \rho\left(r_{kj}\right)\right) \right) \quad (3.22)$$

Essentially, all POIs in s_k are compared against each other and their similarity scores are aggregated using OWA, and then subtracted from 1 to get the dissimilarity. This objective is a maximizing function, hence a more diverse solution is preferred.

- **Popularity**: This objective ensures that the ratio of popular POIs in the solution matches the tourist's degree of interest in popular POIs. It is calculated just as in Equation (3.16) with the slight difference that we don't take the mean across all users. It is formulated as:

$$popularity_{(GA)} = 1 - |\mathit{VPOP}_u - \mathit{RPOP}_u| \quad (3.23)$$

VPOP$_u$ is the same as in Equation (3.14) which is the fraction of top10 POIs visited by the user, but *RPOP*$_u$ in this case is the fraction of top10 POIs in the possible solution. This objective is a maximizing function, hence solutions with a better match are preferred.

- **Preference**: This objective ensures that the POIs in the solution are interesting to the tourist. It is the fraction of relevant POIs in the solution. Relevant

POIs are categorised under types of activities frequently visited by the tourist, obtained with an analysis of the places from which the user has tweeted. It is formulated as:

$$Preference_{(GA)} = \frac{1}{|s_k|} \sum_{r \in s_k} 1_{CAT_{pref}} (\psi(r)) \tag{3.24}$$

Where $1_{CAT_{pref}}$ and $\psi(.)$ are the same as in Equation (3.8). This objective is a maximizing function, hence a solution with more relevant POIs is preferred.

Algorithm Process

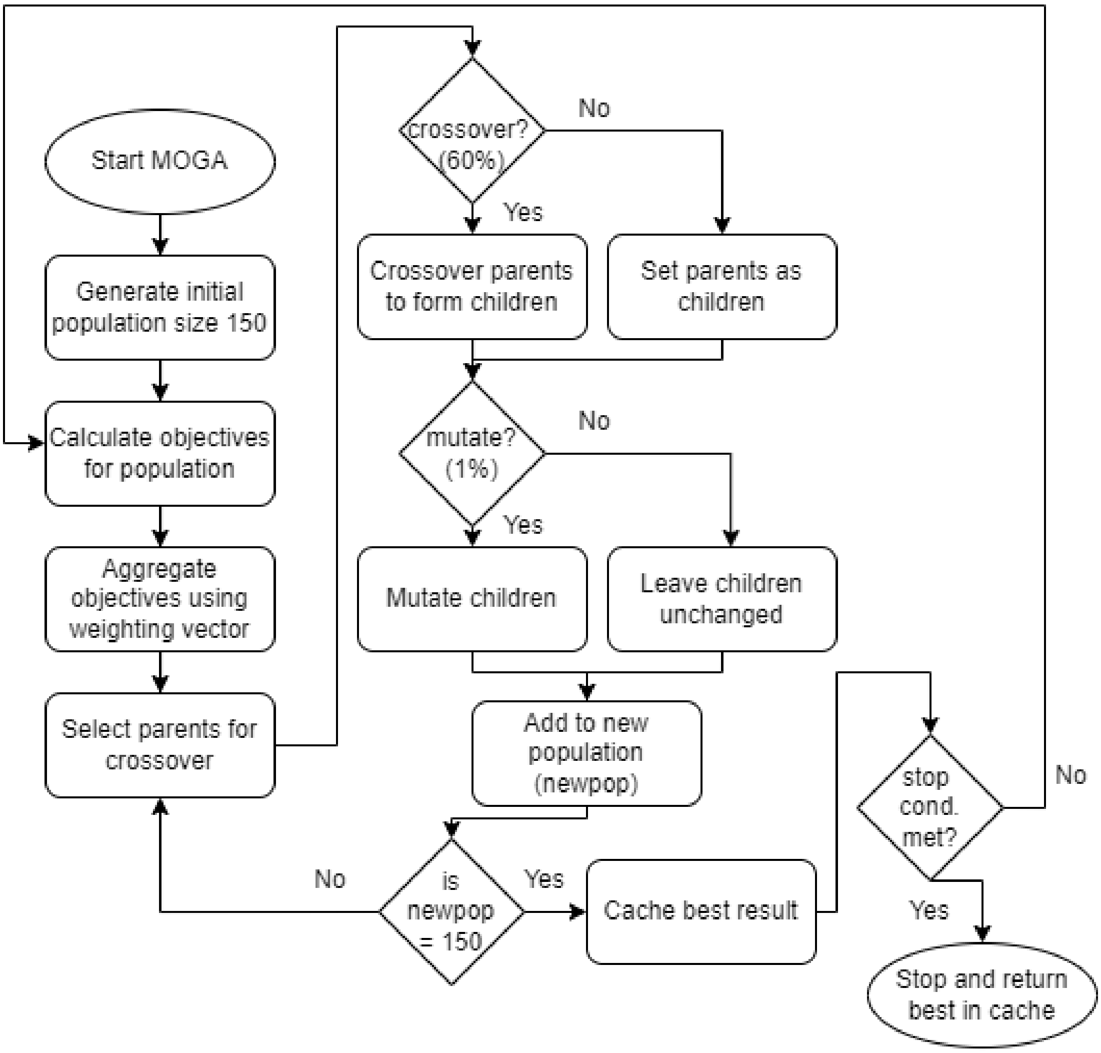

Start MOGA
Generate initial population size 150
Calculate objectives for population
Aggregate objectives using weighting vector
Select parents for crossover
crossover? (60%)
No
Yes
Crossover parents to form children
Set parents as children
mutate? (1%)
No
Yes
Mutate children
Leave children unchanged
Add to new population (newpop)
is newpop = 150
No
Yes
Cache best result
stop cond. met?
No
Yes
Stop and return best in cache

Figure 3.8: Flow chart of MOGA process.

The main steps of the MOGA are the following (as illustrated in Figure 3.8):

Step 1. Generate randomly an initial population, which is a subset of S of size 150.

Step 2. Calculate the objective functions and aggregate them using weighted averaging to get the fitness values of all the members of the population.

Step 3. Select two solutions with a good fitness value (i.e. better objective function score) from the population as *parents*, then *crossover* them to form two *children* according to a crossover rate (60%), and then *mutate* children by swapping the positions of POIs according to a mutation rate (1%). If children are fitter than their parents they are added to the new population, otherwise parents are added to the new population (*weak parent replacement*).

Step 4. Repeat step 3 until the new population is up to size 150, signaling the completion of generation 1. Cache the best solution in the new population.

Step 5. Repeat steps 2 to 4, to move through generations replacing the cached best solution with any better solution. If there is no better solution for 20 generations or 100 generations are completed, stop and return the best solution.

3.5.2 Experiments and Results

Experiment Details

The experiments in this section were an extension of those performed on ReCLARM in Section 3.4.1. The dataset consisted of the 1,140 Twitter users from the test set with their 10 recommended POIs.

The weighting vector [proximity:0.4, diversity:0.3, popularity:0.1, preference:0.2] was used in the fitness function to aggregate our 4 objectives. This vector was constructed to assign higher influence to the proximity and diversity of POIs, and lesser influence to popularity and preference. Proximity and diversity are the most important aspects when visiting a group of POIs within the same day (you wouldn't want to visit similar places or travel long distances). Also, the popularity and preference objectives are already enforced by ReCLARM, so the POIs to select and order already match the user's preferences and interest in popular places.

Each user's POIs were passed through the MOGA and the fitness of the results was compared against the following four baseline algorithms:

1. *First five*: Select the first 5 recommended POIs as the solution (i.e. the ones considered better by ReCLARM).
2. *Random five*: Select randomly 5 of the 10 recommended POIs as solution.
3. *Minimize distance first start (MDFS)*: Select 5 of the 10 recommended POIs that minimize the distance between adjacent POIs using a greedy algorithm with the following steps:
   (a) Add the *first* POI from the 10 recommended POIs to the initial solution.

(b) Add a POI in front of the last POI in solution or behind the first POI of the solution with the least distance to travel.

(c) Repeat step b until the solution contains 5 POIs.

4. *Minimize distance random start (MDRS)*: This algorithm is the same as MDFS but the first POI added to the initial solution is picked randomly.

Additionally, reference values which are the optimal solutions for each user were obtained by sequential checking all the possible permutations of 5 POIs from the user's 10 recommended POIs, and selecting the solution with the best fitness value. These reference values were also included in the comparison of algorithms.

Results

Table 3.7 showcases the fitness value of the MOGA algorithm against all baseline algorithms including reference values, highlighting the minimum, maximum and mean values across all 1,140 users. MOGA performs significantly better than all baseline algorithms, and almost mirrors the reference values.

Table 3.7

Minimum, maximum and mean fitness values for all algorithms, including the reference values for the optimum solutions.

Fitness value	First five	Random five	MDFS	MDRS	MOGA	Reference
Minimum	0.297	0.284	0.305	0.309	0.479	0.480
Maximum	0.665	0.701	0.681	0.689	0.748	0.748
Mean	0.512	0.509	0.528	0.524	0.640	0.641

This result is further enforced in Figure 3.9, which shows a box plot of MOGA, the baseline algorithms, and reference values. The MOGA results outperform the baseline algorithms, as the lower quartile of MOGA sits above the upper quartiles of the rest. It also perfectly matches the Reference plot. A sequential check of all possible solutions is feasible in this case due to a relatively small search space of 30,240 options (which is displayed in our reference values), but a test showed the MOGA to be 27% faster with a runtime of 7 minutes per case compared to 11 when searched sequentially, while obtaining the same optimal result.

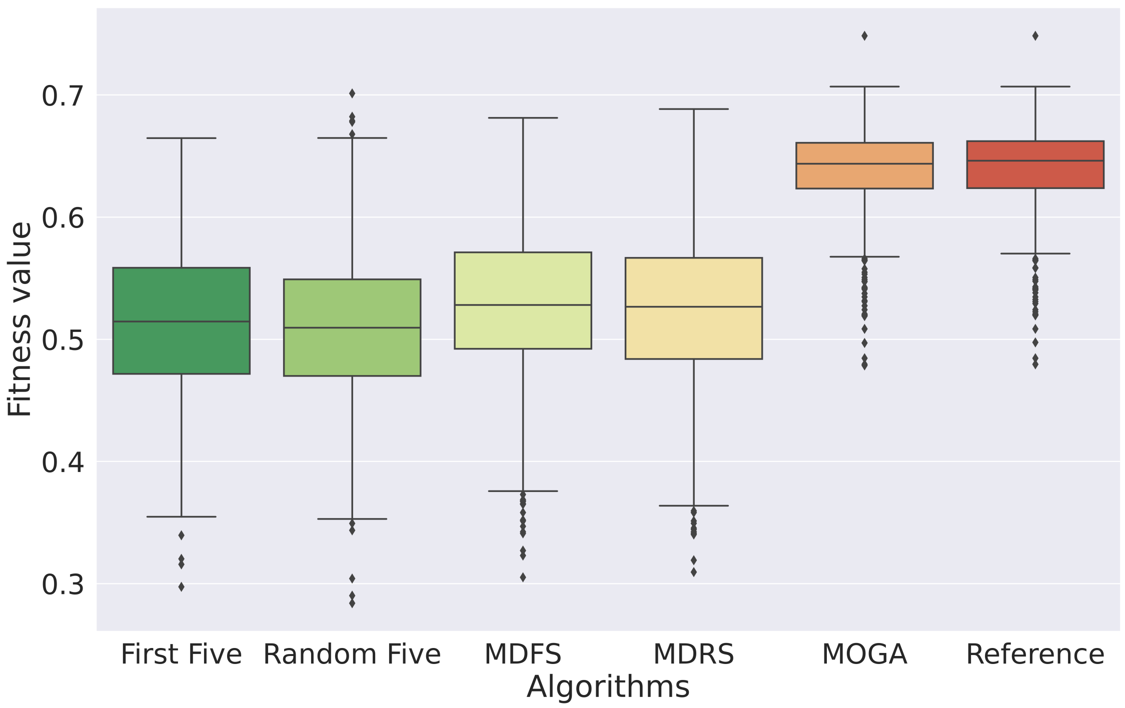


Figure 3.9: Box plots of fitness values for all algorithms including reference values for the optimal solutions.

3.6. Conclusion

In this chapter we have proposed ReCLARM—a cluster-based user profiling technique combined with association rule mining to recommend points of interest to tourists. This technique focuses on user features extracted from social media that represent interests, habits and context. The intention of the clustering step is not to obtain any predefined number of classes of users, but rather to learn abstract generalisations implicit in the data set.

The resulting user profiles are then filtered using association rule mining to recommend touristic activities with high affinity. We evaluated the approach with geolocated tweets from Twitter, posted in the city of Barcelona in 2019, and performed a comparative analysis of the results with and without the clustering technique. We conclude that the clustering approach improves the system's accuracy and its ability to encapsulate a user's interests, by refining the pool of recommendable items to a user with the help of the association rules of his/her associated cluster.

We have also proposed an extension to ReCLARM called ReCLARM-GA, which uses a multi-objective genetic algorithm for selecting and ordering a fixed set of POIs from the larger set of recommended POIs suggested by ReCLARM. It works by balancing four objective functions, *proximity*, *diversity*, *popularity*, and *preference* using weighted averaging.

Overall, the main contributions of this chapter are the following:

- We introduced a method for analysing data from social media to build user profiles that encapsulates their travel preferences and habits.
- We presented insights into the potential benefits of the combination of cluster analysis and association rule mining in tourism recommenders.
- We provided the results of in-depth experiments using a comprehensive set of numerical evaluation metrics to gauge the benefits of social media-based clustering for user profiling.

In the recommendation process, the initial data collection and pre-processing steps were the most challenging, due to gaps in the Twitter data, especially in cases where tweets were not geolocated. This created a partial picture when analysing user preferences which is a recurring problem in social media analysis due to data inconsistency as described in Chapter 1. It was also challenging to find a suitable number of clusters in the user profiling step, since we weren't concerned with the regular segmentation that is achievable using clustering evaluation metrics like *silhouette score*. In addition, the unexpected negative effect of COVID-19 restrictions on tourism left 2019 as the only viable year for the analysis. Finally, we were unable to recommend diverse items without explicitly defining diversity mechanisms in the system.

In the future, it might be necessary to incorporate more user attributes and context into the clustering features, considering categorical features, and experimenting with k-mediods as the clustering algorithm. Finally, it may also be beneficial to test ReCLARM in the real-world to properly fine tune its algorithms.

The next chapter focuses on the more touristic aspects of the analysis in this chapter. It focuses on understanding the types of tourists represented in each cluster in the clustering phase of ReCLARM, and the mobilitity and flow patterns that can be observed from the data of these tourists.

Chapter 4

Identification of Mobility Patterns of Clusters of City Visitors

4.1. Introduction

This chapter focuses on identifying mobility patterns of clusters of tourists. As already shown in Chapter 3, cluster analysis is beneficial in profiling unique groups of tourists to learn their preferences. This is also true in understanding the mobility of tourists at a destination. The resulting clusters of the clustering step described in Chapter 3 could be studied to identify co-visited attractions and the order in which they are visited to analyse tourists flows within a destination.

Leveraging clustering in identifying tourists' mobility patterns enables learning about specific user groups and their travel behaviour. It is then possible to characterise these clusters by the demographic data and visit preferences of their members to discover new insights that are useful to DMOs when suggesting personalised services as described in Section 2.4.2. Additionally, it is also beneficial in battling data sparsity that plagues mobility and sequential studies.

Therefore, in this chapter, we aim to develop a methodology that, through the application of machine learning techniques to geotagged social media data, creates clusters of tourists according to their visiting preferences and travel habits at the destination. The analytical and managerial goal is to know the most visited points of interest and the main flows between them within a destination for each tourist cluster. This will allow the DMOs to uncover their tourists' profiles, their preferences, and their mobility patterns at the destination. This will also let them enhance the visitors experience through the personalisation of tourist packages, services, public transport and also the information offered on each attraction or spot. Finally, this will also enhance the usefulness and efficiency of DMOs to mitigate the array of problems caused by the mobility of visitors towards and around the main points of interest through a better experience design management and interaction with the environment (Anton Clavé, 2019).

The rest of this chapter is structured as follows. Section 4.2 briefly describes the methodology for creating clusters of tourists from social media data. Section 4.3 characterises the visitors present in the generated clusters. Section 4.4 then studies the mobility and flow of the characterised visitors within each cluster. Finally, the discussion, conclusion, and implications of the study are highlighted in Section 4.5.

4.2. Methodology

The methodology is directly linked to the one of Chapter 3. We structured the analysis in this chapter to characterise and study the essence of the clusters generated in Chapter 3. This means the dataset and the clustering process is identical to those used in the experiments in Chapter 3.

In this case, however, we consider all 6,066 users in the clustering process instead of splitting them into a training set and test set. Also, the user's origin is considered after the clustering process to introduce users' demographic data to the analysis. Figure 4.1 shows the percentage user origin in the 25 clusters generated using the clustering process. The top 5 countries of origin (*Spain*, *Italy*, *France*, *United Kingdom (UK)*, *United States of America (USA)*) represented in the dataset are considered, while other countries are classified under *Others*. It can be seen that some clusters are more representative of specific user origins than others.

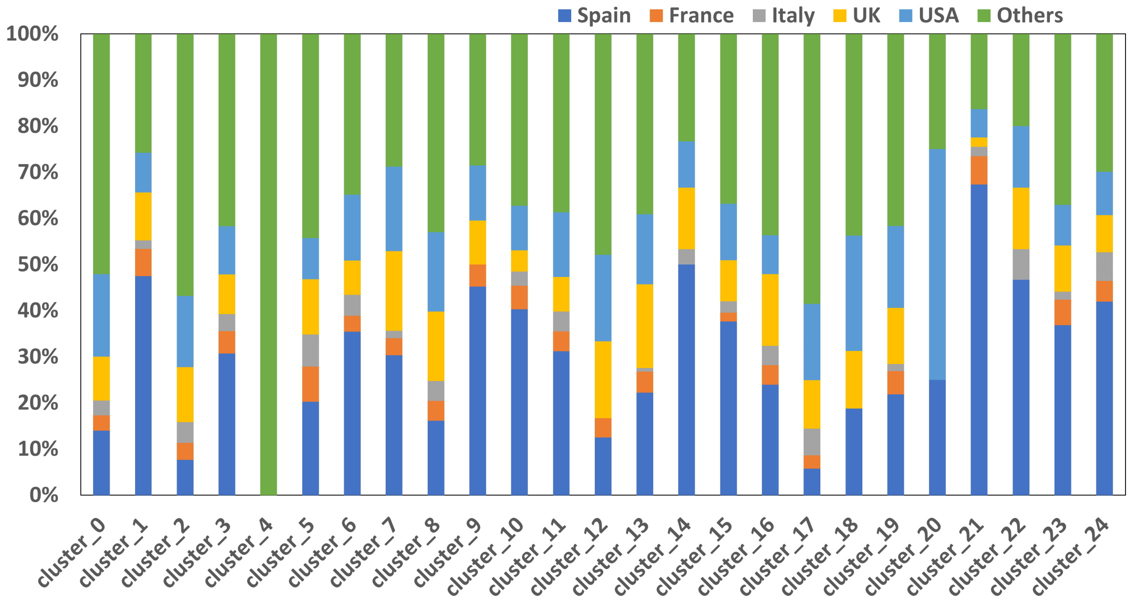


Figure 4.1: Percentage of user country of origin in clusters.

Not all 25 clusters generated in the clustering process qualify for the mobility and flow analysis phase. Five of the 25 clusters had less than 50 users, so they were not considered to be very relevant, and they were dismissed from the posterior mobility analysis. Thus, at the end of the clustering process, there were 20 clusters with a minimum of 50 visitors. By averaging the values of the clustering features (detailed in Section 3.3.2) of the users in each cluster, we obtained a general description of the preferences of the users in that group.

4.3. Characterisation of Visitors in Clusters

In this section, we present a characterisation of the members of the 20 relevant clusters to make a connection between cluster members and their mobilities. Table 4.1 gives a textual description of the common features of the members of each cluster with at least 50 users. The features in **bold** (religious and historic sites) appear in many clusters, showing the big strength of some of the top POIs in Barcelona like Sagrada Família or the Cathedral, and they can also be considered the main tourism themes of Barcelona. Those underlined (recreational facilities, architecture and museums) appear in 3-4 clusters, showing the popularity of places like the Camp Nou football stadium or famous buildings like Casa Batlló or Casa Milà (La Pedrera). The *italized* features are distinctive of that particular cluster (e.g. shopping interest in cluster 6).

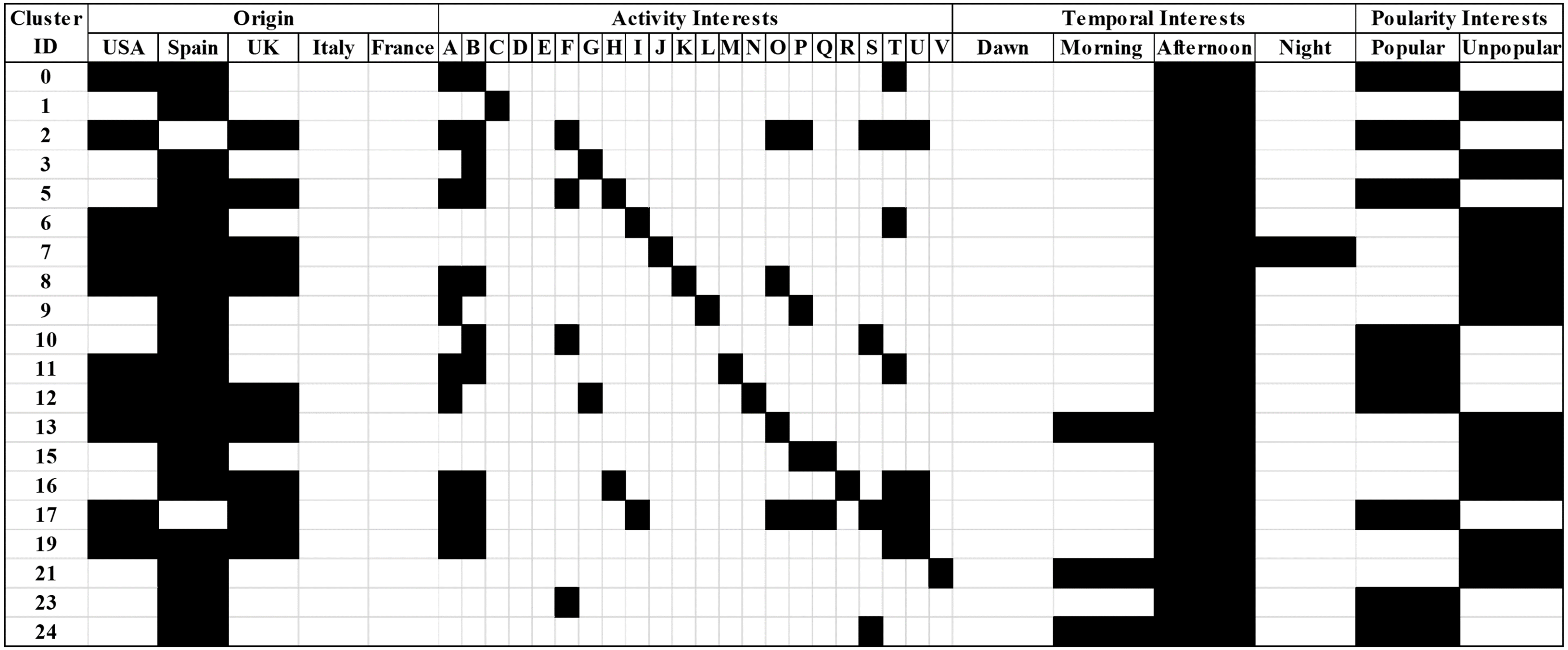
Table 4.1

Cluster description.

Cluster ID	Description of the common tourists in the cluster
0	American and Spanish people interested in **religious** and **historic** sites, that enjoy visiting the most popular places, especially in the afternoons.
1	Spanish people interested in *scenic routes* that enjoy visiting the least popular places, especially in the afternoons.
2	American and British people interested in a wide variety of popular POIs including **religious** sites, **historic** sites, recreational facilities, and *nature*, that enjoy visiting the most popular places, especially in the afternoons.
3	Spanish people very interested in the *beach*, who enjoy visiting the less common places, especially in the afternoons.
5	Spanish and British people very interested in *cultural amenities*, recreational facilities, **historic** and **religious** sites, that enjoy visiting the more popular places, especially in the afternoons.
6	American and Spanish people very interested in *shopping* (malls and markets) and architecture, who enjoy visiting the least popular places, especially in the afternoons.
7	Spanish, British, and American people very interested in *nightlife* (bars, clubs, etc.), that enjoy visiting the least popular places, especially in the afternoons and night.
8	Spanish, British, and American people very interested in *statues* and a bit of **historic** and **religious** POIs, that enjoy visiting the semi-popular and least popular places, especially in the afternoons.

9	Spanish people very interested in *health and care* (beauty salons, etc.), that enjoy visiting the least popular places, especially in the afternoons.
10	Spanish people very interested in recreational facilities (Stadiums), and **historic** sites and museums, who enjoy visiting the more popular places, especially in the afternoons.
11	American and Spanish people very interested in *viewpoints*, but also architecture, **historic** and **religious** sites, that enjoy visiting popular places, especially in the afternoons.
12	Spanish, British, and American people very interested in *sculptures*, also *beaches* and **religious** POIs, that enjoy visiting the more popular places, especially in the afternoons.
13	Spanish, British, and American people very interested in *accommodation* (hotels, resorts, hostels, camp sites etc.), that enjoy visiting the least popular places, especially in the mornings and afternoons.
15	Spanish people very interested in *food* (restaurants and cafes), and a bit of *enotourism*, that enjoy visiting the least popular places, especially in the afternoons.
16	Spanish and British people very interested in *amusement parks*, a bit of *cultural amenities*, **historic** and **religious** POIs, that enjoy visiting the less popular places, especially in the afternoons.
17	American and British people interested in *food*, *enotourism*, *accommodation*, museums, architecture and **religious** sites, that enjoy visiting the least popular places, especially in the afternoons.
19	Spanish, British, and American people very interested in *artworks*, architecture, **historic** and **religious** sites, that enjoy visiting the least popular places, especially in the afternoons.
21	Spanish people very interested in *art centres*, who enjoy visiting the least popular places, especially in the mornings and afternoons.
23	Spanish people very interested in recreational facilities (specifically, Camp Nou), that enjoy visiting the most popular places, especially in the afternoons.
24	Spanish people very interested in museums, who enjoy visiting the most popular places, especially in the mornings and afternoons.

Figure 4.2 provides a graphical representation of the information in Table 4.1, highlighting the main characteristics shared by tourists in each of the 20 clusters, including their origin, the kind of leisure activities they visit, their preferred time of the day, and their interest in different kinds of activities. All clusters are associated with one or more types of activities, which match the activity features used in the clustering process.



Cluster	Origin					Activity Interests																						Temporal Interests				Poularity Interests	
ID	USA	Spain	UK	Italy	France	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	Dawn	Morning	Afternoon	Night	Popular	Unpopular
0	■	■				■	■																		■					■		■	
1		■						■																						■			■
2	■		■			■	■				■									■	■			■	■	■				■		■	
3		■					■					■																		■			■
5		■	■			■	■				■		■																	■		■	
6	■	■												■											■					■			■
7	■	■	■												■															■	■		■
8	■	■	■			■	■									■				■										■			■
9		■				■											■				■									■			■
10		■					■				■													■						■		■	
11	■	■				■	■											■							■					■		■	
12	■	■	■			■						■							■											■		■	
13	■	■	■																	■									■	■			■
15		■																			■	■								■			■
16		■	■			■	■						■										■		■	■				■			■
17	■		■			■	■							■						■	■	■		■	■	■				■		■	
19	■	■	■			■	■																		■	■				■			■
21		■																									■		■	■			■
23		■									■																			■		■	
24		■																						■					■	■		■	

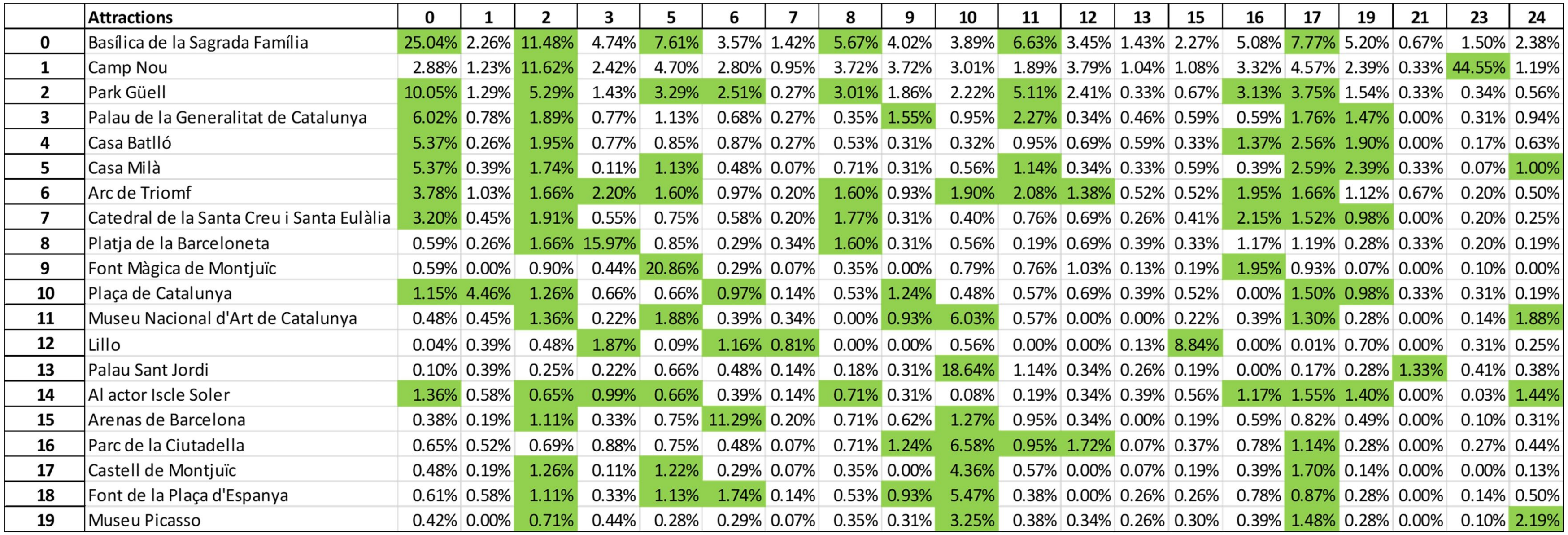
Figure 4.2: Cluster characteristics highlighting tourists' origin, preferred tour time, orientation towards popular POIs, and interests in different activities. Note: A—Religious, B—Historic, C—Routes, D—Nature, E—Art Gallery, F— Recreational Facilities, G—Beach, H—Cultural Amenities, I—Shopping, J—Nightlife, K—Statues, L—Health & Care, M—Viewpoints, N—Sculptures, O—Accommodation, P—Food, Q—Enotourism, R—Amusement Parks, S—Museums, T—Architecture, U—Other Artworks, V—Art Centers.

As it can be seen in Figure 4.2, the historic and religious activities (columns A and B) are the ones associated with most clusters (0, 2, 5, 8, 11, 16, 17, and 19). In most of these clusters, users enjoy visiting the most popular POIs, revealing that the historic and religious attractions are among the most popular in the city of Barcelona. This also proves a direct correlation between tourist interests and their popularity, as it could be expected.

It can also be seen that the individuals in clusters 2 and 17 visit a larger number of popular POIs and have a wider spread of different kinds of activities. These clusters contain mostly non-Spanish tourists, who visit and experience many varieties of POIs that the city offers. On the other hand, clusters characterized mainly by Spanish nationals (clusters 1, 3, 9, 10, 15, 21, 23, and 24) focus on a small set of features and on unpopular POIs, indicating a higher specificity in their favorite POIs and their visits. People in cluster 1, for example, prefer to do scenic routes within the city. However, those in cluster 3 prefer to go to the beach or places near the beach, and some historic sites. Tourists on cluster 9 visit Barcelona mainly for health and care, although they also visit religious sites and cafes, whereas those in cluster 15 prefer food and enotourism, those in cluster 23 are mainly interested in visiting Camp Nou, the football stadium, and people in cluster 24 mainly visit museums.

The mobility analysis based on visitor clusters that was carried out also allows to know the specific POIs visited by the tourists of each cluster. Figure 4.3 shows the percentage of tourists from each cluster who visit the top 20 most visited POIs of the destination, highlighting in each row the values above the average. For example, Camp Nou, the Barcelona football stadium, is visited mostly by tourists from cluster 23, the religious architectures (as the Sagrada Família) and historic monuments (as the Palau de la Generalitat or the Casa Batlló) are mainly visited by tourists of clusters 0, 2, or 17, and beaches (like Barceloneta) or places near them by tourists of cluster 3. This figure also clearly identifies those clusters that are focused on the most popular POIs, and to what degree. Clusters 0 and 2 are heavily focused on popular POIs as they have an above average representation of 55% and 80% (respectively) in the popular POIs. In contrast, clusters 7, 12, 13, 15, 21, and 23 have very little representation in the popular POIs. This shows that the 20 clusters do not favor only popular or unpopular POIs.

In summary, the clusters present unique groups of tourists with different characteristics and preferences derived from the clustering features and allow knowing which are the visitors of the most popular tourist POIs. This information can be very useful for the marketing managers of destinations and tourist attractions because it summarises the visiting preferences of their visitors. Moreover, these data can also be exploited for mobility analysis, as described in the next section.



	Attractions	0	1	2	3	5	6	7	8	9	10	11	12	13	15	16	17	19	21	23	24
0	Basílica de la Sagrada Família	**25.04%**	2.26%	**11.48%**	4.74%	**7.61%**	3.57%	1.42%	**5.67%**	4.02%	3.89%	**6.63%**	3.45%	1.43%	2.27%	5.08%	**7.77%**	5.20%	0.67%	1.50%	2.38%
1	Camp Nou	2.88%	1.23%	**11.62%**	2.42%	4.70%	2.80%	0.95%	3.72%	3.72%	3.01%	1.89%	3.79%	1.04%	1.08%	3.32%	4.57%	2.39%	0.33%	**44.55%**	1.19%
2	Park Güell	**10.05%**	1.29%	**5.29%**	1.43%	**3.29%**	**2.51%**	0.27%	**3.01%**	1.86%	2.22%	**5.11%**	2.41%	0.33%	0.67%	**3.13%**	**3.75%**	1.54%	0.33%	0.34%	0.56%
3	Palau de la Generalitat de Catalunya	**6.02%**	0.78%	**1.89%**	0.77%	1.13%	0.68%	0.27%	0.35%	**1.55%**	0.95%	**2.27%**	0.34%	0.46%	0.59%	0.59%	**1.76%**	**1.47%**	0.00%	0.31%	0.94%
4	Casa Batlló	**5.37%**	0.26%	**1.95%**	0.77%	0.85%	0.87%	0.27%	0.53%	0.31%	0.32%	0.95%	0.69%	0.59%	0.33%	**1.37%**	**2.56%**	**1.90%**	0.00%	0.17%	0.63%
5	Casa Milà	**5.37%**	0.39%	**1.74%**	0.11%	**1.13%**	0.48%	0.07%	0.71%	0.31%	0.56%	**1.14%**	0.34%	0.33%	0.59%	0.39%	**2.59%**	**2.39%**	0.33%	0.07%	**1.00%**
6	Arc de Triomf	**3.78%**	1.03%	**1.66%**	**2.20%**	**1.60%**	0.97%	0.20%	**1.60%**	0.93%	**1.90%**	**2.08%**	**1.38%**	0.52%	0.52%	**1.95%**	**1.66%**	1.12%	0.67%	0.20%	0.50%
7	Catedral de la Santa Creu i Santa Eulàlia	**3.20%**	0.45%	**1.91%**	0.55%	0.75%	0.58%	0.20%	**1.77%**	0.31%	0.40%	0.76%	0.69%	0.26%	0.41%	**2.15%**	**1.52%**	**0.98%**	0.00%	0.20%	0.25%
8	Platja de la Barceloneta	0.59%	0.26%	**1.66%**	**15.97%**	0.85%	0.29%	0.34%	**1.60%**	0.31%	0.56%	0.19%	0.69%	0.39%	0.33%	1.17%	1.19%	0.28%	0.33%	0.20%	0.19%
9	Font Màgica de Montjuïc	0.59%	0.00%	0.90%	0.44%	**20.86%**	0.29%	0.07%	0.35%	0.00%	0.79%	0.76%	1.03%	0.13%	0.19%	**1.95%**	0.93%	0.07%	0.00%	0.10%	0.00%
10	Plaça de Catalunya	**1.15%**	**4.46%**	**1.26%**	0.66%	0.66%	**0.97%**	0.14%	0.53%	**1.24%**	0.48%	0.57%	0.69%	0.39%	0.52%	0.00%	**1.50%**	**0.98%**	0.33%	0.31%	0.19%
11	Museu Nacional d'Art de Catalunya	0.48%	0.45%	**1.36%**	0.22%	**1.88%**	0.39%	0.34%	0.00%	**0.93%**	**6.03%**	0.57%	0.00%	0.00%	0.22%	0.39%	**1.30%**	0.28%	0.00%	0.14%	**1.88%**
12	Lillo	0.04%	0.39%	0.48%	**1.87%**	0.09%	**1.16%**	**0.81%**	0.00%	0.00%	0.56%	0.00%	0.00%	0.13%	**8.84%**	0.00%	0.01%	0.70%	0.00%	0.31%	0.25%
13	Palau Sant Jordi	0.10%	0.39%	0.25%	0.22%	0.66%	0.48%	0.14%	0.18%	0.31%	**18.64%**	1.14%	0.34%	0.26%	0.19%	0.00%	0.17%	0.28%	**1.33%**	0.41%	0.38%
14	Al actor Iscle Soler	**1.36%**	0.58%	**0.65%**	**0.99%**	**0.66%**	0.39%	0.14%	**0.71%**	0.31%	0.08%	0.19%	0.34%	0.39%	0.56%	**1.17%**	**1.55%**	**1.40%**	0.00%	0.03%	**1.44%**
15	Arenas de Barcelona	0.38%	0.19%	**1.11%**	0.33%	0.75%	**11.29%**	0.20%	0.71%	0.62%	**1.27%**	0.95%	0.34%	0.00%	0.19%	0.59%	0.82%	0.49%	0.00%	0.10%	0.31%
16	Parc de la Ciutadella	0.65%	0.52%	0.69%	0.88%	0.75%	0.48%	0.07%	0.71%	**1.24%**	**6.58%**	**0.95%**	**1.72%**	0.07%	0.37%	0.78%	**1.14%**	0.28%	0.00%	0.27%	0.44%
17	Castell de Montjuïc	0.48%	0.19%	**1.26%**	0.11%	**1.22%**	0.29%	0.07%	0.35%	0.00%	**4.36%**	0.57%	0.00%	0.07%	0.19%	0.39%	**1.70%**	0.14%	0.00%	0.00%	0.13%
18	Font de la Plaça d'Espanya	0.61%	0.58%	**1.11%**	0.33%	**1.13%**	**1.74%**	0.14%	0.53%	**0.93%**	**5.47%**	0.38%	0.00%	0.26%	0.26%	0.78%	**0.87%**	0.28%	0.00%	0.14%	0.50%
19	Museu Picasso	0.42%	0.00%	**0.71%**	0.44%	0.28%	0.29%	0.07%	0.35%	0.31%	**3.25%**	0.38%	0.34%	0.26%	0.30%	0.39%	**1.48%**	0.28%	0.00%	0.10%	**2.19%**

Figure 4.3: Top 20 most visited attractions and the percentage of their tweets per cluster, highlighting values above average in green.

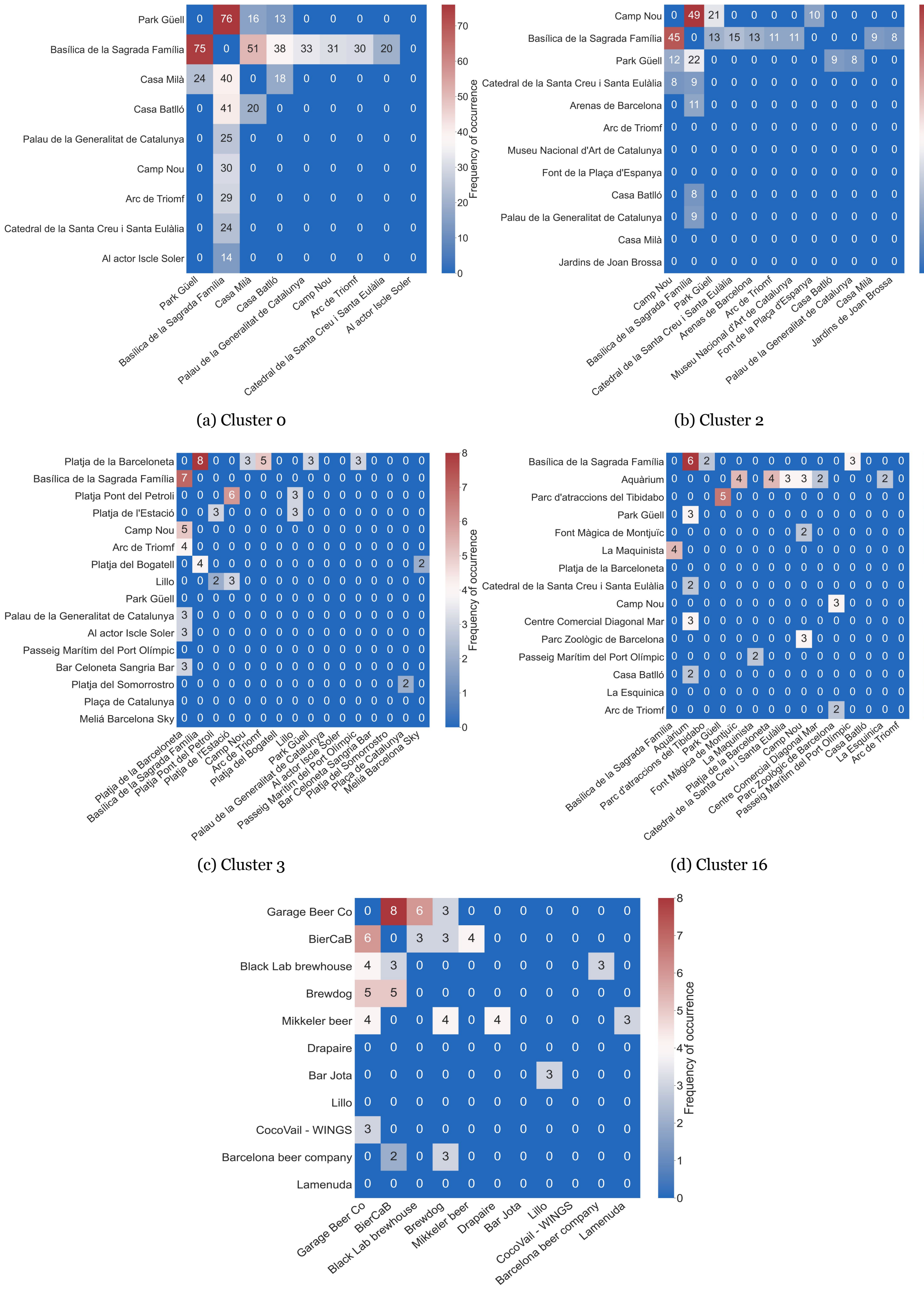
4.4. Tourist Mobility/Flow Analysis of visitor clusters

To analyze the mobility of tourists within each cluster, bigrams (A, B) were extracted from their sequences of visits. These bigrams are *n*-grams of size two that represent the movement of a tourist from A to B (i.e., the user sent a tweet from A and the next one from B). Given a sequence of items, *n*-grams are unique sets of *n* directly adjacent items. For example, a sequence $S = \{a, b, c, d\}$ has the following 2-grams (popularly known as bigrams) $2grams = \{(a, b), (b, c), (c, d)\}$. *N*-grams must maintain the order in the original sequence and must be unique.

After the bigrams were extracted from the sequences of visited places of each tourist in a certain cluster, we then counted their frequency of occurrence (i.e., the number of times a bigram appeared in the cluster). The clusters 0, 2, 3, 7, and 16 were selected to illustrate this analysis because of their diversity (additional charts of the other clusters are shown in Appendix A.2). The top 20 bigrams with a higher frequency in each cluster were taken as the most relevant. Figure 4.4 shows heat map plots of the tourist mobility within clusters, where the color intensity represents the movement between POIs measured by the frequency of occurrence.

As it can be seen from Figure 4.4, *Basílica de la Sagrada Família* acts as a hub in cluster 0, as all other POIs are directly connected to it. In most cases, tourists are visiting the other POIs after visiting *Basílica de la Sagrada Família* because its outflows exceed its inflows from other locations, except *Casa Batlló*, *Catedral de la Santa Creu i Santa Eulàlia*, and *Al actor Iscle Soler*, which might be a result of route preference. The strongest connection can be seen between *Basílica de la Sagrada Família* and *Park Güell* with almost equivalent inflow and outflow between them, which is most probably due to the fact that both attractions feature architecture designed by Antoni Gaudí (a famous Catalan architect). In the case of cluster 2, *Basílica de la Sagrada Família* is once again the most interconnected attraction but with more inflows than in cluster 0, and the strongest connection is between *Basílica de la Sagrada Família* and *Camp Nou*. The other selected clusters (3, 7, and 16) show connections between various attractions with no clear hub.

Figures 4.5 and 4.6 help to further understand tourist mobility with network graphs plotted on the Barcelona city map. Nodes represent POIs, and edges are the bigrams that connect them (the wider is the edge, the more tourists travel between those two locations). It can be seen that proximity plays a role in why *Basílica de la Sagrada Família* acts as a hub in clusters 0 and 2. In cluster 3, the focus is the Mediterranean Sea as most POIs are near the beach, except in the case of *Basílica de la Sagrada Família*, *Park Güell*, and *Camp Nou*, as tourists are willing to travel out of the way to see these POIs. Clusters 7 and 16 show two different kinds of tourists in terms of mobility. The former is focused on bars in close proximity and near the city center, whereas the latter visits many different kinds of places all around the city, including amusement parks, shop malls, and the beach, but also the most popular venues.



(a) Cluster 0

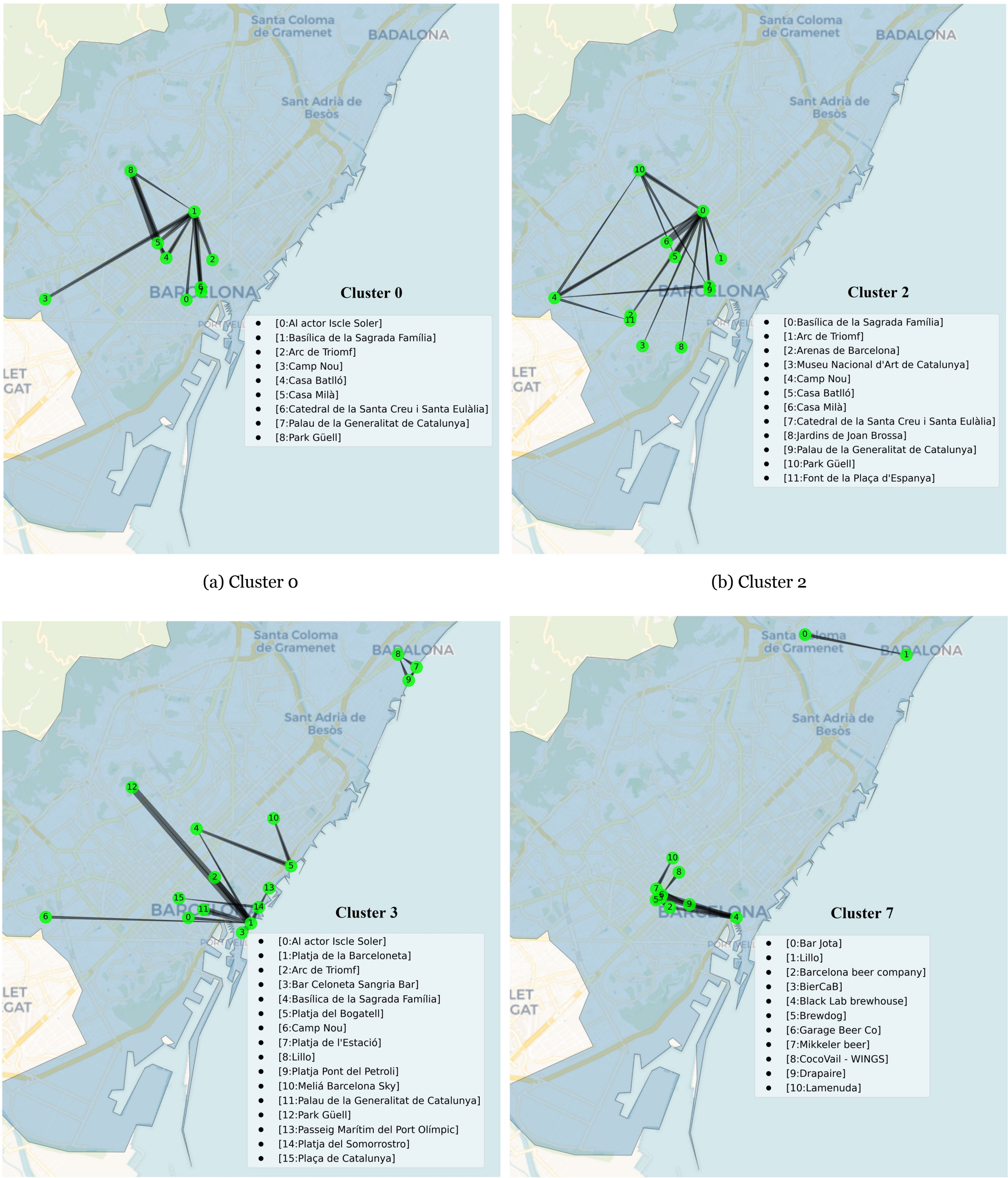
(b) Cluster 2

(c) Cluster 3

(d) Cluster 16

(e) Cluster 7

Figure 4.4: Mobility heat-maps for clusters 0, 2, 3, 16, and 7 representing bigrams with movement from left to bottom.



(a) Cluster 0

(b) Cluster 2

(c) Cluster 3

(d) Cluster 7

Figure 4.5: Tourist mobility patterns between attractions in Barcelona for clusters 0, 2, 3, and 7.

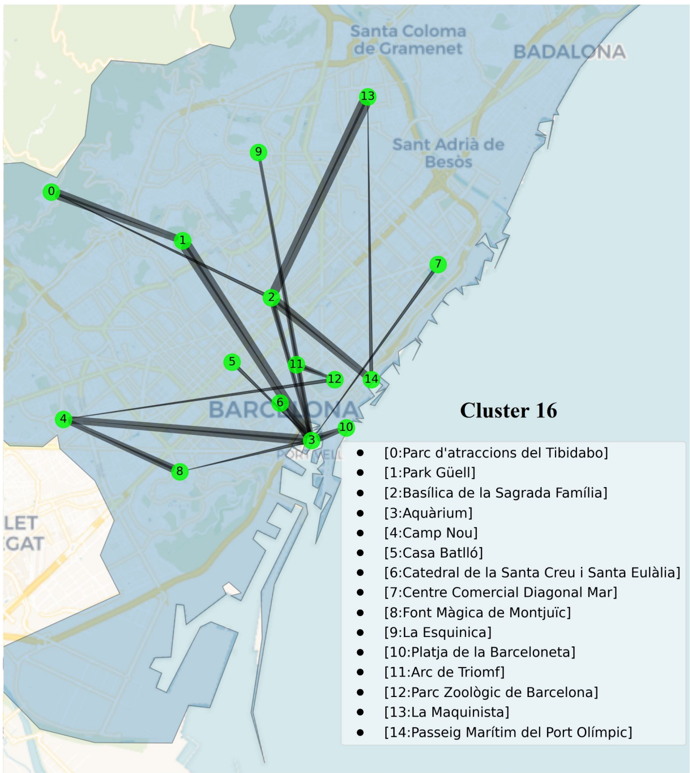

Santa Coloma
de Gramenet
BADALONA
Sant Adrià de
Besòs
BARCELONA
PORT VELL
LET
GAT
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Cluster 16
[0:Parc d'atraccions del Tibidabo]
[1:Park Güell]
[2:Basílica de la Sagrada Família]
[3:Aquàrium]
[4:Camp Nou]
[5:Casa Batlló]
[6:Catedral de la Santa Creu i Santa Eulàlia]
[7:Centre Comercial Diagonal Mar]
[8:Font Màgica de Montjuïc]
[9:La Esquinica]
[10:Platja de la Barceloneta]
[11:Arc de Triomf]
[12:Parc Zoològic de Barcelona]
[13:La Maquinista]
[14:Passeig Marítim del Port Olímpic]

Figure 4.6: Tourist mobility pattern between attractions in Barcelona for cluster 16.

In summary, the analyzed clusters showed inter-connected POIs which are of interest to certain groups of tourists. In some cases, they focus on the popular attractions, but in others they also visit places off the beaten track. The graphic representation in the map of the mobility of different clusters of tourists in a destination provides new knowledge to DMOs, who could take them into account to define new tourist routes, to create targeted marketing campaigns or to optimize transport routes between heavily connected POIs for different types of tourists.

4.5. Conclusions and Implications

The main contribution of the study in this chapter is the introduction of a method for analyzing social media data that creates visitor profiles according to their travel preferences and mobility. To this effect, we generated user profiles through a clustering process identical to that in Chapter 3, and then characterised the clusters by their member's data to understand their visit preferences and habits. We further analysed the mobilities of the members of the clusters showing the connection with their characterisation.

Similarly to Chapter 3, this study corroborates that social media are very useful platforms as sources for big data research in the field of Tourism (Xiang and Fesenmaier, 2017; Mariani et al., 2018), and that they can also be used to study tourist mobility (Chua et al., 2016; Salas-Olmedo et al., 2018; Jin et al., 2018; Önder, 2017). It also improves the knowledge about the different mobility patterns of tourists depending on who they are and what their preferences are (Hawelka et al., 2014). Therefore, the study has shown that clustering visitors by their travel mobility permits uncovering much more information about visitors and their preferences than previous studies. It can also provide complementary information to DMOs, attraction operators, and developers of contextual and next-POI recommender systems (Massimo and Ricci, 2019).

Additionally, the study also reveals the most popular or the most visited points of interest at destinations. Previous studies had also found out the most popular tourist spots or the most visited routes by analyzing geo-tagged social media data (Zhou et al., 2015; Chen et al., 2011), but their analysis did not allow for knowing which were the most visited POIs by the different groups of tourists. Thus, interestingly, the applied method allows for knowing the percentages of tourists in each cluster who visit each attraction the most. This is crucial for the managers of the different tourist attractions that want to know who their majority visitors are as well as their interests; in that way, they will be able to adapt their service and information in an almost personalized way.

Another contribution of the study is to show the mobility of each cluster of tourists between POIs and to see graphically the movement that they make in the map of the destination. Previous studies have shown mobility with place maps and most visited points (Chua et al., 2016). Many studies on GSMD only focused on tourists' mobility patterns (Li et al., 2018b; Orsi and Geneletti, 2013; Gabrielli et al., 2015). However, the difference in mobility patterns between sub-groups or clusters of tourists has not been fully researched (Liu et al., 2018). Therefore, GSMD analysis allows for knowing the dispersion and tourists' movements, the routes they follow, the activities they carry out in a territory (Li et al., 2016; Önder et al., 2016; Orsi and Geneletti, 2013; Vu et al., 2018; Wood et al., 2013; Zhou et al., 2015), and their density of movements (García-Palomares et al., 2015) and flows (Cheng and Edwards, 2015; Miah et al., 2017, 2019). Hence, the resulting information is particularly valuable for city management, since it provides a better knowledge of the connections between points of interest related to different clusters of tourists according to their preferences and behaviors. This information can help DMOs to define new tourist routes, to create targeted marketing campaigns, to optimize transport routes between heavily connected POIs for different types of tourists, or to improve the management of congestion or overcrowding situations.

To sum up, through this application of Artificial Intelligence techniques to social media data creating clusters of tourists, it has been possible to know how to segment them according to their visitor behavior and visit preferences. This information is key to the DMOs and the different service providers of the destinations. The interest of DMOs in analyzing big data and knowing the maximum information about their tourists is based on being able to anticipate their interests and preferences (Miah et al., 2019). This is precisely the information that this study provides. Therefore, the study can have a major impact on the marketing and flows management of tourist destinations. The exploration of the relationship between tourists' profiles, points of interest, and tourist mobility allows for gaining further insight into really concerning debates on tourism pressure in specific locations, and destination carrying capacity. Accordingly, it can be used by local and regional authorities, as well as by planners and urban designers, to deal with urban complexity, especially in successful tourist cities with contradictions and conflicts generated by overtourism (Lew and McKercher, 2006).

From this perspective, the managerial implications of the study are diverse. Using these kinds of analytical tools, DMOs and tourism service providers could be able to offer the most personalized services and information, to attract specific types of tourists to certain points of interest (Liu et al., 2018), to propose the visit of new under-visited attractions to certain market segments, to create new routes or to optimize the existing ones, to enhance public transport services, to develop new POIs or tourist services for the busiest routes (Charles Chancellor, 2012) or to create tourism development plans for the least visited areas (Vu et al., 2018).

In addition, it will allow destinations to encourage smart development, overcoming some of the existing gaps in the level of achievement of their objectives (Femenia-Serra and Ivars-Baidal, 2021). This includes the improvement of smart tourism developments such as the creation of differentiated attractive travel packages (Vu et al., 2020), the adaptation of the marketing and communication tactics to the preferences of visitors (Kotoua and Ilkan, 2017; Lamsfus et al., 2015), the improvement of the satisfaction of tourists (Buonincontri and Micera, 2016; Boes et al., 2015; Molinillo et al., 2019), and the co-creation of a more positive tourist destination image (Jabreel et al., 2018). Finally, from the place management perspective, the detection of visitor flow patterns would help to regulate the carrying capacity of the visitors' points of interest avoiding overcrowding, improving allocation of visitor services and reducing tensions produced by the different tourist and residential uses of the city areas, infrastructures, and services around the points of interest.

Chapter 5

Conclusions and Future Work

The pay-off of becoming a foremost tourism invested country is huge. Trillions of U.S. dollars are spent every year worldwide by tourists visiting new places for entertainment and leisure. This is money to be had by the leading tourism destinations. Approximately 73% of the total travel and tourism GDP was accumulated by the top 20 global tourism leading countries, which amounted to 6.7 trillion U.S. dollars in 2019 (WTTC, 2021a). To remain relevant or become one of the leading tourism destinations, tourism stakeholders create new attractions and experiences to keep tourists interested. This fact contributes to the information overload experienced by tourists visiting a new destination. Tourists are overwhelmed with decisions on where to go, what to do, what is popular, etc. *Recommender systems* are a solution to this problem. As a part of the field of information retrieval, recommender systems have become an active research area in the tourism industry, functioning as intelligent tools designed to aid tourists in their decisions, thereby improving their experience while increasing their consumption of touristic activities. Hence, the goal of this thesis is to develop methods for extracting tourists' preferences and contexts from their online activity, and utilize them to make recommendations of touristic activities. For thus purpose, we implemented an ensemble of machine learning techniques and hand-crafted algorithms to tackle different aspects of the task. The next section summarises the main contributions of this thesis.

5.1. Summary of Contributions

Our first contribution has been an in-depth review of the state of the art of social media analysis in tourism, contextual recommendations in tourism, and flow analysis in tourism (detailed in chapter 2). We reviewed the typical data sources and data retrieval methods used in social media tourism analysis, highlighting the most popular sources, the type of data retrieved, and the parameters used by relevant works for data retrieval. We also reviewed the common analytical methods in social media tourism analysis, detailing the methods used for tourist identification and profiling. We then reviewed context aware recommender systems contrasting them with traditional recommender systems. Furthermore, the contexts and Artificial Intelligence techniques used in tourism recommender systems were reported. The state of art analysis was concluded with a review of tourist flow analysis.

Chapter 3 followed the state of the art analysis with the detailed explanation of our proposed recommender system. The main contribution in this chapter was a tourism recommender system (ReCLARM) that tackles the problem of data sparsity in social media data by combining in a unique way automatic clustering and association rule mining techniques. At first, tourists were identified from social media data using an algorithm based on length of stay and post frequency, and then their experienced activities were identified by considering their location. The identified activities were a main part of the clustering process. A clustering model was then built using extracted features that represented tourists' interests in (popular and unpopular) activities , travel habits, and active periods. The clustering model served to profile tourists and incorporate their contextual features in ReClARM. ReCLARM also contained an association rule mining process to extract frequent POI-sets that modelled the implicit relationship of POIs in the clusters of similar tourists. Finally, recommended POIs were selected with a rule ranking process. ReCLARM was evaluated on tweets downloaded in 2019. The results have shown ReClARM to perform well in several recommender system metrics. The results have also shown the effectiveness of the clustering (profiling) phase by comparing ReCLARM with and without clustering. ReCLARM was further extended to recommend ordered POIs enhancing the experience of tourists. ReCLARM-GA (ReCLARM's extension) used a multi-objective genetic algorithm to optimize the order in which POIs are experienced by a tourist considering his/her preferences and interests in popular or unpopular POIs, and also POI proximity and diversity. ReCLARM-GA was evaluated against some baseline algorithms and shown to perform significantly better.

Chapter 4 focused on developing methods for identifying and understanding the mobility of clusters of tourists. Specifically, methods to be used by destination management organisations for studying tourists flows and city planning. First, tourists were clustered as in chapter 3 and then characterised by their country of origin, degree of interests in activities, interests in popular on unpopular activities, and their active periods during the day. The characterisation of cluster members provided an insight into the different types of tourists and the activities they visit. We have drawn connections between specific countries and the activities they prefer, and also international tourists and locals and the activities they prefer. We have further shown each cluster's degree of interest in the most popular POIs, which allows DMOs to know the types of tourists that flood the popular places. Additionally, tourist mobility within each cluster was extracted and graphed. This showed the most connected POIs in the city and their specific routing needs. An important point drawn from this study is that clustering tourists allows uncovering new information on their preferences and mobilty.

5.2. Future Work

Although the systems proposed in this thesis performs reasonably well in making suitable recommendations and meet the objectives for this dissertation, there is still room for improvement in the algorithms and methods used. For instance, the k-means clustering algorithm is not perfectly suited for the user profiling phase because it cannot handle mixed data variables. It hindered our ability to add more contextual and demographic information in the clustering process. This can be handled using the k-medoids clustering algorithm together with a distance metric that can handle mixed data like the Gower distance, or other suitable techniques (van de Velden et al., 2018). Another point for the integration of contextual information, is that the user profiling only allows for integrating user-focused contexts which leaves out some other important contexts like weather, POI availability, etc. In order to incorporate contexts like these, they will have to be added after profiling.

Another area of our proposed systems that could be improved is the way POI ordering was formulated in section 3.5. The problem was formulated as a combinatorial optimization problem which is suited for the multi-objective genetic algorithm we used. However, this method does not allow for consideration of the operational hours of POIs and the best route with different modes of transport, which may cause inefficiencies in the actual visitation of the POIs. It is possible to formulate the problem as a multi-objective orienteering problem with time windows which allows solving the ordering problem while considering the route and operating times of POIs (Vansteenwegen et al., 2011).

In addition, there are some emerging areas in recommender systems which could be great additions to the research in this thesis work. These are relatively new areas that are currently being research and showing fruitful results.

Fairness in Recommender Systems

Fairness is a responsibility concern in computing systems, among others including accountability, transparency, safety, privacy, and ethics (Ekstrand et al., 2022). It is centered around reducing or eliminating bias in recommender systems. Specifically, fairness asks the question: *Are all users treated indifferently and without bias?* Recommender systems bias may be observed in the gender, age, nationality, etc., of their users. For instance, if a movie streaming app like Netflix is mostly used by people between 18-35, then a movie recommender system built on data from those users might be biased to that demographic and perform poorly in recommendations made to users outside that demographic. It is also possible to experience unfairness in items, that is some items are not given adequate exposure in recommendations (Wang et al., 2022).

Tourism recommender systems are also not immune to unfairness. They could be affected by gender and nationality bias based on the dataset used to train the system. It could also be argued that eliminating bias in tourism recommender systems is more important than in other cases because it could misrepresent a tourist destination (this could be a whole city or country) to a certain category of tourist which is highly detrimental. Also, managers of certain tourist attractions could lose revenue from a lack of exposure as a result of unfairness. Our proposed systems could be extended to properly evaluate for fairness with specialized metrics for this purpose, and facilitate proper representation of POIs and tourists in the dataset used for model construction.

Explainability in Recommender Systems

Explainability is a hot topic in Artificial Intelligence. The current popularity of Artificial Intelligence techniques has seen them applied in numerous fields. This has introduced many AI based systems that are black boxes where the users and sometimes developers do not understand how the outputs are generated. The results from these black box systems may be favourable, but without an understanding of why they are favourable it is difficult to fine tune the systems. As such, research in explainability has ramped up in the deep learning and machine learning fields. Explainable systems allow their results to be scrutinized by users, developers, stakeholders, etc., which leads to more accurate and responsible results.

In explainable recommender systems, recommendations can be explained by user similarity (e.g. *you were recommended this movie because a similar user likes it*), item similarity (e.g. *you were recommended this movie because you watched that movie*), user-item similarity (e.g. *you were recommended this movie because you said you like action movies*), and user opinions (e.g. *you were recommended this movie because you positively reviewed that movie*). Also, item recommendations could be explained by telling why a user is similar to another user (like friendship, consume similar things, etc.), or by presenting a list of the item's features, or by showing product images with highlighted regions-of-interest. It is also possible to create explainable recommender systems by making interpretable algorithms and methods. To make the systems proposed in this thesis explainable, the clustering features in section 3.3.2 could be used to explain user similarity, also the methods for identifying users as tourists could also help a user understand how the models were trained.

To summarise, these research areas are only a few of the concerns in computing systems as mentioned in Ekstrand et al. (2022), that move recommender system design in the direction of responsibility and accountability.

Bibliography

Abbasi-Moud, Z., Hosseinabadi, S., Kelarestaghi, M., and Eshghi, F. (2022). CAFOB: context-aware fuzzy-ontology-based tourism recommendation system. *Expert Systems with Applications*, 199:116877.

Abbasi-Moud, Z., Vahdat-Nejad, H., and Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Systems with Applications*, 167(114324):114324.

Adomavicius, G. and Tuzhilin, A. (2001). Extending recommender systems: A multidimensional approach. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-01)*, pages 4–6, Seattle, WA, USA.

Afsahhosseini, F. and Al-Mulla, Y. (2021). Smart, hybrid and context-aware POI mobile recommender system in tourism in oman. *Journal of Cultural Heritage Management and Sustainable Development*.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487––499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ahas, R., Aasa, A., Mark, Ü., Pae, T., and Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management*, 28(3):898–910.

Ahn, M. J. and McKercher, B. (2015). The effect of cultural distance on tourism: A study of international visitors to Hong Kong. *Asia Pacific Journal of Tourism Research*, 20(1):94–113.

Al-Shamri, M. Y. H. (2016). User profiling approaches for demographic recommender systems. *Knowledge-Based Systems*, 100:175–187.

Almeida, A., Machado, L. P., and Xu, C. (2021). Factors explaining length of stay: Lessons to be learnt from Madeira Island. *Annals of Tourism Research Empirical Insights*, 2(1):100014.

Anton Clavé, S. (2019). Urban tourism and walkability. In *The Future of Tourism*, pages 195–211. Springer International Publishing, Cham.

Arefieva, V., Egger, R., and Yu, J. (2021). A machine learning approach to cluster destination image on Instagram. *Tourism Management*, 85(104318):104318.

Arvianti, Q. R., Baizal, Z. K. A., and Tarwidi, D. (2019). Tourism recommender system using item-based hybrid clustering method (case study: Bandung Raya region). *Journal of Data Science and Its Applications*, pages 95–101.

Bagci, H. and Karagoz, P. (2015). Context-aware location recommendation by using a random walk-based approach. *Knowledge and Information Systems*, 47(2):241–260.

Bahramian, Z., Abbaspour, R. A., and Claramunt, C. (2017). A cold start context-aware recommender system for tour planning using artificial neural network and case based reasoning. *Mobile Information Systems*, 2017:1–18.

Baral, R., Iyengar, S. S., Li, T., and Balakrishnan, N. (2018). CLoSe: Contextualized Location Sequence Recommender. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 470––474, New York, NY, USA. Association for Computing Machinery.

Barchiesi, D., Moat, H. S., Alis, C., Bishop, S., and Preis, T. (2015). Quantifying international travel flows using Flickr. *PLoS One*, 10(7):e0128470.

Batet, M., Moreno, A., Sánchez, D., Isern, D., and Valls, A. (2012). Turist@: Agent-based personalised recommendation of tourist activities. *Expert Systems with Applications*, 39(8):7319–7329.

Batra, A. (2009). Senior pleasure tourists: Examination of their demography, travel experience, and travel behavior upon visiting the Bangkok metropolis. *International Journal of Hospitality and Tourism Administration*, 10(3):197–212.

Bedi, P., Agarwal, S. K., Jindal, V., and Richa (2014). MARST: Multi-agent recommender system for e-tourism using reputation based collaborative filtering. In Madaan, A., Kikuchi, S., and Bhalla, S., editors, *Databases in Networked Information Systems*, pages 189–201. Springer International Publishing.

Beecham, R., Wood, J., and Bowerman, A. (2014). Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 47:5–15.

Benouaret, I. and Lenne, D. (2015). Personalizing the museum experience through context-aware recommendations. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 743–748. IEEE.

Biancalana, C., Gasparetti, F., Micarelli, A., and Sansonetti, G. (2013). An approach to social recommendation for context-aware mobile services. *ACM Transactions on Intelligent Systems and Technology*, 4(1):1–31.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., Beijing, 1 edition.

Boes, K., Buhalis, D., and Inversini, A. (2015). Conceptualising smart tourism destination dimensions. In *Information and Communication Technologies in Tourism 2015*, pages 391–403. Springer International Publishing, Cham.

Borràs, J., Moreno, A., and Valls, A. (2014). Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, 41(16):7370–7389.

Borràs, J., Moreno, A., and Valls, A. (2017). Diversification of recommendations through semantic clustering. *Multimedia Tools and Applications*, 76(22):24165–24201.

Buonincontri, P. and Micera, R. (2016). The experience co-creation in smart tourism destinations: a multiple case analysis of European destinations. *Information Technology and Tourism*, 16(3):285–315.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Bustamante, A., Sebastia, L., and Onaindia, E. (2019). Can tourist attractions boost other activities around? a data analysis through social networks. *Sensors*, 19(11):2612.

Campos, P. G., Díez, F., and Cantador, I. (2014). Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction*, 24(1-2):67–119.

Cao, M.-Q., Liang, J., Li, M.-Z., Zhou, Z.-H., and Zhu, M. (2020). TDIVis: visual analysis of tourism destination images. *Frontiers of Information Technology and Electronic Engineering*, 21(4):536–557.

Cassar, M. L., Caruana, A., and Konietzny, J. (2020). Wine and satisfaction with fine dining restaurants: an analysis of tourist experiences from user generated content on TripAdvisor. *Journal of Wine Research*, 31(2):85–100.

Çetinkaya, M. Y. and Öter, Z. (2016). Role of tour guides on tourist satisfaction level in guided tours and impact on re-visiting intention: a research in Istanbul. *European Journal of Tourism, Hospitality and Recreation*, 7(1):40–54.

Charles Chancellor, H. (2012). Applying travel pattern data to destination development and marketing decisions. *Tourism Planning and Development*, 9(3):321–332.

Chen, J., Becken, S., and Stantic, B. (2022). Assessing destination satisfaction by social media: An innovative approach using Importance-Performance analysis. *Annals of Tourism Research*, 93(103371):103371.

Chen, Y., Sherren, K., Smit, M., and Lee, K. Y. (2021). Using social media images as data in social science research. *New Media and Society*, page 146144482110387.

Chen, Z., Shen, H. T., and Zhou, X. (2011). Discovering popular routes from trajectories. In *2011 IEEE 27th International Conference on Data Engineering*. IEEE.

Cheng, M. and Edwards, D. (2015). Social media in tourism: a visual analytic approach. *Current Issues in Tourism*, 18(11):1080–1087.

Chua, A., Servillo, L., Marcheggiani, E., and Moere, A. V. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management*, 57:295–310.

Ciesielski, M. and Stereńczak, K. (2021). Using Flickr data and selected environmental characteristics to analyse the temporal and spatial distribution of activities in forest areas. *Forest Policy and Economics*, 129(102509):102509.

De Pessemier, T., Dooms, S., and Martens, L. (2013). Context-aware recommendations through context and activity recognition in a mobile environment. *Multimedia Tools and Applications*, 72(3):2925–2948.

Dietz, L. W., Sen, A., Roy, R., and Wörndl, W. (2020). Mining trips from location-based social networks for clustering travelers and destinations. *Information Technology and Tourism*, 22(1):131–166.

Dodds, R. and Butler, R. (2019). The phenomena of overtourism: a review. *International Journal of Tourism Cities*, 5(4):519–528.

Domènech, A., Gutiérrez, A., and Anton Clavé, S. (2020a). Cruise passengers' spatial behaviour and expenditure levels at destination. *Tourism Planning and Development*, 17(1):17–36.

Domènech, A., Mohino, I., and Moya-Gómez, B. (2020b). Using Flickr geotagged photos to estimate visitor trajectories in world heritage cities. *ISPRS International Journal of Geo-Information*, 9(11):646.

Ebrahimpour, Z., Wan, W., Velázquez García, J. L., Cervantes, O., and Hou, L. (2020). Analyzing social-geographic human mobility patterns using large-scale social media data. *ISPRS International Journal of Geo-Information*, 9(2):125.

Edwards, D. and Griffin, T. (2013). Understanding tourists' spatial behaviour: GPS tracking as an aid to sustainable destination management. *Journal of Sustainable Tourism*, 21(4):580–595.

Ekstrand, M. D., Das, A., Burke, R., and Diaz, F. (2022). Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16(1-2):1–177.

Enzensberger, H. M. (1996). A theory of tourism. *New German Critique*, (68):117–135.

Esmaeili, L., Mardani, S., Golpayegani, S. A. H., and Madar, Z. Z. (2020). A novel tourism recommender system in the context of social commerce. *Expert Systems with Applications*, 149(113301):113301.

Fararni, K. A., Nafis, F., Aghoutane, B., Yahyaouy, A., Riffi, J., and Sabri, A. (2021). Hybrid recommender system for tourism based on big data and AI: A conceptual framework. *Big Data Mining and Analytics*, 4(1):47–55.

Farnadi, G., Tang, J., De Cock, M., and Moens, M.-F. (2018). User profiling through deep multimodal fusion. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*, page 171–179, New York, NY, USA. Association for Computing Machinery.

Felfernig, A., Boratto, L., Stettinger, M., and Tkalcic, M. (2018). *Group Recommender Systems*. SpringerBriefs in Electrical and Computer Engineering. Springer International Publishing, Cham, Switzerland, 1 edition.

Femenia-Serra, F. and Ivars-Baidal, J. A. (2021). Do smart tourism destinations really work? the case of Benidorm. *Asia Pacific Journal of Tourism Research*, 26(4):365–384.

Fenza, G., Fischetti, E., Furno, D., and Loia, V. (2011). A hybrid context aware system for tourist guidance based on collaborative filtering. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*. IEEE.

Fogli, A. and Sansonetti, G. (2019). Exploiting semantics for context-aware itinerary recommendation. *Personal and Ubiquitous Computing*, 23(2):215–231.

Forouzandeh, S., Rostami, M., and Berahmand, K. (2022). A hybrid method for recommendation systems based on tourism with an evolutionary algorithm and topsis model. *Fuzzy Information and Engineering*, 14(1):26–50.

Fränti, P., Mariescu-Istodor, R., and Waga, K. (2018). Similarity of mobile users based on sparse location history. In *Artificial Intelligence and Soft Computing*, Lecture notes in Computer Science, pages 593–603. Springer International Publishing, Cham.

Fränti, P., Waga, K., and Khurana, C. (2015). Can social network be used for location-aware recommendation? In *Proceedings of the 11th International Conference on Web Information Systems and Technologies - WEBIST*, pages 558–565. INSTICC, SciTePress.

Fuchs, M., Höpken, W., and Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations – a case from sweden. *Journal of Destination Marketing and Management*, 3(4):198–209.

Gabrielli, L., Furletti, B., Trasarti, R., Giannotti, F., and Pedreschi, D. (2015). City users' classification with mobile phone data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 1007–1012. IEEE.

Garcia, I., Sebastia, L., and Onaindia, E. (2011). On the design of individual and group recommender systems for tourism. *Expert Systems with Applications*, 38(6):7683–7692.

Garcia, I., Sebastia, L., Onaindia, E., and Guzman, C. (2009). A group recommender system for tourist activities. In *E-Commerce and Web Technologies*, pages 26–37. Springer Berlin Heidelberg, Berlin, Heidelberg.

García-Palomares, J. C., Gutiérrez, J., and Mínguez, C. (2015). Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography*, 63:408–417.

Ghermandi, A., Camacho-Valdez, V., and Trejo-Espinosa, H. (2020). Social media-based analysis of cultural ecosystem services and heritage tourism in a coastal region of Mexico. *Tourism Management*, 77(104002):104002.

Giglio, S., Bertacchini, F., Bilotta, E., and Pantano, P. (2019). Using social media to identify tourism attractiveness in six Italian cities. *Tourism Management*, 72:306–312.

Giglio, S., Bertacchini, F., Bilotta, E., and Pantano, P. (2020). Machine learning and points of interest: typical tourist Italian cities. *Current Issues in Tourism*, 23(13):1646–1658.

Gon, M. (2021). Local experiences on Instagram: social media data as source of evidence for experience design. *Journal of Destination Marketing and Management*, 19(100435):100435.

Han, S., Liu, C., Chen, K., Gui, D., and Du, Q. (2021). A tourist attraction recommendation model fusing spatial, temporal, and visual embeddings for Flickr-geotagged photos. *ISPRS International Journal of Geo-Information*, 10(1):20.

Han, S., Ren, F., Du, Q., and Gui, D. (2020). Extracting representative images of tourist attractions from Flickr by combining an improved cluster method and multiple deep learning models. *ISPRS International Journal of Geo-Information*, 9(2):81.

Han, S., Ren, F., Wu, C., Chen, Y., Du, Q., and Ye, X. (2018). Using the TensorFlow deep neural network to classify mainland China visitor behaviours in Hong Kong from check-in data. *ISPRS International Journal of Geo-Information*, 7(4):158.

Hang, L., Kang, S.-H., Jin, W., and Kim, D.-H. (2018). Design and implementation of an optimal travel route recommender system on big data for tourists in Jeju. *Processes*, 6(8):133.

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, 38(1):10–24.

Haruna, K., Akmar Ismail, M., Suhendroyono, S., Damiasih, D., Pierewan, A. C., Chiroma, H., and Herawan, T. (2017). Context-aware recommender system: a review of recent developmental process and future research direction. *Applied Sciences*, 7(12):1211.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

He, R., Kang, W.-C., and McAuley, J. (2017). Translation-based recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*, pages 161––169, New York, NY, USA. Association for Computing Machinery.

Hu, F., Li, Z., Yang, C., and Jiang, Y. (2019). A graph-based approach to detecting tourist movement patterns using social media data. *Cartography and Geographic Information Science*, 46(4):368–382.

Huang, A., Gallegos, L., and Lerman, K. (2017). Travel analytics: Understanding how destination choice and business clusters are connected based on social media data. *Transportation Research Part C: Emerging Technologies*, 77:245–256.

Huang, F., Qiao, S., Peng, J., Guo, B., and Han, N. (2021). STPR: A personalized next point-of-interest recommendation model with spatio-temporal effects based on purpose ranking. *IEEE Transactions on Emerging Topics in Computing*, 9(2):994–1005.

Huang, Q. and Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, 30(9):1873–1898.

Ishanka, U. A. P. and Yukawa, T. (2018). User emotion and personality in context-aware travel destination recommendation. In *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 13–18. IEEE.

Jabreel, M., Huertas, A., and Moreno, A. (2018). Semantic analysis and the evolution towards participative branding: Do locals communicate the same destination brand values as DMOs? *PLoS One*, 13(11):e0206572.

Jabreel, M., Moreno, A., and Huertas, A. (2017). Do local residents and visitors express the same sentiments on destinations through social media? In *Information and Communication Technologies in Tourism 2017*, pages 655–668. Springer International Publishing, Cham.

Jalalimanesh, A., Mansoury, M., and Gandomi, H. (2012). Recommender system based on data mining: Interlibrary case study. In *20th Iranian Conference on Electrical Engineering (ICEE2012)*, pages 806–809. IEEE.

Jiang, W., Xiong, Z., Su, Q., Long, Y., Song, X., and Sun, P. (2021). Using geotagged social media data to explore sentiment changes in tourist flow: A spatiotemporal analytical framework. *ISPRS International Journal of Geo-Information*, 10(3):135.

Jin, C., Cheng, J., and Xu, J. (2018). Using user-generated content to explore the temporal heterogeneity in tourist mobility. *Journal of Travel Research*, 57(6):779–791.

Johnson, I., McMahon, C., Schöning, J., and Hecht, B. (2017). The effect of population and "structural" biases on social media-based algorithms. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, pages 1167––1178, New York, NY, USA. Association for Computing Machinery.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2015). Understanding human mobility from twitter. *PLoS One*, 10(7):e0131469.

Kashevnik, A. M., Ponomarev, A. V., and Smirnov, A. V. (2017). A multimodel context-aware tourism recommendation service: approach and architecture. *Journal of Computer and Systems Sciences International*, 56(2):245–258.

Kolahkaj, M., Harounabadi, A., Nikravanshalmani, A., and Chinipardaz, R. (2020). A hybrid context-aware approach for e-tourism package recommendation based on asymmetric similarity measurement and sequential pattern mining. *Electronic Commerce Research and Applications*, 42:100978.

Korakakis, M., Spyrou, E., Mylonas, P., and Perantonis, S. J. (2017). Exploiting social media information toward a context-aware recommendation system. *Social Network Analysis and Mining*, 7(1):42.

Kotoua, S. and Ilkan, M. (2017). Tourism destination marketing and information technology in Ghana. *Journal of Destination Marketing and Management*, 6(2):127–135.

Kuusik, A., Tiru, M., Ahas, R., and Varblane, U. (2011). Innovation in destination marketing: the use of passive mobile positioning for the segmentation of repeat visitors in Estonia. *Baltic Journal of Management*, 6(3):378–399.

Lalicic, L., Huertas, A., Moreno, A., and Jabreel, M. (2019). Which emotional brand values do my followers want to hear about? an investigation of popular European tourist destinations. *Information Technology and Tourism*, 21(1):63–81.

Lalicic, L., Huertas, A., Moreno, A., and Jabreel, M. (2020). Emotional brand communication on Facebook and Twitter: Are DMOs successful? *Journal of Destination Marketing and Management*, 16(100350):100350.

Lamsfus, C., Martín, D., Alzua-Sorzabal, A., and Torres-Manzanera, E. (2015). Smart tourism destinations: An extended conception of smart cities focusing on human mobility. In *Information and Communication Technologies in Tourism 2015*, pages 363–375. Springer International Publishing, Cham.

Lee, M., Hong, J. H., Chung, S., and Back, K.-J. (2021). Exploring the roles of DMO's social media efforts and information richness on customer engagement: Empirical analysis on Facebook event pages. *Journal of Travel Research*, 60(3):670–686.

Levi, A., Mokryn, O., Diot, C., and Taft, N. (2012). Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*, pages 115––122, New York, NY, USA. Association for Computing Machinery.

Lew, A. and McKercher, B. (2006). Modeling tourist movements. *Annals of Tourism Research*, 33(2):403–423.

Li, D., Zhou, X., and Wang, M. (2018a). Analyzing and visualizing the spatial interactions between tourists and locals: A Flickr study in ten US cities. *Cities*, 74:249–258.

Li, J., Xu, L., Tang, L., Wang, S., and Li, L. (2018b). Big data in tourism research: A literature review. *Tourism Management*, 68:301–323.

Li, X. and Law, R. (2020). Network analysis of big data research in Tourism. *Tourism Management Perspectives*, 33(100608):100608.

Li, Y., Xiao, L., Ye, Y., Xu, W., and Law, A. (2016). Understanding tourist space at a historic site through space syntax analysis: The case of Gulangyu, China. *Tourism Management*, 52:30–43.

Liao, Y. (2020). Hot spot analysis of tourist attractions based on stay point spatial clustering. *Journal of Information Processing Systems*, 16(4):750–759.

Liao, Y., Yeh, S., and Gil, J. (2022). Feasibility of estimating travel demand using geolocations of social media data. *Transportation*, 49(1):137–161.

Liji, U., Chai, Y., and Chen, J. (2018). Improved personalized recommendation based on user attributes clustering and score matrix filling. *Computer Standards and Interfaces*, 57:59–67.

Lim, K. H., Chan, J., Karunasekera, S., and Leckie, C. (2019). Tour recommendation and trip planning using location-based social media: a survey. *Knowledge and Information Systems*, 60(3):1247–1275.

Liu, Q., Wang, Z., and Ye, X. (2018). Comparing mobility patterns between residents and visitors using geo-tagged social media data. *Transactions in GIS*, 22(6):1372–1389.

Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., and Li, X. (2019). Efficient user profiling based intelligent travel recommender system for individual and group of users. *Mobile Networks and Applications*, 24(3):1018–1033.

Lou, N. (2022). Tourism destination recommendation based on association rule algorithm. *Mobile Information Systems*, 2022:1–13.

Lu, W. and Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. *Journal of Hospitality Marketing and Management*, 24(2):119–154.

Lucas, J. P., Luz, N., Moreno, M. N., Anacleto, R., Almeida Figueiredo, A., and Martins, C. (2013). A hybrid recommendation approach for a tourism system. *Expert Systems with Applications*, 40(9):3532–3550.

Luo, Y. and Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: a case study of Yelp. *Sustainability*, 11(19):5254.

Ma, A., Chow, A., Cheung, L., Lee, K., and Liu, S. (2018). Impacts of tourists' sociodemographic characteristics on the travel motivation and satisfaction: The case of protected areas in south china. *Sustainability*, 10(10):3388.

Ma, S. and Kirilenko, A. (2021). How reliable is social media data? Validation of TripAdvisor tourism visitations using independent data sources. In *Information and Communication Technologies in Tourism 2021*, pages 286–293. Springer International Publishing, Cham.

Ma, S. d., Kirilenko, A. P., and Stepchenkova, S. (2020). Special interest tourism is not so special after all: big data evidence from the 2017 great american solar eclipse. *Tourism Management*, 77(104021):104021.

Ma, X., Lu, H., Gan, Z., and Zhao, Q. (2016). An exploration of improving prediction accuracy by constructing a multi-type clustering based recommendation framework. *Neurocomputing*, 191:388–397.

Malik, M., Lamba, H., Nakos, C., and Pfeffer, J. (2021). Population bias in geotagged tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(4):18–27.

Manca, M., Boratto, L., Morell Roman, V., Martori i Gallissà, O., and Kaltenbrunner, A. (2017). Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study. *Online Social Networks and Media*, 1:56–69.

Mariani, M., Baggio, R., Fuchs, M., and Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, 30(12):3514–3554.

Mariescu-Istodor, R., Ungureanu, R., and Fränti, P. (2019). Real-time destination prediction for mobile users. *Advances in Cartography and GIScience of the International Cartographic Association*, 2:1–7.

Massa, P. and Avesani, P. (2007). Trust-aware recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems - RecSys '07*, pages 17––24, New York, New York, USA. Association for Computing Machinery.

Massimo, D. and Ricci, F. (2018). Harnessing a generalised user behaviour model for next-POI recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*, pages 402––406, New York, NY, USA. Association for Computing Machinery.

Massimo, D. and Ricci, F. (2019). Clustering users' POIs visit trajectories for next-POI recommendation. In *Information and Communication Technologies in Tourism 2019*, pages 3–14. Springer International Publishing, Cham.

Massimo, D. and Ricci, F. (2021). Next-POI recommendations matching user's visit behaviour. In *Information and Communication Technologies in Tourism 2021*, pages 45–57. Springer International Publishing, Cham.

Miah, S. J., Vu, H., and Gammack, J. (2019). A big-data analytics method for capturing visitor activities and flows: the case of an island country. *Information Technology and Management*, 20(4):203–221.

Miah, S. J., Vu, H. Q., Gammack, J., and McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information and Management*, 54(6):771–785.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781. Accessed: 2022-08-12.

Mirzaalian, F. and Halpenny, E. (2019). Social media analytics in hospitality and tourism. *Journal of Hospitality and Tourism Technology*, 10(4):764–790.

Mirzaalian, F. and Halpenny, E. (2021). Exploring destination loyalty: Application of social media analytics in a nature-based tourism setting. *Journal of Destination Marketing and Management*, 20(100598):100598.

Missaoui, S., Kassem, F., Viviani, M., Agostini, A., Faiz, R., and Pasi, G. (2019). LOOKER: a mobile, personalized recommender system in the Tourism domain based on social media user-generated content. *Personal and Ubiquitous Computing*, 23(2):181–197.

Molinillo, S., Anaya-Sánchez, R., Morrison, A. M., and Coca-Stefaniak, J. A. (2019). Smart city communication via social media: analysing residents' and visitors' engagement. *Cities*, 94:247–255.

Moreno, A., Valls, A., Isern, D., Marin, L., and Borràs, J. (2013). SigTur/E-Destination: Ontology-based personalized recommendation of tourism and leisure activities. *Engineering Applications of Artificial Intelligence*, 26(1):633–651.

Morgan, A., Wilk, V., Sibson, R., and Willson, G. (2021). Sport event and destination co-branding: Analysis of social media sentiment in an international, professional sport event crisis. *Tourism Management Perspectives*, 39(100848):100848.

Najafabadi, M. K., Mahrin, M. N., Chuprat, S., and Sarkan, H. M. (2017). Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data. *Computers in Human Behavior*, 67:113–128.

Nakayama, M. and Wan, Y. (2019). The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews. *Information and Management*, 56(2):271–279.

Neff, J. C. (1938). Santa Fe and the tourist. *New Mexico Quarterly*, 8(2).

Nguyen, L. V., Hong, M.-S., Jung, J. J., and Sohn, B.-S. (2020a). Cognitive similarity-based collaborative filtering recommendation system. *Applied Sciences*, 10(12):4183.

Nguyen, L. V., Jung, J. J., and Hwang, M. (2020b). OurPlaces: Cross-cultural crowdsourcing platform for location recommendation services. *ISPRS International Journal of Geo-Information*, 9(12):711.

Nilashi, M., Bagherifard, K., Rahmani, M., and Rafe, V. (2017). A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. *Computers & Industrial Engineering*, 109:357–368.

Nilashi, M., Ibrahim, O., Yadegaridehkordi, E., Samad, S., Akbari, E., and Alizadeh, A. (2018). Travelers decision making using online review in social network sites: a case on TripAdvisor. *Journal of Computational Science*, 28:168–179.

Nilashi, M., Samad, S., Ahani, A., Ahmadi, H., Alsolami, E., Mahmoud, M., Majeed, H. D., and Abdulsalam Alarood, A. (2021). Travellers decision making through preferences learning: A case on Malaysian spa hotels in TripAdvisor. *Computers and Industrial Engineering*, 158(107348):107348.

Obembe, D., Kolade, O., Obembe, F., Owoseni, A., and Mafimisebi, O. (2021). Covid-19 and the tourism industry: an early stage sentiment analysis of the impact of social media and stakeholder communication. *International Journal of Information Management Data Insights*, 1(2):100040.

Önder, I. (2017). Classifying multi-destination trips in Austria with big data. *Tourism Management Perspectives*, 21:54–58.

Önder, I., Gunter, U., and Gindl, S. (2020). Utilizing Facebook statistics in Tourism demand modeling and destination marketing. *Journal of Travel Research*, 59(2):195–208.

Önder, I., Koerbitz, W., and Hubmann-Haidvogel, A. (2016). Tracing tourists by their digital footprints. *Journal of Travel Research*, 55(5):566–573.

Orellana, D., Bregt, A. K., Ligtenberg, A., and Wachowicz, M. (2012). Exploring visitor movement patterns in natural recreational areas. *Tourism Management*, 33(3):672–682.

Orlandi, F., Breslin, J., and Passant, A. (2012). Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th International Conference on Semantic Systems - I-SEMANTICS '12*, pages 41–48, New York, New York, USA. Association for Computing Machinery.

Orsi, F. and Geneletti, D. (2013). Using geotagged photographs and GIS analysis to estimate visitor flows in natural areas. *Journal for Nature Conservation*, 21(5):359–368.

Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C., and González, M. C. (2015). Urban magnetism through the lens of geo-tagged photography. *EPJ Data Science*, 4(1).

Pan, H. and Zhang, Z. (2019). Research on context-awareness mobile Tourism e-commerce personalized recommendation model. *Journal of Signal Processing Systems*, 93(2):147–154.

Pandya, S., Shah, J., Joshi, N., Ghayvat, H., Mukhopadhyay, S. C., and Yap, M. H. (2016). A novel hybrid based recommendation system based on clustering and association mining. In *2016 10th International Conference on Sensing Technology (ICST)*, pages 1–6. IEEE.

Pantano, E., Priporas, C.-V., and Stylos, N. (2017). 'you will like it!' Using open data to predict tourists' response to a tourist attraction. *Tourism Management*, 60:430–438.

Pantano, E., Priporas, C.-V., Stylos, N., and Dennis, C. (2019). Facilitating tourists' decision making through open data analyses: a novel recommender system. *Tourism Management Perspectives*, 31:323–331.

Park, S. B., Kim, J., Lee, Y. K., and Ok, C. M. (2020). Visualizing theme park visitors' emotions using social media analytics and geospatial analytics. *Tourism Management*, 80(104127):104127.

Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In *The Adaptive Web*, pages 325–341. Springer Berlin Heidelberg, Berlin, Heidelberg.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2013). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Petutschnig, A., Albrecht, J., Resch, B., Ramasubramanian, L., and Wright, A. (2021). Commuter mobility patterns in social media: Correlating Twitter and LODES data. *ISPRS International Journal of Geo-Information*, 11(1):15.

Phillips, W. J. and Jang, S. (2010). Destination image differences between visitors and non-visitors: a case of New York City. *International Journal of Tourism Research*, 12(5):642–645.

Provenzano, D., Hawelka, B., and Baggio, R. (2018). The mobility network of European tourists: a longitudinal study and a comparison with geo-located twitter data. *Tourism Review*, 73(1):28–43.

Quadrana, M., Cremonesi, P., and Jannach, D. (2019). Sequence-aware recommender systems. *ACM Computing Surveys*, 51(4):1–36.

Raschka, S. (2013). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, 3(24):638.

Renjith, S., Sreekumar, A., and Jathavedan, M. (2020). An extensive study on the evolution of context-aware personalized travel recommender systems. *Information Processing & Management*, 57(1):102078.

Riswanto, E., Robi'in, B., and Suparyanto (2019). Mobile recommendation system for culinary tourism destination using KNN (k-nearest neighbor). *Journal of Physics: Conference Series*, 1201(1):012039.

Roth, C., Kang, S. M., Batty, M., and Barthélemy, M. (2011). Structure of urban movements: polycentric activity and entangled hierarchical flows. *PLoS One*, 6(1):e15923.

Salas-Olmedo, M. H., Moya-Gómez, B., García-Palomares, J. C., and Gutiérrez, J. (2018). Tourists' digital footprint in cities: Comparing big data sources. *Tourism Management*, 66:13–25.

Sarica, S. and Luo, J. (2021). Stopwords in technical language processing. *PLoS One*, 16(8):e0254937.

Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). Collaborative filtering recommender systems. In *The Adaptive Web*, pages 291–324. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sebastia, L., Giret, A., and Garcia, I. (2010). A multi agent architecture for tourism recommendation. In *Advances in Intelligent and Soft Computing*, pages 547–554. Springer Berlin Heidelberg.

Selvi, C. and Sivasankar, E. (2017). A novel optimization algorithm for recommender system using modified fuzzy c-means clustering approach. *Soft Computing*, 23(6):1901–1916.

Sertkan, M., Neidhardt, J., and Werthner, H. (2020). Eliciting touristic profiles: A user study on picture collections. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*, pages 230––238, New York, NY, USA. Association for Computing Machinery.

Shoval, N., McKercher, B., Ng, E., and Birenboim, A. (2011). Hotel location and tourist activity in cities. *Annals of Tourism Research*, 38(4):1594–1612.

Smith, A. E. and Humphreys, M. S. (2006). Evaluation of unsupervised semantic mapping of natural language with leximancer concept mapping. *Behavior Research Methods*, 38(2):262–279.

Sohrabi, B., Raeesi Vanani, I., Nasiri, N., and Ghassemi Rudd, A. (2020). A predictive model of tourist destinations based on tourists' comments and interests using text analytics. *Tourism Management Perspectives*, 35(100710):100710.

Statista (2022). Most popular social networks worldwide as of January 2022, ranked by number of monthly active users. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Accessed: 2022-06-10.

Stieglitz, S., Dang-Xuan, L., Bruns, A., and Neuberger, C. (2014). Social media analytics. *Business and Information Systems Engineering*, 6(2):89–96.

Sugimoto, K., Ota, K., and Suzuki, S. (2019). Visitor mobility and spatial structure in a local urban tourism destination: GPS tracking and network analysis. *Sustainability*, 11(3):919.

Sun, Y., Shao, Y., and Chan, E. H. W. (2020). Co-visitation network in tourism-driven peri-urban area based on social media analytics: A case study in Shenzhen, China. *Landscape and Urban Planning*, 204(103934):103934.

Tenemaza, M., Lujan-Mora, S., Antonio, A. D., and Ramirez, J. (2020). Improving itinerary recommendations for tourists through metaheuristic algorithms: an optimization proposal. *IEEE Access*, 8:79003–79023.

Tiwari, S. and Kaushik, S. (2015). Crowdsourcing based fuzzy information enrichment of tourist spot recommender systems. In *Computational Science and Its Applications – ICCSA 2015*, pages 559–574. Springer International Publishing.

Tlili, T. and Krichen, S. (2021). A simulated annealing-based recommender system for solving the tourist trip design problem. *Expert Systems with Applications*, 186:115723.

Torres-Ruiz, M., Mata, F., Zagal, R., Guzmán, G., Quintero, R., and Moreno-Ibarra, M. (2018). A recommender system to generate museum itineraries applying augmented reality and social-sensor mining techniques. *Virtual Reality*, pages 175––189.

UNWTO (2019). International Tourism 2019 and Outlook for 2020. https://webunwto.s3.eu-west-1.amazonaws.com/s3fs-public/2020-01/Barometro-Jan-2020-EN-pre.pdf. Accessed: 2022-11-02.

UNWTO (2020). UNWTO world tourism barometer and statistical annex, January 2020. *UNWTO World Tourism Barometer*, 18:1–48.

van de Velden, M., D'Enza, A. I., and Markos, A. (2018). Distance-based clustering of mixed data. *WIREs Computational Statistics*, 11(3).

Van der Zee, E. and Bertocchi, D. (2018). Finding patterns in urban tourist behaviour: a social network analysis approach based on TripAdvisor reviews. *Information Technology and Tourism*, 20(1-4):153–180.

Vansteenwegen, P., Souffriau, W., and Oudheusden, D. V. (2011). The orienteering problem: A survey. *European Journal of Operational Research*, 209(1):1–10.

Vecchio, P. D., Mele, G., Ndou, V., and Secundo, G. (2018). Creating value from social big data: implications for smart tourism destinations. *Information Processing and Management*, 54(5):847–860.

Viktoratos, I., Tsadiras, A., and Bassiliades, N. (2018). Combining community-based knowledge with association rule mining to alleviate the cold start problem in context-aware recommender systems. *Expert Systems with Applications*, 101:78–90.

Vu, H. Q., Li, G., and Law, R. (2020). Cross-country analysis of tourist activities based on venue-referenced social media data. *Journal of Travel Research*, 59(1):90–106.

Vu, H. Q., Li, G., Law, R., and Ye, B. H. (2015). Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tourism Management*, 46:222–232.

Vu, H. Q., Li, G., Law, R., and Zhang, Y. (2018). Tourist activity analysis by leveraging mobile social media data. *Journal of Travel Research*, 57(7):883–898.

Wang, L. and Kirilenko, A. P. (2021). Do tourists from different countries interpret travel experience with the same feeling? Sentiment analysis of TripAdvisor reviews. In *Information and Communication Technologies in Tourism 2021*, pages 294–301. Springer International Publishing, Cham.

Wang, Y., Chan, S. C.-F., and Ngai, G. (2012). Applicability of demographic recommender system to tourist attractions: A case study on trip advisor. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 97–101. IEEE.

Wang, Y., Ma, W., Zhang, M., Liu, Y., and Ma, S. (2022). A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*.

Wood, S. A., Guerry, A. D., Silver, J. M., and Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, 3(1):2976.

WTTC (2021a). Economic Impact Reports. https://wttc.org/Research/Economic-Impact. Accessed: 2022-04-28.

WTTC (2021b). Global Economic Impact and Trends 2021. https://wttc.org/Portals/0/Documents/Reports/2021/Global%20Economic%20Impact%20and%20Trends%202021.pdf. Accessed: 2022-04-28.

Wu, Y., Li, Z., Wu, W., and Zhou, M. (2018). Response selection with topic clues for retrieval-based chatbots. *Neurocomputing*, 316:251–261.

Xiang, Z., Du, Q., Ma, Y., and Fan, W. (2018). Assessing reliability of social media data: lessons from mining TripAdvisor hotel reviews. *Information Technology and Tourism*, 18(1-4):43–59.

Xiang, Z. and Fesenmaier, D. R. (2017). Big data analytics, tourism design and smart tourism. In *Analytics in Smart Tourism Design*, pages 299–307. Springer International Publishing, Cham.

Xu, Y., Li, J., Belyi, A., and Park, S. (2021). Characterizing destination networks through mobility traces of international tourists — a case study using a nationwide mobile positioning dataset. *Tourism Management*, 82(104195):104195.

Xue, L. and Zhang, Y. (2020). The effect of distance on tourist behavior: a study based on social media data. *Annals of Tourism Research*, 82(102916):102916.

Yan, Y., Chen, J., and Wang, Z. (2020). Mining public sentiments and perspectives from geotagged social media data for appraising the post-earthquake recovery of tourism destinations. *Applied Geography*, 123(102306):102306.

Yang, E., Kim, J., and Pennington-Gray, L. (2021). Social media information and peer-to-peer accommodation during an infectious disease outbreak. *Journal of Destination Marketing and Management*, 19(100538):100538.

Yochum, P., Chang, L., Gu, T., and Zhu, M. (2020). An adaptive genetic algorithm for personalized itinerary planning. *IEEE Access*, 8:88147–88157.

Yu, Q., Pickering, S., Geng, R., and Yen, D. A. (2021). Thanks for the memories: exploring city tourism experiences via social media reviews. *Tourism Management Perspectives*, 40(100851):100851.

Yuan, C. and Uehara, M. (2019). Improvement of multi-purpose travel route recommendation system based on genetic algorithm. In *2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW)*, pages 305–308. IEEE.

Yuan, Q., Cong, G., Zhao, K., Ma, Z., and Sun, A. (2015). Who, where, when, and what: a nonparametric bayesian approach to context-aware recommendation and search for twitter users. *ACM Transactions on Information Systems*, 33(1):1–33.

Zachlod, C., Samuel, O., Ochsner, A., and Werthmüller, S. (2022). Analytics of social media data – state of characteristics and application. *Journal of Business Research*, 144:1064–1076.

Zanker, M., Fuchs, M., Seebacher, A., Jessenitschnig, M., and Stromberger, M. (2009). An automated approach for deriving semantic annotations of tourism products based on geospatial information. In *Information and Communication Technologies in Tourism 2009*, pages 211–221. Springer Vienna, Vienna.

Zhang, H., van Berkel, D., Howe, P. D., Miller, Z. D., and Smith, J. W. (2021). Using social media to measure and map visitation to public lands in utah. *Applied Geography*, 128(102389):102389.

Zhang, M., Wang, Y., and Olya, H. (2022). Shaping social media analytics in the pursuit of organisational agility: A real options theory perspective. *Tourism Management*, 88(104415):104415.

Zhang, X., Yang, Y., Zhang, Y., and Zhang, Z. (2020). Designing tourist experiences amidst air pollution: A spatial analytical approach using social media. *Annals of Tourism Research*, 84(102999):102999.

Zheng, X., Luo, Y., Sun, L., Yu, Q., Zhang, J., and Chen, S. (2021). A novel multi-objective and multi-constraint route recommendation method based on crowd sensing. *Applied Sciences*, 11(21):10497.

Zhou, X., Xu, C., and Kimmons, B. (2015). Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Computers, Environment and Urban Systems*, 54:144–153.

Appendix A

Additional data and Figures

A.1. Activity Tree

The complete activity tree is presented below. Main categories are in bold, subcategories are italicised, and specific activities as leaves are normal.

```
Activities
|-- Routes
| |-- SportsRoutes
| | |-- route_bicycle
| | |-- route_canoe
| | |-- route_hiking
| | |-- route_running
| | |-- route_ski
| | |-- route_horse
| | |-- route_mtb
| | +-- route_piste
| |-- RelaxationRoutes
| | +-- route_foot
| |-- NatureRoutes
| | +-- MountainRoutes
| | |-- climbing_route
| | +-- climbing_route_bottom
| |-- TownRoutes
| | |-- highway_pedestrian
| | +-- highway_footway
| +-- CultureRoutes
| |-- route_historic
| |-- historic_path
| +-- historic_way
|-- Sports
| |-- AquaticSports
| | |-- sport_water_polo
| | |-- sport_canoe
| | |-- sport_scuba_diving
| | |-- sport_kitesurfing
| | |-- sport_surfing
| | |-- sport_water_ski
| | |-- sport_cliff_diving
| | |-- sport_sailing
| | +-- sport_rowing
| |-- AirSports
| | |-- sport_parachuting
| | |-- sport_paragliding
| | +-- sport_free_flying
| |-- Climbing
| | |-- sport_climbing
| | +-- sport_climbing_adventure
| |-- MotorSports
| | +-- sport_karting
| |-- ShootingSports
| | |-- sport_shooting
| | |-- sport_archery
| | +-- sport_paintball
| +-- OtherSports
| |-- sport_bullfighting
| |-- sport_cycling
| |-- sport_golf
| |-- sport_9pin
| |-- sport_10pin
| |-- sport_ice_skating
| |-- sport_fishing
| +-- sport_skiing
|-- Gastronomy
| |-- Food
| | |-- amenity_bbq
| | |-- amenity_biergarten
| | |-- amenity_cafe
| | |-- amenity_restaurant
| | |-- amenity_food_court
| | |-- amenity_fast_food
| | |-- amenity_ice_cream
| | +-- craft_bakery
| +-- Enotourism
| |-- craft_winery
| |-- shop_brewing_supplies
| |-- landuse_vineyard
| |-- landuse_orchard
| +-- craft_brewery
```

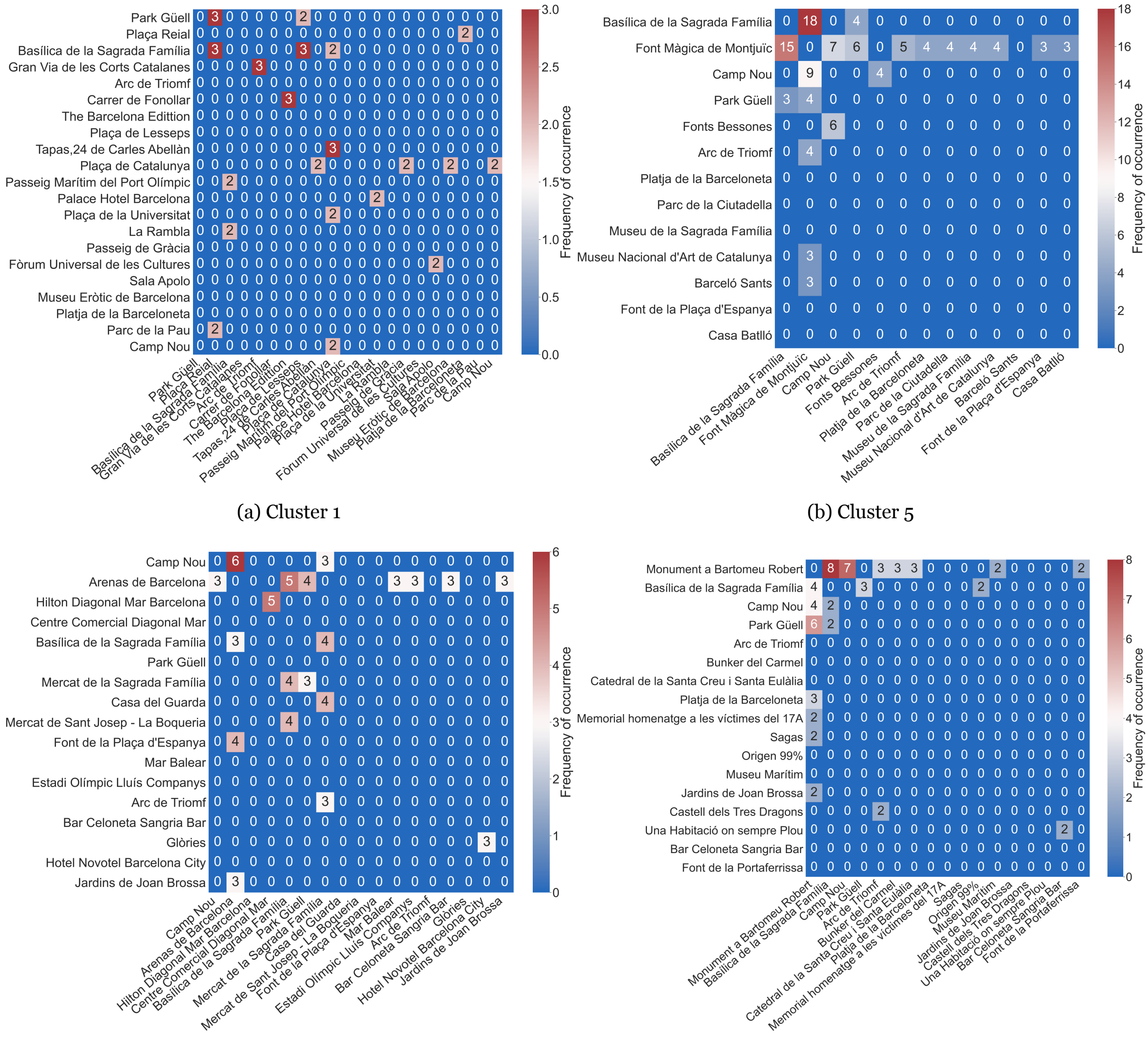
```
|-- Leisure
| |-- Parks&Recreation
| | |-- AmusementParks
| | | |-- tourism_zoo
| | | |-- tourism_theme_park
| | | |-- leisure_water_park
| | | +-- tourism_aquarium
| | +-- RecreationFacilities
| | |-- leisure_sports_centre
| | |-- amenity_cinema
| | |-- amenity_theatre
| | |-- leisure_stadium
| | |-- leisure_playground
| | |-- amenity_casino
| | |-- amenity_gambling
| | |-- building_stadium
| | |-- leisure_pitch
| | |-- leisure_amusement_arcade
| | |-- leisure_miniature_golf
| | |-- leisure_swimming_pool
| | |-- leisure_swimming_area
| | |-- leisure_ice_rink
| | |-- leisure_golf_course
| | |-- leisure_disk_golf_course
| | |-- leisure_bowling_alley
| | |-- leisure_horse_riding
| | |-- leisure_fishing
| | |-- leisure_garden
| | |-- leisure_park
| | +-- tourism_picnic_site
| |-- Beach
| | +-- natural_beach
| |-- Health&Care
| | |-- leisure_sauna
| | |-- shop_beauty
| | |-- shop_cosmetics
| | +-- shop_massage
| |-- NightLife
| | |-- amenity_nightclub
| | |-- amenity_pub
| | |-- amenity_stripclub
| | |-- amenity_bar
| | +-- amenity_brothel
| +-- Shopping
| |-- amenity_marketplace
| +-- shop_mall
```

```
|-- Accommodation
|  |-- tourism_hotel
|  |-- tourism_hostel
|  |-- building_hotel
|  |-- tourism_motel
|  |-- tourism_guest_house
|  |-- tourism_apartment
|  |-- tourism_chalet
|  |-- tourism_alpine_hut
|  |-- amenity_camping
|  |-- tourism_camp_site
|  |-- leisure_beach_resort
|  +-- leisure_resort
|-- Transportation
|  |-- aeroway_aerodrome
|  |-- building_train_station
|  |-- public_transport_station
|  |-- building_transportation
|  |-- aerialway_station
|  +-- railway_station
|-- Nature
|  |-- Landscape
|  |  |-- Landform
|  |  |  |-- natural_cliff
|  |  |  |-- natural_cave_entrance
|  |  |  |-- natural_peak
|  |  |  |-- natural_glacier
|  |  |  |-- natural_volcano
|  |  |  |-- natural_wood
|  |  |  |-- natural_grassland
|  |  |  |-- natural_heath
|  |  |  |-- natural_sand
|  |  |  |-- natural_rock
|  |  |  |-- natural_mountain_range
|  |  |  |-- natural_valley
|  |  |  |-- natural_ridge
|  |  |  |-- natural_desert
|  |  |  +-- natural_tree
|  |  |-- CoastalAreas
|  |  |  |-- natural_bay
|  |  |  |-- natural_coastline
|  |  |  +-- natural_reef
|  |  +-- InlandWaters
|  |  |-- waterway_stream
|  |  |-- waterway_waterfall
|  |  |-- waterway_canal
| | |-- waterway_river
| | |-- natural_water
| | |-- natural_spring
| | +-- natural_hot_spring
| +-- ProtectedAreas
| |-- leisure_nature_reserve
| |-- boundary_national_park
| +-- boundary_protected_area
+-- Culture
|-- Museums
| |-- tourism_museum
| |-- amenity_arts_centre
| |-- tourism_gallery
| +-- tourism_artwork
|-- Monuments
| |-- Religious
| | |-- building_cathedral
| | |-- building_chapel
| | |-- building_church
| | |-- historic_monastery
| | |-- historic_church
| | |-- building_temple
| | |-- amenity_monastery
| | |-- historic_wayside_cross
| | +-- amenity_place_of_worship
| +-- Historic
| |-- historic_fort
| |-- historic_battlefield
| |-- historic_cannon
| |-- historic_citywalls
| |-- historic_ruins
| |-- historic_archaeological_site
| |-- historic_tower
| |-- historic_aqueduct
| |-- historic_city_gate
| |-- historic_castle
| |-- historic_monument
| |-- historic_wayside_shrine
| |-- historic_memorial
| |-- historic_manor
| |-- historic_pillory
| |-- historic_heritage
| +-- historic_tomb
|-- Viewpoint
| +-- tourism_viewpoint
+-- CulturalAmenities
```

```
|-- amenity_fountain
|-- barrier_city_wall
|-- amenity_planetarium
|-- amenity_grave_yard
+--- amenity_crypt
```

A.2. Complementary charts for chapter 4

A.2.1 Mobility heatmap charts for other relevant clusters



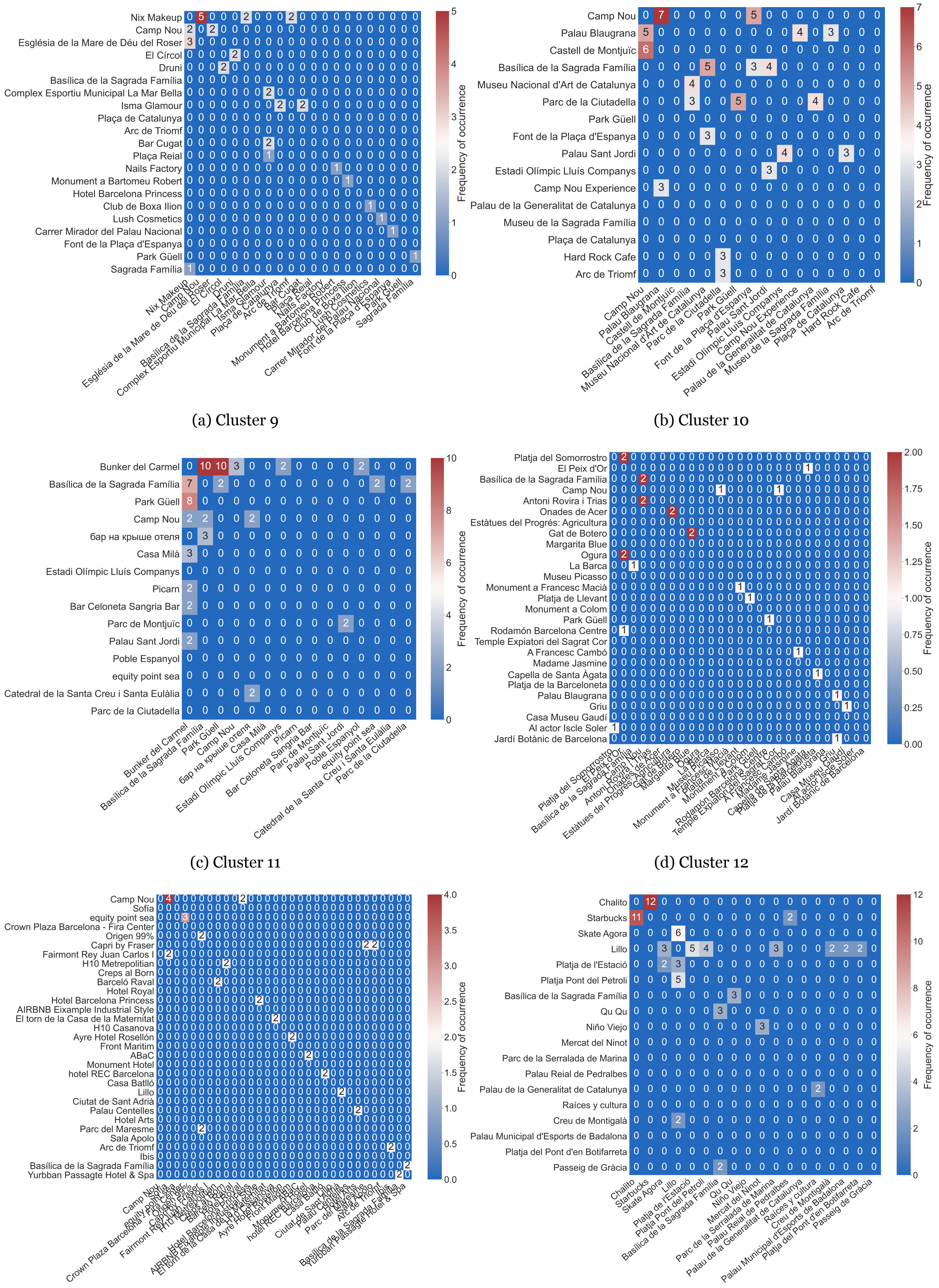
(a) Cluster 1

(b) Cluster 5

(c) Cluster 6

(d) Cluster 8

Figure A.1: Mobility heat-maps for clusters 1, 5, 6, and 8 representing bigrams with movement from left to bottom.



(a) Cluster 9

(b) Cluster 10

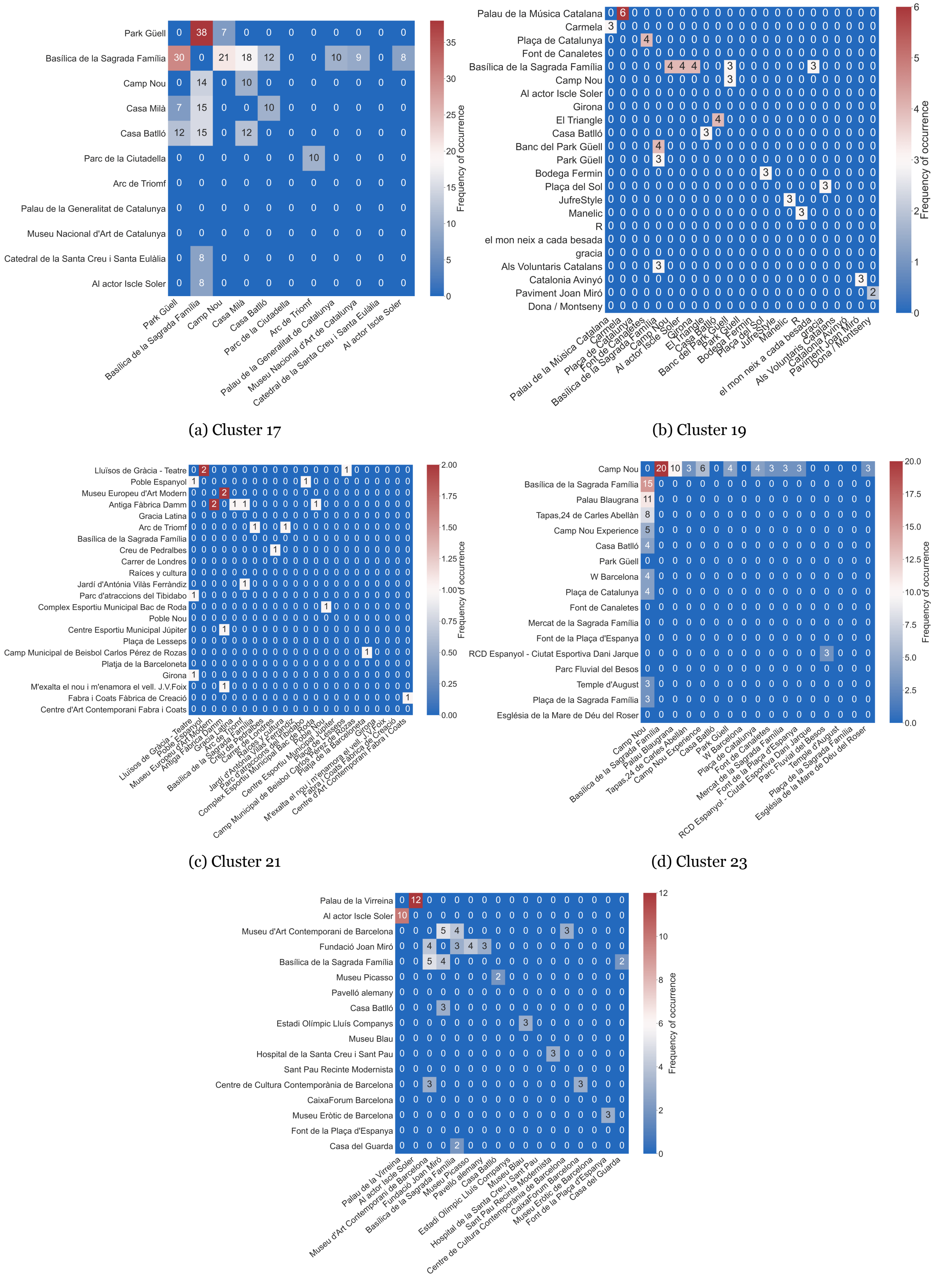
(c) Cluster 11

(d) Cluster 12

(e) Cluster 13

(f) Cluster 15

Figure A.2: Mobility heat-maps for clusters 9, 10, 11, 12, 13, and 15 representing bigrams with movement from left to bottom.



(a) Cluster 17

(b) Cluster 19

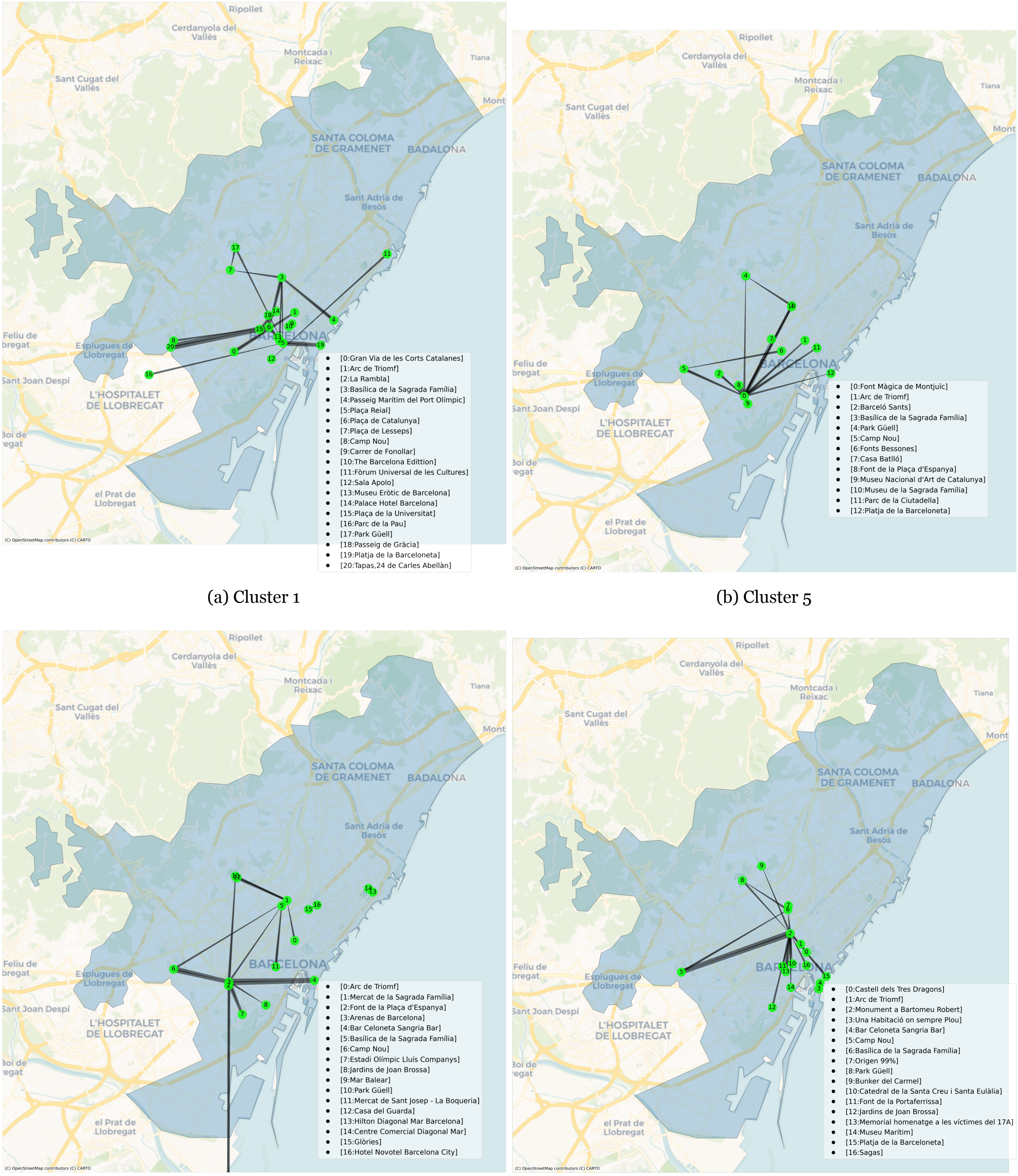
(c) Cluster 21

(d) Cluster 23

(e) Cluster 24

Figure A.3: Mobility heat-maps for clusters 17, 19, 21, 23, and 24 representing bigrams with movement from left to bottom.

A.2.2 Tourist mobility pattern charts for other relevant clusters


[0:Gran Via de les Corts Catalanes]
[1:Arc de Triomf]
[2:La Rambla]
[3:Basílica de la Sagrada Família]
[4:Passeig Marítim del Port Olímpic]
[5:Plaça Reial]
[6:Plaça de Catalunya]
[7:Plaça de Lesseps]
[8:Camp Nou]
[9:Carrer de Fonollar]
[10:The Barcelona Edittion]
[11:Fòrum Universal de les Cultures]
[12:Sala Apolo]
[13:Museu Eròtic de Barcelona]
[14:Palace Hotel Barcelona]
[15:Plaça de la Universitat]
[16:Parc de la Pau]
[17:Park Güell]
[18:Passeig de Gràcia]
[19:Platja de la Barceloneta]
[20:Tapas,24 de Carles Abellàn]
[0:Font Màgica de Montjuïc]
[1:Arc de Triomf]
[2:Barceló Sants]
[3:Basílica de la Sagrada Família]
[4:Park Güell]
[5:Camp Nou]
[6:Fonts Bessones]
[7:Casa Batlló]
[8:Font de la Plaça d'Espanya]
[9:Museu Nacional d'Art de Catalunya]
[10:Museu de la Sagrada Família]
[11:Parc de la Ciutadella]
[12:Platja de la Barceloneta]
[0:Arc de Triomf]
[1:Mercat de la Sagrada Família]
[2:Font de la Plaça d'Espanya]
[3:Arenas de Barcelona]
[4:Bar Celoneta Sangria Bar]
[5:Basílica de la Sagrada Família]
[6:Camp Nou]
[7:Estadi Olímpic Lluís Companys]
[8:Jardins de Joan Brossa]
[9:Mar Balear]
[10:Park Güell]
[11:Mercat de Sant Josep - La Boqueria]
[12:Casa del Guarda]
[13:Hilton Diagonal Mar Barcelona]
[14:Centre Comercial Diagonal Mar]
[15:Glòries]
[16:Hotel Novotel Barcelona City]
[0:Castell dels Tres Dragons]
[1:Arc de Triomf]
[2:Monument a Bartomeu Robert]
[3:Una Habitació on sempre Plou]
[4:Bar Celoneta Sangria Bar]
[5:Camp Nou]
[6:Basílica de la Sagrada Família]
[7:Origen 99%]
[8:Park Güell]
[9:Bunker del Carmel]
[10:Catedral de la Santa Creu i Santa Eulàlia]
[11:Font de la Portaferrissa]
[12:Jardins de Joan Brossa]
[13:Memorial homenatge a les víctimes del 17A]
[14:Museu Marítim]
[15:Platja de la Barceloneta]
[16:Sagas]
(C) OpenStreetMap contributors (C) CARTO

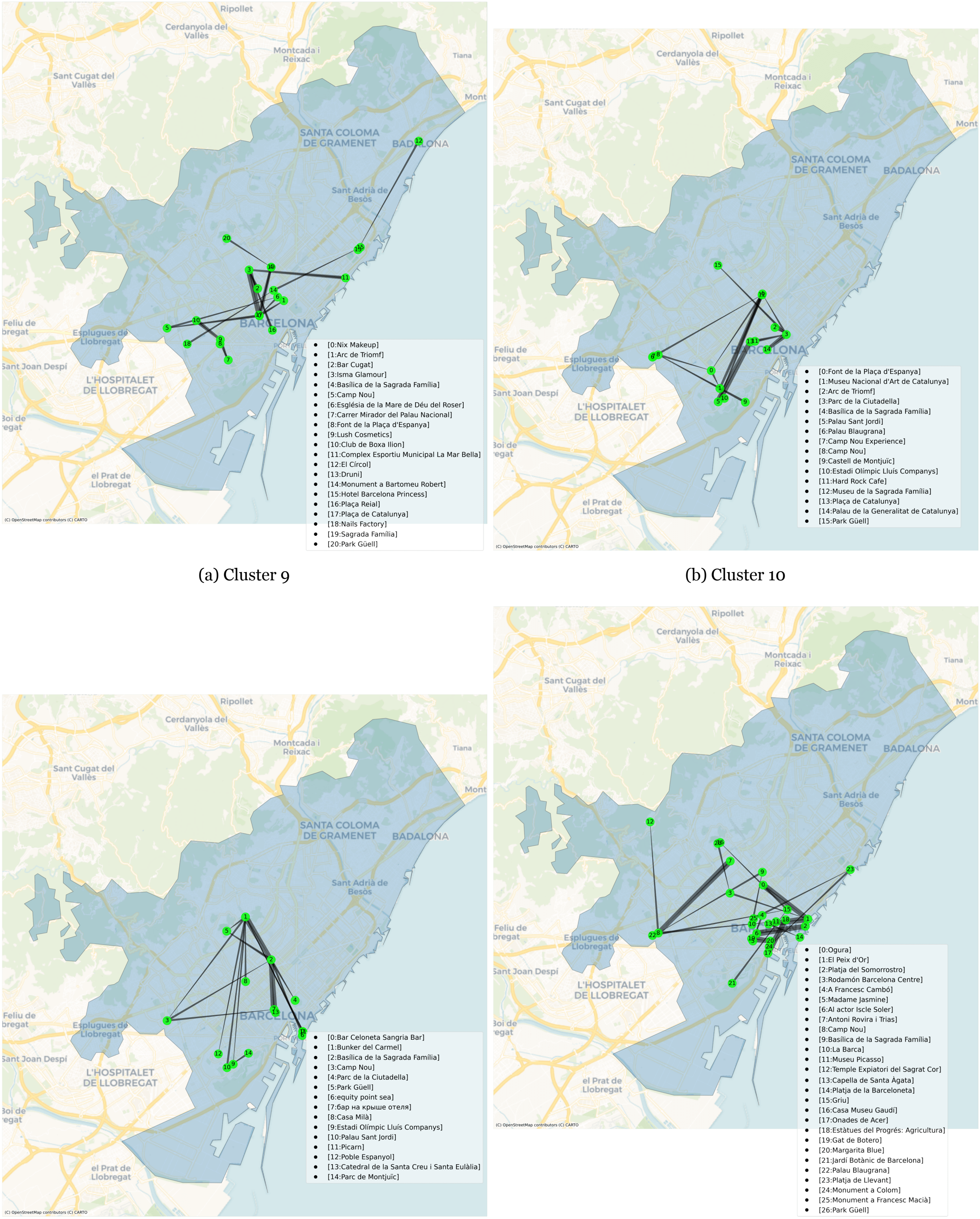
(a) Cluster 1

(b) Cluster 5

(c) Cluster 6

(d) Cluster 8

Figure A.4: Tourist mobility pattern between attractions in Barcelona for clusters 1, 5, 6, and 8.


[0:Nix Makeup]
[1:Arc de Triomf]
[2:Bar Cugat]
[3:Isma Glamour]
[4:Basílica de la Sagrada Família]
[5:Camp Nou]
[6:Església de la Mare de Déu del Roser]
[7:Carrer Mirador del Palau Nacional]
[8:Font de la Plaça d'Espanya]
[9:Lush Cosmetics]
[10:Club de Boxa Ilion]
[11:Complex Esportiu Municipal La Mar Bella]
[12:El Círcol]
[13:Druni]
[14:Monument a Bartomeu Robert]
[15:Hotel Barcelona Princess]
[16:Plaça Reial]
[17:Plaça de Catalunya]
[18:Nails Factory]
[19:Sagrada Família]
[20:Park Güell]
[0:Font de la Plaça d'Espanya]
[1:Museu Nacional d'Art de Catalunya]
[2:Arc de Triomf]
[3:Parc de la Ciutadella]
[4:Basílica de la Sagrada Família]
[5:Palau Sant Jordi]
[6:Palau Blaugrana]
[7:Camp Nou Experience]
[8:Camp Nou]
[9:Castell de Montjuïc]
[10:Estadi Olímpic Lluís Companys]
[11:Hard Rock Cafe]
[12:Museu de la Sagrada Família]
[13:Plaça de Catalunya]
[14:Palau de la Generalitat de Catalunya]
[15:Park Güell]
[0:Bar Celoneta Sangria Bar]
[1:Bunker del Carmel]
[2:Basílica de la Sagrada Família]
[3:Camp Nou]
[4:Parc de la Ciutadella]
[5:Park Güell]
[6:equity point sea]
[7:бар на крыше отеля]
[8:Casa Milà]
[9:Estadi Olímpic Lluís Companys]
[10:Palau Sant Jordi]
[11:Picarn]
[12:Poble Espanyol]
[13:Catedral de la Santa Creu i Santa Eulàlia]
[14:Parc de Montjuïc]
[0:Ogura]
[1:El Peix d'Or]
[2:Platja del Somorrostro]
[3:Rodamón Barcelona Centre]
[4:A Francesc Cambó]
[5:Madame Jasmine]
[6:Al actor Iscle Soler]
[7:Antoni Rovira i Trias]
[8:Camp Nou]
[9:Basílica de la Sagrada Família]
[10:La Barca]
[11:Museu Picasso]
[12:Temple Expiatori del Sagrat Cor]
[13:Capella de Santa Àgata]
[14:Platja de la Barceloneta]
[15:Griu]
[16:Casa Museu Gaudí]
[17:Onades de Acer]
[18:Estàtues del Progrés: Agricultura]
[19:Gat de Botero]
[20:Margarita Blue]
[21:Jardí Botànic de Barcelona]
[22:Palau Blaugrana]
[23:Platja de Llevant]
[24:Monument a Colom]
[25:Monument a Francesc Macià]
[26:Park Güell]
(C) OpenStreetMap contributors (C) CARTO

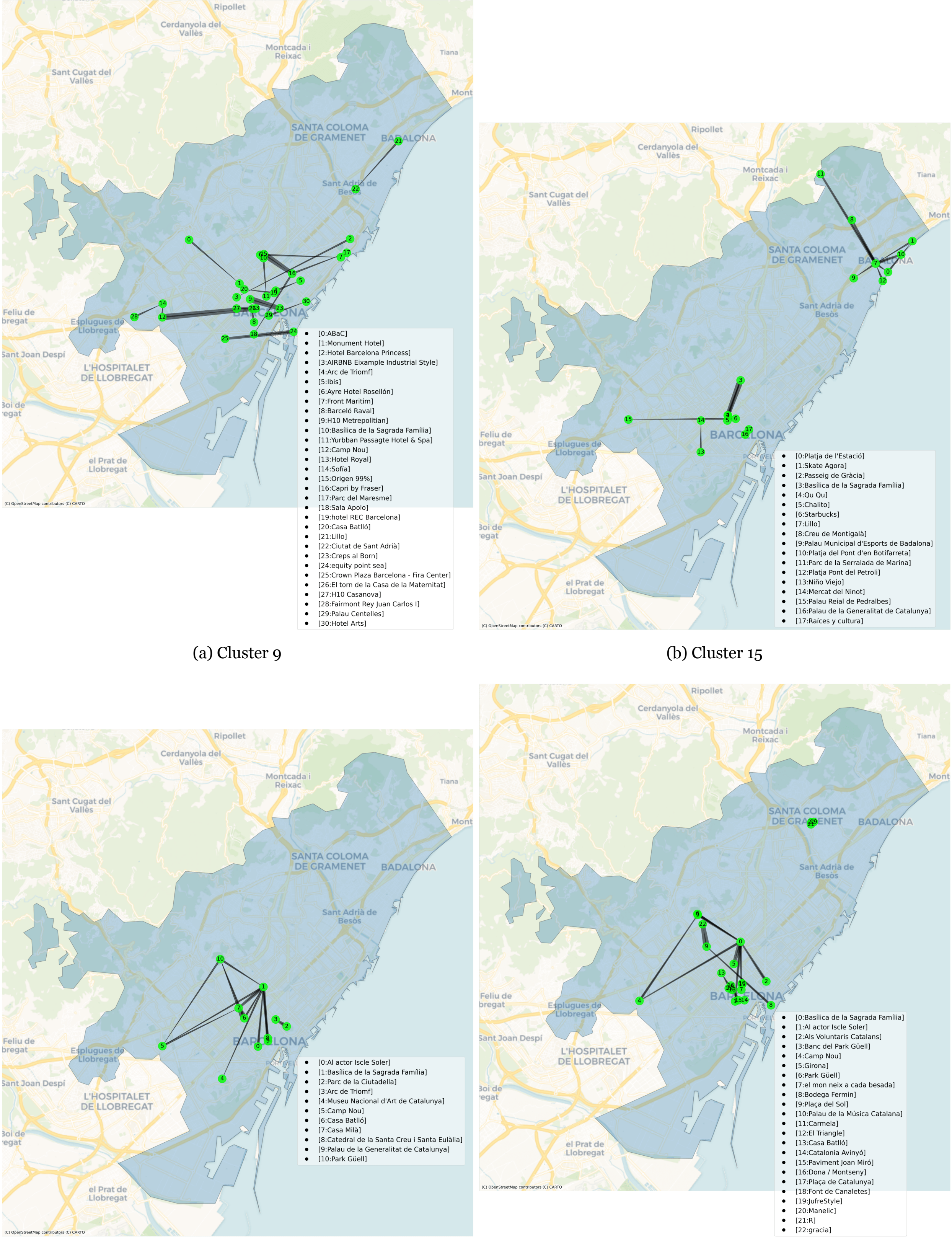
(a) Cluster 9

(b) Cluster 10

(c) Cluster 11

(d) Cluster 12

Figure A.5: Tourist mobility pattern between attractions in Barcelona for clusters 9, 10, 11, and 12.


[0:ABaC]
[1:Monument Hotel]
[2:Hotel Barcelona Princess]
[3:AIRBNB Eixample Industrial Style]
[4:Arc de Triomf]
[5:Ibis]
[6:Ayre Hotel Rosellón]
[7:Front Maritim]
[8:Barceló Raval]
[9:H10 Metrepolitian]
[10:Basílica de la Sagrada Família]
[11:Yurbban Passagte Hotel & Spa]
[12:Camp Nou]
[13:Hotel Royal]
[14:Sofía]
[15:Origen 99%]
[16:Capri by Fraser]
[17:Parc del Maresme]
[18:Sala Apolo]
[19:hotel REC Barcelona]
[20:Casa Batlló]
[21:Lillo]
[22:Ciutat de Sant Adrià]
[23:Creps al Born]
[24:equity point sea]
[25:Crown Plaza Barcelona - Fira Center]
[26:El torn de la Casa de la Maternitat]
[27:H10 Casanova]
[28:Fairmont Rey Juan Carlos I]
[29:Palau Centelles]
[30:Hotel Arts]
(C) OpenStreetMap contributors (C) CARTO

(a) Cluster 9

[0:Platja de l'Estació]
[1:Skate Agora]
[2:Passeig de Gràcia]
[3:Basílica de la Sagrada Família]
[4:Qu Qu]
[5:Chalito]
[6:Starbucks]
[7:Lillo]
[8:Creu de Montigalà]
[9:Palau Municipal d'Esports de Badalona]
[10:Platja del Pont d'en Botifarreta]
[11:Parc de la Serralada de Marina]
[12:Platja Pont del Petroli]
[13:Niño Viejo]
[14:Mercat del Ninot]
[15:Palau Reial de Pedralbes]
[16:Palau de la Generalitat de Catalunya]
[17:Raíces y cultura]
(C) OpenStreetMap contributors (C) CARTO

(b) Cluster 15

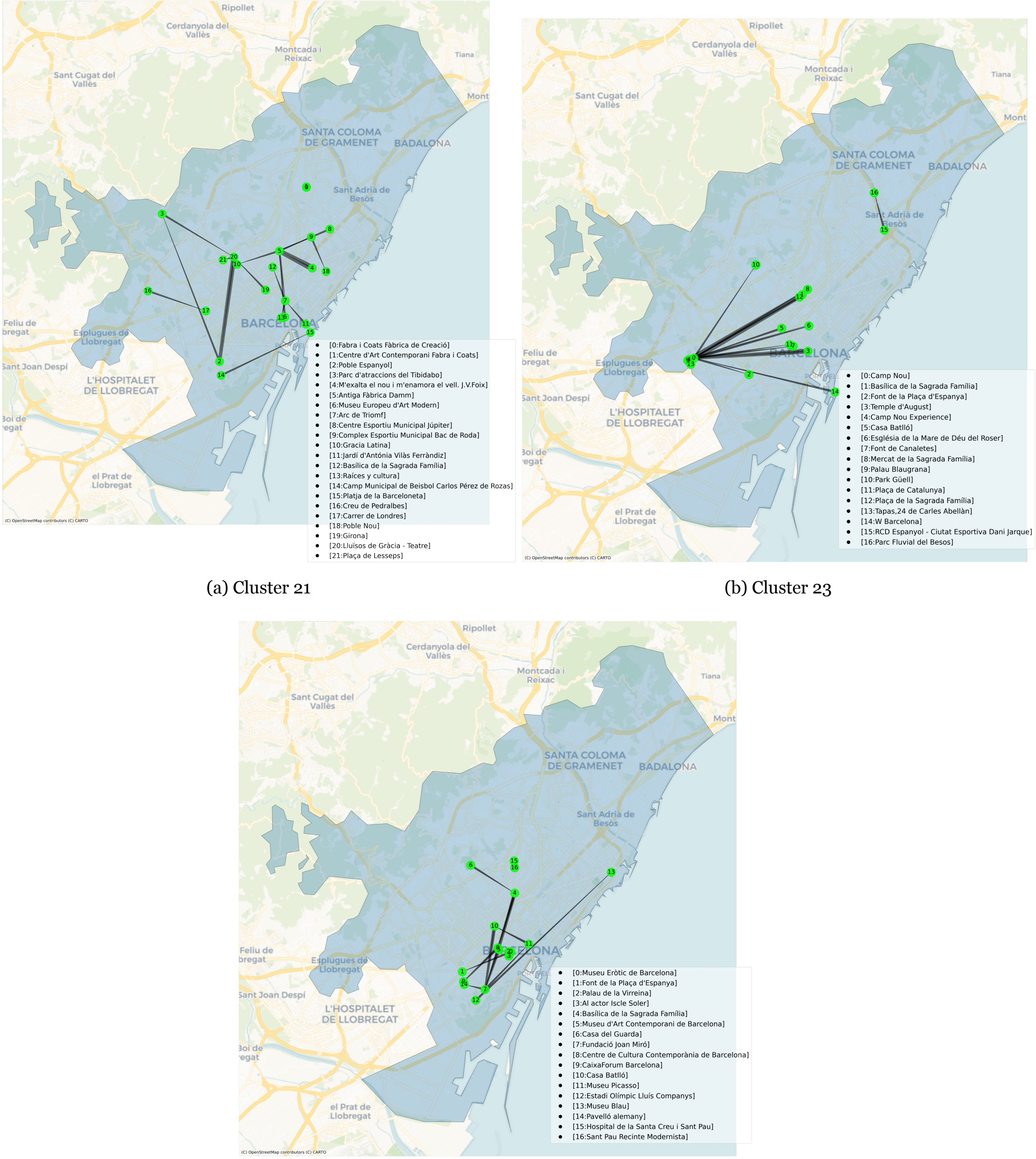
[0:Al actor Iscle Soler]
[1:Basílica de la Sagrada Família]
[2:Parc de la Ciutadella]
[3:Arc de Triomf]
[4:Museu Nacional d'Art de Catalunya]
[5:Camp Nou]
[6:Casa Batlló]
[7:Casa Milà]
[8:Catedral de la Santa Creu i Santa Eulàlia]
[9:Palau de la Generalitat de Catalunya]
[10:Park Güell]
(C) OpenStreetMap contributors (C) CARTO

(c) Cluster 17

[0:Basílica de la Sagrada Família]
[1:Al actor Iscle Soler]
[2:Als Voluntaris Catalans]
[3:Banc del Park Güell]
[4:Camp Nou]
[5:Girona]
[6:Park Güell]
[7:el mon neix a cada besada]
[8:Bodega Fermin]
[9:Plaça del Sol]
[10:Palau de la Música Catalana]
[11:Carmela]
[12:El Triangle]
[13:Casa Batlló]
[14:Catalonia Avinyó]
[15:Paviment Joan Miró]
[16:Dona / Montseny]
[17:Plaça de Catalunya]
[18:Font de Canaletes]
[19:JufreStyle]
[20:Manelic]
[21:R]
[22:gracia]
(C) OpenStreetMap contributors (C) CARTO

(d) Cluster 19

Figure A.6: Tourist mobility pattern between attractions in Barcelona for clusters 13, 15, 17, and 19.


[0:Fabra i Coats Fàbrica de Creació]
[1:Centre d'Art Contemporani Fabra i Coats]
[2:Poble Espanyol]
[3:Parc d'atraccions del Tibidabo]
[4:M'exalta el nou i m'enamora el vell. J.V.Foix]
[5:Antiga Fàbrica Damm]
[6:Museu Europeu d'Art Modern]
[7:Arc de Triomf]
[8:Centre Esportiu Municipal Júpiter]
[9:Complex Esportiu Municipal Bac de Roda]
[10:Gracia Latina]
[11:Jardí d'Antónia Vilàs Ferràndiz]
[12:Basílica de la Sagrada Família]
[13:Raíces y cultura]
[14:Camp Municipal de Beisbol Carlos Pérez de Rozas]
[15:Platja de la Barceloneta]
[16:Creu de Pedralbes]
[17:Carrer de Londres]
[18:Poble Nou]
[19:Girona]
[20:Lluïsos de Gràcia - Teatre]
[21:Plaça de Lesseps]
(C) OpenStreetMap contributors (C) CARTO

(a) Cluster 21

[0:Camp Nou]
[1:Basílica de la Sagrada Família]
[2:Font de la Plaça d'Espanya]
[3:Temple d'August]
[4:Camp Nou Experience]
[5:Casa Batlló]
[6:Església de la Mare de Déu del Roser]
[7:Font de Canaletes]
[8:Mercat de la Sagrada Família]
[9:Palau Blaugrana]
[10:Park Güell]
[11:Plaça de Catalunya]
[12:Plaça de la Sagrada Família]
[13:Tapas,24 de Carles Abellàn]
[14:W Barcelona]
[15:RCD Espanyol - Ciutat Esportiva Dani Jarque]
[16:Parc Fluvial del Besos]
(C) OpenStreetMap contributors (C) CARTO

(b) Cluster 23

[0:Museu Eròtic de Barcelona]
[1:Font de la Plaça d'Espanya]
[2:Palau de la Virreina]
[3:Al actor Iscle Soler]
[4:Basílica de la Sagrada Família]
[5:Museu d'Art Contemporani de Barcelona]
[6:Casa del Guarda]
[7:Fundació Joan Miró]
[8:Centre de Cultura Contemporània de Barcelona]
[9:CaixaForum Barcelona]
[10:Casa Batlló]
[11:Museu Picasso]
[12:Estadi Olímpic Lluís Companys]
[13:Museu Blau]
[14:Pavelló alemany]
[15:Hospital de la Santa Creu i Sant Pau]
[16:Sant Pau Recinte Modernista]
(C) OpenStreetMap contributors (C) CARTO

(c) Cluster 24

Figure A.7: Tourist mobility pattern between attractions in Barcelona for clusters 21, 23, and 24.