TUM School of Computation, Information and Technology
Technische Universität München

Data-driven Destination Recommender Systems

Linus W. Dietz

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:
apl. Prof. Dr. Georg Groh

Prüfer*innen der Dissertation:

1. Prof. Dr.-Ing. Jörg Ott
2. Prof. Dr. Alejandro Bellogín Kouki

Die Dissertation wurde am 05. 09. 2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 05. 01. 2023 angenommen.

Acknowledgments

First and foremost, I want to thank Wolfgang Wörndl for the inspiration and advising on the research topic of this thesis. It was delightful to work at the Chair of Connected Mobility at the Technical University of Munich which, is due to the great colleagues and Jörg Ott, who established an excellent research environment. One of my greatest joys at university was working with brilliant students by advising their research projects. The quality of the theses was outstanding and many were subsequently published in international research outlets. Especially, I would like to thank Rinita Roy, Avradip Sen, Saadi Myftija, and Sameera Thimbiri Palage, whose projects contributed to this thesis.

My gratitude extends to the RecSys community, which I am happy to be part of and be able to contribute to. Of the many colleagues, I want to thank Alejandro Bellogín, as he generously hosted me for a research visit in the summer of 2022 and subsequently served as second examiner for this thesis. Furthermore, I would like to thank the external co-authors of the publications embedded in this thesis, Julia Neidhardt, Mete Sertkan, and Pablo Sánchez, for the great collaboration and discussions.

Finally, I would like to thank my parents, who supported me throughout my studies, and Ilka, who has been by my side at all times.

Abstract

Deciding where to travel is a complex, emotionally involving, and financially relevant decision which people face relatively infrequently. Some aspects of tourist recommendations, such as point-of-interest recommendation, hotel recommendations, or restaurant recommendations, are commercially well established, whereas there are few successful recommender systems for individual travel destinations. In this thesis, we present several contributions in the context of destination recommendation covering traveler mobility analysis, destination characterization, and conversational recommender systems.

Understanding traveler mobility forms the basis for more personalized recommendations. We propose methods to analyze global traveler mobility from location-based social networks to learn which data sources are suitable for analyses in this domain, how people travel around the world, and which types of travelers can be observed. Our cluster analyses of trips and travelers reveal distinct groups, which can serve as an initial preference elicitation step, but we could also show that the common practice of evaluating point-of-interest recommendations without differentiating these groups leads to misleading results. Furthermore, we use the mined trips to construct a specialized map of hierarchical travel regions and to recommend the personalized duration of stay at destinations.

To correctly match traveler preferences with destinations in the content-based recommendation domain, we investigate which data sources are suitable for characterizing destinations. Constructing 18 data models and eliciting the concept of touristic experience in cities using an expert study, we determine that textual data sources, e.g., Wikipedia articles, do a good job of emulating the touristic experience using rank agreement metrics. Additionally, we are able to optimize data sources with explicit features to be competitive by learning the importance of feature weights using black-box learning.

Finally, we present the CityRec conversational destination recommender system. Since users often struggle to verbalize their true preferences and might have unrealistic expectations about destinations, we propose a novel conversational paradigm, “Navigation by Revealing Trade-offs”, to overcome the wishful-thinking problem and inform users of the trade-offs involved in choosing one destination over another. The seamless integration of user interface and algorithms was evaluated using a large-scale user study with 600 participants.

Zusammenfassung

Die Entscheidung, wohin eine Reise gehen soll, ist eine komplexe, emotional fordernde und finanziell bedeutsame Entscheidung, vor der Menschen allerdings relativ selten stehen. Einige relevante Bereiche von Reisempfehlungen, wie z.B. Empfehlungen von Sehenswürdigkeiten, Hotels oder Restaurants sind kommerziell gut etabliert. Allerdings existieren – bedingt durch die immense Komplexität von Reiseempfehlungen – noch kaum erfogreiche Empfehlungssysteme für individuelle Reiseplanung. Hier knüpft die vorlegende Doktorarbeit an und schafft durch verschiedene Beiträge zur individuellen Reiseplanung Grundlagen für verbesserte Systeme.

Im ersten Teil schlagen wir Methoden zur Analyse der globalen Reisemobilität aus standortbasierten sozialen Netzwerken heraus vor, um zu bestimmen, (1) welche Datenquellen sich für Analysen in diesem Bereich eignen, (2) wie Menschen die Welt bereisen und (3) welche Arten von Reisenden beobachtet werden können. Über Cluster-Analysen von Reisen und Reisenden werden unterschiedliche Gruppen ermittelt, die als Ausgangspunkt zur Präferenzerhebung in Empfehlungsdiensten dienen können. Mithilfe der Unterscheidung dieser Gruppen können wir die gängige Praxis in der Evaluation von Empfehlungsalgorithmen für Sehenswürdigkeiten, die keine Gruppen unterscheidet als wenig zielführend kennzeichnen. Darüber hinaus verwenden wir die gesammelten Reisedaten, um eine spezielle Karte von hierarchisch gegliederten Reiseregionen zu erstellen und eine personalisierte Aufenthaltsdauer bezogen auf die jeweiligen Reisezielen zu empfehlen.

Im zweiten Teil dieser Arbeit untersuchen wir, welche Datenquellen geeignet sind, um Reiseziele für merkmals-basierte Empfehlungsdienste zu charakterisieren. Durch die Erstellung von 18 Datenmodellen und die Erfassung des touristischen Erlebnisses in Städten stellen wir in einer Expertenstudie fest, dass textbasierte Datenquellen wie z.B. Wikipedia-Artikel, das gewünschte Konzept mit Hilfe von Rangkorrelationskoeffizienten gut abbilden. Wir sind darüber hinaus in der Lage, Datenquellen mit expliziten Merkmalen so zu optimieren, dass sie wettbewerbsfähig werden, indem wir die Gewichtung der Bedeutung von Merkmalen mithilfe von Black-Box-Lernen optimieren.

Zuletzt stellen wir das CityRec-System zur dialogbasierten Empfehlung von Reisezielen vor. Da Nutzer oft Schwierigkeiten haben, ihre tatsächlichen Präferenzen auszudrücken und möglicherweise unrealistische Erwartungen an Reiseziele haben, schlagen wir mit „Navigation by Revealing Trade-offs“ (Navigation durch Offenlegung von Kompromissen) ein neuartiges Konversationsparadigma vor, um das Problem des Wunschdenkens zu überwinden und die Nutzer über die mit ihren Entscheidungen verbundenen Kompromisse zu informieren. Die nahtlose Integration der Nutzeroberfläche und darauf abgestimmten Algorithmen wurde in einer groß angelegten Nutzerstudie mit 600 Teilnehmern evaluiert.

Contents

List of Figures

List of Tables

List of Abbreviations

HCI human-computer interaction

LBSN location-based social network

MRR Mean Reciprocal Rank

nDCG Normalized Discounted Cumulative Gain

OSM OpenStreetMap

POI point of interest

RS recommender system

Publications

This thesis consists of an overview and of the following publications, which are referred to in the text by their numeral prefixed with P. All publications are subject to a full peer-review process.

[P1] Linus W. Dietz, Avradip Sen, Rinita Roy, and Wolfgang Wörndl. "Mining Trips from Location-based Social Networks for Clustering Travelers and Destinations." In: *Information Technology & Tourism* 22.1 (Mar. 2020), pp. 131–166. ISSN: 1098-3058. DOI: 10.1007/s40558-020-00170-6.

[P2] Pablo Sánchez and Linus W. Dietz. "Travelers vs. Locals: The Effect of Cluster Analysis in Point-of-Interest Recommendation." In: *30th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP'22. New York, NY, USA: ACM, July 2022, pp. 132–142. DOI: 10.1145/3503252.3531320.

[P3] Linus W. Dietz and Wolfgang Wörndl. "How Long to Stay Where? On the Amount of Item Consumption in Travel Recommendation." In: *ACM RecSys 2019 Late-breaking Results*. Sept. 2019, pp. 31–35.

[P4] Linus W. Dietz, Mete Sertkan, Saadi Myftija, Sameera Thimbiri Palage, Julia Neidhardt, and Wolfgang Wörndl. "A Comparative Study of Data-driven Models for Travel Destination Characterization." In: *Frontiers in Big Data* 5 (Apr. 2022). ISSN: 2624-909X. DOI: 10.3389/fdata.2022.829939.

[P5] Linus W. Dietz, Saadi Myftija, and Wolfgang Wörndl. "Designing a Conversational Travel Recommender System Based on Data-driven Destination Characterization." In: *ACM RecSys Workshop on Recommenders in Tourism*. Sept. 2019, pp. 17–21.

[P6] Linus W. Dietz, Sameera Thimbiri Palage, and Wolfgang Wörndl. "Navigation by Revealing Trade-offs for Content-based Recommendations." In: *Information and Communication Technologies in Tourism*. Ed. by Jason L. Stienmetz, Berta Ferrer-Rosell, and David Massimo. Cham: Springer, Jan. 2022, pp. 149–161. ISBN: 978-3-030-94751-4. DOI: 10.1007/978-3-030-94751-4_14.

A full reprint of the embedded publications can be found in Appendix A.

1 Introduction

Recommender systems are ubiquitous in today's digital media. Given the enormous quantity of content on the Web, recommendation technologies help users to overcome information overload they would otherwise be suffering. With few exceptions, most content shown on online platforms is determined by some kind of recommendation algorithm, making it *"one of the most compelling success stories of AI"* [72]. Users greatly benefit from this development, since they are presented with content that is of relevance to them instead of having to browse millions of products, songs, or videos. Even traditional media, such as news organizations who heavily relied on curation in the past, need to employ recommendation technology to satisfy the information needs of their heterogeneous audiences [73, 115]. At the same time, the advances in recommender systems put all market players under immense pressure to roll out high-quality recommendations in order to not loose against competitors with superior quality of personalized recommendations. This leads to a high interest in recommendation technology from industry and enables a competitive and fast-growing research community[1].

Today's rating-based recommendation algorithms have reached such high maturity level that it becomes increasingly hard to independently evaluate the actual progress in these fields [32, 31]. Since the performance of common rating-based recommendation algorithms is so high, the competitiveness of real-world systems increasingly depends on the user experience such as the presentation of items [77, 70]. Furthermore, many recommendation scenarios cannot be addressed with rating-based algorithms, but require custom solutions with a seamless integration of algorithms and user interfaces.

1.1 Motivation

In this doctoral thesis, we analyze and resolve various problems concerning travel recommender systems. Following the success of e-commerce, e-tourism has transformed the way how people plan and book their vacations [42, 57, 150]. A significant difference to recommendations in the e-commerce domain is that each trip is different in a sense that when people visit a destination, the experience will be determined by various factors, such as the visited attractions, the cultural experience, the climate during the visiting period, the costs involved, and many more. To compute meaningful recommendations, a destination recommender would need to approximate a scoring function over all aspects that are relevant to the current user. Furthermore, when planning a trip to an unknown destination, users' expectations might be influenced by biased information [101, 86], which motivated us to develop unbiased, data-driven methods that yield results that can directly be used in future travel recommender systems.

From a recommender systems research perspective, traveling is a relatively rare activity compared to watching a movie or buying a product, which in conjunction with the fuzziness of destinations as items, makes collaborative filtering approaches

[1] `https://recsys.acm.org/statistics/`

infeasible [19]. Thus, the recommendations need to based on features describing the destinations, which can also be used to familiarize the users users with potential destinations. This interaction with potential items within an e-tourism recommender system, can be helpful for the users to come to a decision, since there there is a high emotional involvement in the decision making in travel planning [150]. These challenges offer a huge opportunity to recommender systems that can provide users with relevant and reliable information about destinations to visit.

Analyzing the current state of digital travel information systems in the e-tourism industry and scientific community, it is striking how lacking the support for individual travel planning is. While there are various commercial platforms for hotel and point of interest (POI) recommendations, gathering reliable information about destinations is tedious as the information is scattered and becomes outdated quickly. Tourism boards and destination marketing organizations naturally promote their own destinations while commercial travel websites for booking hotels or airfares feature destination recommendations only as a side product as these platforms are currently mostly consulted when the user has already decided which destination to visit. One rare example would be booking.com's efforts to make suitable recommendations in a trip continuation scenario [52]. Other than that, commercial platforms present non-personalized destination recommendations for the current travel season, typically based on a popularity recommendation strategy. In academia, research contributions usually focus on a narrow aspect of the problem, disregarding much of the complexity [22]. Furthermore, existing approaches in literature are often heavily reliant on costly expert knowledge, which might be feasible to show the merit of a research idea, but would become impractical in real-world applications in terms of scalability, costs, and quality [48, 103].

1.2 Problem Statement and Contributions

There is enormous potential of improving the current state of e-tourism applications using data that can be gathered and processed in an automated way. The title of this thesis starts with "data-driven" to emphasize that there is huge potential to utilize available information from LBSNs, such as Foursquare or Twitter, open data initiatives such as Wikipedia or OpenStreetMap, and public data platforms of companies and other institutions. The main focus is on the destination recommendation problem, however, the implications of this work can be applied outside of this scope.

Figure 1.1 describes the structure of the contributions of this thesis. Traveling is a mobile activity, thus, in the first part of the thesis, we want to understand the mobility of travelers around the world to use these insights to improve travel recommender systems. On the left side, one can see that the traveler mobility analysis enables many use cases, e.g., uncovering traveler types that can aid to improve POI recommender systems, determining travel regions, and recommending the duration of stay. The other two chapters 4 and 5 on the right side of Figure 1.1 are about the characterization of destinations within content-based recommendations and how destinations can be recommended in a conversational recommender system while informing the users of the trade-offs involved in their choices, respectively.

In the following, we describe the three main topic areas of this thesis by identifying various challenges around destination recommender systems and describe how we resolve them.

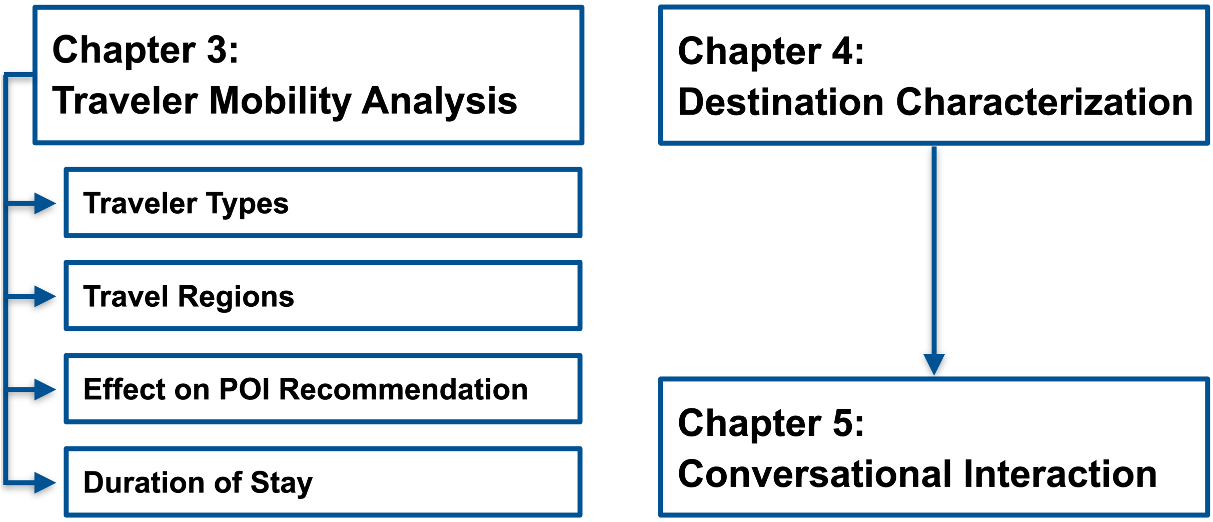

Chapter 3:
Traveler Mobility Analysis
Traveler Types
Travel Regions
Effect on POI Recommendation
Duration of Stay
Chapter 4:
Destination Characterization
Chapter 5:
Conversational Interaction

Figure 1.1: Structure of the contributions

1.2.1 Traveler Mobility Analysis Using LBSN Data

The first part of this thesis deals with analyzing global mobility of travelers from LBSN data. LBSN data is so useful for understanding the mobility of travelers around the world since traditional data sources are usually very limited in scope: for example, call data records from mobile phones are limited to one country making international travel impossible to track while the GPS location information from a single app typically has way fewer users compared to LBSN data [8]. The drawback is that the location of a user is only known when she actively checks in to venues or makes a geotagged posts. We identify the following challenges of using LBSN data for traveler mobility analysis and applying the obtained information for improving recommender systems in the travel and tourism domain.

Challenge 1: How can LBSN data be used to determine the behavior of domestic and international travelers? To address this challenge, we develop software solutions to analyze global traveler mobility based on LBSN data, which allows us to collect international trips and to assess the quality of the obtained data depending on the use case. We develop the `tripmining` library[2], which combines the geotagged check-in stream of users into trips, which are then quantified regarding their underlying mobility as well as the reliability of incomplete check-in data [34]. This helps us to evaluate which types of LBSNs are suited for understanding global travel and the respective use cases [38].

Challenge 2: Given the behavior of travelers; which groups can be identified? The main result of this global traveler mobility analysis is to identify different types of trips solely based on the mobility metrics. This results in a data-driven characterization of users that can be employed to make more personalized recommendation in cold-start recommendation scenarios, i.e., as a model of the long-term preferences when short term preferences need to be elicited [159, 151, 141].

Challenge 3: What is the effect of establishing different groups of users on POI recommendation performance? To analyze the impact of improved personalization through the identification of different types of users, we quantify the effect of the traveler type identification on the performance of POI recommender algorithms [126].

[2] `https://github.com/LinusDietz/tripmining`

Challenge 4: How does the map of global travel regions look like? In many regions, such as the Schengen Region in Europe travel is not impeded by administrative boundaries. This also means there is a mismatch between administrative regions and the users' understanding of travel regions. For example, if a traveler wants to do a hiking vacation in the European Alps, the national borders will be mostly irrelevant. To understand this phenomenon, we propose an alternative hierarchical map of travel regions based on traveler mobility instead of administrative regions [127]. This shows which national borders are relevant for defining a travel destination.

Challenge 5: What are the recommended durations of stay at one specific destination? Another application of the traveler mobility analysis is that it can be used to recommend the duration of stay at specific destinations [41]. For this, we propose a statistical method which uses the distribution of the duration of stays in a target city as well as the past travel behavior of the current user.

1.2.2 Destination Characterization

Given the absence of reliable rating data, destination recommendation is typically done using the content-based recommendation paradigm [19]. However, the decision of which features to include for the characterization of items in content-based recommendation can be challenging if the features represent latent concepts leading to ad-hoc decisions [155]. This topic is worth investigating and we identified four major challenges that need to be addressed to make a positive impact on content-based destination recommender systems.

Challenge 6: What data sources are suitable to characterize destinations? To answer this, we construct 18 data models to characterize 140 cities using online data sources of three categories: textual data, factual data, and data models based on the distributions of venues within a destination.

Challenge 7: How can a latent concept such as the touristic experience of a destination be elicited? To obtain a ground truth for evaluation, we create a web system and invite experts on travel and tourism to give their opinion on this concept.

Challenge 8: How can data models be evaluated against the expert opinion? The outcome of the expert study was a list of most similar cities to a base city. To enable a direct comparison between the data models and the target concept, we present variants of established rank agreement metrics to measure the distance between a complete permutation and a top-k list. The results show that data models used in academic recommender systems indeed fall behind textual online sources in terms of how well they emulate the experts opinion on what manifests the touristic experience.

Challenge 9: To what extent can data models describing destinations be optimized to better capture the touristic experience? For data model which had explicit features, we are able to show that by learning the weights, a better alignment of the recommendation function to the desired concept can be achieved [39].

1.2.3 Navigation by Revealing Trade-offs

When deciding where to travel, the decision making is influenced by various factors, such as not knowing what possible destinations exist, having a false image of certain destinations, and having a relatively high financial and emotional investment [86, 150]. Furthermore, traveling is about exploration, so the value of the recommender system lies in recommending fitting destinations a user is not already familiar with. At the same time, users often struggle to express their preferences directly [168], which favors conversational recommendation approaches over one-shot recommendations in this domain [71]. Finally, users can fall victim to the "wishful-thinking" problem, i.e., expecting very high quality items at a very low price tag, which do not exist in reality.

To directly address these challenges, we developed the "Navigation by Revealing Trade-offs" paradigm, which allows users to interact with concrete items, and refine their preferences in a conversational interaction. The exploration of the search space is supported by a visualization of the trade-offs involved of choosing one destination over another, i.e., how much more expensive a city will be if one wants to get a better cultural experience.

Challenge 10: How can a user interface of a conversational recommender system be designed to enable exploration and visualize the trade-offs involved in choosing one destination over another? We present CityRec, a conversational destination recommender system which features a user interface that operates with example items and visualizes the change in feature values when selecting items.

Challenge 11: How can the exploration of the search space be directed so that the users are able understand the extent of the search space, with a gradual convergence towards the user's preferences? CityRec uses a utility function to determine which candidate items should be presented to the user. With time, the utility function learns which features are important to the user and gradually converges towards the target values.

This paradigm, "Navigation by Revealing Trade-offs," was evaluated in a large-scale user study with 600 participants on a 140 destinations data set. The destinations from all over the world have been characterized along six dimensions relevant to travelers [38].

1.3 Structure of this Thesis

In the upcoming Chapter 2, we give an overview of the scientific literature in the relevant areas. Chapter 3 describes the methodology of how to derive trips all around the world from LBSN data. Furthermore, it showcases various applications of traveler mobility analysis to improve travel recommender systems. In Chapter 4, we present the study on destination characterization for content-based recommender systems. Chapter 5 describes the CityRec system which uses the "Navigation by Revealing Trade-offs" paradigm for guiding the user towards fitting recommendations through the search space. Finally, we draw our conclusions and point out future research directions in Chapter 6.

2 Fundamentals

Destination recommendation is a comparatively under-researched discipline within modern recommender systems research. This is mainly due to the lack of commercial application of recommending destinations within the current e-tourism ecosystem: while there are various successful platforms whose core business is to recommend hotels or restaurants in exchange for a commission, travel destinations are not business entities. Even though local tourism boards aim to attract travelers to their regions, there are no significant online commercial destination recommendation platforms where users can receive destination recommendations. Instead, prospective travelers need to inform themselves to come to a decision or use the services of a travel agency. An essential difference to product recommendation is that a travel destination has several aspects which altogether manifest the experience of visiting it. Thus, no meaningful ratings are available, making traditional recommendation algorithms less suited [19, 150]. This “fuzziness” of destinations as items to be recommended motivated this research to improve various aspects of this challenging recommendation domain.

Starting with an overview of travel recommender systems, we highlight the fundamental differences in recommending products or media. One important difference to traditional recommender system domains is that destinations do not come with reliable ratings; thus, all rating-based algorithms, such as collaborative filtering, cannot be employed [47]. Instead, recommendations are typically driven using the items’ features using content-based filtering. Furthermore, there is a significant cold-start challenge in travel recommender systems, which can be overcome with defining tourist roles or personas a user can identify with in the preference elicitation phase. Finally, conversational recommender systems are an established method to aid users in expressing and refining their preferences. This is especially useful for this domain since it is known that users struggle to express their real preferences when deciding where to travel to [168].

A significant part of this work is based on the analysis of traveler mobility. Using LBSN data, we mine trips of users to understand their travel behavior and propose methods to automatically group users based on their mobility. Working with the limitations of check-in-based mobility data requires close monitoring of the data quality.

Finally, we systematically tackle the open question of how travel destinations can be characterized effectively and efficiently.

2.1 Travel Recommender Systems

Individual tourism is a challenging domain for recommender systems due to the substantial complexities of planning an independent trip and the enormous economic importance of the travel and tourism industry [22]. Back in 2014, Borràs, Moreno, and Valls identify four different tasks concerning tourism recommender systems [12]: recommending travel destinations or travel packages [81], suggesting attractions [88, 125], planning trips [49, 40], and accounting for social aspects [56]. Since then, however, the most active areas have shifted towards hotel recommendations [50, 9], trip recommendations [61, 63, 121], and POI recommendation [125, 123, 87].

While the recommendation of *defined* items, i.e., hotels, POIs, and restaurants mainly relied on interaction data, such as ratings or visits [125], destination recommender systems needed to fall back to the content-based paradigm [85]. Werthner and Ricci identify the following challenges in recommending destinations: the intangibility of the recommended item, the high consumption costs, and the high emotional involvement in the decision making [150]. This makes travel recommender systems less suited toward standard off-the-shelf solutions and justifies more sophisticated systems and interaction paradigms to support the decision-making of the individual users.

2.1.1 Content-Based Recommendation

The cold start problem in recommender systems is defined as the phase when a new item or user is introduced to the system; thus, no interaction data is available to compute personalized recommendations. A purely collaborative filtering system cannot operate in the absence of ratings. Hence, the content-based paradigm is a commonly used fall-back mechanism to perform recommendations in cold-start situations [14, 17]. To facilitate the content-based paradigm, users and items need to be embedded into the same feature space that allows for calculating a similarity metric. Finding features to describe items is key, as they solely determine the recommendation outcome [85], however, where this is possible, content-based recommendations have been shown to overcome the item cold-start problem using hybridization [17, 20, 23]. In some scenarios, it is unattainable to get sufficient interaction data, which means that all recommendations are based on content-based algorithms [135, 109].

Nevertheless, content-based recommendations only work if the users' preferences can be embedded into the same feature space as the items. Fortunately, this can be done by, e.g., characterizing the users based on their past activities [8, 144, 40], or other more sophisticated preference elicitation approaches [14] including preference-based navigation [113] or conversational recommender systems [71, 25].

When selecting features, it is not trivial to ensure that the selection is ideal for the task at hand. Yao and Harper identify the fundamental problem that a data model naturally diverges a bit from the domain, and it is hard to capture a ground truth for item similarity [155]: what are the movies most similar to "Fight Club"? Which cities are most similar to Munich? We as humans might have an intuition about such similarity concepts, but it is asses how well the recommendation algorithms emulate such, possibly latent, concepts. Evaluating this is an under-researched challenge in content-based recommendation, especially in the travel and tourism domain, where items such as destinations or travel packs are often not as clearly defined as consumer products [102].

2.1.2 Destination Characterization

As previously mentioned, the items in content-based recommender systems need to be characterized along various features. In literature, this has been done using ad-hoc methods, expert knowledge, or based on existing data sets.

For example, Herzog and Wörndl developed a region recommender for personalized inter-continental travel [62]. The destinations were characterized along various travel interests, such as *nature & wildlife*, *beaches*, or *winter sports*. The underlying data for determining the suitability of the regions for these activities came from several online and offline information sources, which must be incorporated and updated manually using the expert knowledge of the authors.

When it comes to cities as destinations, there are various approaches that characterize different districts for intangible concepts. Quercia et al. used LBSN data and Google Street View imagery to determine intangible concepts such as the smell, the soundscape [3], and general happiness [111] on a street granularity [112]. Analogously, street imagery can also be reliably used to measure distributions of income, education, unemployment, housing, living environment, health, and crime, as Suel et al. have demonstrated [140]. Analyzing LBSN data to characterize cities and their districts has been an active topic in previous years [134]. It has been shown that such data is quite helpful to unveil characteristics of certain districts within a city [78], as McKenzie and Adams have done using kernel density estimation models of check-ins to identify thematic areas within a city [93].

In destination recommender systems, data models have been constructed using various data sets derived through aggregating multiple information sources, including proprietary tourism data sets. Sertkan, Neidhardt, and Werthner characterized a vast data set of 16,950 destinations based on 26 motivational ratings and 12 geographical attributes within the Seven Factor Model of tourism motifs [129]. They proposed a cluster analysis and regression analysis to map the destinations to the vector space of the Seven Factor Model [103]. This framework was recently also used by Grossmann et al. to elicit preferences of prospective tourists using pictures of destinations and an underlying ontology to characterize the image contents. While modeling the user's interests using travel-related pictures has been shown to be possible, obtaining a representative set of images of global destinations in an automated fashion is an open research problem [131, 105].

2.1.3 Tourist Roles

Another frequently used method to mitigate the cold-start problem is to assign users into groups, to have an initial filtering based on the assumed group characteristics. The characterization of tourists has been an active field in tourism research over the last decades. One of the first works by Cohen established four different social roles of tourists: the “organized mass tourist”, “individual mass tourist”, “explorer”, and “drifter” [28]. Pearce used fuzzy set theory to define 15 different travel roles [110], while McKercher used a cultural sciences approach to classify tourists based on the importance of cultural motives when deciding which destination to visit and the depth of cultural experience gathered by the tourist [94]. Finally, Yiannakis and Gibson took a sociological perspective to determine 17 roles that are enacted by people when they travel while also associating these with different psychological needs [156].

In light of this diversity of tourist categorizations in the literature, the best grouping of tourists and their preferences remains unclear. More importantly, none of the existing categorizations have been validated with observational data [102], so it is unclear to what extent the categories apply to real travelers. To address this challenge, Neidhardt et al. developed the *Seven Factor Model* of tourist behavioral patterns [102] based on the *Big Five Factor Model* [90] from psychology and a factor analysis of the 17 tourist roles proposed by Yiannakis and Gibson [156, 51]. With a destination recommender system in mind, they elicited user preferences through an image classification task, where the users are to pick the most appealing travel-related photos from a collection. The classification of these pictures along the *Seven Factors* has been previously determined using a questionnaire. This framework of the seven travel behavioral patterns was used in various subsequent articles, both to group users [130] as well as to characterize destinations within the seven-factor model using tourism data sets [128, 129] or image data of destinations [131].

2.2 Conversational Recommender Systems

As opposed to one-shot recommendation, conversational recommender systems augment the preference elicitation by giving the user a chance to provide feedback on the currently recommended items [71]. We scope our survey toward custom user interface concepts and algorithms, thereby excluding approaches based on natural language processing, such as chat bots [108].

Critiquing is a popular approach to eliciting and refining user preferences in a conversational manner. It is usually associated with content-based filtering, although there is some research incorporating collaborative approaches [152] or even unstructured item descriptions [113]. One of the early systems, FindMe [18] introduced the concept of unit critiquing, which can be seen as the start of the conversational exploration of the search space in recommender systems research. The static unit critiquing was quite successful in several domains [18, 16], but there is opportunity to perform a smarter exploration of the item space [91]. For example, McCarthy et al. [89] proposed dynamic critiquing to show how compound critiques can be generated dynamically, cycle-by-cycle, by mining the feature patterns of the remaining products.

The evolution of dynamic compound critiques is the multi-attribute utility theory (MAUT) [161], which introduced a utility function to rank a list of multi-attribute products. Once the user selects a critique, the corresponding product is set as the current preference product in the user model, and a new set of critiques is generated using a utility function. The MAUT was successfully evaluated against dynamic critiquing [89], thereby reducing the number of critiquing cycles. Chen and Pu extended the MAUT-based approach and called it “preference-based organization interfaces” [24]. In their approach, the authors organized all potential critiques in a trade-off vector showing whether the features were compromised or improved compared to the current recommendation. That enabled them to determine useful compound critiques and successfully evaluate their systems using a computer configuration data set. However, we feel that such an approach is more suited for products with clear specifications since, in tourism, relative differences between the feature values of items are of higher importance.

One major issue with critiquing is the divergence of the intended direction of exploration. McGinty and Smythstudied selection strategies for recommending items in critiquing [92]. Their Adaptive Selection approach resulted in a reduction in critiquing cycles, and they could prove that their critiquing-based approaches would converge faster than preference-based approaches. Another important insight of their work was that the user should not lose progress, i.e., the previous recommendation should be included in the upcoming cycle.

2.3 Traveler Mobility Analysis

The analysis of human mobility gives insights into various aspects of everyday life. Before the advent of online social networks on GPS-enabled devices, data sources like mobile phone communication records [54], Wi-Fi usage [162], and raw GPS trajectories [167] were used to analyze individual human mobility. Given today's availability of LBSN data that enriches a pure location trace with further information, such as user-posted content and the user's social network, much research has been done analyzing mobility using data from Twitter, Foursquare, and other platforms [64].

2.3.1 Predictability

A prominent research objective is the predictability of human mobility based on past behavior. Song et al. found that individual mobility patterns followed reproducible scaling laws [136] and described the limits of the extent to which human mobility can be predicted [137]. More recently, Ouyang et al. have analyzed mobility data to predict travel trajectories using a deep learning framework [107]. Similar approaches to predicting the next visited place exist for tourists as well [163]. The correlation of locations with a social activity that can be studied with LBSN data promises interesting insights into social behavior. Cheng et al. found recurring daily and weekly patterns of activity [27] and Wang et al. found a positive pairwise correlation between social connectedness [149], i.e., the strength of interactions and mobility. Noulas et al. analyze activity patterns of Foursquare users, such as the spatial and temporal distances between two check-ins [106]. They discover place transitions that could well be used to predict or recommend the future locations of users.

2.3.2 Application in Recommender Systems

Contrary to the approaches in the previous section, destination recommender systems should generally not predict the user's mobility but provide the user with recommendations that she will like but would not have discovered otherwise. Given the richness of LBSN mobility data, various approaches used such data to improve recommender systems [8]. Zheng and Xie studied spatial co-occurrences that can also be used to identify similar users and generate implicit ratings for collaborative filtering algorithms [166]. Bao, Zheng, and Mokbel matched the travelers in a foreign city to local experts based on their respective home behaviors to improve the accuracy of a point of interest recommender [7]. LBSN data has also been used to capture cross-border movement [10]. The authors demonstrate how the movement dynamics of people in a country can be analyzed, however, this study is not about tourists and is limited to one country, Kenya. Hsieh, Li, and Lin used past LBSN data to recommend traveling paths [65], while Zheng et al. proposed heuristics to approximate the similarity of tourist trips [164]. For this, they present solutions to derive the popularity, the proper time of day to visit, the transit time between venues, and the best order to visit the places.

2.4 Rank Agreement Metrics

To compare data models of destination recommender systems, we rely on rank agreement metrics of ordered lists. In literature, one can find various methods to compute the agreement of two ranked lists. They are also known as rank "correlation" methods and essentially capture a notion of similarity between the ordering of items within two lists. For complete permutation groups, i.e., both lists have the same items and the same length, there are several established metrics, such as Kendall's Tau Distance [75], Spearman's Footrule Distance [139], and Spearman's ρ [138]. Based on these measures, countless other methods have been proposed to cater to the needs of more specialized domains and different assumptions.

However, the problem of assessing two incomplete lists, also called top-k lists, makes the matter more complicated, and the choice depends on the application area. Critchlow was first to establish a theoretical basis for such rank agreement metrics [30], assuming a fixed domain of items D. One of the most comprehensive papers on the rank agreement of top-k lists is the one of Fagin et al. [46]. Unlike Critchlow, they did not assume a fixed domain of items and, thus, proposed very general distance measures for top-k lists that are not directly useful to our scenario. The authors also proved that, in the general case, the measures for top-k lists reside in the same equivalence class and showcased further applications of these measures in the context of the rank aggregation problem [43, 80].

An important property of Kendall's Tau and Spearman's Footrule is that all ranks are treated equally, i.e., they do not take the potentially non-uniform relevancy of top-ranked or bottom-ranked into account. In many domains, the assumption of uniform relevancy does not hold, thus, several other measures have been proposed. Iman and Conover proposed a concordance measure that prioritizes rank agreements at the top of the rankings [67], while Shieh proposed a weighted variant of Kendall's Tau, where the analyst can prioritize either low-ranked or high-ranked items [133]. The Average Precision Correlation is another important measure in information retrieval that more heavily penalizes differences of top-ranked items compared to Kendall's Tau [157]. Deciding which method to use can be challenging, given the often subtle differences within these metrics. The decision should be based on either analytical insights from the domain or on the metrics' performance in modeling business success indicators.

2.5 Summary

In this chapter, we described the domain of travel recommender systems and the scientific foundations for the methods and contributions of this thesis. Given how decision-making is done in touristic travel, we concluded the conversational recommendation paradigm to be well-suited for prospective travelers to contrast possible alternatives of where to travel. Since the medium in the conversational interaction is characterized items and features, i.e., users giving the system feedback that a certain aspect of the current recommendation should be refined in a certain way, it is important that the data models closely reflect reality. Unfortunately, it is hard to elicit a ground truth for latent concepts, such as how the touristic experience will be at a destination, which makes the quality of competing data models and features hard to judge. In a purely content-based recommender system, the recommendations are based on a distance metric and some features describing all items. Thus, one can generate ranked lists based on the distance metric over the item features and assess the overall similarity of the data models using rank agreement metrics on the resulting ranked lists.

Since traveling leaves mobility traces, mobility analysis of traveler behavior can provide relevant insights into the domain and users' preferences. Identifying tourist roles has been of long-standing interest within the research community but has not been done using the global mobility of travelers. Furthermore, capturing the behavior of travelers can be useful to shape the outcome of recommendations, e.g., which destinations are frequently traveled together and how long one should stay at a specific destination.

3 Traveler Mobility Analysis Using Location-based Social Networks

Analyzing the mobility of travelers reveals valuable information about their behavior, preferences, and visited destinations. We argue that it is important to capture how people actually travel the world and use this information to shape the outcome of the recommendations. The destination recommendation domain is in dire need of analyses capturing real user behavior since real-world empirical user studies of such recommender systems are infeasible due to the high costs of traveling. Furthermore, traveler mobility analysis can reveal information about the popularity of destinations, give valuable insights to destination marketers, and can be a valuable tool to characterize users based on their past trips automatically.

Traveler mobility can be observed in different ways; however, nowadays, LBSN data has been the most accessible and useful resource for global mobility research. Examples for LBSNs are Gowalla, Brightkite, and Foursquare, where users check in into venues or traditional online social networks such as Twitter, Facebook, or Instagram, if users enrich their posts with a geolocation [118]. Additionally, there are other platforms, such as Flickr, where a location trace can be constructed using geographic metadata. The advantage of LBSN data over analyzing the number of accommodation bookings in a city, tracking ticket sales of flights or trains, is that we obtain individual mobility traces instead of aggregate travel patterns. However, access to such data has been increasingly restricted on many platforms due to business [68] and privacy reasons [98].

The basic idea of our approach is to chronologically sort all of a user's geotagged content into a stream of check-ins and to segment it into periods of being at home and of travel. Consecutive check-ins outside the user's home are combined into a trip that will be characterized using different metrics regarding the quality of the data and the underlying mobility. We describe the algorithms and the design decisions of the `tripmining` library in Section 3.1.

The collected trips capture much information about individual travelers, revealing common behavior and preferences over whole populations. In Section 3.2, we present two studies to reveal which groups of trips can be discerned in a cluster analysis. The outcome of the cluster analysis is shaped by the input features, which makes it interesting to compare the clusters of a purely mobility-based approach using Twitter data with a data set from Foursquare that also includes additional information about the activities of the users.

Another application of this cluster analysis method of traveler behavior is to compare the recommendation outcome of different groups of locals and travelers in an offline evaluation study in the POI recommendation domain. Section 3.3 summarizes this approach, which combines the cluster analysis of traveler types [38] with a beyond-accuracy analysis of the performance of different groups in POI recommendation [124].

Aggregating the mined trips and transforming them into a mobility graph establishes a unique view of global traveler mobility. We use this global mobility graph to propose an alternative, hierarchical travel region map of the world without any information on existing administrative and national boundaries. By employing graph community detection algorithms, we detect which cities are frequently traveled together and, thus, form coherent travel regions in Section 3.4.

Finally, we can use the determined stays to compute personalized recommendations regarding the durations of stays. This approach, described in Section 3.5, can be useful for personalized travel planning applications.

3.1 Mining Trips from Location-Based Social Networks

As opposed to collecting trajectories from GPS trajectories with continuous recordings, deriving trips from the users' geotagged posts on a LBSN requires careful consideration due to the relatively low frequency of the data. Despite this incomplete view of a user's mobility, LBSN users leave spatio-temporal traces, which can be characterized whether the quality of the trace is sufficient for the mobility analysis to be done. For this task, we developed an algorithm to derive trips from users' timelines on LBSNs and metrics to quantify the quality of the trips. The project was open-sourced on Github as a Python module under the name `tripmining`[1].

3.1.1 Tripmining Algorithm

Throughout all analyses, the locations of the users are geocoded to a city or municipality. The chronologically sorted list of the check-ins of each user is segmented into periods of being at home and periods of travel. For this, the home location needs to be determined reliably, for which several strategies exist [74]. We use the plurality strategy, i.e., choosing the city with the highest number of check-ins. Kariryaa et al. show that this simple heuristic has a very high accuracy which is on par with more sophisticated and computationally intensive methods such as the geometric median [74]. In addition, we propose to use a threshold for the number of stays at the most frequent city to prevent false classifications for users who predominately use social media when traveling. In the forthcoming analyses, we discard all travelers whose check-ins at home are fewer than 50%.

Having determined the home city, the actual trip mining starts: check-ins at home are discarded, while contiguous check-ins outside the hometown are combined into trips. Since we are interested in the underlying mobility of the user, we further aggregate consecutive check-ins at the same location into blocks. Thus, the result is a collection of trips, each having at least one block. This approach is so far straightforward, but it is worthwhile to mention some aspects: first, the segmentation of the trips and blocks is triggered based on a change in the location of the user, whether it is a newly visited city or returning home. Second, the algorithm does not aim to infer any information that is not backed by evidence: for example, if there is a change of location, the transition time between the two check-ins is not counted towards the adjacent blocks. Finally, we characterize all trips regarding the reliability of the check-in stream. Depending on the use case, the analyst can decide to drop trips whose quality is insufficient for the task at hand.

To visualize the approach, Figure 3.1 exemplifies a check-in stream of a user from Munich. The user's check-in stream starts on day 0 (d_0) in Munich and is followed by a

[1] `https://github.com/LinusDietz/tripmining`

$$\overbrace{\text{Munich } d_0}^{Home} \rightarrow_2 \underbrace{\overbrace{d_3 \text{ Paris } d_{11}}^{\text{Block}}}_{\text{Trip 1}} \rightarrow_0 \overbrace{d_{12} \text{ Munich } d_{100}}^{Home} \rightarrow_0$$

$$\rightarrow_0 \underbrace{\overbrace{d_{101} \text{ Paris } d_{101}}^{\text{Block}} \rightarrow_0 \overbrace{d_{101} \text{ New York City } d_{105}}^{\text{Block}} \rightarrow_3 \overbrace{d_{109} \text{ Washington, D.C. } d_{118}}^{\text{Block}}}_{\text{Trip 2}} \rightarrow_1$$

$$\rightarrow_1 \overbrace{d_{120} \text{ Munich}}^{Home} \rightarrow \dots$$

Figure 3.1: Example of a user's check-in stream with two trips.

block of nine days (d_3–d_{11}) in Paris. In this case, the block is terminated by a check-in in Munich on the next day (d_{12}). Since Munich is the user's home location, the first trip is considered completed with only one block.

Staying home for 83 days, the user is then observed checking-in in Paris on d_{101} and a few hours later in New York City. Since she was located in Munich the day before, it seems quite probable that she traveled from Munich to New York City with a stopover in Paris. The check-in stream shows several check-ins in New York City until d_{105} and continues with check-ins in Washington, D.C. from d_{109}–d_{118} before the trip is again terminated by the return to Munich on d_{120}. Thus, this trip lasts 18 days and consists of three blocks.

3.1.2 Quality Assurance Metrics

Using the algorithm above to concatenate check-ins into trips, one would potentially get many trips comprising only a single check-in. For this reason, we characterize trips using quality metrics, which act as filters to ensure that the determined trips are reliable enough for the task at hand. For example, one could only analyze trips with a minimum duration of seven days to filter out typical business trips.

Furthermore, to approximate the actual mobility of the user, a relatively steady check-in behavior during travel is required. For this, we propose and discuss various metrics to capture the quality of a user's check-in stream.

3.1.2.1 Check-in Frequency

The check-in frequency shown in Equation 3.1 is not robust against a multitude of check-ins on one day, which makes it unsuitable for assessing the reliability of the check-in stream.

$$\text{check-in frequency} = \frac{\text{check-ins}}{\text{days}} \tag{3.1}$$

3.1.2.2 Check-in Density

In this regard, the better measure is the check-in density (Equation 3.2), as it captures the fraction of days with a check-in during a trip. Thus, it captures how steady the check-in stream is, which is more important than having several check-ins at the same location in one day. The minimum value of check-in density should be chosen depending on the use case.

$$\text{check-in density} = \frac{\text{days with check-in}}{\text{days}} \tag{3.2}$$

3.1.2.3 Maximum Transition Speed

The transition speed between two locations is the covered distance divided by the transition time. It can be useful to detect irregularities in the check-in stream, such as multiple people sharing one account, which can result in simultaneous check-ins from distant locations. We typically use it as a quality metric, i.e., discarding all users with a transition speed above 1100km/h [145], but it could also be used to select certain means of travel, e.g., airfares.

$$\text{maximum transition speed} = \text{argmax}_{t \in \text{Transitions}}(\frac{t_{\text{distance}}}{t_{\text{time}}}) \tag{3.3}$$

3.1.3 Mobility Metrics

While the metrics from the previous section were about the quality of the users' check-in stream, the following metrics capture the underlying mobility of the users.

3.1.3.1 Trip Duration

The trip duration is the number of calendar days between the first and the last check-in of the trip.

3.1.3.2 Number of Locations, Blocks, and Countries

We also analyzed the number of distinct locations, blocks, and countries within a trip. The number of locations is naturally lower than the number of blocks since one location can be visited in several noncontiguous blocks of a trip. Furthermore, we typically discriminate between international and domestic trips.

3.1.3.3 Radius of Gyration

The Radius of Gyration is a measure of how far the users traveled within a trip [54]. In simple terms, the radius of gyration measures the mean distance between the center location of the trip to all other check-ins. Thus, it is more robust against skewed check-in distributions than the distance between consecutive check-ins.

3.1.3.4 Displacement

Displacement measures the distance between the user's home location and the mean position of the places visited during the trips. It is conceptually similar to the Radius of Gyration; the only difference is the reference location from which the distance is calculated.

3.1.3.5 Venue Information

In some cases, there is additional information available for the visited venues. Concretely, when working with Foursquare data, it is known which type of establishment the user has checked into, e.g., "Outdoors," "Nightlife," or "Arts & Entertainment." This can reveal further information about the activities a user has done during the trip.

3.1.4 Summary

The proposed approach makes it possible to mine trips from the check-in stream of LBSN users. The logic of constructing trips from user timelines is relatively straightforward, thus, the emphasis of the contribution lies within the derived metrics for the quality of the trips and the mobility of the travelers. The tripmining library provides analysts high flexibility to select the suitable thresholds given the use case at hand. It is implemented as a Python module, published under the permissive MIT License on Github[2]. It provides functionality for parsing various data sets and can be extended for parsers of other data sets in about 30–50 lines of Python code.

Analyzing various LBSNs data sets, we found that they are not equally suited for all analyses, which comes from how this data was collected. For example, trips derived from Foursquare or geotagged Tweets are often of sufficient quality to understand the mobility between different destinations; however, the data from the check-in stream is typically too infrequent to reliably capture the mobility within a city. Data from the image sharing platform Flickr is even less suited for most analyses since we observe that few people post geotagged pictures of their home, making the home detection infeasible on a city level, and the number of posted pictures is way sparser compared to LBSN posts.

The mined trips can serve as starting points for various improvements to recommender systems. First of all, they reveal many analytic insights into global travel behavior, such as the relative popularity of cities throughout the year. This can be used to increase the diversity of recommendations and, thus, avoid peak season visits for travelers sensitive to mass tourism. Furthermore, it reveals patterns of destinations often visited together, which can serve as input for approaches to resolving the tourist trip design problem [61, 121]. In the following, we present various applications of the trip mining methodology on LBSN data.

3.2 Cluster Analysis to Discover Trip Types

The first application of the mined trips for touristic information systems is a cluster analysis. By revealing different kinds of tourist trips, we can offer insights into the general characteristics of different types of travelers using unsupervised machine learning. While this is an analytic result in itself, it can be directly used as part of user modeling within travel recommender systems. To reveal which types of trips are undertaken in which quantity, we run the tripmining algorithm on two data sources: A self-collected Twitter data set and a publicly available Foursquare data set [154]. In the former, we have a pure mobility trajectory with a check-in granularity of cities. In contrast, in the latter, the check-ins are attributed to specific venues with further information about the categorization of the venue.

[2] https://github.com/LinusDietz/tripmining

3.2.1 Clustering Method

In both analyses, we run the users' check-in streams through the `tripmining` library, discard users with an uncertain home, and also drop all trips that do not fulfill the quality criterion of a check-in density of at least 0.2. To eliminate short weekend trips and typical business travel, the analysis is only based on trips that lasted at least 7 calendar days. Since the goal is to capture the underlying phenomena of the users' travel behavior, we only use the metrics from Section 3.1.3, which capture the users' mobility instead of the data quality.

Using all trips, we perform a correlation analysis to remove redundant features with a Pearson Correlation Coefficient of > 0.75. The reason for this exclusion is that highly correlated features will not improve the segregation in the clustering algorithm. Instead, it biases the results towards one phenomenon. Since we use the K-means clustering algorithms with the Euclidean distance, we further normalize all features using min-max normalization to avoid bias toward metrics with generally high values, such as the displacement from home.

Finally, we run the clustering algorithm with a different number of clusters and systematically assess the quality of the determined clusters depending on the number of clusters k in terms of the average silhouette width [120]. The silhouette width measures how well a data object fits into its labeled cluster as opposed to all other clusters. Therefore, it is a robust and easy-to-interpret method that gives a broad overview of the overall solution quality, as well as information about each data object.

3.2.2 Case Study on Twitter Trips

The first trip clustering study was on trips from Twitter. We were able to collect almost $100,000$ trips; however, due to memory limitations, we drew a random sample of $40,000$ trips to run the clustering using K-means clustering. Since no venue information was available, this case study is about the pure mobility of travelers. Applying the aforementioned correlation analysis method, we retained five mobility metrics: the duration of the trip, the number of locations, the number of blocks, the radius of gyration, and the displacement from home.

Analyzing the clustering results from $k = 2$ to $k = 7$ clusters, there is always one dominant cluster of domestic trips and several smaller, more specialized ones. According to the silhouette width, a solution with two or three clusters would be acceptable. Analyzing the resulting clusters, we chose the clustering result of $k = 3$ as our final segmentation tabulated in Table 3.1.

Since we do not discriminate between international and domestic trips, unsurprisingly, most trips (87.1%) are in the *"Domestic"* cluster, with a low mean displacement of 565.57 km and only 1.22 countries on average. This imbalanced result is potentially an outcome of most Twitter users residing in the USA. The two other clusters are smaller in number and more specialized. The *"Globetrotters"* travel further, visit the most countries, and display the highest radius of gyration. With 35 days duration, these trips are also the longest. Finally, the *"Distant Vacationers"* travel furthest from home but are not as active during their travel. Their radius of gyration is only one-third of the Globetrotters, despite visiting nearly as many distinct locations.

Table 3.1: Twitter: resulting clusters. Mean value/standard deviation. *The number of countries was not an input feature of the clustering algorithm.

	Domestic	Globetrotters	Distant Vacationers
Relative Size	87.1%	6.4%	6.5%
Silhouette	0.83	0.25	0.38
Duration	19.65/53.45	34.99/96.03	25.05/59.18
Locations	5.13/6.6	9.73/12.35	8.18/7.92
Blocks	1.46/1.83	4.89/6.93	3.34/4.07
Countries*	1.22/0.6	3.04/2.01	2.39/1.48
Radius of Gyration	329.2/541.36	5,172.25/2,435.53	1,677.74/1,428.58
Displacement	565.57/887.83	5,262.2/2,323.11	8,733.89/2,834.18

3.2.3 Case Study on Foursquare Trips

Unlike the Twitter study, the data set from Foursquare [154] enabled us to analyze what clusters are formed when taking social aspects of the travelers into account. Again, we use the mobility features of the trips but enrich them with the type of venues the travelers checked into using the Foursquare venue categories. This results in the following features: trip duration, countries visited, the displacement, the radius of gyration, and the number of respective check-ins in the categories Food, Nightlife, Arts & Entertainment, and Outdoors & Recreation. No features had to be removed after the correlation analysis.

Table 3.2: Foursquare: resulting clusters. Mean value/standard deviation

	Party	City	Foreign	World	Domestic Long	Domestic Short
Relative Size	1.6%	14.7%	3.2%	1.2%	3.2%	76.2%
Silhouette	0.08	0.28	0.01	0.16	0.23	0.65
Duration	16.01/11.85	11.46/5.06	12.87/6.72	17.72/14.89	40.76/22.41	10.3/3.87
Locations	4.36/2.81	3.7/1.8	4.89/2.49	6.48/3.88	6.07/3.52	2.78/1.47
Blocks	3.24/2.8	2.53/1.8	4.14/2.26	5.71/3.67	5.5/6.06	2.07/1.45
Countries	1.06/0.24	1.02/0.14	1.85/0.54	2.39/1.05	1.03/0.17	1.01/0.1
Radius of Gyration	107.65 /289.05	47.01 /138.24	1,304.51 /963.4	3,987.41 /2,725.93	75.12 /216.55	27.85 /100.1
Displacement	192.69 /596.07	65.08 /204.52	1,692.77 /1,184.4	7,224.18 /2,678.58	95.05 /232.25	39.5 /146.52
Food	1.92/6.64	0.97/1.6	1.55/2.1	1.94/2.34	3.22/5.07	0.94/1.24
Arts	0.48/0.94	0.33/0.67	0.35/0.79	0.8/1.31	0.9/2.15	0.28/0.66
Outdoors	0.72/1.99	0.45/0.94	0.54/0.96	0.87/1.47	2.28/3.54	0.46/0.92
Nightlife	3.67/1.13	1.23/0.42	0.21/0.49	0.49/1	0.26/0.56	0/0

With this data set, the results were more nuanced, and using the silhouette width, we determined 6 as the optimal number of clusters. Analyzing the results summarized in Table 3.2 more closely, again, a dominant cluster of *“Short Domestic”* trips arises with 76% of all trips residing in this group. These trips are, on average, the shortest, have the smallest radius of gyration, and are almost exclusively in the traveler’s home country since the displacement is, on average, as small as 40 km. The other clusters are low in number and highly specialized:

The *“Party”* trips are about two-week long trips that visit around four cities in one country. They are distinguished by their high number of food check-ins and very high number of nightlife check-ins.

The *"City"* trips are similar to the domestic short trips, however, they span more cities, and these destinations are more distant from their home town.

The *"Foreign"* trips are about two-week long trips to several cities located about 1,700 km away from home. People travel quite extensively, as the radius of gyration of about 1,300 km indicates.

The *"World"* trips are similar to the *"Globetrotters"* from Twitter. They visit the most locations, travel the farthest, and have the highest radius of gyration with nearly 4,000 km.

Finally, there is the cluster of the *"Long Domestic"* trips that last about six weeks, roughly corresponding to the summer holiday duration in many countries. The small radius of gyration and the high number of outdoors and food check-ins indicate that these trips might be monothematic vacations, e.g., at a beach resort during summer holidays.

3.2.4 Summary

This section described a method for revealing tourist types using LBSN mobility data. We presented two case studies of international and domestic trips stemming from Twitter and Foursquare. On Twitter, only three clusters emerged, whereas, on Foursquare, most trips resided in two clusters, with four more very specialized ones.

Besides the analytic value of the analysis, the results can be useful for user modeling within various tourism recommender systems. Without any user interaction required, a system can automatically derive the preferences of a user from data about their past trips. This can be achieved through an app permission by which the user grants access to their timeline on an LBSN that they have been using, e.g., through a third-party Facebook or Twitter application. Also, large online travel platforms such as Booking.com, Tripadvisor, and AirBnB could leverage a similar approach with additional features to classify their user and, thereby, establish a foundation for providing personalized recommendations, which can then be further refined using explicit preference elicitation techniques.

It must be noted that the cluster analysis results strongly depend on the input data. Developers of recommender systems should carefully evaluate only to include features useful for the preference elicitation and the recommendation outcome. Otherwise, the approach is at risk of overfitting the data and results in outlier groups, as can already be observed in the Foursquare case study.

3.3 Impact of Traveler Type Analysis on Point-of-Interest Recommendation

A related domain to destination recommendation is point of interest (POI) recommendation. This is an interesting challenge, as travelers are often in need of such recommendations when they arrive at a specific city or region [12, 143]. Similar to the mobility analyses of the previous section, much of the POI recommendation data sets stem from LBSNs, such as Foursquare, Gowalla, or Yelp [8]. Despite the richness and availability of LBSN data, POI recommendation has specific aspects that differ from the conventional recommendation of movies, books, or music that affect the recommenders' performance: there are various influences, such as social dynamics, temporal variations, and sequential patterns, i.e., people visiting venues in a particular order. Most importantly, the geographic location of venues has an important influence since users tend to visit nearby locations over distant ones [79, 82]. Finally, the sparsity of the interaction data is more severe than in traditional recommendation scenarios, such as books or movies [125].

Revisiting the literature on POI recommendation, we observe that all users are treated in the same way; i.e., most POI recommender systems studies report their accuracy metrics such as Precision or the nDCG averaged over all users irrespective of their age, gender, or even the city where they reside in. In many recommendation domains, it has been shown that recommendation models exhibit various biases toward certain users [95, 45, 44]. Yet, most studies on POI recommendation do not even differentiate between users being in their home city or on travel.

Motivated by this observation, we analyze the extent to which the performance of POI recommendation algorithms differs among different groups of users. We analyze the effect of subdividing groups of travelers and locals on the recommendation outcome in five well-known cities: Istanbul, Mexico City, Tokyo, New York, and London. To discover the groups within these two categories, we characterize the users based on the behavior they exhibit using various features, thereby focusing on mobility patterns and the types of venues they visited. We obtain different user groups through a cluster analysis following the general procedure to the one proposed in Section 3.2. Then we analyze the performance of different recommendation algorithms in each of the obtained subclusters in terms of ranking accuracy, novelty, and diversity.

3.3.1 User Behavior Characterization and Cluster Analysis

In this first step, we aim to find coherent groups of users that can be discriminated based on information that is relevant to POI recommendation and can be extracted from LBSNs. When performing a cluster analysis, the features selected shape the outcome, thus, it is imperative to compute features that actually help to define the user characteristics. Again, we use the global-scale check-in data set from Foursquare [154], which, in contrast to Twitter, contains information about the venues. We aim to determine expressive features to characterize travelers and locals independently since the behavior of these groups differs significantly, depending on whether a user is at home in a city or she is on a visit. Consequently, different features capture the user behavior of the two categories of users.

Starting from the complete data set of 33M check-ins in 415 different cities, we performed a data cleaning step, which included the removal of duplicate check-ins of a user in the same venue. We enforced a 10-core for users and POIs, i.e., ensured that ultimately all users have at least ten interactions and each POI has at least ten visits. Finally, we split the data set following a temporal partition in which 80% of the most ancient interactions are sent to the training set, whereas the other 20% is used as the test set.

Using the information in the training set only, we perform two cluster analyses independently for locals and travelers. To perform this study correctly, it is essential to know a user's home because only check-ins of the home city of a user should be used to compute their behavior as a local; likewise, a user's travel behavior should solely be characterized using check-ins outside of the home city. For this, we use the home detection strategy of the `tripmining` library with a threshold of 50% of the check-ins needed to be performed in the most frequent city. This step excludes another 8,548 (6.20%) users with an unclear home from the training data, resulting in 129,294 valid users in the training set.

3.3.1.1 Local Users Behavior Cluster Analysis

To discover distinct groups of user activity in their home town, we exclusively analyzed check-ins users have performed in their home cities and computed various features, including mobility metrics, such as the radius of gyration, the mean distance from the city center, and the mean distance between consecutive check-ins. Further features describe the activity of the users, e.g., the mean time between check-ins, the activity period, the number of check-ins, and the number of unique POIs visited. Finally, we also count check-ins in relevant categories, such as visiting POIs labeled with "Arts & Entertainment," "Outdoors & Recreation," "Food," "Nightlife Spot," and "Shops & Services."

Having obtained these features, we perform a correlation analysis similar to the one in Section 3.2.1 to remove redundant and orthogonal features that would bias or deteriorate the quality of the formed clusters. This step resulted in eliminating the following metrics: mean check-ins per day, the total number of check-ins, and the number of check-ins in "Colleges & Universities." Using the K-means algorithm, we systematically analyzed the outcome of the algorithm on the min-max normalized features using the Euclidean Distance and Observing the quality of the resulting clusters using different values for k, we observed that the quality of the segmentation to be very low, despite having performed the relevant steps of the prior correlation analysis. Experimenting with different feature combinations, the silhouette width ranged in the area of 0.3 for 3–4 clusters and further dropped with a higher k. However, when dropping the mobility features (radius of gyration, mean check-in distance, and mean distance to the city center), we obtained clearly better results and finally chose the optimal configuration of a silhouette width of 0.57 for $k = 3$, which can be seen in Table 3.3.

Table 3.3: Cluster results of the 129,294 locals. In the absence of mobility features, the segmentation is mostly driven by the users' activity level. Values represent the mean/standard deviation.

	L1	L2	L3
Name	Low	Medium	High
Ratio	25.3%	28.0%	46.6%
Activity Duration	79.74/40.47	205.65/38.98	341.86/30.75
Unique POIs	14.36/ 9.89	20.63/12.68	26.03/16.66
Arts & Entertainment	1.30/2.70	2.11/3.49	3.58/5.54
Outdoors & Recreation	4.18/ 7.43	5.87/10.01	6.55/12.23
Food	6.65/ 9.03	10.45/12.40	13.67/18.48
Nightlife Spot	1.42/3.60	2.38/4.83	3.73/7.38
Shops & Service	4.43/ 6.27	6.43/ 8.01	8.74/11.61

There are three clusters, two of which respectively make up about a quarter of the users and a larger one containing the remaining 46.6% of the locals. We interpret the fact that the mobility features, such as the radius of gyration and the distance to the city center, prevented the algorithm from finding an acceptable segmentation of the locals as a clear indication that these features are unsuitable for distinguishing different resident groups in the data set at hand. This may be for several reasons: residents might be more active in their respective districts, making it hard to characterize their behavior with metrics in relation to the entire city. In addition, commuting introduces noise, which is difficult to eliminate given the volatile usage of LBSNs during leisure and work time. Finally, the mobility metrics to characterize residents of five different cities might need more careful deliberation: cultural and geographic circumstances could be too different to find universal clusters across all cities. This means that the clustering result of the locals is mainly influenced by the user activity level.

3.3.1.2 Travelers Behavior Cluster Analysis

Similar to the locals, we analyzed the behavior of the users when traveling outside of their home city. The processing was entirely performed using the `tripmining` library, which resulted in $64,316$ trips of $38,903$ travelers. We aggregated all trips of a traveler as their traveler profile and again used the same method used for the locals to select the features for the cluster analysis. This aggregation of multiple trips of one user to a traveler profile is a notable difference from the clustering approach in Section 3.2, where we analyzed the features of the trips independently. Due to a high correlation to the number of trips, we eliminated the number of stays in cities (non-distinct) and the number of "Food" check-ins. The final features lead to a clustering result of four clusters with a silhouette width of 0.68.

Table 3.4: Cluster results of the 38,903 travelers. The discovered groups shed light on the preferred type of trips the users did. Values are the mean/standard deviation.

	T1	T2	T3	T4
Name	Foreign Cities	Active Vacationers	Domestic	Globetrotters
Ratio	11.1%	5.0%	80.6%	3.2%
Ratio Domestic Trips	0.09%	55.36%	99.94%	0.46%
Displacement	1324.58/1086.56	1746.56/1806.27	503.19/ 673.21	7599.87/2715.97
Radius of Gyration	263.34/ 556.04	1783.14/2029.40	108.96/ 285.24	1968.35/2818.53
Number of Trips	1.33/1.01	2.77/1.20	1.64/1.44	1.30/0.89
Unique Cities	1.64/0.96	3.45/1.54	1.58/0.94	2.30/1.44
Arts & Entertainment	0.44/0.85	0.89/1.29	0.38/0.81	0.72/1.35
Outdoors & Recreation	0.75/1.33	1.53/2.18	0.95/1.92	0.80/2.31
Nightlife Spot	0.27/0.76	0.68/1.25	0.44/1.12	0.37/1.93
Shops & Service	0.90/1.62	1.65/2.22	0.98/1.80	0.90/1.62

The four traveler clusters tabulated in Table 3.4 show comparable groups to the studies presented in Section 3.2 despite the results in Tables 3.1 and 3.2 comprising trips. In this study, we aggregated the trip metrics per traveler before clustering. With approximately 81%, T3 (Domestic) is the largest group comprising travelers whose trips were almost exclusively domestic and quite near their home cities. T1 (Foreign Cities) are infrequent travelers with only 1.33 trips that are almost exclusively international trips, where the users were relatively stationary at their destination, which can be seen in the low radius of gyration. T4 (Globetrotters) is somewhat similar; however, this group of intercontinental travelers was more into POIs of the "Arts & Entertainment" category than Foreign Cities. The high radius of gyration in Globetrotters can be an artifact of airfare stopovers because such check-ins are also included in the travel behavior. Finally, T2 (Active Vacationers) is a likewise small cluster, but it has the most active travelers with 2.77 trips visiting many unique cities both in their own country and abroad.

In summary, our independent characterization of the users' check-in behavior in their home city and during travel allowed us to discover three and four distinct groups of locals and travelers, respectively. Our main takeaway from the cluster analysis is that the mobility metrics explored in our work seem unsuitable for characterizing locals in our LBSN check-in data set, as the clustering algorithms struggle to find distinct groups using these features. This is an important observation since it implies that by mixing travelers and locals when evaluating POI recommendation algorithms, we will likely observe disparate results due to the fact that we may not adapt well to the interests of any of them. To systematically investigate the effect of having established these groups, we evaluate the performance of POI recommender systems in each group.

3.3.2 Experimental Settings

For the recommendation experiments, we performed the same preprocessing steps temporal splitting as for the cluster analysis, cf. Section 3.3.1. Since the five cities of the study (Istanbul, Mexico City, Tokyo, New York, and London) have different characteristics, especially geographic properties that are exploited by some of the recommendation models, we train and test the algorithms separately for each city. Table 3.5 summarizes relevant statistics of the individual cities: note that repetitive check-ins constitute a large part of the interactions (the percentage of unique check-ins reach at most 60%), which makes it difficult to recommend new POIs to users. Since we follow the "TrainItems" methodology [122], we furthermore remove all visited venues of a user in the training set as candidates for recommendation in the test set for this user. We firmly believe that this approach is suitable because, as opposed to repeated consumption of items, in, e.g., the music domain, the inherent value of POI recommendation is to suggest new places for users to be discovered.

Table 3.5: Statistics of the data set and cities used in the experiments. $|\mathbf{U}|$, $|\mathbf{V}|$, $\mathbf{C}$, and $\frac{|\mathbf{C}|}{|\mathbf{U}|\cdot|\mathbf{V}|}\%$ represent the number of users, venues, check-ins, and the density, respectively. As in LBSNs, some users may check-in in the same venue more than once, we also report in column $|\mathbf{C}|_u$ the number of unique check-ins and $\frac{|\mathbf{C}|_u}{|\mathbf{U}|\cdot|\mathbf{V}|}\%$ represents the density with the unique check-ins.

City	Split	$\lvert\mathbf{U}\rvert$	$\lvert\mathbf{V}\rvert$	$\lvert\mathbf{C}\rvert$	$\lvert\mathbf{C}\rvert_u$	$\frac{\lvert\mathbf{C}\rvert}{\lvert\mathbf{U}\rvert\cdot\lvert\mathbf{V}\rvert}\%$	$\frac{\lvert\mathbf{C}\rvert_u}{\lvert\mathbf{U}\rvert\cdot\lvert\mathbf{V}\rvert}\%$
Filtered data set	Full	139,270	251,115	9,266,149	4,354,336	0.02650	0.01245
	Training	137,842	248,692	7,412,919	3,596,596	0.02162	0.01049
	Test	108,213	196,945	1,853,230	1,134,909	0.00870	0.00532
Istanbul	Full	29,307	20,366	1,569,015	821,683	0.26288	0.13767
	Training	26,894	19,976	1,189,646	645,536	0.22144	0.12016
	Test	21,780	17,226	379,369	248,157	0.10112	0.06614
Mexico City	Full	5,944	7,978	286,638	147,850	0.60445	0.31178
	Training	5,690	7,948	237,188	125,675	0.52447	0.27789
	Test	4,018	6,442	49,450	32,616	0.19104	0.12601
Tokyo	Full	6,631	5,543	227,391	122,814	0.61866	0.33414
	Training	6,213	5,534	186,248	103,768	0.54169	0.30180
	Test	4,194	4,831	41,143	28,211	0.20306	0.13924
New York	Full	8,170	3,557	109,611	68,988	0.37718	0.23739
	Training	7,238	3,548	92,790	59,342	0.36133	0.23108
	Test	3,319	2,867	16,821	12,728	0.17677	0.13376
London	Full	4,235	1,612	43,794	26,472	0.64150	0.38776
	Training	3,520	1,607	35,516	21,697	0.62786	0.38357
	Test	1,749	1,361	8,278	6,108	0.34776	0.25660

3.3.2.1 Algorithms

The algorithms used in our experiments are representative of the state of the art in POI recommendation [125] and can be categorized into classic and POI-specific algorithms. For their exact formulations, we refer the reader to the respective references.

- Classic recommendation algorithms:
  - Rnd: performs recommendations of venues randomly.
  - Pop: recommends to the target user the venues that have been visited by the largest number of users.
  - UB/IB: non-normalized user and item-based neighborhood approaches [104, 4].
  - HKV: Matrix Factorization (MF) algorithm that uses Alternate Least Squares for optimization (from [66]).
  - BPRMF: Bayesian Personalized Ranking (a pairwise personalized ranking loss optimization algorithm) using an MF approach (from [116]). We used the version from MyMedialite's[3] library.
- Specific algorithms for POI recommendation:
  - IRENMF: Weighted MF method from [84]. This method incorporates geographical information in two different ways: instance-level influence (users tend to visit neighboring locations) and region-level influence (they assume that the user preferences are shared in the same geographical region).
  - GeoBPR: Geographical Bayesian Personalized Ranking. A POI recommender optimized using BPR [160]. It analyzes the POIs visited by the target user and assumes that she will prefer to visit new POIs that are close to the ones she visited previously.
  - FMFMGM: Probabilistic MF with multi-center Gaussian model. It is a hybrid approach proposed by [26] that combines Probabilistic MF (PMF) with a Multi-center Gaussian Model technique (MGM).
  - RankGeo-FM: a ranking-based MF approach model proposed by Li et al. [79]. They model the geographical influence by exploiting the neighboring POIs (by geographical distance) with respect to the target POI using an additional latent matrix for the users.
  - PGN: popularity, geographical, and user-based neighborhood. A hybrid approach that combines the popularity algorithm (Pop), user-based neighborhood (UB), and a geographical recommender that recommends to the target user the venues closer to the average geographical position of all the venues visited by the user. The final score is an aggregation of every item score provided by each recommender after normalizing its values by the maximum score of each method.

To ensure fair competition between the algorithms, we perform a hyperparameter tuning step for all recommenders [31]. Table 3.6 lists the tested configurations. For the subsequent experiments, we selected the respective configuration that maximized nDCG@5.

[3] MyMedialite library: `http://www.mymedialite.net/`

Table 3.6: Parameters tuning in the recommenders. The best configurations are selected by maximizing nDCG@5.

Rec	**Parameters**
UB/IB/PGN	Sim = {Vector Cosine, Set Jaccard}, $k = \{20, 40, 60, 80, 100, 120\}$
HKV	Iter = 20, Factors = $\{10, 50, 100\}$, $\lambda = \{0.1, 1, 10\}$, $\alpha = \{0.1, 1, 10, 100\}$
BPRMF	Factors = $\{10, 50, 100\}$, BiasReg = $\{0, 0.5, 1\}$, LearnRate = 0.05, Iter = 50, RegU = RegI = $\{0.0025, 0.001, 0.005, 0.01, 0.1\}$, RegJ = RegU/10
IRENMF	Factors = $\{50, 100\}$, geo-$\alpha = \{0.4, 0.6\}$, $\lambda_3 = \{0.1, 1\}$, clusters = {5, 50}
FMFMGM	Factors = $\{50, 100\}$, $\alpha = \{0.2, 0.4\}$, $\theta = \{0.02, 0.1\}$, dist = 15, iter = 30, $\alpha_2 = \{20, 40\}$, $\beta = 0.2$, sigmoid = False, LearnRate = 0.0001
RankGeo-FM	Factors = $\{50, 100\}$, $\alpha = \{0.1, 0.2\}$, c = 1, $\epsilon = 0.3$, neighs = $\{10, 50, 100, 200\}$ iter = 120, decay = 1, boldDriver = True, learnRate = 0.001

3.3.2.2 Evaluation Metrics

As we mentioned above, we will not only measure the performance of the recommendations in terms of nDCG, but also we take the novelty (in terms of EPC), and the diversity (in terms of Aggregate Diversity, or Item Coverage, IC) into account. All metrics are reported with a cutoff of 5.

- **Normalized Discounted Cumulative Gain (nDCG)**: is the most prevalent metric to measure accuracy in information retrieval and recommender systems [60].

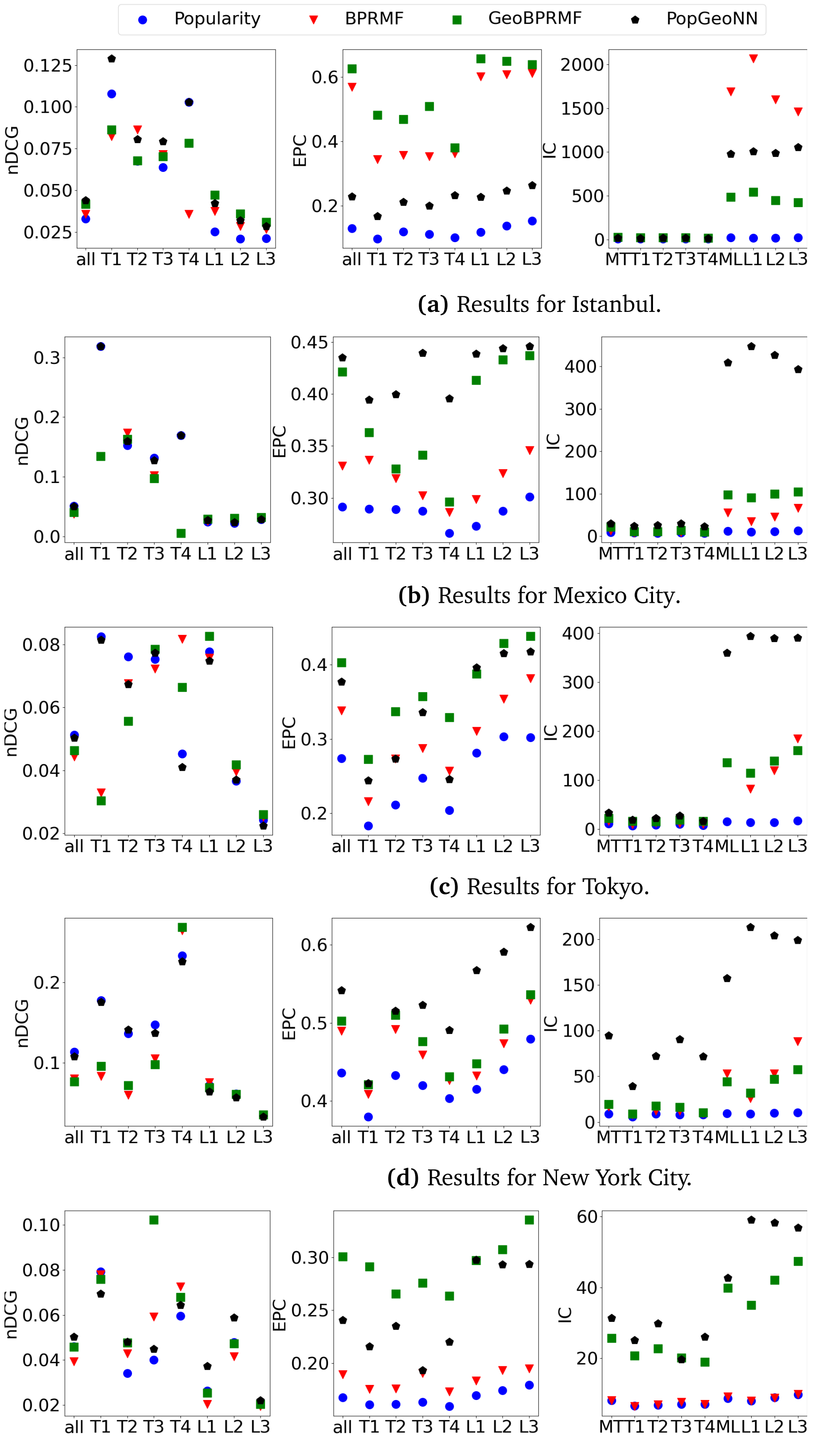
- **Expected Popularity Complement (EPC)**: a novelty metric that rewards recommending less popular items [146]. We report the normalized EPC value by applying the min-max normalization.

- **Item Coverage (IC)**, also known as Aggregate Diversity: a diversity metric that measures the number of different items an algorithm is able to recommend [60].

3.3.3 Performance of Recommenders in Specific User Groups

To visualize the results of the recommenders, we use a scatter plot that combines all cities, the three evaluation metrics, and report the performance for four exemplary algorithms, Pop and BPRMF from the classical recommenders and GeoBPR and PGN from the POI recommenders, as they are the algorithms which generally obtain the best results in terms of nDCG. We report the overall performance of all users in the test set along with the respective clusters of the travelers (denoted with T1, T2, T3, and T4) and the locals (L1, L2, and L3) in Figure 3.2.

To report comparable values for the Item Coverage with clusters of different sizes, we compute this metric by performing random subsamples. We select the cluster of the travelers and locals with the smallest number of users and then compute the IC values with this amount of random users. The reported values are the mean values after repeating the sampling $1,000$ times, thereby making the values comparable. This is why instead of the “all” label, we use two additional ones when representing the IC metric, “MT” (mean of travelers) and “ML” (mean of locals), which would be used for selecting a subsample of all users with the size of the lowest traveler and local group, respectively.



(a) Results for Istanbul.

(b) Results for Mexico City.

(c) Results for Tokyo.

(d) Results for New York City.

(e) Results for London.

Figure 3.2: Outcome of the recommendation for the different cities, metrics, algorithms, and groups.

3.3.3.1 Locals versus Travelers Results

The results reveal some interesting effects: surprisingly, travelers generally obtain better values than locals in terms of accuracy in most cities, despite that in each city, we have way more training data for locals than for travelers. Notably, travelers generally have a slightly lower novelty than locals, indicating that they tend to receive recommendations of more popular POIs. This makes sense because when a tourist visits a city, she is more likely to visit the most popular venues than if she was a local. Furthermore, frequently and repetitively visited venues such as airports and train stations can usually not be recommended to the locals since they have a higher chance of having already visited them during the training period. This also leads to the effect that locals tend to receive recommendations of more diverse POIs. By contrast, most travelers will visit a city for the first time during the evaluation period; thus, it is more probable that they actually visit one of the recommended popular POIs, which results in decreased novelty and diversity scores. Finally, because there are far fewer travelers than locals, it is expected that despite having computed the IC metric using the subsamples, we obtain much lower results for travelers than for locals, making a direct comparison impossible. In general, these results tend to support the findings of Sánchez and Bellogín [124]. However, we performed a different data preprocessing, splitting methodology, and also a different analysis and characterization of travelers and locals.

3.3.3.2 Cluster Results

In addition to the analysis performed for travelers and locals, it is also interesting to study the behavior of the recommendation models among the different types of travelers and locals, i.e., the clusters derived in Section 3.3.1.

Regarding the travelers, there is no common behavior in the different cities. For example, T4 obtains the highest values in nDCG for New York in all recommenders, whereas, in other cities, such as Istanbul and Mexico City, some models obtain very low values for these users. Regarding novelty and diversity, T1 obtains the worst results in the cities of Tokyo and New York, whereas, in London, it is one of the best groups in both dimensions. Despite these discrepancies among the travelers, we can observe some common behavior, such as T2 and T3 generally obtaining comparable results. This may be explained by the features shown in Table 3.4, where we can observe that these two groups have the highest ratio of domestic trips, whereas T1 and T4 tend to make more abroad travels, visiting more popular POIs as we can observe in the performance in both nDCG and EPC metrics. In fact, except for Mexico City, the Pop algorithm consistently achieves higher values in EPC (and hence, more novel items) for both T2 and T3.

Regarding the locals, in all cities, except Mexico City, L3 (highest activity level) is the cluster that obtains the lowest levels in terms of accuracy but obtains higher values of novelty than the other clusters. In this cluster, all recommenders have similar performance in terms of ranking accuracy, whereas the other groups show higher variations. We argue that this has to do with the larger size of this cluster and the larger number of venues the users of this cluster visit compared to the other clusters. Hence, it is more probable to recommend these users less popular venues given the probability that they have visited more venues before than the other two groups with a lower activity level, making it more difficult to recommend both novel and relevant venues to them.

3.3.4 Summary

Unsurprisingly, by segmenting the Foursquare users into different clusters of travelers and locals, we are able to observe well-differentiated behaviors and recommendation outcomes. When analyzing the recommendations, we found that despite travelers being fewer in numbers, they tend to get higher values in terms of accuracy and lower values in terms of novelty and diversity compared to the recommendations for the locals with more available training data. Our results once more emphasize the role of the popularity bias in POI recommendation. However, we believe this bias would be worth analyzing more in-depth for this domain, as it has been done in other traditional recommendation scenarios [1, 11]. It should also not go unnoticed that for both travelers and locals, the performance of the recommendation algorithms is rather low, and sometimes the best performing algorithm in the basic popularity recommender, which is a previously observed tendency in this domain [124, 125].

A relevant insight of this study is that by assessing the quality of the clustering results, it is imperative to use different features to derive the clusters of travelers and locals. We note that the geographical information was especially relevant for the travelers, as we found four highly differentiated groups according to the ratio of domestic trips and geographic displacement. For locals, we found that the most important features were regarding the activity level, especially in terms of activity duration and the number of unique POIs visited. Interestingly, that mobility features derived from such an LBSN data set are unsuited to obtain high-quality clusters for the locals.

Most importantly, we could show that different user groups exhibit very different behavior; therefore, it would be misleading to measure the performance of recommendation algorithms for all users as a whole. Especially when the recommendations should be tailored to specific groups, a "one-size-fits-it-all" algorithm, which seemingly produces good recommendations, might fail in a particular user group. Concretely, we could measure differences of $> 400\%$ between different user groups in terms of nDCG, such as in London and New York using GeoBPR recommender or Mexico City using the PGN algorithm. Besides, we also observed some important differences between user groups when measuring EPC, in the case of Tokyo for the PGN and BPRMF algorithms, although the difference was less severe in comparison to the accuracy. Finally, in view of the analyses and results obtained, we would like to raise concerns that this Foursquare data set may not generally be appropriate for use in the travel and tourism domain because the vast majority of them have barely checked-in in more than one city, as can be seen in Tables 3.3, and 3.4. Although this particular data set seems to be ill-suited for obtaining general conclusions about the decision-making of travelers in a real-world environment, we do believe LBSN data is a useful resource to tourism applications, including recommending venues to users exploiting the interactions of users in their home city [114].

3.4 Discovering Regions Using Graph-based Community Detection of Tourist Trips

The mobility of travelers manifested in the trips can be used for further insights into global travel behavior. In this section, we describe a methodology to obtain a map of the world's travel regions that are entirely based on tourist travel behavior instead of political and administrative regions. With this approach, we aim to uncover implicit tourist regions independent of administrative boundaries, e.g., in areas where travel can occur irrespective of national borders, such as the Schengen Area of Europe. To achieve this, we construct a graph of flows from the trips and use a community detection algorithm to cluster single destinations into coherent travel regions [127]. Again, the trips from the self-collected Twitter data set serve as the basis for the analysis.

3.4.1 Global Mobility Graph Creation and Community Detection

To capture the global mobility in its entirety, graph-based structures are a common choice. For the transformation of individual trips into a mobility graph, we performed the following steps: each node corresponds to a city, and the weight of the edges should model the *traveled-together* relation, i.e., that two cities have been visited within the same trip, as closely as possible. Thus, the flow between two cities is computed by summing up the co-occurrences of the two nodes in a clique formed by all cities in a trip, over all trips. For example, if a trip consisted of travel from Munich to Berlin via Nuremberg, we would also count the transitive flow from Munich to Berlin. This makes the undirected graph denser, but we argue that including the transitive links of a trip models the phenomenon better than discarding this information. Finally, we determine the weight of each edge as the flow divided by the Euclidean Distance between the two cities. Including the distance in the edge weight reduces the noise in the flow graph introduced by distant traffic hubs, such as airports.

Transforming the mobility patterns into this graph-based representation enables us to run community detection algorithms to see which cities form coherent clusters with only global mobility as input. The Infomap algorithm is a graph community detection algorithm that is designed to discover the underlying structure of the nodes and edges [119], which can yield multi-level hierarchies for communities. Using a random walk strategy, the algorithm optimizes community segmentation by selecting groups so that the inter-community flows are maximized while the flows between groups are minimized. Since it uses a probabilistic model for community detection, we re-run the algorithm ten times to reduce the probability of obtaining a local minimum, which is monitored using the description length [119].

In our approach, a community corresponds to a set of cities that form a region. Since the results of Infomap are hierarchical, it will return a tree of regions and subregions, depending on a graph-theoretic termination criterion. This is especially useful since choosing the right granularity of regions depends on the use case and allows for a flexible application of the method.

3.4.2 Results

The resulting graph consists of 14,558 nodes and 3,624,909 undirected edges. The degree distribution is long-tailed with very few high degree nodes and 87% of nodes having a degree of less than 1,000. The graph density of 0.034 indicates a sparse graph.

The communities computed by the Infomap algorithm show four top-level regions that align well with existing continental boundaries, cf. Table 3.7 and Figure 3.3. These are further subdivided into a region hierarchy of up to four levels. Level 2 regions roughly align with national boundaries; however, as discussed below, we can observe some interesting exceptions to this rule. Level 3 regions are the most numerous and comprise travel regions, which are of primary relevance for a destination recommender system. In most cases, the community detection algorithm terminates after level 3, but

Table 3.7: Numerical description of the four top-level regions.

Region name	Nodes	2nd-level regions	3rd-level regions	4th-level regions
World	15,858	53	942	476
South America	1,873	9	193	19
North & Central America	4,193	17	254	145
Europe & West Africa	6,591	14	381	116
Asia & Oceania	3,201	13	114	196

in some areas, with more Twitter data, these regions are further subdivided into level 4 regions, which can even be individual districts of cities.

In the following, we will exemplify the regions formed on each level and discuss some interesting artifacts.

3.4.2.1 Level 1 – Continents

On this level, the division between the Americas is a perfect cut between North and Central America and South America. Africa is under-represented in the data because it has only a few check-ins in Morocco, Algeria, Ghana, Nigeria, Kenya, and South Africa. These countries are merged in the European cluster except Kenya, which is in the Asian region. The European cluster is merged with all of Russia, Turkey, and the Arabian Peninsula. The Asian region comprises the Indian subcontinent, South East Asia, South Korea, Japan, Australia, and New Zealand.

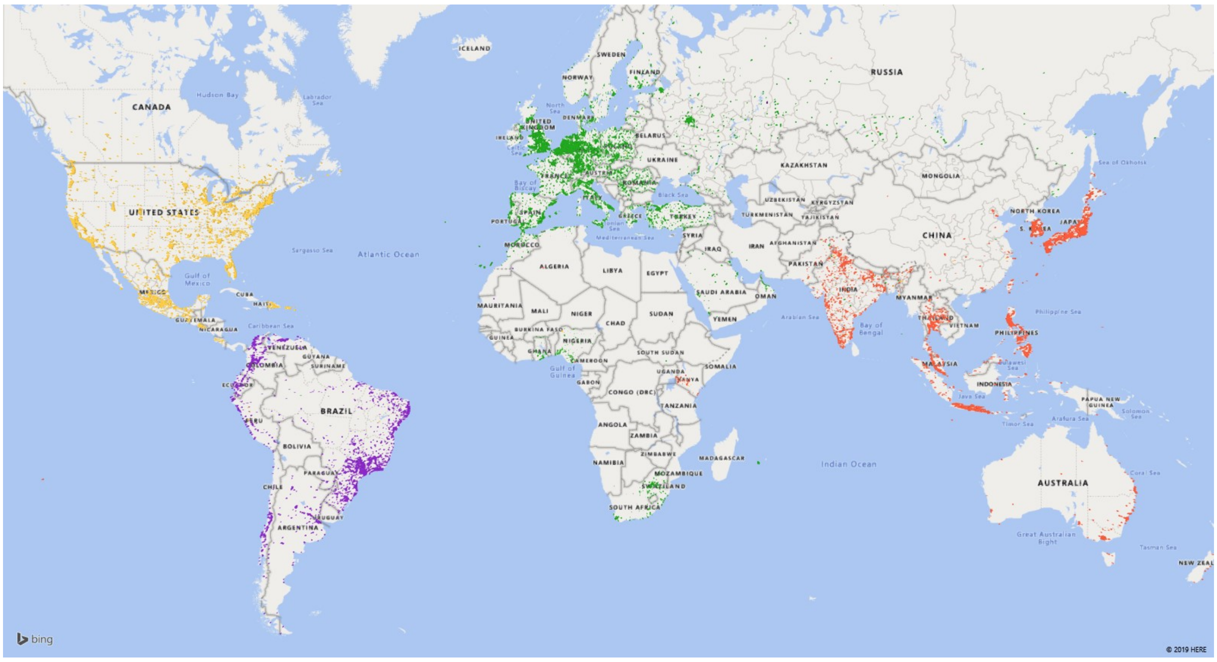


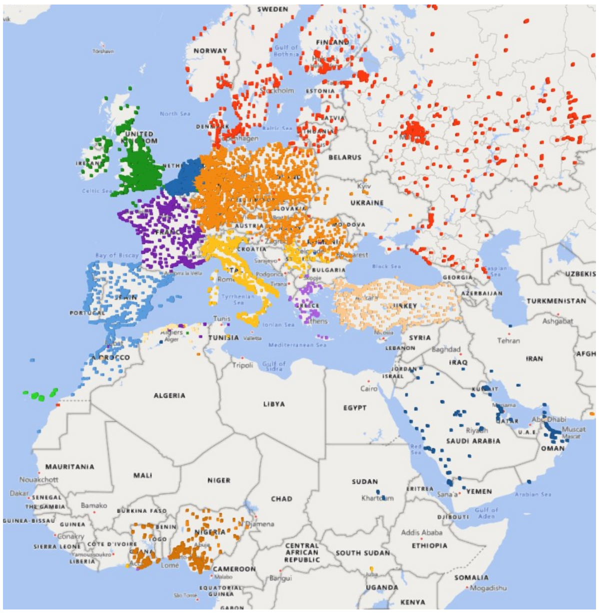
Figure 3.3: The top-level regions

We observe that on this level, the destinations' geography and accessibility play a dominant role. For example, the Arab countries are united with the European cluster due to the major aviation hubs. Unfortunately, the lack of data from Africa and some countries in Central Asia hinders the formation of additional clusters in these areas.

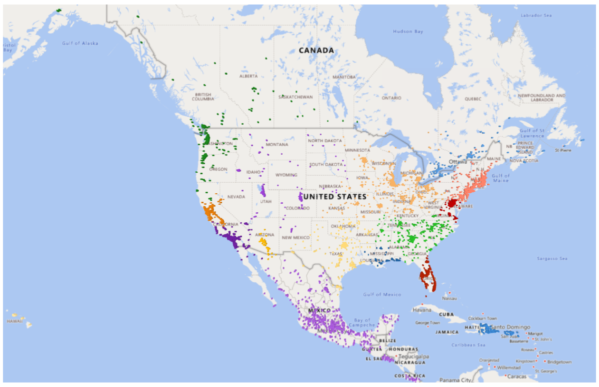
3.4.2.2 Level 2 – Countries

At the second level of the region hierarchy, we found that many regions align well with national boundaries; however, with some notable exceptions.

In Europe, a large second-level cluster is found spanning the countries of Germany, Austria, Switzerland, Hungary, the Czech Republic, Poland, and Romania (cf. Figure 3.4a). The Scandinavian countries are clustered together with Russia and the Baltic countries. Italy forms one region with Serbia, however, the unavailability of data from Croatia possibly influenced this result in an unpredictable manner. The Iberian countries are clustered with Morocco, which could be attributed to immigration patterns and the very cheap flight and ferry prices between these countries. Belgium and the Netherlands form another region, and also the British Isles are clustered together. On the other hand, France, Turkey, and Greece form regions identical to their national borders.



(a) The second-level community structure of Europe



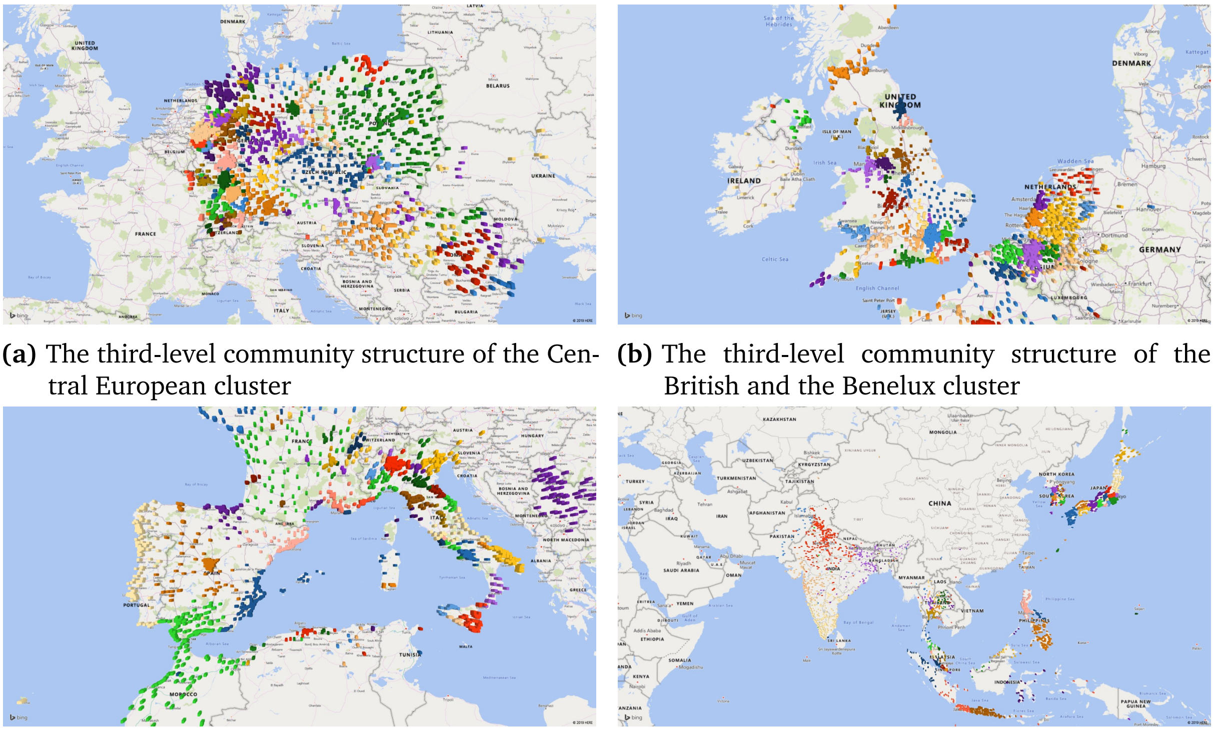
(b) The second-level communities of North America

The second-level regions formed in North America in Figure 3.4b mostly disregard national boundaries. Mexico is in one region with other Central American countries, while the USA and Canada are divided into fourteen clusters. The western Canadian states are merged with Oregon and Washington, while California is split into two major clusters, with the southern cluster expanding down to Tijuana and Mexicali in Mexico. Mexican cities on the borders of Arizona and Texas are also members of predominantly American clusters. Several other well-known regions, such as the New England states, Florida, and the Great Lakes area, form their own clusters.

The second-level regions reveal that there are some countries that are traveled to exclusively, but other countries are visited together in one trip. In extensive countries such as Brazil and the USA, we observe a subdivision into multiple domestic subregions at the second level. We see this as an indication that the approach is well suited for discovering domestic tourism regions.

3.4.2.3 Level 3 – Destinations

The regions formed at the third level of the hierarchy comprise tourism destinations; however, the results show varying granularities in different parts of the world, with some regions containing further subregions.



(a) The third-level community structure of the Central European cluster

(b) The third-level community structure of the British and the Benelux cluster

(c) The third-level community structure of the Iberian and Italian clusters

(d) The third-level community structure of South Asia

The third-level clusters of the big Central European cluster in Figure 3.5a vary in terms of cities' size and density. The dense regions are typically very contiguous and are centered around a major city. For example, the brown region containing Munich, Germany, is comparatively large and includes southern Bavaria, Germany. Large areas of the Czech Republic, Poland, and Hungary form homogeneous clusters with no further subregions. Figure 3.5b shows that Belgium forms three main regions at the third level with another shared region in the partly German-speaking areas in the east of the country, spanning over to North-Rhine Westphalia. The Netherlands is divided into six regions that align well with the local administrative divisions. Similar contiguous subdivisions are found in the British Isles, Spain, and Italy. The clustering of Morocco, cf. Figure 3.5c, with Gibraltar, UK and Andalusia, Spain is curious, and we have no convincing theory why this is the case besides very accessible ferry connections and cultural similarities due to a shared history.

In Asia, Pakistan and India are separated at this level, with India forming four subregions (Figure 3.5d). In Thailand, one region is formed by places along the touristically very active coast, while the inland regions are divided into numerous smaller regions. The regions formed in Japan are similar to the administrative prefectures.

The third-level hierarchical result generally provides regions that can be seen as coherent tourist destinations. At this level, they become small enough to visit them exhaustively within a few days, and most do not contain further subregions. Thus, for applicability in a global destination recommender system, the level 3 regions are a very good fit, despite some influence of the administrative regions. This is acceptable since administrative borders naturally influence travel behavior, and there are many examples where our detected regions either combine several administrative regions or subdivide them.

3.4.2.4 Further Levels

Interestingly, some regions are further subdivided, which we discuss using one striking example from New York City, NY, USA. In the third-level region of New York State, a fourth-level region with the Burroughs Manhattan, Bronx, Brooklyn, Queens, Staten Island, and Jersey City is formed (cf. Figure 3.6). Long Island contains two more regions, while four other regions surround New York City.

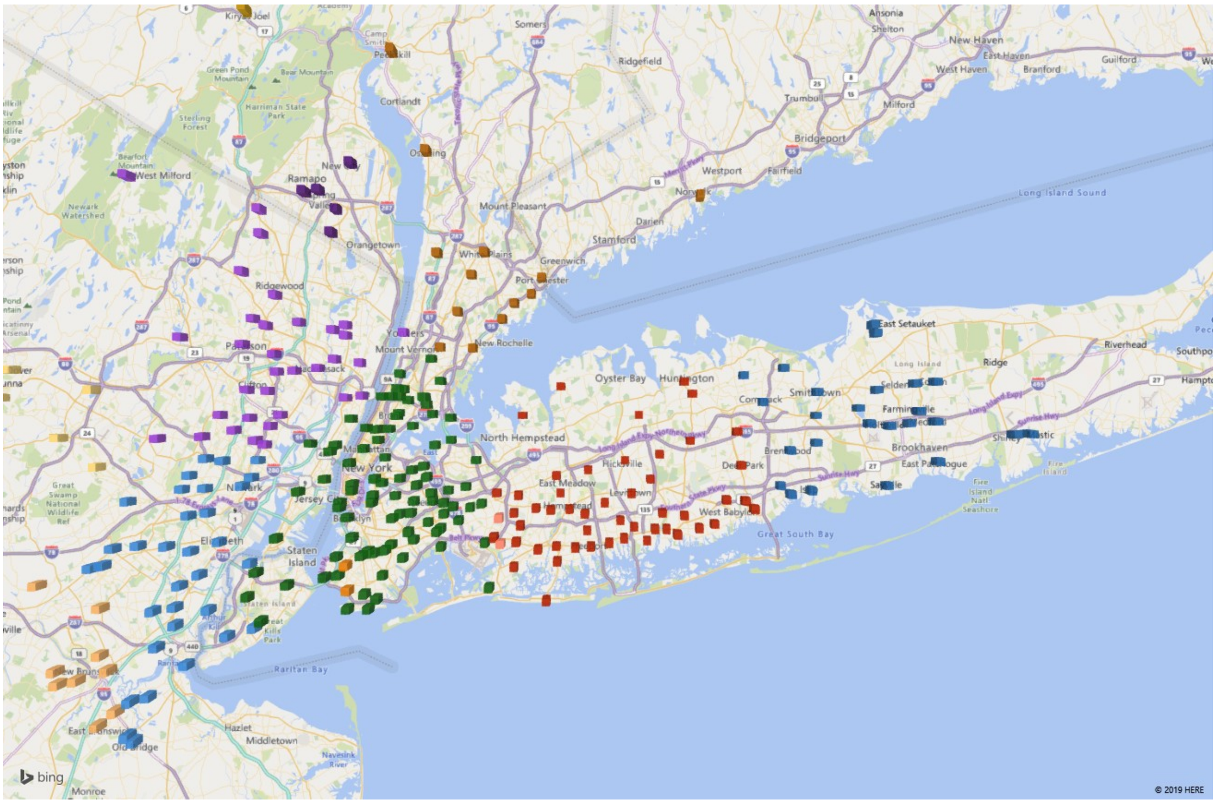


Figure 3.6: The fourth-level community structure within New York, NY, USA

This shows that given that the Infomap algorithm obtains sufficient data, it is capable of discovering very fine-grained regions even within cities. This example of New York City is an artifact of the high-population density, the municipality structure, and a large amount of Twitter data in this area.

3.4.3 Summary

The first three levels of the cluster hierarchy roughly align to continents, countries, and travel regions. The fourth level gives an even finer granularity of destinations, however, in most regions, there is insufficient data for the community detection algorithm to descend to this level.

Generally, the influence of national borders is still observable, however, other factors such as the language (e.g., France, Greece, British Isles, and the USA) seem to be relevant to retaining national boundaries, while other regions are clustered together, as can be seen in central Europe. Thus, migratory patterns also play a role, as can be seen in Central Europa as well as the merging of the Baltic countries together with Russia; an effect that we attribute to the significant Russian communities living in the Baltic countries.

An important limitation of this approach is missing data. If the underlying data source is missing check-ins from a country for any reason, the algorithm does not have reasonable means to counter this. In the case of China, where Twitter is a target of censorship [5], independent clusters simply form around the large country. In the case of small countries with missing data, such as Croatia or Belarus, the algorithm can ignore the missing data resulting in clusters that encompass the area.

In conclusion, this approach provides a fine-grained map of touristic travel regions solely based on traveler mobility. Since the region model is hierarchical, its application scenarios are flexible, and developers can pick the hierarchy that suits their needs best. To make the region model usable, Sharaki has developed a software library to handle such hierarchical region models[4] and also create visualizations[5] [132].

In the next section, we show another application of collecting trips from LBSN data, namely recommending the personalized duration of stay.

3.5 Recommending the Duration of Stay

Mainstream recommender systems research is mainly concerned with predicting the ratings for items of an active user, which results in a subset of top-ranked items that are to be presented to the user in an appealing way. This challenge of finding the “best” items according to any metric is essential across all recommender system domains. In the destination recommendation domain, it is not only important to visit the right region but to plan the trip duration in a way that personal preferences, costs, and the value of staying longer at a destination are balanced. This is especially relevant when travel packages [153] or composite trips are recommended [121].

We examine the problem of determining the personalized duration of stay using both the past trips of the active user and the overall duration of stays at the specific cities. In Section 3.2, we already determined the typical overall trip duration of groups of travelers using a cluster analysis, however, in this section, we focus on individual stays at a destination. For this, we used the mined trips from the Foursquare data [41].

3.5.1 Algorithm

Our proposed solution addresses the question of making personalized recommendations regarding the duration of a tourist’s visit to a city by considering two factors: the typical time that all tourists spend in that city and the current user’s average duration of stay at previously visited destinations. Thus, the first step is to compute the distribution of the durations of people’s stay at a destination since there can be substantial differences between destinations regarding how long one needs to explore it. For example, a smaller city can be covered within a day or two, whereas a major metropolis might require more time. The second aspect is the pace at which the particular traveler visits destinations. Some tourists want to immerse themselves deeply into a culture and therefore stay at

[4] `https://github.com/osharaki/travel_regions`

[5] `https://github.com/osharaki/travel_regions_visualizer`

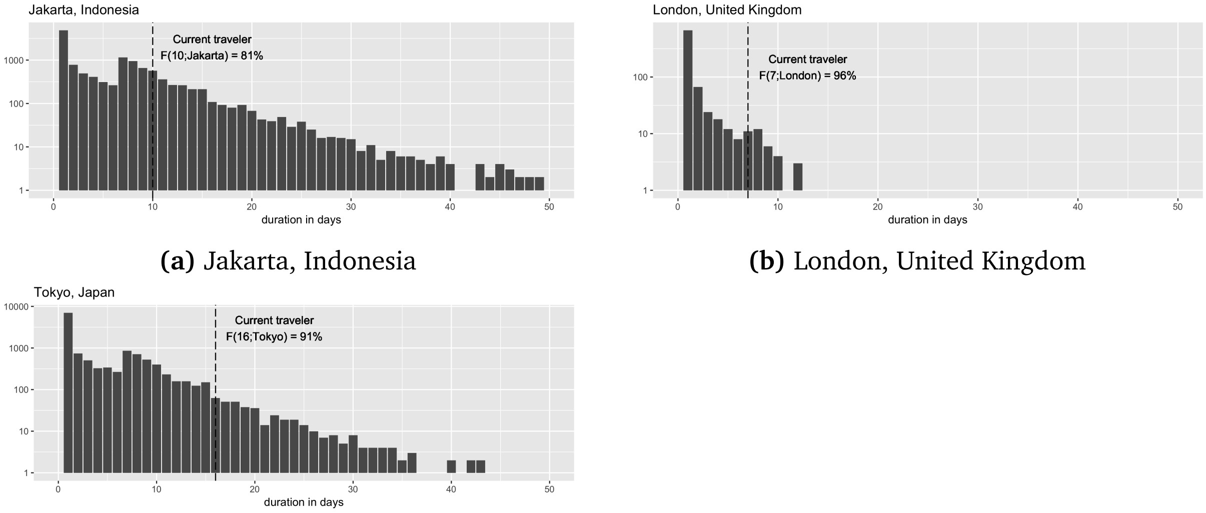
each location for a longer time, whereas others want to visit as many different places as possible during their holidays. This past behavior can be used to personalize the recommended duration of stays.

Using the `tripmining` library, we mine $223,688$ domestic and $10,963$ international trips from the Foursquare data set [154] with a total of 690,897 blocks for further analysis. We compute the distributions of the durations of stay for all towns with more than $15,000$ inhabitants.

The next step is to determine the pace at which the current typically travels, i.e., the distribution of the duration of the individual traveler's past blocks. To obtain this information, we can either ask the traveler to provide some information about past trips directly, or we can request access to the individual's mobility patterns from her profile on a LBSN. Once we have this information about past trips, we can derive the user's pace by comparing it to the quantiles of all other travelers who have visited the same destinations. This essentially establishes a collaborative filtering method to derive the duration of stays from actual user behavior.

3.5.2 Example

To visualize our approach, we show how the algorithm would calculate the personalized duration of a sample user's visit to Washington, D.C. To that end, we calculate the quantiles of the previously visited cities. In our example, the user made three previous visits, spending 16 days in Tokyo, ten days in Jakarta, and seven days in London. We have visualized the distributions of the durations of blocks in the three cities in Figures 3.7c, 3.7a, and 3.7b. The duration of these trips reveals that our user is a relatively slow-paced traveler compared to others, as her lengths of stay are toward the right side of the distributions.



(a) Jakarta, Indonesia

(b) London, United Kingdom

(c) Tokyo, Japan

Figure 3.7: Distributions of the stay durations of the current user's past trips.

The trip to Tokyo was at 91%, the stay in Jakarta at 81%, and the visit to London at 96% of the cumulative distribution function. To aggregate the user's pace over the previous trips, we can calculate the mean percentile, which is 91%. We can then find that percentile in the distribution of visits to Washington, D.C., where the 91st percentile of the distribution is at ten days (see Figure 3.8). Therefore, this would be the recommended duration of stay.

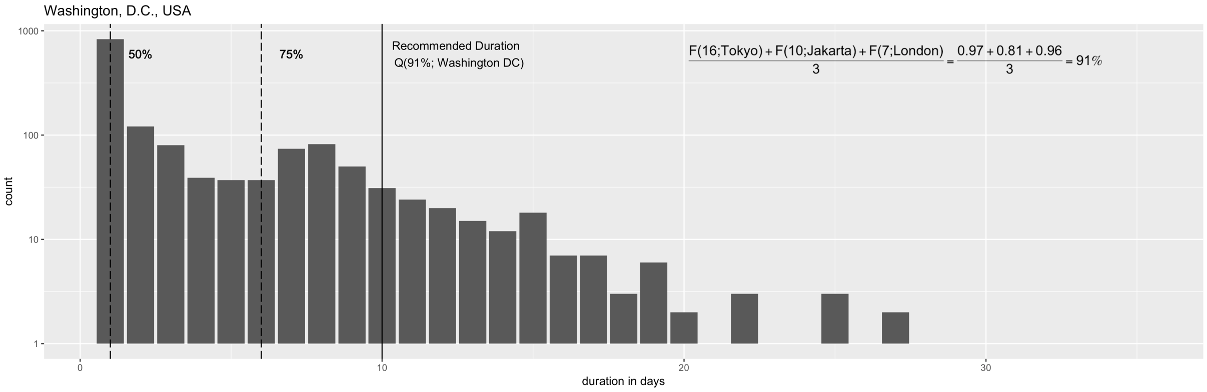


Figure 3.8: Distribution of the durations of blocks of tourist stays in Washington, D.C., USA

3.5.3 Summary

The recommended duration of stay at a specific destination can be determined using additional information about the domain and the user's past behavior. We showcased an approach based on the analysis of global mobility data from location-based social networks. The underlying method is, however, generalizable to similar problems, given the availability of appropriate data. We argue that such data is indeed often available, especially in commercial recommender systems. In the tourism sector, airlines and hotel portals have a long history of user data, which they could easily leverage when making recommendations [53]. After all, the proposed approach can be used in any recommender systems domain, where the amount of the recommendation matters and where information about the distribution of the quantity is available for both all users and the particular user of interest.

Needless to say, this research idea should be systematically evaluated trips from different LBSNs and compared to other machine learning baselines [2].

3.6 Summary

Capturing the mobility of travelers can improve travel recommender system in meaningful ways. Since traveling is a global phenomenon, the popularity of LBSNs provided us with great data sources to analyze how people travel the world compared to the incomplete view one would obtain from traditional data sources used in mobility research. The foundation of this chapter is a method to derive individual trips from check-in-based data and assess it regarding the reliability of the data. This enabled us to characterize the behavior of users, e.g., using a cluster analysis of trip types.

Different groups of users might have different needs, however, the evaluation practice in POI recommender systems is to report performance metrics aggregated over all users. By discovering groups of local users and travelers based on their behavior using a similar cluster analysis, we found that there are significant differences in the performance of POI recommendation models for the different groups. This finding calls for more rigorous evaluation practices and helps to understand the differences in recommendation performance in different populations.

In Section 3.4, we presented an approach to compute an alternative map of the world by clustering cities into travel regions based on which cities are commonly traveled together. The clustering is based on a community detection algorithm that takes a global mobility graph as input, which we constructed using the mined trips from Twitter. Lastly, we proposed a statistical method to derive the personalized duration of stay for uses based on their past trips and the distribution of stay lengths at a destination.

This range of applications underlines the value of traveler mobility analysis for improving travel recommender systems and enabling use cases that would otherwise not be achievable. With minor adaptations, all these contributions can be used to improve an individual trip planning application, whether be it the region model for destinations of different scopes, kick-starting the preference elicitation by analyzing the users' past trips, or recommending how long they should stay at the individual destinations to make most of the limited vacation period. When it comes to recommending where to travel, we will analyze the quality of data models in content-based destination recommender systems in the following Chapter 4.

4 Destination Characterization

The performance of data-driven systems is inherently determined by the underlying quality of data. In content-based recommendation, there are usually many competing ways to characterize items to compute recommendations. As recommendations are based on the items' features, it is imperative to capture each entity as close as possible with respect to the user's conception of the domain. This leads to the challenge of determining which instantiation of the available data is the *best* to emulate a possibly latent concept, i.e., what the recommendations are about. Concretely, a destination recommender system should characterize the items in a way that similar destinations concerning the touristic experience should also be located close to each other in the information space in which they are embedded. Such an information space is defined by the features used to characterize all items, but how does one know whether this information space reflects the users' conception of the domain? In principle, similar items in the physical world should also be similar in the information space, despite the loss of information and granularity. In some domains, the mapping of features is obvious, for example, in the case of recommending a computer configuration: the feature values, such as available memory or number of USB outlets, have a clear meaning and can be easily interpreted by the users and the algorithms [161]. In other domains, however, the ground truth for items similarity is hard to capture, which is a fundamental problem [155]: what are the movies most similar to "Fight Club?" Which cities are most similar to Munich? We as humans might have an intuition about such similarity concepts, but it is hard to show that recommendation algorithms actually emulate such latent concepts well.

To the best of our knowledge, this two-fold challenge of choosing accurate data sources to characterize items, as well as determining which features to incorporate in a content-based recommender system, has not been analyzed in a systematic way [155]. We propose a toolbox of methods to compare data sources with each other and also with respect to what is important in the domain of such a recommender system. In this study, we analyze these problems within the particularly challenging domain of content-based destination recommendation [78, 83]. For this, we introduce 18 destination characterization methods for 140 cities, which we have collected from literature or constructed ourselves. Using well-established rank correlation methods [75], we compute their pairwise similarities, thereby revealing families of similar data sources. To evaluate the data sources with respect to how tourists experience a destination, we conduct an expert study to elicit this latent concept. Using variants of established top-k rank agreement methods, we are able to assess the quality of the data sources by their similarity to the expert opinions. The choice of rank agreement methods [75] on ranked lists generated using the similarity measure of the recommender system has the advantage that our methods are straightforward to apply in other domains.

4.1 Data Sources for Destination Characterization

To characterize destinations, we collected data from various online data sources tabulated in Table 4.1. These data sources belong to three major categories: data models based on venues, textual sources, and fact-based data sets. Overall, we characterized 140 cities in Europe, Asia, Africa, and the Americas. To perform the analyses, it was necessary to characterize all cities with all data sources, which led to the exclusion of small and obscure destinations. All analyzed data sources have some touristic relevancy, meaning that they are commonly used to learn about destinations or that they are already part of travel-related information systems.

Table 4.1: Overview of the data sources for characterizing cities

Type	Name	Category	Data Objects	Number of Objects	Acronym
Venue Data	Foursquare	LBSN	Venues	2,468,736	FSQ
	OpenStreetMap	Collaborative Map	Map entities	3,106,856	OSM
Textual	Wikipedia	Collaborative Encyclopedia	Documents	1,150,719 words	WP
	Wikitravel	Travel-related Wiki	Documents	984,777 words	WT
	Google Travel	Travel Information	Documents	56,499 words	GT
Factual	Webologen	Travel Information Provider	City Features	49 tourism facts/city	TF
	Nomad List	Collaborative Travel Information	City Features	8 features / city	Nomadlist
	Seven Factor Model	Scientific Characterization	Derived Factors	7 factors / city	7FM-2018
	Geographic Location	Geographic Location	Latitude, Longitude	1 coordinate pair / city	GEO

4.1.1 Venue-based

The intuition behind this class of data sources is that the variety of venues one can visit at a destination would reflect the experience of a traveler. The following characterization methods rely on the assumption that the larger the variety of, e.g., restaurants or cultural sites of a city is, the better the score should be in these categories. Thus, these methods do not aim to assess the quality of the venues since most venues do not come with quality indications such as ratings.

The two data sources were Foursquare and OpenStreetMap. In both cases, we queried all venues within a destination that were categorized into a tourism-related category. To establish a multi-dimensional vector space of these categories, the number of venues in each category was normalized to make large and small cities comparable, as well as accounting for some categories being more frequent than others. The resulting features are, thus, based on the distribution of venues in a destination and take a normalized score in $[0, 1]$ [35].

Using this method, we constructed two data models from Foursquare, thereby leveraging the category hierarchy of the venues. The FSQ-TOP model comprises the venues of four top-level categories: “Arts & Entertainment,” “Food,” “Outdoors & Recreation,” and “Nightlife,” whereas the FSQ-2nd level uses the 337 second-level categories as aggregation targets.

Similarly, we could use the hierarchy of the OpenStreetMap (OSM) map features to create a top-level model of categories such as “tourism,” “leisure,” “historic,” “natural,” “sport,” venue count, and area. As opposed to the Foursquare characterization, OSM also provided us with exact city boundaries so that we could compute the area of the destination. The OSM-2nd model likewise comprises 14 tourism-related subcategories, as well as the venue count and the area.

4.1.2 Textual

The following family of data sources to assess the similarity of destinations are text documents describing the cities. Text as a medium is an efficient way to describe what a destination is about and what its attractions are. We selected three online resources which we argue that travelers commonly use to inform themselves about destinations: Wikipedia, Wikitravel, and Google Travel.

Wikipedia[1] articles of these prominent cities have converged to a similar structure providing both background information of the cultural history of a destination as well as brief descriptions of the city’s main attractions. Thus, it is commonly used to get a first impression of a city. Wikitravel[2], on the other hand, is a collaborative travel guide that offers more concrete information about possible activities, recommended restaurants, and general advice for traveling. The target audience of these two data sources is, thus, slightly different, with Wikitravel being more oriented towards travelers seeking concrete advice for coping in a destination, whereas Wikipedia is more aimed at prospective travelers learning about a destination.

Another popular travel information system is Google Travel[3]. Based on actual traveler visits and local insights, the platform provides a list of the most iconic attractions. To characterize the destinations, we combined all descriptions of these attractions into one document.

After applying the same pre-processing steps to the HTML text files, the pairwise document similarity was computed using the Jaccard Distance, Word2Vec embeddings, and a transformers-based approach with BERT. For the Jaccard models, a document term matrix was constructed, and the similarity between the cities was determined using the Jaccard Distance. For the embeddings-based models, Word2Vec and BERT, we used pre-trained models as zero-shot encoders to embed the documents. In the case of the Word2Vec-based models, we aggregated the pre-trained word embeddings using mean-pooling to obtain the document embedding and used the cosine similarity to compute the similarity. Given that the BERT-based sentence encoder was already pre-trained on Wikipedia articles, we did not need to perform further fine-tuning of (hyper-) parameters. After the embedding step, we again used the cosine similarity to determine the rankings. Thus, we obtained nine textual ranked lists for further evaluation: three data sources × three similarity measures.

[1] https://en.wikipedia.org

[2] https://wikitravel.org

[3] https://www.google.com/travel

4.1.3 Factual

The third category is factual data with a focus on travel and tourism. This group comprises data sources that provide facts about destinations, such as rated features or geo-social features relevant to travelers. This does not imply that the quality of the data is beyond scrutiny.

4.1.3.1 Webologen Tourism Facts

The former German eTourism start-up Webologen compiled a data set of 30,000 cities, which are described by 22 geographical attributes and 27 “motivational” ratings. With this multitude of features, this data set provides a very detailed image of a tourism destination, which makes it an interesting source. The similarity is computed using the Gower Distance [55], as the features are both binary and interval scaled.

4.1.3.2 Seven Factor Model

This data model is a mapping between Neidhardt et al.’s Seven Factor Model and the tourism facts of the Webologen data set [129]. Using the Seven Factor representations for each destination in our data set, we use the Euclidean Distance to compute the city similarity rankings on the seven-dimensional vectors.

4.1.3.3 Nomad List

Nomad List[4], a platform aimed toward digital nomads, employed a mixture of own data modeling and crowdsourcing to characterize cities for their suitability for digital nomadism. Built as a specialized platform for this community, it offers rich information about destinations, but it should also be noted that the target audience is not tourists. Since these features were already available in a normalized interval format, we used the Euclidean Distance to compute the city similarity rankings.

4.1.3.4 Geographical Distance

The geographic position of a city might not provide much insight into the characterization of destinations. However, it still serves as a simple baseline for assessing the similarities of other methods. We used the Haversine Distance [117] to compute this distance.

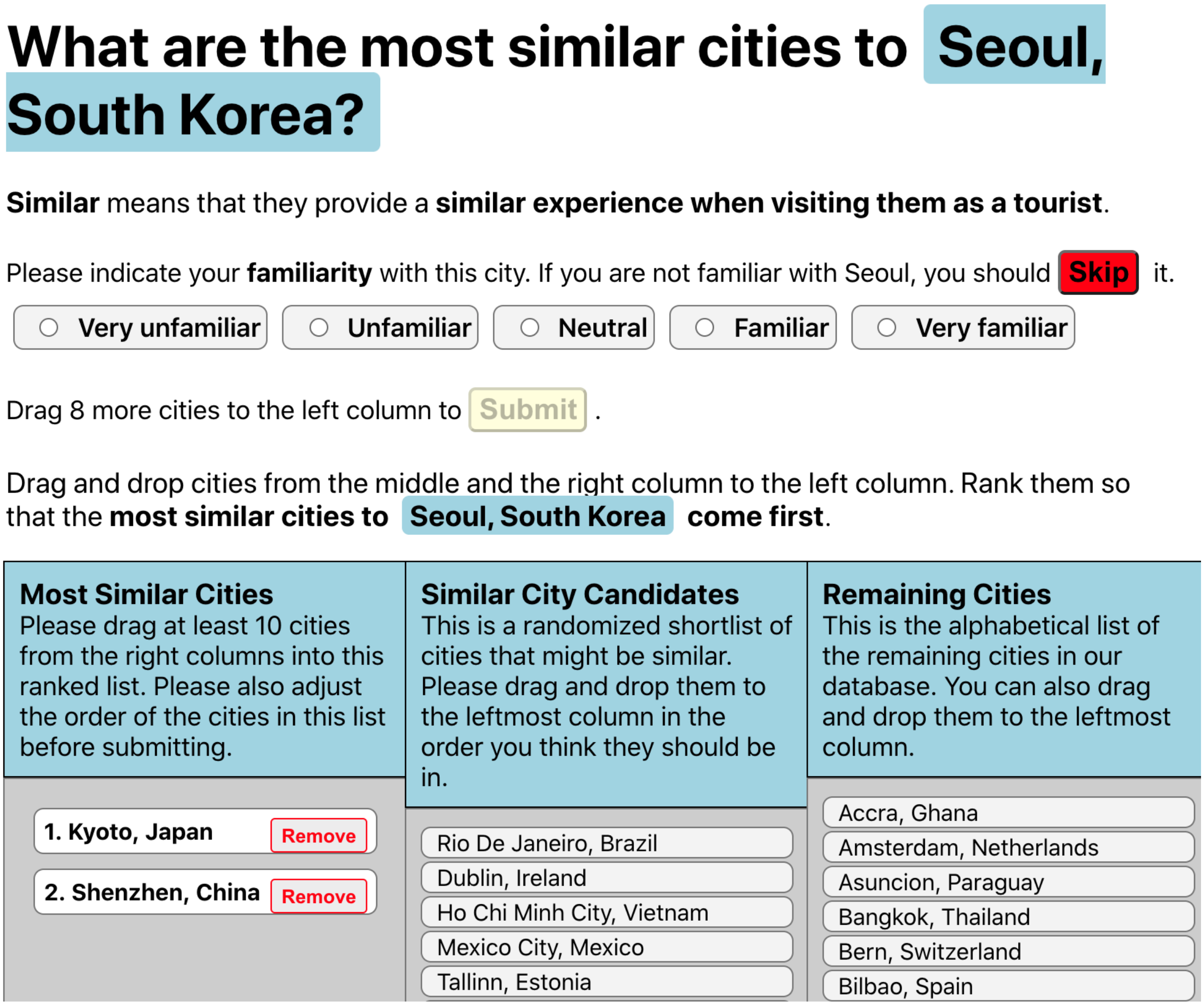
4.2 Eliciting a Desired Concept Through Expert Opinions

To find out which data source is best suited for the domain of destination recommendation, we developed a web-based expert study to capture a specific concept we are interested in: *“Similar experience when visiting cities as a tourist.”* To make this latent concept explicit, we asked experts from the travel and tourism domain to give their opinion on this matter by selecting the most similar destinations to a given city.

[4] `https://nomadlist.com`

4.2.1 Expert Survey Instrument

Eliciting such a latent concept is not trivial, as different people might have varying views on it. Additionally, ranking cities by their similarity requires much travel experience and is still a challenging task even if one has visited them. For this reason, we decided to use a web-based expert study among experts in the travel and tourism domain, as well as representatives of local tourism boards. To trade off the number of expected available experts, the difficulty of the task, and the time each expert would be willing to spend, our design choices were as follows: the study was delivered as a web page to allow easy access and no further software requirements besides a web browser. Furthermore, we focused the evaluation on 50 very prominent destinations for which the experts should determine and rank the most similar cities.



What are the most similar cities to Seoul, South Korea?

Similar means that they provide a **similar experience when visiting them as a tourist**.

Please indicate your **familiarity** with this city. If you are not familiar with Seoul, you should **Skip** it.

○ Very unfamiliar ○ Unfamiliar ○ Neutral ○ Familiar ○ Very familiar

Drag 8 more cities to the left column to Submit .

Drag and drop cities from the middle and the right column to the left column. Rank them so that the **most similar cities to Seoul, South Korea come first**.

Most Similar Cities Please drag at least 10 cities from the right columns into this ranked list. Please also adjust the order of the cities in this list before submitting.	**Similar City Candidates** This is a randomized shortlist of cities that might be similar. Please drag and drop them to the leftmost column in the order you think they should be in.	**Remaining Cities** This is the alphabetical list of the remaining cities in our database. You can also drag and drop them to the leftmost column.
1. Kyoto, Japan Remove	Rio De Janeiro, Brazil	Accra, Ghana
2. Shenzhen, China Remove	Dublin, Ireland	Amsterdam, Netherlands
	Ho Chi Minh City, Vietnam	Asuncion, Paraguay
	Mexico City, Mexico	Bangkok, Thailand
	Tallinn, Estonia	Bern, Switzerland
		Bilbao, Spain

Figure 4.1: User interface of the expert survey

The user interface was designed using a drag-and-drop metaphor, where the experts were asked to drag the cities to the left "Most Similar Cities" column. All other cities were available in the middle "Similar City Candidates" and right "Remaining Cities" column. Introducing the column with a shortlist of 30 cities was necessary since going through an unordered list of 139 items is not practical for human experts, as it would have taken a long time depleting their concentration. For this reason, we added this shortlist of 30 cities which were the most similar to the base city according to the aggregation of all methods in a randomized order to ease the task for the experts without introducing bias in favor of a specific data source. Finally, the experts nevertheless had all destinations available with the right column containing all remaining 109 cities in alphabetical order. Figure 4.1 contains a screenshot of the application.

When the experts finished dragging at least ten cities to the left column, they were asked to reorder their choices to emphasize the list semantics before they could submit their final ranking. The minimum number of 10 cities was chosen to give the partial rank agreement methods sufficient information to compute meaningful results and to limit the time it takes for the experts to complete a city ranking. It also corresponds to the reality of information retrieval or recommender systems, where only a few highly relevant items are important.

By recruiting the experts from the tourism research community as well as the local experts from the tourism boards, we aimed to obtain a heterogeneous group of experts that are sufficiently knowledgeable for the task. Contacting the experts in the local tourism boards was done based on the assumption that those are the ones who know their competition best. We are confident that this rigorous sampling method ensured that both the quality and the quantity of the responses were very high, despite being an online study conducted during the Covid-19 Pandemic.

4.2.2 Data Analysis of the Expert Survey

In total, we received 164 destination rankings from the survey instrument, which we analyzed for quality and to confirm that the experts' behavior is in line with our research goals.

Despite being an expert study, the link could still be distributed over the Internet. Thus we took various precautions to protect the data quality against potential low-effort and spam submissions: we excluded responses that were completed in less than one minute and those that did not adjust the internal ordering within the results column. Furthermore, we removed responses where the experts indicated their familiarity with the city on the "Very unfamiliar" or "Unfamiliar" levels. After removing cities with only one ranking, the final data set for evaluation comprised 28, with a total of 88 rankings coming from 37 different IP addresses. The top five most characterized cities were London, UK; New York City, NY, USA; Miami, FL, USA; Barcelona, Spain; and Nice, France.

To verify that the survey worked as intended, we first analyzed the agreement of the cities chosen by the experts. Since traditional inter-rater reliability metrics operate on ratings instead of ranked lists, we choose to compute the agreement as the pairwise size of the intersection over the union (in %) of two expert rankings for the same city. To make the results of cities with a different number of expert ratings comparable, we looked at the mean value over all pairs. The agreement ranges between 11% in the case of Brussels, Mumbai, and Osaka, while it reaches up to 54% in the case of San Diego. On average, the experts' lists had an overlap of about 25%, which we consider quite good, given that the experts at most choose 10–14 cities out of 139. Agreeing on about one-fourth of the most similar destinations both shows that there is clear common ground among the experts, but also that an intangible concept such as the touristic experience cannot be determined in a purely objective way.

The second aspect we evaluated was the influence of the shortlist in the middle column (cf. Figure 4.1). Overall, 79.82% of the selected cities came from the shortlist, which we see as a confirmation that the recruited experts were serious about their task and did not only follow the ranking provided by the shortlist.

To summarize, we developed an easy-to-use web-based tool to elicit a concept we argue that a generic destination recommender system should follow. To create a high-quality data set, we recruited domain experts and analyzed the outcome of the study to confirm that our design choices for the elicitation were reasonable. The outcome is a collection of 88 rankings of the 10–14 most similar destinations to 28 cities.

4.3 Rank Agreement Metrics for Incomplete Rankings

In their original definition, the rank agreement methods introduced in Section 2.4 such as Kendall's Tau Distance [75], or Spearman's Footrule Distance [139] are defined over two complete permutations of the same finite list. This assumption does not generally hold since we were unable to characterize all cities with all data sources resulting in missing characterization of cities. We sidestepped this problem by considering only a subset of destinations that could be characterized with all methods. However, the outcome of the expert study is a collection of top-k lists, with each list containing $k >= 10$ most similar cities to the city the expert characterized. To find out which data model is the most similar to the experts' opinions, we need to modify the rank agreement methods to cope with this scenario. Concretely, this means that we need to compute the rank agreement between an expert's top-k ranking τ, where $10 \leq k \leq 139$ and a complete permutation of length 140.

4.3.1 Proposed Methods

In light of the complexities of comparing the agreement of two top-k lists, we tighten the assumptions about the two lists to obtain deterministic solutions for our concrete problem. Since we have a fixed domain of items and one of our lists is always the complete permutation, our problem is less complex than the general case as it is systematically discussed in the work of Fagin, Kumar, and Sivakumar [46, Section 3.1].

We use the following notation: each *ranked list* is a *permutation* of the set of permutations S_D of D. $rl(i)$ denotes the rank of a city i in the ranked list rl. $rl(1)$ is always the city based on which the model was created.

With our problem of comparing the agreement of a top-k list with the permutation of D, we only need to discriminate three cases, omitting the case when two items are present in one top-k list, but their relative ranking is unknown in the other list. Due to the fixed domain assumption, this case cannot happen, as the other list is always the full permutation. This results in a simpler problem without any room for uncertainty that might arise from having items that are in one top-k list but not in the other.

- Case 1: $i \in \tau$, and $rl(i) \leq k$ (the item is in the top-k list and the rank of the item in the permutation is at most k)
- Case 2: $i \in \tau$, but $rl(i) > k$ (the item is in the top-k list but the rank of the item in the permutation is greater than k)
- Case 3: $i \notin \tau$ (the item is not in the top-k list)

Using this insight, we propose variants to Kendall's Tau and Spearman's Footrule distance for top-k lists.

- **Modified Spearman's Footrule Distance**

  If a city is in the top-k list (Case 1 & 2), we can compute the distance between $\tau(i)$ and $rl(i)$ as before since all information is still available. In Case 3, we do not add any penalty since we have no information about which penalty should be applied. Thus, $F'(\tau, rl)$ is simply the footrule distance between all elements of τ and the corresponding elements in rl.

$$F'(\tau, rl) = \sum_{i=1}^{k} |\tau(i) - rl(i)|$$

- **Modified Kendall's Tau Distance**

  For a modified Kendall's Tau Distance, we again count the number of discordant pairs between τ and rl. This situation is similar to the modified footrule distance, as only the penalties from the elements of τ are applied.

$$T'(\tau, rl) = \sum_{i,j \in P} \bar{T}_{i,j}(\tau, rl),$$

  where $P = \{\{i, j\} | i \neq j \text{ and } i, j \in D\}$, and $\bar{T}_{i,j}(\tau, rl) = 1$ if i and j are in the opposite order, and 0 otherwise.

4.3.2 Example

To make the approach concrete, we show the experts' rankings for the city of Munich in Table 4.2. The first column shows the first ten cities of the Wikipedia-jaccard list. To obtain the score of a data model concerning the expert's opinion, we compute the two modified rank agreement metrics between the ranked list and the experts' partial rankings. The overall score is the mean value of the rank agreement metric of all experts and all cities. The lower part of the table shows individual values and the aggregation: in this example, the opinion of Expert 1 is quite close to the ranked list according to both metrics. According to the modified Kendall's Tau, Expert 2's ranking is closer than Expert 3, however, the modified Footrule distance is lower for Expert 3. This is due to the potentially exotic choices of Expert 2 including Dubai (rank 48 in the ranked list), Vancouver (rank 54), and Boston (rank 84), which are heavily penalized in Footrule distance.

The final score of a data model according to one of the metrics is computed by the mean value of all expert rankings over all cities.

4.4 Results

We evaluate our work in three ways: first, we conduct an exploratory analysis using pairwise comparisons of the ranked lists to capture commonalities between them. Second, we compare each data model against the top-k lists that encode the expert-elicited concept, and finally, we present the results of a black-box optimization of selected data sources against the expert-elicited concept.

Table 4.2: Three expert opinions in the city of Munich are contrasted with the WP-jaccard ranked lists. The ranking of Expert 1 is closer to the ranked list than the two others.

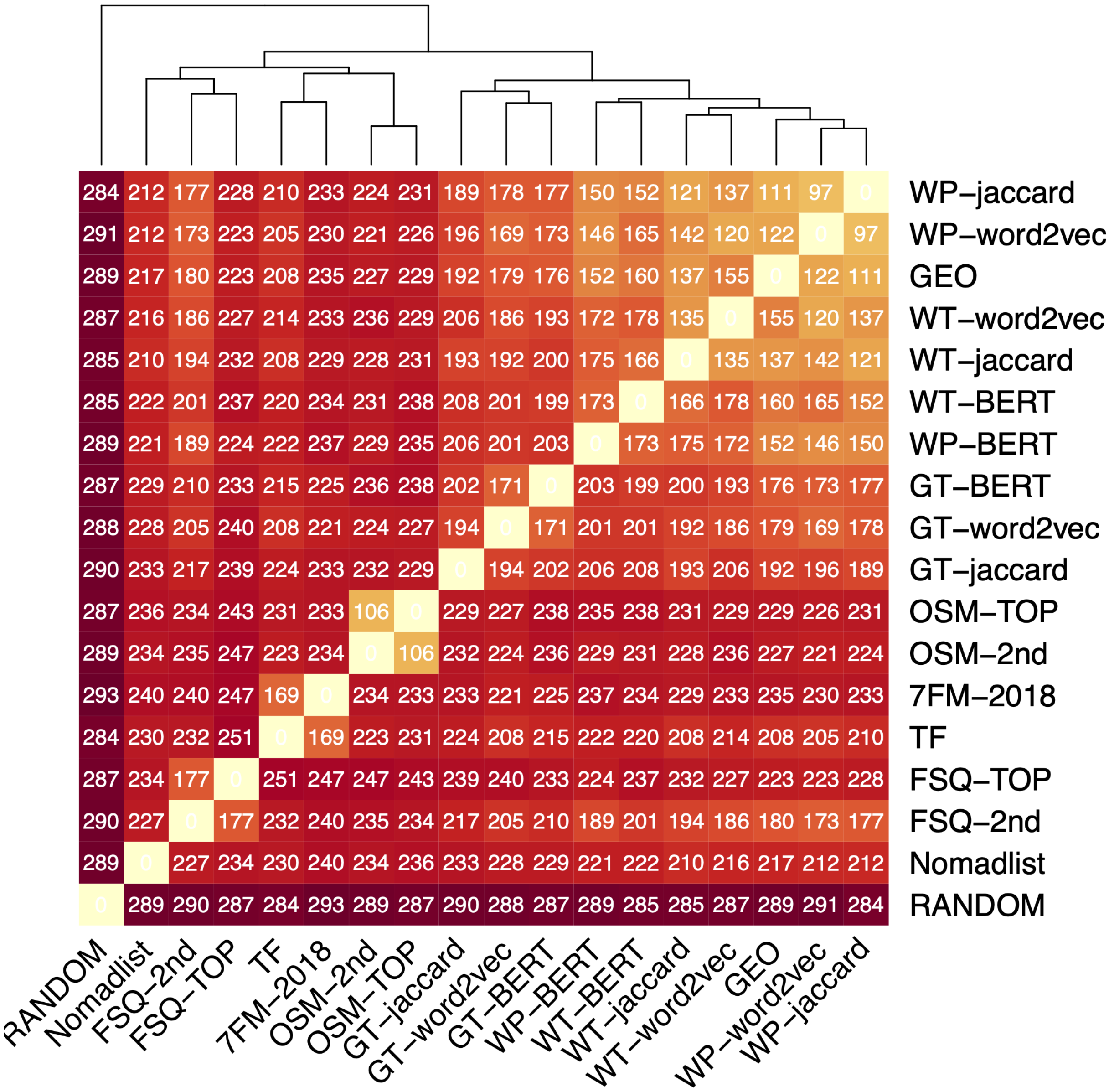
	WP-jaccard	**Expert 1**	**Expert 2**	**Expert 3**
1	Vienna	Salzburg	Vienna	Frankfurt
2	Dusseldorf	Vienna	Milan	Brussels
3	Leipzig	Cologne	Dusseldorf	Heidelberg
4	Berlin	Graz	Paris	Budapest
5	Frankfurt	Milan	Boston	Hamburg
6	Heidelberg	Edinburgh	Luxembourg	Barcelona
7	Cologne	Dusseldorf	Berlin	Vienna
8	Nuremberg	Hamburg	Cologne	Prague
9	Salzburg	Amsterdam	Vancouver	Berlin
10	Copenhagen	Brussels	Dubai	Rome
$F'(\tau, m)$	$\overline{x} = 233.67$	146	292	263
$T'(\tau, m)$	$\overline{x} = 16.33$	14	15	20

4.4.1 Assessing the Similarity of Data Sources

As an initial comparison of our data sources with an emphasis on detecting commonalities, we compare the data sources for each city against each other using Kendall's Tau in the original formulation for full permutations. The result is visualized as a heat map in Figure 4.2, here the color is determined by the mean pairwise distances among all ranked lists derived from the data models, and the sort order is adjusted using hierarchical clustering using the Euclidean Distance, which is also the basis of the dendrogram on the top. The values in the cells are Kendall's Tau distance, rounded to integers.

At first glance, one can see RANDOM being clearly separated from all other data models since it has no correlation to any of them. The first group is the family of textual data sources together with GEO. The very close grouping of GEO with all models stemming from Wikipedia and Wikitravel – irrespective of the text processing method – can be explained with the amount of geographic information that is encoded within the articles describing the cities. Nomad List seems to be unrelated to any other data source in particular but, unlike RANDOM, still has a low correlation to all other data models. Recall that the Nomad List platform is about digital nomadism, which is a form of travel, but certainly, the platform's features are not aimed at ordinary tourists resulting in a somewhat orthogonal result. The high agreement between the Webologen features (TF) and 7FM-2018 is interesting because it shows that the tourism facts are still manifested in the Seven Factor Model representation of the destinations.

This analysis helped to get a broad overview of the data sources and their commonalities. The hierarchical clustering grouped the data models in well-comprehensible families; however, the pairwise comparisons also revealed that some models that one could have expected to be quite similar, such as the top-level and second-level aggregation of Foursquare, are indeed not that similar, likely due to the large branching factor of the category tree. The benefit of this analysis is that an analyst can quickly recognize whether data models capture similar concepts to make a decision if they can be interchanged in case them being highly correlated.



	RANDOM	Nomadlist	FSQ–2nd	FSQ–TOP	TF	7FM–2018	OSM–2nd	OSM–TOP	GT–jaccard	GT–word2vec	GT–BERT	WP–BERT	WT–BERT	WT–jaccard	WT–word2vec	GEO	WP–word2vec	WP–jaccard
WP–jaccard	284	212	177	228	210	233	224	231	189	178	177	150	152	121	137	111	97	0
WP–word2vec	291	212	173	223	205	230	221	226	196	169	173	146	165	142	120	122	0	97
GEO	289	217	180	223	208	235	227	229	192	179	176	152	160	137	155	0	122	111
WT–word2vec	287	216	186	227	214	233	236	229	206	186	193	172	178	135	0	155	120	137
WT–jaccard	285	210	194	232	208	229	228	231	193	192	200	175	166	0	135	137	142	121
WT–BERT	285	222	201	237	220	234	231	238	208	201	199	173	0	166	178	160	165	152
WP–BERT	289	221	189	224	222	237	229	235	206	201	203	0	173	175	172	152	146	150
GT–BERT	287	229	210	233	215	225	236	238	202	171	0	203	199	200	193	176	173	177
GT–word2vec	288	228	205	240	208	221	224	227	194	0	171	201	201	192	186	179	169	178
GT–jaccard	290	233	217	239	224	233	232	229	0	194	202	206	208	193	206	192	196	189
OSM–TOP	287	236	234	243	231	233	106	0	229	227	238	235	238	231	229	229	226	231
OSM–2nd	289	234	235	247	223	234	0	106	232	224	236	229	231	228	236	227	221	224
7FM–2018	293	240	240	247	169	0	234	233	233	221	225	237	234	229	233	235	230	233
TF	284	230	232	251	0	169	223	231	224	208	215	222	220	208	214	208	205	210
FSQ–TOP	287	234	177	0	251	247	247	243	239	240	233	224	237	232	227	223	223	228
FSQ–2nd	290	227	0	177	232	240	235	234	217	205	210	189	201	194	186	180	173	177
Nomadlist	289	0	227	234	230	240	234	236	233	228	229	221	222	210	216	217	212	212
RANDOM	0	289	290	287	284	293	289	287	290	288	287	289	285	285	287	289	291	284

Figure 4.2: Crosswise-analysis of all data models using Kendall's Tau method. The entries are sorted using hierarchical clustering; the dendrogram reveals families of data sources.

4.4.2 Comparison with the Touristic Experience

Finally, we get to answer which characterization method would be most suited to use within a content-based information retrieval system such as a destination recommender. Having elicited the concept of "*similar experience when visiting cities as a tourist*" with the expert study, we can now compare the partial rankings of the experts with our characterization methods. We use the two modified metrics from Section 4.3.1 that compare a full permutation with a top-k list and also tabulate the MRR and Precision as comparison baselines. Note that the MRR and Precision do not capture the internal rankings provided by the experts. To compute them, we treated the rankings provided by the experts as a set and aggregated the metrics over all cities included in the lists, treating each element as an individual query.

Generally, the results in Table 4.3 confirm the picture of Figure 4.2: the versions that have shown to be similar there also rank similarly in the comparison to the expert ranking.

Table 4.3: Ranking of the different data sources using the modified rank agreement methods for top-k lists as well as MRR and Precision.

Spearman's FR top-k		Kendall's Tau top-k		Mean Reciprocal Rank		Precision@5		Precision@10	
WP-jaccard	297.011	GEO	18.284	WP-jaccard	0.101	WP-word2vec	0.186	WP-jaccard	0.304
GEO	304.091	WP-jaccard	18.750	WP-word2vec	0.101	WT-word2vec	0.182	GEO	0.302
WP-word2vec	318.080	WP-word2vec	19.068	WT-word2vec	0.100	WP-jaccard	0.178	WT-jaccard	0.297
WT-word2vec	322.489	WT-jaccard	19.227	FSQ-2nd	0.094	GEO	0.162	WT-word2vec	0.297
WT-jaccard	330.057	WP-BERT	19.568	WP-BERT	0.093	WT-jaccard	0.161	WP-word2vec	0.291
FSQ-2nd	330.420	GT-word2vec	19.852	GEO	0.093	FSQ-2nd	0.154	WP-BERT	0.279
WP-BERT	343.307	WT-word2vec	20.011	WT-jaccard	0.093	WP-BERT	0.147	FSQ-2nd	0.264
GT-word2vec	346.955	FSQ-2nd	20.625	GT-word2vec	0.088	GT-word2vec	0.139	GT-word2vec	0.263
TF	395.841	WT-BERT	20.818	WT-BERT	0.081	WT-BERT	0.139	WT-BERT	0.243
GT-BERT	396.375	TF	21.409	TF	0.075	TF	0.113	TF	0.231
WT-BERT	402.159	GT-BERT	21.477	GT-BERT	0.074	GT-BERT	0.103	GT-BERT	0.202
GT-jaccard	408.943	GT-jaccard	21.864	GT-jaccard	0.067	OSM-2nd	0.099	OSM-2nd	0.195
7FM-2018	457.909	OSM-2nd	22.375	7FM-2018	0.065	Nomadlist	0.092	GT-jaccard	0.187
Nomadlist	461.830	Nomadlist	22.420	OSM-2nd	0.065	OSM-TOP	0.090	7FM-2018	0.187
FSQ-TOP	506.500	FSQ-TOP	22.864	Nomadlist	0.063	GT-jaccard	0.087	OSM-TOP	0.180
OSM-TOP	516.114	7FM-2018	22.966	OSM-TOP	0.063	7FM-2018	0.086	Nomadlist	0.169
OSM-2nd	521.273	OSM-TOP	23.045	FSQ-TOP	0.054	FSQ-TOP	0.060	FSQ-TOP	0.122
RANDOM	649.398	RANDOM	23.341	RANDOM	0.039	RANDOM	0.033	RANDOM	0.073

The textual data models derived from Wikipedia, Wikitravel, and Google Travel, as well as the geographic location, performed best, followed by the 2nd-level aggregation of Foursquare and the factual ones. OSM and the Foursquare top-level categories conclude the ranking with the random model, unsurprisingly, performing worst.

The general stability of the ranking among the rank agreement metrics is high, but the absolute values of the data sources and the random baseline are quite close in some metrics, which we attribute to the small signal-to-noise ratio in the data: the rankings have only been computed based on 10 – 14 items out of 140.

Depending on the perspective, there are many nuances in the results that are noteworthy, which we elaborated on in the original publication [39]. Our main take on this is that there are notable differences between the data sources, and developers of recommender systems should choose ones that best emulate the recommender system's domain. The results of our study reveal that data models of previously deployed destination recommender systems [129, 100] are outperformed concerning the concept of touristic experience. This is a concern, as it reveals a mismatch between the features and the user's understanding of the destinations, whereas, e.g., textual data sources seem to do a better job of encoding the characteristics of a city.

To summarize, the proposed rank agreement metrics for top-k lists have been successfully employed in determining the quality of the data sources with respect to the expert-elicited concept. They produce comparable rankings as established information retrieval metrics, such as MRR and Precision. The advantage is that the rank agreement metrics operate on ranked lists instead on sets, making them conceptually more adequate to employ than MRR, Precision, or similar metrics.

4.4.3 Optimization of Data Models

With this tooling established, there is now potential to refine existing data models based on tourism facts and the venue distributions. The idea is that by learning the importance of the respective features, the target concept can be better emulated. By assigning different weights to the features based on their importance in computing similarity metrics, rich models with several features can be fine-tuned towards the expert-elicited concept. This is useful since standard similarity metrics in content-based recommendation, such as the Euclidean Distance, give the same weight to all features. Naturally, one can only adjust weights if the features are explicitly present. For this reason, the textual data sources and the geographic distance cannot be optimized since there are no directly usable features that can be weighted differently.

Given the combinatorial explosion of the search space for weights, we have used black-box learning, namely Simulated Annealing [76], for tuning the weights $[0, 1]$ of the data sources with explicit features. The proprietary TF and 7FM-2018 sources were only provided to us as rankings; thus, we could not optimize those.

Table 4.4: Optimization towards the Expert Opinion using Spearman's Footrule top-k

Model	Unoptimized	Optimized	Improvement
Nomadlist	461.83	426.27	7.70%
FSQ-TOP	506.50	503.33	0.63%
FSQ-2nd	330.42	312.47	5.43%
OSM-TOP	516.11	508.38	1.50%
OSM-2nd	521.27	490.33	5.94%

The optimization tabulated in Table 4.4 works better with more features, as can be seen with Nomadlist, FSQ-2nd, and OSM-2nd. We attribute the small relative changes in FSQ-TOP and OSM-TOP to the fact that they capture slightly orthogonal concepts to the expert-elicited baseline, and due to their smaller number of features, they are harder to optimize toward this concept. However, since the domain has a high signal-to-noise ratio, these small relative improvements become relevant in the overall comparison. Concretely, the optimized version of FSQ-2nd would be the third most competitive data model in Table 4.3.

Furthermore, even with minor contributions to the overall performance, the learned weights for the features give further analytic insight into their importance. Features with a very low weight could be dropped, while the feature selection of a potential combined data model of several data sources should be guided by the learned weights.

The results of this in-depth analysis of the weights are certainly quite specialized with respect to the target concept and the intricacies of the respective data sources. Thus, we do not further elaborate on the other optimized models but refer the reader to the original publication [39]. Finally, we want to emphasize the generalizability of the methods to further domains, where the data source's features are known, and a baseline exists in the form of ranked lists.

4.5 Summary

This chapter raises concerns about the underlying data used in content-based destination recommender systems. Motivated by the question of model choice in destination recommender systems, we proposed methods to make such data models of destinations comparable against each other and against a – potentially latent – concept that the recommender system should emulate when computing content-based recommendations.

In our evaluation, we collected data from various online data sources and instantiated 18 variants of possible city characterizations. We evaluated the commonalities of the data sources, through which it became apparent that, for example, articles about destinations on Wikipedia and Wikitravel encode much geographic information.

Furthermore, we conducted an expert study that provided us with partial rankings of similar destinations, which we could use to assess how well each data model approximates the touristic experience of a city using variants of rank agreement metrics. According to the expert opinion, the touristic experience was best approximated using the textual similarities from Wikipedia, Wikitravel, and the geographic location. This means that when simply retrieving the most similar destinations according to the touristic experience, one can choose one of the top-ranked entries from Table 4.3.

Finally, we could show that it is possible to optimize the distance metric of a content-based recommender system towards a desired concept if the features characterizing the city are available.

From a recommender systems research perspective, the results show that existing destination recommender systems do not necessarily use data models that capture the concept of similar touristic experience very well. This might be intentional if the system's purpose is to capture a different concept or possibly due to the previous lack of a concrete instantiation of the concept. A limitation of the top-ranked textual or geographic characterizations is that they do not come with specific features the user can interact with. This is a drawback since it means that they cannot directly place the user's preferences and the items in a common vector space to perform content-based recommendation as frequently done in travel recommender systems [19]. Furthermore, standard recommendation techniques such as critiquing [25], i.e., giving a system feedback about the features of a suggested item, are only possible if a fixed number of features characterize the items.

Our work has provided the community with adequate tools to optimize feature-based data models towards a desired concept, such as the similar touristic experience. The methodological contribution is not limited to recommender systems in the tourism domain but can be applied in other domains similarly as the proposed metrics operate on ranked lists. Latent similarity concepts are prevalent in many domains such as music [158] or leisure activities [15]; generally anywhere, where the accuracy of the information retrieval system depends on the embeddings of items in a search space.

In the next chapter, we present a content-based destination recommender system that relies on the conversational interaction of the user with explicit features describing the destinations. Thus, this section is an important basis for determining which data model to choose in such a system.

5 Navigation by Revealing Trade-offs

Nowadays, much of the online content served to users is the outcome of some more or less personalized recommendation algorithm. Irrespective of the web page or application, providers try to optimize the shown items to improve online metrics, such as click-through, user retention, or other, possibly very specialized goals [59]. Inspecting and consuming the items, users are typically unaware of what these recommendations are based on and have little means to adjust the presented items. From an algorithmic perspective, such systems have gained impressive performance in finding suitable items for a user from sparse rating matrices. Therefore, they are widely used in recommender systems for online content, such as news, products, or music. These personalized algorithms help show relevant items to users, thereby hiding the plethora of items that are likely not of interest. While this mitigates information overload and decreases the need for expensive content curation, there are various scenarios where such algorithms are not well suited for satisfying users' needs.

In this chapter, we investigate a different type of recommender systems, where the users explicitly interact with the system to fulfill an information need to make a conscious decision about which items to choose. The aforementioned classic recommendation algorithms rely on the availability of interaction data, which is not always available. In such a case, these systems typically fall back to various non-personalized baseline strategies, such as recommending popular items or curated content. Thus, in a complex domain, such as travel and tourism, the algorithmic advances of the previous decades are of lesser value. When recommending destinations, there are no clear interaction signals simply because items are not so well defined in terms of their scope. As we have discussed in Section 3.4, the reality of travel regions is that they do not necessarily match administrative boundaries. Furthermore, the travel market is segmented into regional tourism boards, each promoting the interests of their own members. The consequence is that destination marketing information is biased toward influencing users to visit their own destination. This federated structure on the business side also means that there is no source for meaningful ratings for destinations. How can a hypothetical rating of 4.5 of one city be compared with a 3.6 in another continent?

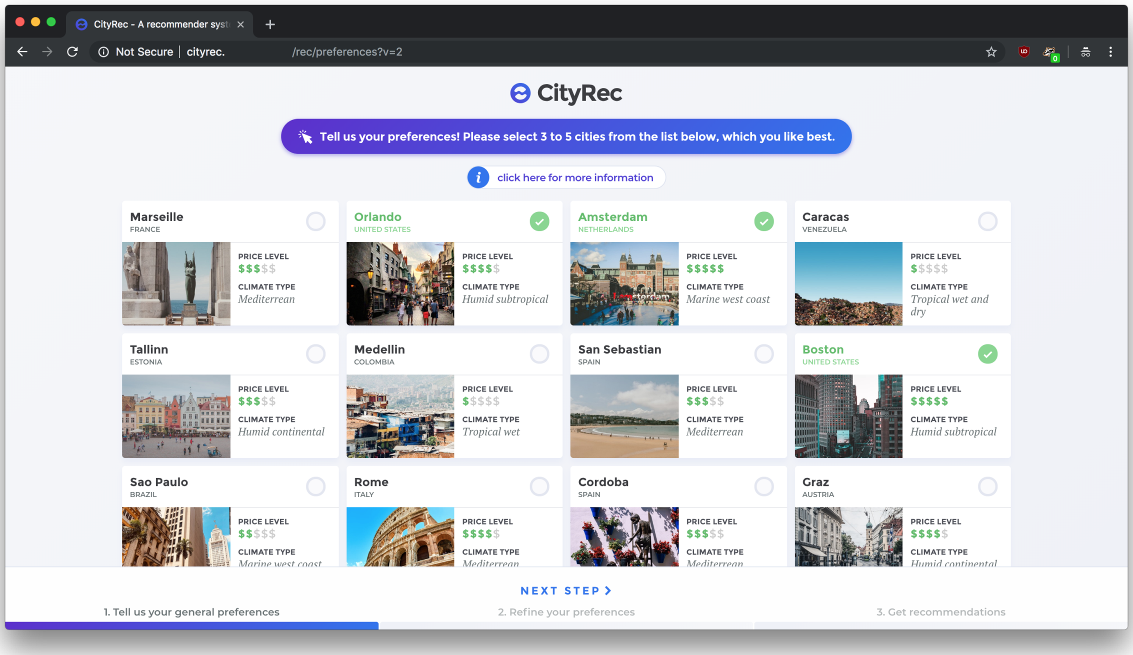
Given that traveling is a relatively rare phenomenon and a high-stakes recommendation domain [19], it has also been shown that users demonstrate different decision-making behavior compared to purchasing physical products [150]. These challenges necessitate employing sophisticated preference elicitation strategies, and content-based recommendations have been shown to be more useful in assessing the suitability of an item. Since users often struggle to declare their true preferences [130], the system should provide users with the opportunity to familiarize themselves with the items in the domain and support the decision-making process by allowing them to actively refine their preferences. Thus, the problem of recommending a city for travel is a perfect fit for the conversational, content-based recommendation paradigm.

Throughout this chapter, we describe the CityRec system, a conversational recommender system (RS) that performs content-based recommendations of destinations based on features evaluated in Chapter 4. Following up on the observation that critiques with concrete examples can be useful [147], we compare traditional unit critiquing [18] with a more sophisticated approach, which informs users about the trade-offs involved in their critiquing choices.

The primary motivation is to overcome the “wishful thinking” problem in such exploratory information systems: having an image of a “dream vacation” on their mind, users will specify their preferences accordingly; however, these possibly naïve specifications can easily define the empty set. Thus, a system should show users the various trade-offs involved, such as popularity and crowdedness, quality and price. To overcome this problem, we present a novel concept to navigate the item space that we call *“Navigation by Revealing Trade-offs.”* Unlike much of recommender systems research which either focuses on the interface or the algorithms, this paradigm requires a seamless integration of the user interface and corresponding recommendation algorithms.

5.1 User Interaction

The user interface is designed to assist the user in expressing and refining their travel preferences using concrete destinations and features that characterize these items.



CityRec - A recommender syst
Not Secure | cityrec. /rec/preferences?v=2
CityRec

Tell us your preferences! Please select 3 to 5 cities from the list below, which you like best.

click here for more information

Marseille ○
FRANCE
PRICE LEVEL $$$$$
CLIMATE TYPE *Mediterrean*

Orlando ✓
UNITED STATES
PRICE LEVEL $$$$$
CLIMATE TYPE *Humid subtropical*

Amsterdam ✓
NETHERLANDS
PRICE LEVEL $$$$$
CLIMATE TYPE *Marine west coast*

Caracas ○
VENEZUELA
PRICE LEVEL $$$$$
CLIMATE TYPE *Tropical wet and dry*

Tallinn ○
ESTONIA
PRICE LEVEL $$$$$
CLIMATE TYPE *Humid continental*

Medellin ○
COLOMBIA
PRICE LEVEL $$$$$
CLIMATE TYPE *Tropical wet*

San Sebastian ○
SPAIN
PRICE LEVEL $$$$$
CLIMATE TYPE *Mediterrean*

Boston ✓
UNITED STATES
PRICE LEVEL $$$$$
CLIMATE TYPE *Humid subtropical*

Sao Paulo ○
BRAZIL
PRICE LEVEL $$$$$
CLIMATE TYPE *Marine west coast*

Rome ○
ITALY
PRICE LEVEL $$$$$
CLIMATE TYPE *Mediterrean*

Cordoba ○
SPAIN
PRICE LEVEL $$$$$
CLIMATE TYPE *Mediterrean*

Graz ○
AUSTRIA
PRICE LEVEL $$$$$
CLIMATE TYPE *Humid continental*

NEXT STEP ›
1. Tell us your general preferences
2. Refine your preferences
3. Get recommendations

Figure 5.1: The initial preference elicitation screen.

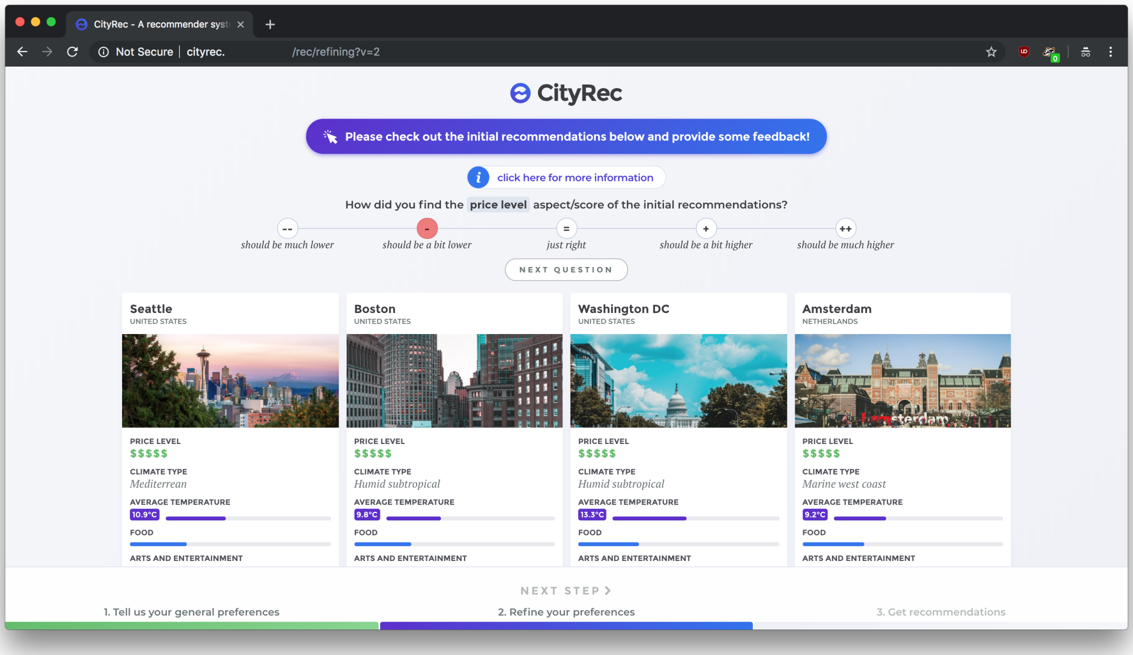
5.1.1 Initial Preference Elicitation

In all our experimental conditions, we used an initial screen that showed 12 cities from the item space (cf. Figure 5.1). This initial seed of 12 destinations is not random but a diverse representation of the data set. Thus, it is a first step for the user to get familiar with the items available for selection. The diversity is guaranteed since the cities are selected from groups derived from a cluster analysis on the available features [35]. The optimal result of the cluster analysis was obtained with five groups, which are used to generate numerous, diverse, but equivalent shortlists because each cluster is represented at least twice. To get an initial user profile, we ask the user to choose three to five that best reflect their preferences.

This selection provided all subsequent algorithms with a starting point for the conversational exploration of the search space.

5.1.2 Unit Critiquing

The original CityRec system [35] would now show a set of four destinations, which are closest to the current user model according to the similarity metric for the cities (see Section 5.2.1). The users have control over their preference profile by critiquing the features one after another with different intensities from *"much lower"* — *"lower"* — *"just right"* — *"higher"* to *"much higher."*


CityRec - A recommender syst
Not Secure | cityrec.
/rec/refining?v=2
CityRec
click here for more information
NEXT QUESTION
NEXT STEP ›
1. Tell us your general preferences
2. Refine your preferences
3. Get recommendations

Please check out the initial recommendations below and provide some feedback!

How did you find the price level aspect/score of the initial recommendations?

○ -- *should be much lower* ● - *should be a bit lower* ○ = *just right* ○ + *should be a bit higher* ○ ++ *should be much higher*

Seattle
UNITED STATES
PRICE LEVEL
$$$$$
CLIMATE TYPE
Mediterrean
AVERAGE TEMPERATURE
10.9°C
FOOD
ARTS AND ENTERTAINMENT

Boston
UNITED STATES
PRICE LEVEL
$$$$$
CLIMATE TYPE
Humid subtropical
AVERAGE TEMPERATURE
9.8°C
FOOD
ARTS AND ENTERTAINMENT

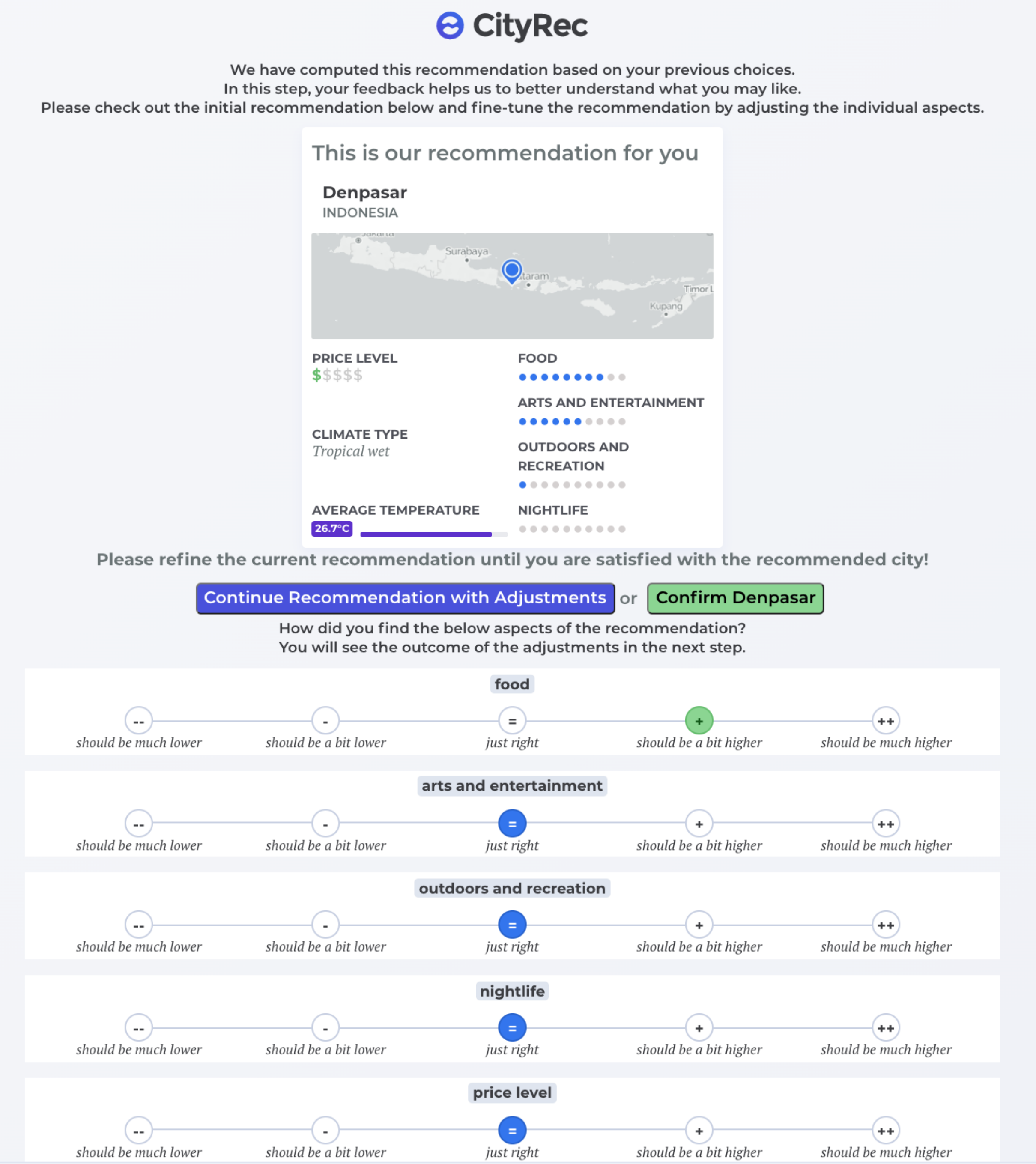
Washington DC
UNITED STATES
PRICE LEVEL
$$$$$
CLIMATE TYPE
Humid subtropical
AVERAGE TEMPERATURE
13.3°C
FOOD
ARTS AND ENTERTAINMENT

Amsterdam
NETHERLANDS
PRICE LEVEL
$$$$$
CLIMATE TYPE
Marine west coast
AVERAGE TEMPERATURE
9.2°C
FOOD
ARTS AND ENTERTAINMENT

Figure 5.2: The original refinement screen. [35]

As seen in Figure 5.2, the user now has more information about the cities, enabling her to make informed decisions. Using this feedback, the system would update the user profile scores by -0.2, -0.1, 0, 0.1, or 0.2 and show an updated set of cities in the next step.

In the initial publication [35], this unit critiquing system was evaluated against a non-critiquing baseline. For the main evaluation, we further developed it to be comparable with the other conversation algorithms: we removed all images of cities, as they might introduce a bias by shifting the users' attention away from the features. Furthermore, it is very hard to find images that adequately capture the multitude of aspects a city has [131].



CityRec

We have computed this recommendation based on your previous choices.
In this step, your feedback helps us to better understand what you may like.
Please check out the initial recommendation below and fine-tune the recommendation by adjusting the individual aspects.

This is our recommendation for you

Denpasar
INDONESIA

PRICE LEVEL
$$$$$

FOOD
●●●●●●●●○○

ARTS AND ENTERTAINMENT
●●●●●●○○○○

CLIMATE TYPE
Tropical wet

OUTDOORS AND RECREATION
●○○○○○○○○○

AVERAGE TEMPERATURE
26.7°C

NIGHTLIFE
○○○○○○○○○○

Please refine the current recommendation until you are satisfied with the recommended city!

Continue Recommendation with Adjustments or Confirm Denpasar

How did you find the below aspects of the recommendation?
You will see the outcome of the adjustments in the next step.

food

-- *should be much lower* | - *should be a bit lower* | = *just right* | [✓ + *should be a bit higher*] | ++ *should be much higher*

arts and entertainment

-- *should be much lower* | - *should be a bit lower* | [✓ = *just right*] | + *should be a bit higher* | ++ *should be much higher*

outdoors and recreation

-- *should be much lower* | - *should be a bit lower* | [✓ = *just right*] | + *should be a bit higher* | ++ *should be much higher*

nightlife

-- *should be much lower* | - *should be a bit lower* | [✓ = *just right*] | + *should be a bit higher* | ++ *should be much higher*

price level

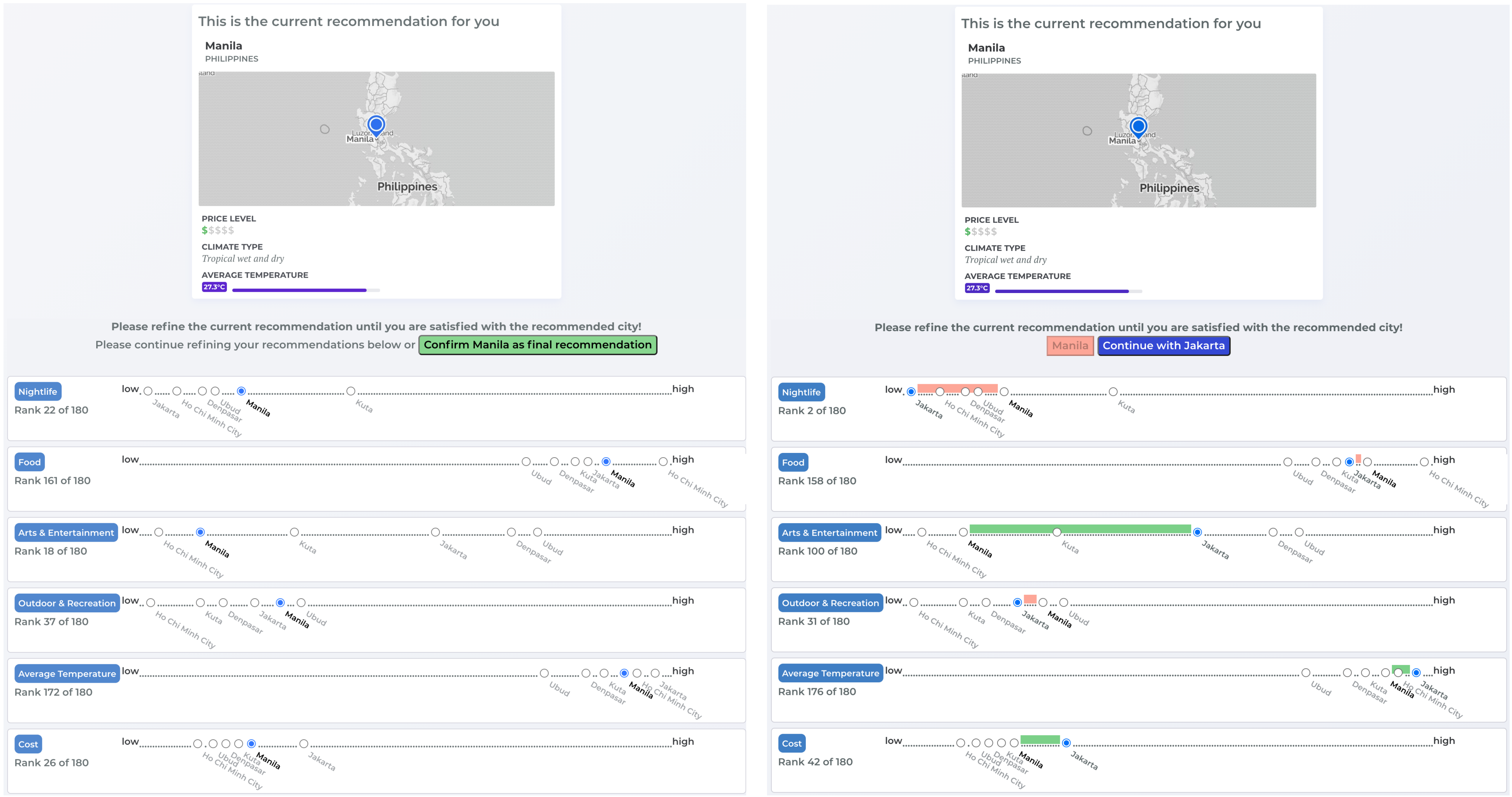
-- *should be much lower* | - *should be a bit lower* | [✓ = *just right*] | + *should be a bit higher* | ++ *should be much higher*

Figure 5.3: The refinement screen in the user study. [36]

Figure 5.3 shows the consolidated unit user interface for the large-scale user study. Now only one city is shown, and each feature can be critiqued simultaneously. Furthermore, the user can take unlimited critiquing cycles until she is satisfied with the outcome.

5.1.3 Revealing Trade-offs

Figure 5.4 shows the interface element for our conversational *"Navigation by Revealing Trade-offs"* approach. At the top of the page, the currently recommended city is shown; below is the novel user interface. This component shows the current city along with five other cities recommended based on the utility function. For each feature, the five candidate items are shown in an ordered list from low to high, depending on the



This is the current recommendation for you

Manila
PHILIPPINES

PRICE LEVEL
$$$$$

CLIMATE TYPE
Tropical wet and dry

AVERAGE TEMPERATURE
27.3°C

Please refine the current recommendation until you are satisfied with the recommended city!
Please continue refining your recommendations below or Confirm Manila as final recommendation

Nightlife
Rank 22 of 180

Food
Rank 161 of 180

Arts & Entertainment
Rank 18 of 180

Outdoor & Recreation
Rank 37 of 180

Average Temperature
Rank 172 of 180

Cost
Rank 26 of 180

(a) Preference Refining Page. This is shown to the user at the beginning of each conversational cycle. The current city is marked in bold, and five alternatives are displayed on the spectrum of each feature.

This is the current recommendation for you

Manila
PHILIPPINES

PRICE LEVEL
$$$$$

CLIMATE TYPE
Tropical wet and dry

AVERAGE TEMPERATURE
27.3°C

Please refine the current recommendation until you are satisfied with the recommended city!
Manila Continue with Jakarta

Nightlife
Rank 2 of 180

Food
Rank 158 of 180

Arts & Entertainment
Rank 100 of 180

Outdoor & Recreation
Rank 31 of 180

Average Temperature
Rank 176 of 180

Cost
Rank 42 of 180

(b) Trade-off Visualization. The green and red shades indicate the trade-off involved should the user choose Jakarta instead of Manila. The user explores various alternatives before continuing.

Figure 5.4: User Interface of Navigation by Revealing Trade-offs

score. Users can select an item to see the feature value differences in all feature spaces compared to the currently recommended city. An increase in feature value is indicated using a green shade; a decrease is shown in red.

If the user is satisfied with the current recommendation, the user can choose not to continue with refining but to confirm the current recommendation. In this case, the user is forwarded to the final recommendation page.

5.2 Algorithms

Having discussed the front-end of CityRec, we now turn our attention to the algorithmic solutions that drive the exploration of the search space.

5.2.1 Content-based Filtering

In content-based RS, all recommendations are computed based on the current user model and the features that characterize the respective items. The typical distance metric is the Euclidean Distance based on the vector space defined by the destination features. The Navigation by Revealing Trade-offs approach also operates in the same vector space, however, the goal of this interface concept is to present the user with various candidates that drive the exploration through the search space in a more efficient and effective way. The following section describes the proposed candidate selection strategies to progress the search for suitable recommendations.

5.2.2 Candidate Selection Strategy

One issue with a purely content-based recommendation strategy is that it does not consider the user preference variations during the refinement. Furthermore, a simple distance metric will return the most similar items to a query, which hinders the exploration of the search space. Thus, we propose a candidate selection strategy, which we call the "Variance Bi-distribution" utility function, to enable the exploration of the search space towards items the user is interested in. This utility function (Equation 5.3) is defined by two normal distributions per feature, each representing an increase or decrease of feature value. The two normal distributions are given as $\backsim N(\mu_1, \sigma^2)$ and $\backsim N(\mu_2, \sigma^2)$, where μ_1 and μ_2 define the position of the bell curves on the normalized value range of the feature, and σ defines the shape of the curve.

The distance between the currently selected reference item, ref_k, and the respective bell curves are computed by adding or subtracting an offset computed in Equation 5.1. This offset between ref_k and μ is the standard deviation of each feature value f of all previous items in the conversational history H by the number of previous conversational iterations n moderated by a constant C_m, which can be empirically determined for each data set. The value of C_m depends on the size of the data set as well as the distribution of the feature values. Intuitively, it can be regarded as a moderator for the step size similar to the learning rate in machine learning [69].

The behavior of this aspect of the algorithm can be summarized as follows: the mean of the normal distribution is farther from the currently selected items if the variance of a feature is higher, however, this effect becomes less with each conversational cycle.

$$\mu_1 = ref_k - \frac{\sqrt{Var(f \in H)} \cdot C_m}{n} \quad \mu_2 = ref_k + \frac{\sqrt{Var(f \in H)} \cdot C_m}{n} \tag{5.1}$$

The second parameter of the normal distributions, σ, is computed in a similar way (cf. Equation 5.2). This has the effect that with higher variance, we obtain a flatter distribution and, thus, a lower impact of this feature on the utility score.

$$\sigma = \frac{\sqrt{Var(f \in H)} \cdot C_s}{n} \tag{5.2}$$

The intuition behind this is that if the user has a strong preference regarding a feature having a certain value and consistently picks cities with a high temperature, the system is quite certain of this user's preference toward temperature and, thus, should put high weight on this feature. Conversely, suppose a user has selected cities with another feature having both low and high values resulting in a high variance. In that case, it can be seen as a signal that the user has no specific preferences toward the feature as it is unimportant to the user. Thus, the impact of such a high-variance feature should be smaller than a low-variance feature. Over time, we increase this effect by dividing by the number of previous iterations n. This further helps the algorithm converge.

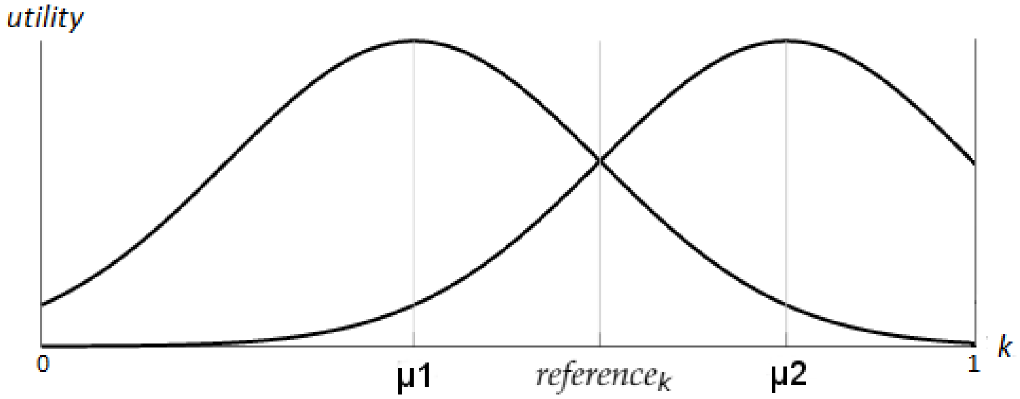


Figure 5.5: Variance Bi-Distribution

The maximum score of the two distribution functions for a given item feature is taken as the utility score of the respective feature. This is visualized in Figure 5.5: the two bell curves have their maxima shifted to the left and right from the current city (reference$_k$), giving cities whose feature values are closer to μ_1 and μ_2 a higher utility weight for this feature.

We then compute the overall utility of each item as the sum of all feature scores of the utility function.

$$\text{utility} = \sum_{f \in F} s(f) \tag{5.3}$$

The effect of this utility function is that it balances fast exploration in the beginning and fine-tuning in later stages of the search. If a feature variance is high and the number of iterations small model adjusts μ_1 and μ_2 further away from the reference item, with a higher σ resulting in a flatter distribution of the feature's utility function. In this case, items far away in the feature space also would get higher utility scores, ensuring users are presented with cities more spread across the feature space. With a larger number of iterations, the user preferences for particular features are converging, i.e., the user will be presented with an increasingly narrower band of feature values to refine the preferences. As a result, μ_1 and μ_2 are closer to the feature value of the current recommended item, and with a smaller σ, items with similar feature values receive substantially higher utility scores than the items with dissimilar feature values. However, if a feature's variance is still high, the curve will stay quite flat, giving this feature less weight, thus recognizing that the user is rather indifferent toward this feature. This convergence behavior can be observed in Figure 5.4 back on page 55. After some iterations, the algorithm determined that the user has a clear preference for high scores in the food and temperature aspects and low scores in nightlife, outdoor & recreation, and cost. Thus, the refining candidates are quite close to each other, whereas they are spread along the spectrum of the "Arts & Entertainment feature."

5.2.3 Elimination of Candidates

To further improve the convergence, we propose a variant that eliminates items whose feature values have been refined in a contrary way. The reasoning behind this elimination of candidates is that if a user refines a feature of an item, it becomes explicit information that the value of the feature is unsatisfactory and should take only values toward the direction of the refinement. Thus, we can compute candidates just as before, however, items with a lower (or higher) value than the original item ref_k are removed from the search space. For example, if the user refines the value of *Arts & Entertainment* of Manila in Figure 5.4b in favor of Jakarta, the system will assume that all cities that have a lower value in *Arts & Entertainment* than Manila should be excluded from future suggestions, which can quickly reduce the search space.

5.3 Evaluation & Results

A user study was set up to compare the three systems: Unit Critiquing, Trade-off Refining, and Trade-off Refining with the Elimination Variant. The experiment was conducted in December 2020 with 600 participants on the online experimentation platform Prolific[1]. After quality checks, 419 valid submissions were analyzed regarding the dependent variables, which comprised the ResQue Questionaire and the interaction behavior with the system.

5.3.1 Quantitative Analysis

Regarding the number of conversational cycles, we observed that all sessions using the Trade-off interface were finished by the users within six cycles, with a mean value of 2.38/2.46, whereas the baseline unit critiquing interface needed more cycles with a mean value of 4.44 cycles. Thus, the Trade-off UI reduced the iterations by 46.4% (44.6% in the Elimination Variant), which is a significant reduction when testing the hypothesis using a t-test.

For the survey items, we computed cross-wise Wilcoxon rank sum tests for independent populations using the three independent variables. The null hypotheses were that there

[1] https://www.prolific.co/

Table 5.1: Hypothesis testing of the dependent variables between the baseline unit critiquing and the two variants of the Trade-off Refinement. The mean values of the survey items coded as integers from 1 to 5 are for informative purposes only.

	Baseline	**Trade-offs**			**Trade-offs w. Elim.**		
Variable	mean	mean	p	w	mean	p	w
(Q1) Interest match	3.81	**4.12**	**0.002**	7378	**4.07**	**0.005**	8727.5
(Q2) Better than friend	3.26	3.25	0.939	9053.5	3.26	0.749	10212.5
(Q3) Cities are familiar	4.09	4.14	0.605	8794.5	4.22	0.187	9572
(Q4) Rec. cities are attractive	4.06	4.18	0.314	8524	4.05	0.538	10816.5
(Q5) Discover new Cities	3.66	3.76	0.42	8608	3.71	0.711	10179
(Q6) Adequate layout	**3.78**	3.45	**0.003**	10917.5	3.56	**0.044**	11765
(Q7) Easy to modify preferences	**4.14**	3.59	**< 0.001**	11846.5	3.66	**< 0.001**	13235
(Q8) Became familiar quickly	**4.19**	3.67	**< 0.001**	11681	3.60	**< 0.001**	14125
(Q9) Influenced selection	3.44	**3.64**	**0.043**	9104.5	**3.63**	**0.044**	9104.5
(Q10) Overall satisfaction	3.82	3.84	0.743	8910.5	3.77	0.534	10824.5
Number of Conversational Cycles	4.44	**2.38**	**< 0.001**	-	**2.46**	**< 0.001**	-

is no differences in the medians of the responses. Since we could not find significant differences between the Trade-off Refining and Trade-off Refining with the Elimination Variant, we only tabulated the outcomes in Table 5.1 with respect to the baseline unit critiquing. Besides the analysis of the number of conversational cycles, we could refute the null hypothesis in favor of the Trade-off Variants in (Q1) and (Q9), while the baseline received better responses in (Q6), (Q7), and (Q8). This mixed result can be summarized in a way that the Trade-off interface had superior perceived recommendation accuracy at the expense of the users' perceived ease of use.

5.3.2 Discussion

The superior perceived accuracy measured by (Q1) at about 45% fewer conversational cycles underlines the merit of our proposed user interface. However, the subjects rated the usability-related metrics of the unit critiquing system higher (Q6 – Q8). We suspect that this is because unit critiquing has already been employed in various RSs, so it is quite possible that many users were already familiar with this concept. Dealing with a new refinement interface involving reasoning about trade-offs certainly involves more cognitive effort and, thus, might need more familiarization (Q8) than only one session. The study was designed in a way that users could only submit the survey once, and we did not familiarize the users with the system before their session to avoid learning effects. The significant difference in (Q9) "This recommender system influenced my selection of cities." in favor of the Trade-off interface is likely an artifact of the comparative lengthy search in the unit critiquing since both values are in the center of the Likert Scale. Interestingly, there were no significant differences in any dependent variables between the Trade-off Refinement and its Elimination Variant. We attribute this to the low number of conversational cycles needed to come up with a satisfactory result. In the given data set of 180 cities, the elimination of candidates was probably not necessary, as the utility function was able to recommend attractive items after two or three cycles. Nevertheless, we are confident that eliminating parts of the search space based on the users' choices could be useful, and we plan to analyze the merit of the Elimination Variant with larger item sets of over 1000 items.

5.4 Summary

The success of modern recommender systems depends on the seamless integration of algorithms and user interface elements. Given that existing critiquing systems have often neglected to inform users about the trade-offs of the critiquing actions explicitly, we developed the Navigation by Revealing Trade-offs system, which integrates a user interface concept with a utility function to compute refinement candidates. The evaluation shows that perceived accuracy is better than the unit critiquing baseline at similar reductions in the number of conversational cycles as other advanced critiquing approaches have demonstrated [92, 161].

6 Conclusions

In this dissertation, we motivated, analyzed, and resolved various challenges within the field of destination recommendation. The three major parts revolved around traveler mobility analysis, destination characterization, and a conversational destination recommender system. In the following, we revisit the identified challenges from the beginning of this thesis to highlight how we resolved them, particularize potential limitations of our solutions, and indicate future research directions should be taken based on our work.

6.1 Traveler Mobility Analysis Using LBSN Data

Traditionally, researchers were limited in their abilities to analyze the individual mobility of a large number of travelers on a global scale due to the lack of suitable data. With the proliferation of location-based social networks in the last two decades, a new data source emerged which enables the analysis of global mobility patterns.

Challenge 1: How can LBSN data be used to determine the behavior of domestic and international travelers? The advantage of analyzing mobility from LBSNs is that the data is both available in sufficient quantity and the possibility to capture traveler mobility on a planet-scale. Due to the low frequency of the check-in-based data, different methods need to be used than for analyzing GPS-based data or call data records. We developed the `tripmining` library, which combines geotagged posts into trips and assesses these trips regarding their data quality and the underlying mobility characteristics of the user (Section 3.1). We were able to determine which LBSNs are suitable for which analysis and established guidelines for the quality parameters for different use cases. The library was used with various LBSN data sets within this thesis (Sections 3.2, 3.3, 3.4, 3.5) as well in further publications [34, 33, 37, 127, 121, 148]

Challenge 2: Given the behavior of travelers; which groups can be identified? In Section 3.2, we performed two cluster analyses on data sets from Twitter and Foursquare to segment trips into distinct groups using a cluster analysis approach. The results reveal three clusters of trips in the Twitter data set and six in the Foursquare data set, which enriches the mobility features with the type of venues the travelers have visited. The determined groups are well-distinguishable by their feature values and provide domain insights into how frequent different types of trips are within the observed populations. The number of clusters seems to depend on the amount of information the algorithm can use to discriminate groups. We conclude that the higher number of clusters in the Foursquare study is due to the additional social aspects available through the visited venues since the mobility features were similar in the two studies.

Performing a cluster analysis on the users' past mobility behavior can be very useful to obtain an automatic characterization of a user and, thus, increase the personalization possibilities of a travel recommender system in cold-start scenarios without requiring user interaction. Based on our findings, we encourage collecting more features for the clustering since it gives opportunity to refine the characterization towards more specialized groups.

Challenge 3: What is the effect of establishing different groups of users on POI recommendation performance? Using a similar approach as in the previous challenge, we used a cluster analysis to identify different sub-groups within groups of travelers and local users, respectively. Note that in this study, the local users were not derived using the tripmining library since this group comprised people within their home city. Guided by the hypothesis that different groups of locals and travelers who visit a city should show different behavior in visiting POIs, we quantified the effect of the user type identification on the accuracy and fairness of POI recommendations (Section 3.3). Due to the popularity bias in this domain, it turns out that it is easier to make relevant recommendations to groups of travelers even though way less training data is available for these groups of users compared to the different types of locals.

This is an important insight since it questions the validity of standard evaluation procedures in POI recommendation: we argue that travelers visiting a city have different recommendation needs compared to people who are home in a city. Yet, studies evaluating POI recommendation approaches treat all these users in the same way, reporting aggregated results over all users. By discovering sub-groups within the two main categories of travelers and locals, we were able to determine notable differences in the recommendation performance between these sub-groups.

Although this study only used one data set and analyzed five major cities, it clearly questions current practices of POI evaluation. If the problem can be confirmed with other data sets, it would be imperative to develop more rigorous evaluation practices in this domain based on our findings. One direct implication of our study is that at least locals and visitors need to be discriminated in future POI recommendation studies.

Challenge 4: How does the map of global travel regions look like? To answer this question, we transformed the mined trips into a mobility graph and ran the Infomap community detection algorithm [119] to cluster the world's cities into hierarchical travel regions (Section 3.4). We argue that using these travel regions in travel recommender systems would be more useful than using administrative regions. Since the discovered travel regions reflect which cities are often visited in combination, they align better with the user's understanding of travel destinations, where national or administrative borders are of lesser importance. The obtained world map reveals in which areas national borders are relevant and where international travel is common.

This map could be further refined with more data: we observe that more data is advantageous to form more detailed regions, and the absence of data from, e.g., China entirely prevents forming regions in such areas. In the future, it would be interesting to observe how this travel region world map evolves over time since tourism is an ever-changing phenomenon – which should be reflected in such a data-driven product.

Challenge 5: What are the recommended durations of stay at one specific destination? The mined trips also reveal how long people stay at a specific destination. By computing the distributions of stay durations, it is possible to recommend a personalized duration of stay for a specific city based on the user's past trips. Section 3.5 describes a statistical approach for this problem using LBSN data; however, this problem is worth evaluating on a larger scale with various other methods, including supervised learning.

With the development of the tripmining library, we were able to analyze various LBSN data sources regarding the mobility of the users. Although we were able to show that not all data sources are suitable, it is possible to determine this using the quality metrics of the library and make informed decisions on whether the data quality is adequate for the current use case. Indeed, it would have been interesting to analyze some of the most popular online social networks. Unfortunately, data from Facebook, Instagram, or TikTok is kept sealed by the respective companies. At the same time, the more open policies of Twitter, Foursquare, and Flickr allowed us to analyze the mobility of these platforms' users.

We showcased three major applications of collecting and analyzing trips from LBSNs. First, a cluster analysis method to find distinctive groups of trips and travelers, which in itself gives relevant domain insights and can provide the opportunity for personalization in cold-start recommendation scenarios. We used the same method to evaluate the impact of discriminating groups of users within the POI recommendation domain. Second, we proposed a method to propose an alternative map of the World's travel regions using a community detection algorithm based on the global mobility graph between 14,558 nodes. Third, we proposed a method to utilize the number of stays at different destinations to derive recommendations regarding the personalized duration of stay at a destination. The fact that the analysis of only a few available data sets could reveal so much information about travel showed that global LBSN mobility analysis is an excellent method to learn about the travel domain, and we were able to use this information to make substantial contributions to future travel recommender systems. Since we were limited to publicly available data, we expect that our methods can yield even better data if used in conjunction with proprietary large-scale data sets of travel organizations or global internet platforms that constantly track the locations of the users.

6.2 Destination Characterization

There are various approaches to characterize destinations, yet it was entirely unclear which one is the best from a domain perspective. We argue that the data model of a content-based recommender system should emulate the domain as closely as possible to enable transparent and unbiased decision-making. To the best of our knowledge, we are first to analyze this problem in the travel recommendation domain and establish a ground truth data set in the area of destination recommendation.

Challenge 6: What data sources are suitable to characterize destinations? In Section 4.1, we describe how we constructed 18 data models to characterize 140 cities using online data sources of three categories: textual data, factual data, and data models based on the distribution of venues within a destination. Our approach to compare the data models using rank agreement metrics revealed commonalities encoded within different data sources, e.g., articles about cities on Wikipedia and Wikitravel include much geographic information.

We found it surprisingly simple to construct data models to characterize destinations using open data; however, there is no systematic way of determining which features to include in a data model and which ones to omit. It would be very interesting to explore additional data sources to characterize cities in an unbiased way. Especially pictorial data and topological features would be promising to explore. Furthermore, it would again be interesting to capture the evolution of destinations as they increase or decline in popularity or in the audience they attract.

Challenge 7: How can a latent concept such as the touristic experience of a destination be elicited? As it was necessary to obtain a ground truth for evaluating the performance of the data models, we invited experts in travel and tourism to give their opinion on this concept. Since it is very difficult to elicit a ground truth for similarity of items, the proposed web-based system required careful design to support the experts in completing the difficult task without biasing them in a systematic way. Since, to the best of our knowledge, only one previous study has approximated a ground truth for item similarity in the movie domain [155], this system constitutes an important contribution to evaluation techniques in recommender systems. We argue that due to that outcome in our system are ranked lists of similar items, the approximation is more adequate than the approach with pairwise comparisons in the work of Yao and Harper [155]. We were successful in collecting 164 destination rankings, processing them, and also quantifying potentially involved biases to ensure a reliable response quality (Section 4.2).

In the future, this system could be adapted to different domains and established as a generic tool to capture ground truth information about latent concepts using skilled experts and large-scale crowd intuition.

Challenge 8: How can data models be evaluated against the expert opinion? Recall that the outcome of the expert study was a list of most similar cities to a base city. To address this special case of measuring the distance between a complete permutation and a top-k list, we proposed variants of established rank agreement metrics to enable a direct comparison between the data models and the target concept. The advantage of these metrics is that, in contrast to traditional information retrieval metrics such as Precision or Mean Reciprocal Rank, the internal ordering of the rankings is considered. Interestingly, our evaluation in Section 4.4.2 indicates that our textual sources outperform the data models collected from previous destination recommender system studies. We see this as an indication that these documents are already well targeted to inform prospective travelers about the touristic experience.

These results can also be seen as a mandate for data scientists to improve the current state of data models of content-based destination recommender systems. Generally, the quality of data models in content-based recommender systems needs to be evaluated to whether the features of the items sufficiently approximate the desired concept of the domain.

Challenge 9: To what extent can data models describing destinations be optimized to better capture the touristic experience? Without altering the input data, an obvious way to improve the data models with explicit features is to adjust the distance function of the recommender system. Commonly used distance metrics in content-based recommender systems such as the Euclidean Distance use all features with the same weight. We argue that this is unrealistic, as naturally some aspects of a city, e.g., the quality of restaurants, are more relevant for the touristic experience than others. In Section 4.4.3, we were able to show that by learning the weights, a better alignment of the recommendation function to the desired concept can be achieved, making some feature-based models competitive.

This optimization could be enhanced by selecting high-quality features from different data sources to construct a combined data model. Since we aimed to determine which data models are best, this was out of scope in our study, however, we are confident that this would be a fruitful approach in future works.

Constructing data models for content-based recommendation should not be based on intuition or by simply including all available domain features. By collecting and constructing different data models for characterizing destinations, we analyze this problem systematically using rank agreement metrics that capture the similarity of the resulting rankings. Since we were able to elicit the latent concept of the touristic experience, we could assess how well the data models from existing recommender systems and online sources perform in direct comparison. The results show that common information for travelers, such as Wikipedia articles about destinations, better captures the touristic experience; however, it is possible to learn the relative importance of respective features to make existing data models of destination recommender systems competitive.

Since there are so far few studies related to the ground truth of the data and recommendation models in recommender systems [155], it is hard to predict how much potential there is in improving data and algorithms to capture the respective domains better. Certainly, users get confused if there is a divergence of certain aspects of items with reality. Thus, mitigating such discrepancies are important in recommender systems, and we provide a toolbox of methods to do so.

6.3 Navigation by Revealing Trade-offs

We developed a novel paradigm for conversational recommendation to support prospective travelers in deciding where to travel. This paradigm, "Navigation by Revealing Trade-offs," was motivated by the "wishful-thinking" problem and helped to inform the users about the trade-offs involved in choosing one item over another. Since users directly interact with features of the items, the characterization of the destinations from Chapter 4 served as the foundation for the used data sets. The proposed paradigm is based on the seamless integration of user interface elements and algorithms to drive the algorithmic convergence towards recommendations that fulfill the users' needs.

Challenge 10: How can a user interface of a conversational recommender system be designed to enable exploration and visualize the trade-offs involved in choosing one destination over another? The user interface of the CityRec conversational recommender system informs users about the relevant features of candidate cities (Section 5.1). When users determine the direction in which they want to refine the current recommendations, the changes in the feature value are visualized so they can consider the potential downsides of choosing one item over another. Since the refining is always towards a possible item, the user can not establish a configuration that is not available in reality.

Challenge 11: How can the exploration of the search space be directed so that the users are able to understand the extent of the search space, with a gradual convergence towards the user's preferences? As described in Section 5.2, we proposed a utility function in CityRec to determine which candidate items should be presented to the user. The design of the utility function was chosen in a way that it initially allows the exploration of the search space but gradually converges towards the user's preferences with an increasing number of conversational iterations. Furthermore, it learns which features are important to the user and decreases the impact of features that are irrelevant to the user.

To evaluate the "Navigation by Revealing Trade-offs" paradigm, we conducted a large-scale user study with 600 participants on a 140 destinations data set. The results of our study show that the perceived recommendation accuracy is better than the unit critiquing baseline; however, the less complex unit critiquing system receives better usability ratings. Thus, we can conclude that the ideas that shaped the prototype have merit, which should be further investigated in subsequent studies. Results from different domains can help to understand the trade-offs of a more complex system with better accuracy against a less accurate one with better usability.

These promising results we could obtain in a between-subject design show that the paradigm combining algorithms and user interface elements works. Nevertheless, to deploy this in a commercial system, the individual aspects, especially in the user interactions, need more detailed human-computer interaction (HCI) studies to understand in what scenarios they work best and how they can be adapted to balance the business goals with the user intentions.

6.4 Perspectives on Destination Recommender Systems

Travel and tourism are among the largest industries, with many regions being economically dependent on the tourism sector. Countless destinations all around the world compete for visitors, and in today's age of digitalization, peoples' habits drift towards planning their trips online, marking a declining influence of traditional travel agencies [21, 13]. Consequently, the online ecosystem for travel and tourism provides different services to support travel-minded users. Destination marketing organizations typically promote their own regions, whereas commercial online platforms for travelers focus on tourism products that they can make a profit on through commissions such as transportation, accommodation, and venue recommendation. Despite some efforts of these platforms to support users earlier in the travel planning and booking funnel, such features have not gained widespread adoption, and recommendations are typically not personalized. This is not surprising, given the enormous complexities in travel planning and decision-making. In this thesis, we proposed various building blocks to overcome these complexities by utilizing the mobility of users, for which we could showcase various data-driven applications. We evaluated the quality of different data models for destination characterization and designed a conversational recommender system that helps users to familiarize themselves with destinations, thus, empowering them to choose travel destinations that fit their interests, even though they have not been there.

We envision that in the near future, travel planning applications will help users design personalized trips in a similar way as travel agencies have done. Since this is an immensely complex decision-making problem, it can only be solved through cooperative interaction between users and computer systems. Herein, the challenge is to win the users' trust by being transparent with respect to how the system operates [29], e.g., by providing explanations of the recommendations [142] and educating users about the domain using visualizations [99]. As shown in the example of our CityRec system, the success of such systems depends on the seamless integration of user interface aspects and algorithms. Since tourism is a dynamic field, the information powering such systems must be constantly updated; thus, highly automated solutions must be achieved to allow scaling. Our contributions showcase that data-driven solutions can be realized for many aspects of travel recommender systems, such as characterizing traveler mobility, determining travel regions, recommending the duration of stay, and characterizing destinations. With minor adaptations, they can be deployed within travel platforms to aid users in planning their trips and recommend fitting items such as accommodations and attractions.

Commercial trip planning systems will face additional challenges, such as ensuring that the recommendations on such platforms give a fair amount of exposure to individual items [6] and balance the interests of multiple stakeholders [165], such as mitigating the negative effects of overtourism [96]. Given the extent to which recommender systems influence the behavior of users, such platforms will have a high impact on tourism, which means that they must embrace their ethical responsibility [97]. In the context of travel recommender systems, this includes efforts to mitigate the climate crisis by providing users with recommendations that are similar in experience but minimize greenhouse gas emissions.

Bibliography

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. “Managing Popularity Bias in Recommender Systems with Personalized Re-ranking.” In: *The Thirty-Second International Flairs Conference*. FLAIRS’2019. 2019, pp. 413–418.

[2] Abhishek Agarwal and Linus W. Dietz. “Recommending the Duration of Stay in Personalized Travel Recommender Systems.” In: *RecSys Workshop on Recommenders in Tourism*. Vol. 3219. RecTour’2022. CEUR Workshop Proceedings (CEUR-WS.org), Sept. 2022.

[3] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Francesco Aletta. “Chatty Maps: Constructing Sound Maps of Urban Areas from Social Media Data.” In: *Royal Society Open Science* 3.3 (Mar. 2016), pp. 1–19. DOI: 10.1098/rsos.150690.

[4] Fabio Aiolli. “Efficient Top-n Recommendation for Very Large Scale Binary Rated Datasets.” In: *7th ACM Conference on Recommender Systems*. ACM, Oct. 2013, pp. 273–280. DOI: 10.1145/2507157.2507189.

[5] David Bamman, Brendan O’Connor, and Noah Smith. “Censorship and Deletion Practices in Chinese Social Media.” In: *First Monday* 17.3 (Mar. 2012). ISSN: 1396-0466. DOI: 10.5210/fm.v17i3.3943.

[6] Ashmi Banerjee, Gourab K. Patro, Linus W. Dietz, and Abhijnan Chakraborty. “Analyzing ‘Near Me’ Services: Potential for Exposure Bias in Location-based Retrieval.” In: *International Workshop on Fair and Interpretable Learning Algorithms*. FILA’20. IEEE, Dec. 2020. DOI: 10.1109/bigdata50022.2020.9378476.

[7] Jie Bao, Yu Zheng, and Mohamed F. Mokbel. “Location-based and Preference-aware Recommendation Using Sparse Geo-social Networking Data.” In: *20th International Conference on Advances in Geographic Information Systems*. SIGSPATIAL’12. New York, NY, USA: ACM, 2012, pp. 199–208. DOI: 10.1145/2424321.2424348.

[8] Jie Bao, Yu Zheng, David Wilkie, and Mohamed Mokbel. “Recommendations in Location-based Social Networks: A Survey.” In: *GeoInformatica* 19.3 (Feb. 2015), pp. 525–565. ISSN: 1384-6175. DOI: 10.1007/s10707-014-0220-8.

[9] Lucas Bernardi, Themistoklis Mavridis, and Pablo Estevez. “150 Successful Machine Learning Models: 6 Lessons Learned at Booking.com.” In: *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD ’19. Anchorage, AK, USA: ACM, 2019, pp. 1743–1751. ISBN: 9781450362016. DOI: 10.1145/3292500.3330744.

[10] Justine I. Blanford, Zhuojie Huang, Alexander Savelyev, and Alan M. MacEachren. “Geo-located Tweets. Enhancing Mobility Maps and Capturing Cross-border Movement.” In: *PLOS ONE* 10.6 (June 2015). Ed. by Renaud Lambiotte, pp. 1–16. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0129202.

[11] Ludovico Boratto, Gianni Fenu, and Mirko Marras. “The Effect of Algorithmic Bias on Recommender Systems for Massive Open Online Courses.” In: *41st European Conference on IR Research*. Springer, Apr. 2019, pp. 457–472. DOI: 10.1007/978-3-030-15712-8_30.

[12] Joan Borràs, Antonio Moreno, and Aida Valls. “Intelligent Tourism Recommender Systems: A Survey.” In: *Expert Systems with Applications* 41.16 (Nov. 2014), pp. 7370–7389. DOI: 10.1016/j.eswa.2014.06.007.

[13] David Boto-García, Emma Zapico, Marta Escalonilla, and José F. Baños Pino. “Tourists’ Preferences for Hotel Booking.” In: *International Journal of Hospitality Management* 92 (Jan. 2021). ISSN: 0278-4319. DOI: 10.1016/j.ijhm.2020.102726.

[14] Matthias Braunhofer, Mehdi Elahi, and Francesco Ricci. “Techniques for Cold-starting Context-aware Mobile Recommender Systems for Tourism.” In: *Intelligenza Artificiale* 8.2 (2014), pp. 129–143. ISSN: 1724-8035. DOI: 10.3233/IA-140069.

[15] Marcelo Darío Rodas Brítez. “A Content-based Recommendation System for Leisure Activities.” PhD thesis. University of Trento, 2019.

[16] Robin D. Burke. “Interactive Critiquing for Catalog Navigation in E-commerce.” In: *Artificial Intelligence Review* 18.3 (Dec. 2002), pp. 245–267. ISSN: 1573-7462. DOI: 10.1023/A:1020701617138.

[17] Robin D. Burke. “Hybrid Web Recommender Systems.” In: *The Adaptive Web: Methods and Strategies of Web Personalization*. Ed. by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Berlin, Heidelberg: Springer, 2007, pp. 377–408. ISBN: 978-3-540-72079-9. DOI: 10.1007/978-3-540-72079-9_12.

[18] Robin D. Burke, Kristian J. Hammond, and Benjamin C. Young. “The FindMe Approach to Assisted Browsing.” In: *IEEE Expert* 12.4 (July 1997), pp. 32–40. ISSN: 0885-9000. DOI: 10.1109/64.608186.

[19] Robin D. Burke and Maryam Ramezani. “Recommender Systems Handbook.” In: ed. by Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. Boston, MA, USA: Springer, Oct. 2011. Chap. Matching Recommendation Technologies and Domains, pp. 367–386. ISBN: 978-0-387-85820-3.

[20] Erion Çano and Maurizio Morisio. “Hybrid Recommender Systems: A Systematic Literature Review.” In: *Intelligent Data Analysis* 21.6 (2017), pp. 1487–1524. ISSN: 1571-4128.

[21] José I. Castillo-Manzano and Lourdes López-Valpuesta. “The Decline of the Traditional Travel Agent Model.” In: *Transportation Research Part E: Logistics and Transportation Review* 46.5 (Sept. 2010), pp. 639–649. ISSN: 1366-5545. DOI: 10.1016/j.tre.2009.12.009.

[22] Kinjal Chaudhari and Ankit Thakkar. “A Comprehensive Survey on Travel Recommender Systems.” In: *Archives of Computational Methods in Engineering* 27.5 (Oct. 2019), pp. 1545–1571. ISSN: 1886-1784. DOI: 10.1007/s11831-019-09363-7.

[23] Jen-Hsiang Chen, Kuo-Ming Chao, and Nazaraf Shah. “Hybrid Recommendation System for Tourism.” In: *10th International Conference on e-Business Engineering*. IEEE, Sept. 2013, pp. 156–161. DOI: 10.1109/icebe.2013.24.

[24] Li Chen and Pearl Pu. "Preference-based Organization Interfaces: Aiding User Critiques in Recommender Systems." In: *User Modeling*. Springer, 2007, pp. 77–86. DOI: 10.1007/978-3-540-73078-1_11.

[25] Li Chen and Pearl Pu. "Critiquing-based Recommenders: Survey and Emerging Trends." In: *User Modeling and User-Adapted Interaction* 22.1 (Apr. 2012), pp. 125–150. ISSN: 1573-1391. DOI: 10.1007/s11257-011-9108-6.

[26] Chen Cheng, Haiqin Yang, Irwin King, and Michael R. Lyu. "Fused Matrix Factorization with Geographical and Social Influence in Location-based Social Networks." In: *26th AAAI Conference on Artificial Intelligence*. AAAI, 2012. DOI: 10.1609/aaai.v26i1.8100.

[27] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. "Exploring Millions of Footprints in Location Sharing Services." In: *Fifth International Conference on Weblogs and Social Media*. ICWSM '11. Palo Alto, CA, USA: AAAI, July 2011, pp. 81–88.

[28] Erik Cohen. "Towards a Sociology of International Tourism." In: *Social Research* 39.1 (1972), pp. 164–182.

[29] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. "The Effects of Transparency on Trust in and Acceptance of a Content-based Art Recommender." In: *User Modeling and User-Adapted Interaction* 18.5 (Aug. 2008), pp. 455–496. DOI: 10.1007/s11257-008-9051-3.

[30] Douglas Critchlow. *Metric Methods for Analyzing Partially Ranked Data*. New York, NY, USA: Springer, 1985. 232 pp.

[31] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. "A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research." In: *ACM Transactions on Information Systems* 39.2 (Apr. 2021), pp. 1–49. DOI: 10.1145/3434185.

[32] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches." In: *13th ACM Conference on Recommender Systems*. RecSys'19. Copenhagen, Denmark: ACM, Sept. 2019, pp. 101–109. ISBN: 9781450362436. DOI: 10.1145/3298689.3347058.

[33] Linus W. Dietz. "Data-Driven Destination Recommender Systems." In: *26th Conference on User Modeling, Adaptation and Personalization*. UMAP '18. New York, NY, USA: ACM, July 2018, pp. 257–260. DOI: 10.1145/3209219.3213591.

[34] Linus W. Dietz, Daniel Herzog, and Wolfgang Wörndl. "Deriving Tourist Mobility Patterns from Check-in Data." In: *WSDM Workshop on Learning from User Interactions*. Los Angeles, CA, USA, Feb. 2018.

[35] Linus W. Dietz, Saadi Myftija, and Wolfgang Wörndl. "Designing a Conversational Travel Recommender System Based on Data-driven Destination Characterization." In: *ACM RecSys Workshop on Recommenders in Tourism*. Sept. 2019, pp. 17–21.

[36] Linus W. Dietz, Sameera Thimbiri Palage, and Wolfgang Wörndl. "Navigation by Revealing Trade-offs for Content-based Recommendations." In: *Information and Communication Technologies in Tourism*. Ed. by Jason L. Stienmetz, Berta Ferrer-Rosell, and David Massimo. Cham: Springer, Jan. 2022, pp. 149–161. ISBN: 978-3-030-94751-4. DOI: 10.1007/978-3-030-94751-4_14.

[37] Linus W. Dietz, Rinita Roy, and Wolfgang Wörndl. "Characterisation of Traveller Types Using Check-in Data from Location-based Social Networks." In: *Information and Communication Technologies in Tourism*. Ed. by Juho Pesonen and Julia Neidhardt. Cham: Springer, Dec. 2018, pp. 15–26. ISBN: 978-3-030-05940-8.

[38] Linus W. Dietz, Avradip Sen, Rinita Roy, and Wolfgang Wörndl. "Mining Trips from Location-based Social Networks for Clustering Travelers and Destinations." In: *Information Technology & Tourism* 22.1 (Mar. 2020), pp. 131–166. ISSN: 1098-3058. DOI: 10.1007/s40558-020-00170-6.

[39] Linus W. Dietz, Mete Sertkan, Saadi Myftija, Sameera Thimbiri Palage, Julia Neidhardt, and Wolfgang Wörndl. "A Comparative Study of Data-driven Models for Travel Destination Characterization." In: *Frontiers in Big Data* 5 (Apr. 2022). ISSN: 2624-909X. DOI: 10.3389/fdata.2022.829939.

[40] Linus W. Dietz and Achim Weimert. "Recommending Crowdsourced Trips on wOndary." In: *ACM RecSys Workshop on Recommenders in Tourism*. Vancouver, BC, Canada, Oct. 2018, pp. 13–17.

[41] Linus W. Dietz and Wolfgang Wörndl. "How Long to Stay Where? On the Amount of Item Consumption in Travel Recommendation." In: *ACM RecSys 2019 Late-breaking Results*. Sept. 2019, pp. 31–35.

[42] Sara Dolnicar. "Market Segmentation for e-Tourism." In: *Handbook of e-Tourism*. Ed. by Zheng Xiang, Matthias Fuchs, Ulrike Gretzel, and Wolfram Höpken. Cham: Springer, Feb. 2021, pp. 1–15. DOI: 10.1007/978-3-030-05324-6_53-1.

[43] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. "Rank Aggregation Methods for the Web." In: *10th International Conference on World Wide Web*. WWW'21. New York, NY, USA: ACM, 2001, pp. 613–622. DOI: 10.1145/371920.372165.

[44] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. "FaiRecSys: Mitigating Algorithmic Bias in Recommender Systems." In: *International Journal of Data Science and Analytics* 9.2 (Mar. 2020), pp. 197–213. DOI: 10.1007/s41060-019-00181-5.

[45] Michael D. Ekstrand and Maria Soledad Pera. "The Demographics of Cool: Popularity and Recommender Performance for Different Groups of Users." In: *RecSys Late-breaking Results*. Vol. 1905. CEUR Workshop Proceedings. CEUR-WS.org, 2017.

[46] Ronald Fagin, Ravi Kumar, and D. Sivakumar. "Comparing Top K Lists." In: *SIAM Journal on Discrete Mathematics* 17.1 (Jan. 2003), pp. 134–160. DOI: 10.1137/s0895480102412856.

[47] Daniel R Fesenmaier, Karl W Wöber, and Hannes Werthner. *Destination Recommendation Systems: Behavioral Foundations and Applications*. Cabi, 2006.

[48] Ángel García-Crespo, José Luis López-Cuadrado, Ricardo Colomo-Palacios, Israel González-Carrasco, and Belén Ruiz-Mezcua. “Sem-fit: A Semantic Based Expert System to Provide Recommendations in the Tourism Domain.” In: *Expert Systems with Applications* 38.10 (Sept. 2011), pp. 13310–13319. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2011.04.152.

[49] Damianos Gavalas, Charalampos Konstantopoulos, Konstantinos Mastakas, and Grammati Pantziou. “A Survey on Algorithmic Approaches for Solving Tourist Trip Design Problems.” In: *Heuristics* 20.3 (June 2014), pp. 291–328. ISSN: 1572-9397. DOI: 10.1007/s10732-014-9242-5.

[50] Marie Al-Ghossein. “Context-aware Recommender Systems for Real-world Applications.” PhD thesis. Université Paris-Saclay, Feb. 2019.

[51] Heather Gibson and Andrew Yiannakis. “Tourist Roles: Needs and the Lifecourse.” In: *Annals of Tourism Research* 29.2 (Apr. 2002), pp. 358–383.

[52] Dmitri Goldenberg, Kostia Kofman, Pavel Levin, Sarai Mizrachi, Maayan Kafry, and Guy Nadav. “Booking.com Wsdm Webtour 2021 Challenge.” In: *ACM WSDM Workshop on Web Tourism*. WebTour’21. New York, NY, USA: ACM, Mar. 2021.

[53] Dmitri Goldenberg and Pavel Levin. “Booking.com Multi-destination Trips Dataset.” In: *44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, July 2021. DOI: 10.1145/3404835.3463240.

[54] Marta C. González, César A. Hidalgo, and Albert-László Barabási. “Understanding Individual Human Mobility Patterns.” In: *Nature* 453.7196 (June 2008), pp. 779–782. ISSN: 0028-0836. DOI: 10.1038/nature06958.

[55] J. C. Gower. “A General Coefficient of Similarity and Some of Its Properties.” In: *Biometrics* 27.4 (Dec. 1971), pp. 857–871. ISSN: 0006341X, 15410420. DOI: 10.2307/2528823.

[56] Ulrike Gretzel. “Intelligent Systems in Tourism: A Social Science Perspective.” In: *Annals of Tourism Research* 38.3 (July 2011), pp. 757–779. ISSN: 0160-7383. DOI: 10.1016/j.annals.2011.04.014.

[57] Ulrike Gretzel, Matthias Fuchs, Rodolfo Baggio, Wolfram Hoepken, Rob Law, Julia Neidhardt, Juho Pesonen, Markus Zanker, and Zheng Xiang. “E-tourism beyond Covid-19: A Call for Transformative Research.” In: *Information Technology & Tourism* 22.2 (May 2020), pp. 187–203. DOI: 10.1007/s40558-020-00181-3.

[58] Wilfried Grossmann, Mete Sertkan, Julia Neidhardt, and Hannes Werthner. “Pictures As a Tool for Matching Tourist Preferences with Destinations.” In: *Personalized Human-Computer Interaction*. Munich, Germany: De Gruyter, Sept. 2019, pp. 183–200. DOI: 10.1515/9783110552485-007.

[59] Andreas Grün and Xenija Neufeld. “Challenges Experienced in Public Service Media Recommendation Systems.” In: *Fifteenth ACM Conference on Recommender Systems*. New York, NY, USA: ACM, Sept. 2021, pp. 541–544. ISBN: 9781450384582. DOI: 10.1145/3460231.3474618.

[60] Asela Gunawardana and Guy Shani. “Evaluating Recommender Systems.” In: *Recommender Systems Handbook*. Springer, 2015, pp. 265–308. DOI: 10.1007/978-1-4899-7637-6_8.

[61] Daniel Herzog, Linus W. Dietz, and Wolfgang Wörndl. “Tourist Trip Recommendations – Foundations, State of the Art and Challenges.” In: *Personalized Human-Computer Interaction*. Ed. by Miriam Augstein, Eelco Herder, and Wolfgang Wörndl. Berlin, Germany: de Gruyter Oldenbourg, Sept. 2019, pp. 159–182. ISBN: 978-3-11-055247-8. DOI: 10.1515/9783110552485-006.

[62] Daniel Herzog and Wolfgang Wörndl. “A Travel Recommender System for Combining Multiple Travel Regions to a Composite Trip.” In: *CBRecSys*. 2014, pp. 42–48.

[63] Daniel Andreas Herzog. “A User-centered Approach to Solving the Tourist Trip Design Problem for Individuals and Groups.” PhD thesis. Technische Universität München, 2020.

[64] Andrea Hess, Karin Anna Hummel, Wilfried N. Gansterer, and Günter Haring. “Data-driven Human Mobility Modeling.” In: *ACM Computing Surveys* 48.3 (Dec. 2015), pp. 1–39. ISSN: 0360-0300. DOI: 10.1145/2840722.

[65] Hsun-Ping Hsieh, Cheng-Te Li, and Shou-De Lin. “Exploiting Large-scale Check-in Data to Recommend Time-sensitive Routes.” In: *ACM SIGKDD International Workshop on Urban Computing*. UrbComp ’12. Beijing, China: ACM, 2012, pp. 55–62. ISBN: 978-1-4503-1542-5. DOI: 10.1145/2346496.2346506.

[66] Yifan Hu, Yehuda Koren, and Chris Volinsky. “Collaborative Filtering for Implicit Feedback Datasets.” In: *Eighth IEEE International Conference on Data Mining*. IEEE, Dec. 2008, pp. 263–272. DOI: 10.1109/ICDM.2008.22.

[67] Ronald L. Iman and W. J. Conover. “A Measure of Top-down Correlation.” In: *Technometrics* 29.3 (Aug. 1987), pp. 351–357. DOI: 10.2307/1269344.

[68] Jim Isaak and Mina J. Hanna. “User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection.” In: *Computer* 51.8 (Aug. 2018), pp. 56–59. DOI: 10.1109/MC.2018.3191268.

[69] Robert A. Jacobs. “Increased Rates of Convergence through Learning Rate Adaptation.” In: *Neural Networks* 1.4 (Jan. 1988), pp. 295–307. ISSN: 0893-6080. DOI: 10.1016/0893-6080(88)90003-2.

[70] Dietmar Jannach and Christine Bauer. “Escaping the McNamara Fallacy: Towards More Impactful Recommender Systems Research.” In: *AI Magazine* 41.4 (Dec. 2020), pp. 79–95. DOI: 10.1609/aimag.v41i4.5312.

[71] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. “A Survey on Conversational Recommender Systems.” In: *ACM Computing Surveys* 54.5 (2021), pp. 1–36. DOI: 10.1145/3453154.

[72] Dietmar Jannach, Pearl Pu, Francesco Ricci, and Markus Zanker. “Recommender Systems: Trends and Frontiers.” In: *AI Magazine* 43.2 (June 2022), pp. 145–150. DOI: 10.1002/aaai.12050.

[73] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. “News Recommender Systems – Survey and Roads Ahead.” In: *Information Processing & Management* 54.6 (Nov. 2018), pp. 1203–1227. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2018.04.008.

[74] Ankit Kariryaa, Isaac Johnson, Johannes Schöning, and Brent Hecht. "Defining and Predicting the Localness of Volunteered Geographic Information Using Ground Truth Data." In: *Conference on Human Factors in Computing System*. CHI'18. ACM, 2018. DOI: 10.1145/3173574.3173839.

[75] Maurice Kendall. *Rank Correlation Methods*. London, UK: Griffin, 1970. ISBN: 0852641990.

[76] Scott Kirkpatrick, Daniel Gelatt, and Mario. P. Vecchi. "Optimization by Simulated Annealing." In: *Science* 220.4598 (May 1983), pp. 671–680. DOI: 10.1126/science.220.4598.671.

[77] Joseph A. Konstan and John Riedl. "Recommender Systems: From Algorithms to User Experience." In: *User Modeling and User-Adapted Interaction* 22.1-2 (Apr. 2012), pp. 101–123. ISSN: 0924-1868. DOI: 10.1007/s11257-011-9112-x.

[78] Geraud Le Falher, Aristides Gionis, and Michael Mathioudakis. "Where Is the Soho of Rome? Measures and Algorithms for Finding Similar Neighborhoods in Cities." In: *9th International AAAI Conference on Web and Social Media*. ICWSM'15. Palo Alto, CA, USA: AAAI, May 2015, pp. 228–237.

[79] Xutao Li, Gao Cong, Xiao-Li Li, Tuan-Anh Nguyen Pham, and Shonali Krishnaswamy. "Rank-geofm: A Ranking Based Geographical Factorization Method for Point of Interest Recommendation." In: *38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Aug. 2015, pp. 433–442. DOI: 10.1145/2766462.2767722.

[80] Shili Lin and Jie Ding. "Integration of Ranked Lists Via Cross Entropy Monte Carlo with Applications to mRNA and microRNA Studies." In: *Biometrics* 65.1 (May 2008), pp. 9–18. DOI: 10.1111/j.1541-0420.2008.01044.x.

[81] Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. "Personalized Travel Package Recommendation." In: *IEEE 11th International Conference on Data Mining*. ICDM '11. Vancouver, BC, Canada: IEEE, Dec. 2011, pp. 407–416. DOI: 10.1109/icdm.2011.118.

[82] Yiding Liu, Tuan-Anh Pham, Gao Cong, and Quan Yuan. "An Experimental Evaluation of Point-of-interest Recommendation in Location-based Social Networks." In: *VLDB Endowment* 10.10 (June 2017), pp. 1010–1021. DOI: 10.14778/3115404.3115407.

[83] Yiding Liu, Kaiqi Zhao, and Gao Cong. "Efficient Similar Region Search with Deep Metric Learning." In: *24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. London, United Kingdom: ACM, 2018, pp. 1850–1859. ISBN: 9781450355520. DOI: 10.1145/3219819.3220031.

[84] Yong Liu, Wei Wei, Aixin Sun, and Chunyan Miao. "Exploiting Geographical Neighborhood Characteristics for Location Recommendation." In: *23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, Nov. 2014, pp. 739–748. DOI: 10.1145/2661829.2662002.

[85] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. "Recommender Systems Handbook." In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. Boston, MA, USA: Springer US, Oct. 2011. Chap. Content-based Recommender Systems: State of the Art and Trends, pp. 73–105. ISBN: 978-0-387-85820-3.

[86] Qiuju Luo and Dixi Zhong. "Using social network analysis to explain communication characteristics of travel-related electronic word-of-mouth on social networking sites." In: *Tourism Management* 46 (Feb. 2015), pp. 274–282. ISSN: 0261-5177. DOI: 10.1016/j.tourman.2014.07.007.

[87] David Massimo. "Exploiting Explainable and Generalized Human Behaviour Models for Recommendation." PhD thesis. Free University of Bozen-Bolzano, 2020.

[88] David Massimo and Francesco Ricci. "Clustering Users' Pois Visit Trajectories for Next-poi Recommendation." In: *Information and Communication Technologies in Tourism*. Ed. by Juho Pesonen and Julia Neidhardt. Cham: Springer, Dec. 2018, pp. 3–14.

[89] Kevin McCarthy, James Reilly, Lorraine McGinty, and Barry Smyth. "On the Dynamic Generation of Compound Critiques in Conversational Recommender Systems." In: *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Berlin, Heidelberg: Springer, 2004, pp. 176–184.

[90] Robert R. McCrae and Oliver P. John. "An Introduction to the Five-factor Model and Its Applications." In: *Personality* 60.2 (June 1992), pp. 175–215. ISSN: 0022-3506. DOI: 10.1111/j.1467-6494.1992.tb00970.x.

[91] Lorraine McGinty and James Reilly. "On the Evolution of Critiquing Recommenders." In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. Boston, MA, USA: Springer, Oct. 2011, pp. 419–453. DOI: 10.1007/978-0-387-85820-3_13.

[92] Lorraine McGinty and Barry Smyth. "Adaptive Selection: An Analysis of Critiquing and Preference-based Feedback in Conversational Recommender Systems." In: *Electronic Commerce* 11.2 (Dec. 2006), pp. 35–57. DOI: 10.2753/jec1086-4415110202.

[93] Grant McKenzie and Benjamin Adams. "Juxtaposing Thematic Regions Derived from Spatial and Platial User-generated Content." In: *13th International Conference on Spatial Information Theory*. Ed. by Eliseo Clementini, Maureen Donnelly, May Yuan, Christian Kray, Paolo Fogliaroni, and Andrea Ballatore. Vol. 86. Leibniz International Proceedings in Informatics. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, pp. 1–14. ISBN: 978-3-95977-043-9. DOI: 10.4230/LIPIcs.COSIT.2017.20.

[94] Bob McKercher. "Towards a Classification of Cultural Tourists." In: *International Journal of Tourism Research* 4.1 (Jan. 2002), pp. 29–38. DOI: 10.1002/jtr.346.

[95] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. "Investigating Gender Fairness of Recommendation Algorithms in the Music Domain." In: *Information Processing & Management* 58.5 (Sept. 2021), p. 102666. DOI: 10.1016/j.ipm.2021.102666.

[96] Claudio Milano, Marina Novelli, and Joseph M. Cheer. "Overtourism and Tourismphobia: A Journey Through Five Decades of Tourism Development, Planning and Local Concerns." In: *Travel and Tourism in the Age of Overtourism*. Routledge, Feb. 2021, pp. 1–5. DOI: 10.4324/9781003140610-1.

[97] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. "Recommender systems and their ethical challenges." In: *AI & Society* 35.4 (Feb. 2020), pp. 957–967. DOI: 10.1007/s00146-020-00950-y.

[98] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. “Unique in the Crowd: The Privacy Bounds of Human Mobility.” In: *Scientific Reports* 3.1 (Mar. 2013), pp. 1–5. DOI: 10.1038/srep01376.

[99] Belgin Mutlu, Eduardo Veas, Christoph Trattner, and Vedran Sabol. “VizRec: A Two-Stage Recommender System for Personalized Visualizations.” In: *20th International Conference on Intelligent User Interfaces Companion*. IUI Companion ’15. Atlanta, Georgia, USA: ACM, Mar. 2015, pp. 49–52. ISBN: 9781450333085. DOI: 10.1145/2732158.2732190.

[100] Saadi Myftija and Linus W. Dietz. “CityRec — a Data-driven Conversational Destination Recommender System.” In: *e-Review of Tourism Research* 17.5 (2020), pp. 808–816. ISSN: 1941-5842.

[101] Yeamduan Narangajavana, Luis José Callarisa Fiol, Miguel Ángel Moliner Tena, Rosa María Rodríguez Artola, and Javier Sánchez García. “The influence of social media in creating expectations. An empirical study for a tourist destination.” In: *Annals of Tourism Research* 65 (July 2017), pp. 60–70. ISSN: 0160-7383. DOI: https://doi.org/10.1016/j.annals.2017.05.002.

[102] Julia Neidhardt, Rainer Schuster, Leonhard Seyfang, and Hannes Werthner. “Eliciting the Users’ Unknown Preferences.” In: *8th ACM Conference on Recommender Systems*. RecSys ’14. New York, NY, USA: ACM, 2014, pp. 309–312. ISBN: 978-1-4503-2668-1. DOI: 10.1145/2645710.2645767.

[103] Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. “A Picture-based Approach to Recommender Systems.” In: *Information Technology & Tourism* 15.1 (Mar. 2015), pp. 49–69. ISSN: 1943-4294. DOI: 10.1007/s40558-014-0017-5.

[104] Xia Ning, Christian Desrosiers, and George Karypis. “A Comprehensive Survey of Neighborhood-based Recommendation Methods.” In: *Recommender Systems Handbook*. Boston, MA, USA: Springer, 2015, pp. 37–76. DOI: 10.1007/978-1-4899-7637-6_2.

[105] Lyndon Nixon. “Do Dmos Promote the Right Aspects of the Destination? A Study of Instagram Photography with a Visual Classifier.” In: *Information and Communication Technologies in Tourism 2022*. Ed. by Jason L. Stienmetz, Berta Ferrer-Rosell, and David Massimo. Cham: Springer, Jan. 2022, pp. 174–186. ISBN: 978-3-030-94751-4.

[106] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. “An Empirical Study of Geographic User Activity Patterns in Foursquare.” In: *Fifth International Conference on Weblogs and Social Media*. ICWSM ’11. Palo Alto, CA, USA: AAAI, July 2011, pp. 570–573.

[107] Xi Ouyang, Chaoyun Zhang, Pan Zhou, and Hao Jiang. “Deepspace: An Online Deep Learning Framework for Mobile Big Data to Understand Human Mobility Patterns.” In: *CoRR* abs/1610.07009 (2016).

[108] Sunjeong Park and Lim Youn-kyung. “Design Considerations for Explanations Made by a Recommender Chatbot.” In: *IASDR Conference 2019*. IASDR. 2019.

[109] Michael J. Pazzani and Daniel Billsus. “Content-based Recommendation Systems.” In: *The Adaptive Web: Methods and Strategies of Web Personalization*. Ed. by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Berlin, Heidelberg: Springer, 2007, pp. 325–341. ISBN: 978-3-540-72079-9. DOI: 10.1007/978-3-540-72079-9_10.

[110] Philip L. Pearce. *The Social Psychology of Tourist Behavior*. International Series in Experimental Social Psychology, 1982.

[111] Daniele Quercia, Neil Keith O’Hare, and Henriette Cramer. “Aesthetic Capital: What Makes London Look Beautiful, Quiet, and Happy?” In: *17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW ’14. Baltimore, Maryland, USA: ACM, 2014, pp. 945–55. ISBN: 9781450325400. DOI: 10.1145/2531602.2531613.

[112] Daniele Quercia, Rossano Schifanella, Luca Maria Aiello, and Kate McLean. “Smelly Maps: The Digital Life of Urban Smellscapes.” In: *Ninth International AAAI Conference on Web and Social Media*. ICWSM’15. Palo Alto, CA, USA: AAAI, 2015, pp. 327–336.

[113] Arpit Rana and Derek Bridge. “Navigation-by-preference – a New Conversational Recommender with Preference-based Feedback.” In: *25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: ACM, Mar. 2020, pp. 155–165. DOI: 10.1145/3377325.3377496.

[114] Logesh Ravi and Subramaniyaswamy Vairavasundaram. “A Collaborative Location Based Travel Recommendation System through Enhanced Rating Prediction for the Group of Users.” In: *Computational Intelligence and Neuroscience* 2016 (Mar. 2016), pp. 1–28. DOI: 10.1155/2016/1291358.

[115] Shaina Raza and Chen Ding. “News Recommender System: A Review of Recent Progress, Challenges, and Opportunities.” In: *Artificial Intelligence Review* 55.1 (July 2021), pp. 749–800. DOI: 10.1007/s10462-021-10043-x.

[116] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. “BPR: Bayesian Personalized Ranking from Implicit Feedback.” In: *UAI*. AUAI Press, 2009, pp. 452–461.

[117] C. C. Robusto. “The Cosine-haversine Formula.” In: *The American Mathematical Monthly* 64.1 (Jan. 1957), p. 38. DOI: 10.2307/2309088.

[118] Oliver Roick and Susanne Heuser. “Location Based Social Networks - Definition, Current State of the Art and Research Agenda.” In: *Transactions in GIS* 5.17 (May 2013), pp. 763–784. DOI: 10.1111/tgis.12032.

[119] M. Rosvall, D. Axelsson, and C. T. Bergstrom. “The Map Equation.” In: *The European Physical Journal Special Topics* 178.1 (Nov. 2009), pp. 13–23. ISSN: 1951-6401. DOI: 10.1140/epjst/e2010-01179-1.

[120] Peter J. Rousseeuw. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” In: *Computational and Applied Mathematics* 20.1987 (Nov. 1987), pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.

[121] Rinita Roy and Linus W. Dietz: “Triprec – a Recommender System for Planning Composite City Trips Based on Travel Mobility Analysis.” In: *ACM WSDM Workshop on Web Tourism*. WebTour’21. New York, NY, USA: ACM, Mar. 2021.

[122] Alan Said and Alejandro Bellogín. "Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks." In: *8th ACM Conference on Recommender Systems*. RecSys'14. ACM, 2014, pp. 129–136. DOI: 10.1145/2645710.2645746.

[123] Pablo Sánchez. "Exploring Attributes, Sequences, and Time in Recommender Systems: From Classical to Point-of-interest Recommendation." PhD thesis. Universidad Autónoma de Madrid, 2021.

[124] Pablo Sánchez and Alejandro Bellogín. "On the Effects of Aggregation Strategies for Different Groups of Users in Venue Recommendation." In: *Information Processing & Management* 58.5 (2021), p. 102609. DOI: 10.1016/j.ipm.2021.102609.

[125] Pablo Sánchez and Alejandro Bellogín. "Point-of-interest Recommender Systems Based on Location-based Social Networks: A Survey from an Experimental Perspective." In: *ACM Computing Surveys* (Jan. 2022). ISSN: 0360-0300. DOI: 10.1145/3510409.

[126] Pablo Sánchez and Linus W. Dietz. "Travelers vs. Locals: The Effect of Cluster Analysis in Point-of-Interest Recommendation." In: *30th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP'22. New York, NY, USA: ACM, July 2022, pp. 132–142. DOI: 10.1145/3503252.3531320.

[127] Avradip Sen and Linus W. Dietz. "Identifying Travel Regions Using Location-based Social Network Check-in Data." In: *Frontiers in Big Data* 2.12 (June 2019). DOI: 10.3389/fdata.2019.00012.

[128] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. "Mapping of Tourism Destinations to Travel Behavioural Patterns." In: *Information and Communication Technologies in Tourism*. Ed. by Brigitte Stangl and Juho Pesonen. Cham: Springer, Dec. 2017, pp. 422–434. ISBN: 978-3-319-72923-7.

[129] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. "What Is the "personality" of a Tourism Destination?" In: *Information Technology & Tourism* 21.1 (Mar. 2019), pp. 105–133. ISSN: 1943-4294. DOI: 10.1007/s40558-018-0135-6.

[130] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. "Eliciting Touristic Profiles: A User Study on Picture Collections." In: *28th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP '20. Genoa, Italy: ACM, 2020, pp. 230–38. ISBN: 9781450368612. DOI: 10.1145/3340631.3394868.

[131] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. "PicTouRe - a Picture-based Tourism Recommender." In: *14th ACM Conference on Recommender Systems*. RecSys'20. Virtual Event, Brazil: ACM, Sept. 2020, pp. 597–599. ISBN: 9781450375832. DOI: 10.1145/3383313.3411526.

[132] Omar Sharaki. "Developing a Software Library for Processing and Evaluating Travel Region Models." MA thesis. Technical University of Munich, Mar. 2021.

[133] Grace S. Shieh. "A Weighted Kendall's Tau Statistic." In: *Statistics & Probability Letters* 39.1 (July 1998), pp. 17–24. DOI: 10.1016/s0167-7152(98)00006-6.

[134] Thiago H. Silva, Aline Carneiro Viana, Fabrício Benevenuto, Leandro Villas, Juliana Salles, Antonio Loureiro, and Daniele Quercia. "Urban Computing Leveraging Location-based Social Network Data: A Survey." In: *ACM Computing Surveys* 52.1 (Feb. 2019), pp. 1–39. ISSN: 0360-0300. DOI: 10.1145/3301284.

[135] Barry Smyth. “Case-based Recommendation.” In: *The Adaptive Web: Methods and Strategies of Web Personalization*. Ed. by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Berlin, Heidelberg: Springer, 2007, pp. 342–376. ISBN: 978-3-540-72079-9. DOI: 10.1007/978-3-540-72079-9_11.

[136] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. “Modelling the Scaling Properties of Human Mobility.” In: *Nature Physics* 6.10 (Sept. 2010), pp. 818–823. ISSN: 1745-2473. DOI: 10.1038/nphys1760.

[137] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. “Limits of Predictability in Human Mobility.” In: *Science* 327.5968 (Feb. 2010), pp. 1018–1021. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1177170.

[138] Charles Spearman. “The Proof and Measurement of Association between Two Things.” In: *The American Journal of Psychology* 15.1 (Jan. 1904), pp. 72–101. ISSN: 0002-9556. DOI: 10.2307/1412159.

[139] Charles Spearman. “‘Footrule’ for Measuring Correlation.” In: *British Journal of Psychology* 2.1 (July 1906), pp. 89–108. DOI: 10.1111/j.2044-8295.1906.tb00174.x.

[140] Esra Suel, John W. Polak, James E. Bennett, and Majid Ezzati. “Measuring Social, Environmental and Health Inequalities Using Deep Learning and Street Imagery.” In: *Scientific Reports* 9.1 (Apr. 2019). ISSN: 2045-2322. DOI: 10.1038/s41598-019-42036-w.

[141] Ke Sun, Tieyun Qian, Tong Chen, Yile Liang, Quoc Viet Hung Nguyen, and Hongzhi Yin. “Where to Go Next: Modeling Long- and Short-term User Preferences for Point-of-interest Recommendation.” In: *AAAI Conference on Artificial Intelligence* 34.01 (Apr. 2020), pp. 214–221. DOI: 10.1609/aaai.v34i01.5353.

[142] Nava Tintarev and Judith Masthoff. “A Survey of Explanations in Recommender Systems.” In: *23rd International Conference on Data Engineering Workshop*. IEEE, Apr. 2007, pp. 801–810. DOI: 10.1109/icdew.2007.4401070.

[143] Christoph Trattner, Alexander Oberegger, Leandro Balby Marinho, and Denis Parra. “Investigating the Utility of the Weather Context for Point of Interest Recommendations.” In: *Information Technology & Tourism* 19.1-4 (2018), pp. 117–150. DOI: 10.1007/s40558-017-0100-9.

[144] Chieh-Yuan Tsai, Gerardo Paniagua, Yu-Jen Chen, Chih-Chung Lo, and Liguo Yao. “Personalized Tour Recommender through Geotagged Photo Mining and LSTM Neural Networks.” In: *MATEC Web of Conferences* 292.1003 (Sept. 2019). Ed. by N. Mastorakis, V. Mladenov, and A. Bulucea. DOI: 10.1051/matecconf/201929201003.

[145] Levente Varga, András Kovács, Géza Tóth, István Papp, and Zoltán Néda. “Further We Travel the Faster We Go.” In: *PLOS ONE* 11.2 (Feb. 2016). Ed. by Daqing Li, pp. 1–9. DOI: 10.1371/journal.pone.0148913.

[146] Saúl Vargas and Pablo Castells. “Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems.” In: *5th ACM Conference on Recommender Systems*. RecSys’11. ACM, 2011, pp. 109–116. DOI: 10.1145/2043932.2043955.

[147] Paolo Viappiani, Boi Faltings, and Pearl Pu. “Preference-based Search Using Example-critiquing with Suggestions.” In: *Journal of Artificial Intelligence Research* 27.1 (Dec. 2006), pp. 465–503. ISSN: 1076-9757.

[148] Lukas Vorwerk and Linus W. Dietz. “An Interactive Dashboard for Traveler Mobility Analysis.” In: *ACM WSDM Workshop on Web Tourism*. WebTour’21. ACM, Mar. 2021.

[149] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási. “Human Mobility, Social Ties, and Link Prediction.” In: *17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD’11. New York, NY, USA: ACM, Aug. 2011, pp. 1100–1108. DOI: 10.1145/2020408.2020581.

[150] Hannes Werthner and Francesco Ricci. “E-commerce and Tourism.” In: *Communications of the ACM* 47.12 (Dec. 2004), pp. 101–105. ISSN: 0001-0782. DOI: 10.1145/1035134.1035141.

[151] Yuxia Wu, Ke Li, Guoshuai Zhao, and Xueming Qian. “Long- and Short-term Preference Learning for Next POI Recommendation.” In: *28th ACM International Conference on Information and Knowledge Management*. CIKM’19. Beijing, China: ACM, Nov. 2019, pp. 2301–2304. ISBN: 9781450369763. DOI: 10.1145/3357384.3358171.

[152] Haoran Xie, Debby D. Wang, Yanghui Rao, Tak-Lam Wong, Lau Y. K. Raymond, Li Chen, and Fu Lee Wang. “Incorporating User Experience into Critiquing-based Recommender Systems: A Collaborative Approach Based on Compound Critiquing.” In: *Machine Learning and Cybernetics* 9.5 (May 2018), pp. 837–852. ISSN: 1868-808X. DOI: 10.1007/s13042-016-0611-2.

[153] Min Xie, Laks V. S. Lakshmanan, and Peter T. Wood. “Composite Recommendations: From Items to Packages.” In: *Frontiers of Computer Science* 6.3 (June 2012), pp. 264–277. ISSN: 1673-7466. DOI: 10.1007/s11704-012-2014-1.

[154] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. “NationTelescope: Monitoring and Visualizing Large-scale Collective Behavior in LBSNs.” In: *Journal of Network and Computer Applications* 55 (Sept. 2015), pp. 170–180. DOI: 10.1016/j.jnca.2015.05.010.

[155] Yuan Yao and F. Maxwell Harper. “Judging Similarity: A User-centric Study of Related Item Recommendations.” In: *12th ACM Conference on Recommender Systems*. RecSys ’18. Vancouver, British Columbia, Canada: ACM, 2018, pp. 288–296. ISBN: 978-1-4503-5901-6. DOI: 10.1145/3240323.3240351.

[156] Andrew Yiannakis and Heather Gibson. “Roles Tourists Play.” In: *Annals of Tourism Research* 19.2 (Jan. 1992), pp. 287–303. ISSN: 0160-7383. DOI: 10.1016/0160-7383(92)90082-z.

[157] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. “A New Rank Correlation Coefficient for Information Retrieval.” In: *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’08. Singapore, Singapore: ACM, 2008, pp. 587–594. ISBN: 9781605581644. DOI: 10.1145/1390334.1390435.

[158] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. “Hybrid Collaborative and Content-based Music Recommendation Using Probabilistic Model with Latent User Preferences.” In: *7th International Conference on Music Information Retrieval*. ISMIR’06. Victoria, Canada, 2006, pp. 296–301.

[159] Zeping Yu, Jianxun Lian, Ahmad Mahmoody, Gongshen Liu, and Xing Xie. “Adaptive User Modeling with Long and Short-term Preferences for Personalized Recommendation.” In: *28th International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 4213–4219. DOI: 10.24963/ijcai.2019/585.

[160] Fajie Yuan, Joemon M. Jose, Guibing Guo, Long Chen, Haitao Yu, and Rami Suleiman Alkhawaldeh. “Joint Geo-spatial Preference and Pairwise Ranking for Point-of-interest Recommendation.” In: *28th International Conference on Tools with Artificial Intelligence:* IEEE, Nov. 2016, pp. 46–53. DOI: 10.1109/ictai.2016.0018.

[161] Jiyong Zhang and Pearl Pu. “A Comparative Study of Compound Critique Generation in Conversational Recommender Systems.” In: *Adaptive Hypermedia and Adaptive Web-Based Systems*. Ed. by Vincent P. Wade, Helen Ashman, and Barry Smyth. Berlin: Springer, 2006, pp. 234–243. ISBN: 978-3-540-34697-5. DOI: 10.1007/11768012_25.

[162] Yan Zhang, Lin Wang, Yi-Qing Zhang, and Xiang Li. “Towards a Temporal Network Analysis of Interactive WiFi Users.” In: *Europhysics Letters* 98.6 (June 2012). DOI: 10.1209/0295-5075/98/68002.

[163] Weimin Zheng, Xiaoting Huang, and Yuan Li. “Understanding the Tourist Mobility Using GPS: Where Is the Next Place?” In: *Tourism Management* 59 (Apr. 2017), pp. 267–280. ISSN: 0261-5177. DOI: 10.1016/j.tourman.2016.08.009.

[164] Weimin Zheng, Rui Zhou, Zhemin Zhang, Yihui Zhong, Surui Wang, Zhao Wei, and Haipeng Ji. “Understanding the Tourist Mobility Using GPS: How Similar Are the Tourists?” In: *Tourism Management* 71 (Apr. 2019), pp. 54–66. ISSN: 0261-5177. DOI: 10.1016/j.tourman.2018.09.019.

[165] Yong Zheng. “Multi-Stakeholder Recommendations: Case Studies, Methods and Challenges.” In: *13th ACM Conference on Recommender Systems*. RecSys’19. Copenhagen, Denmark: ACM, Sept. 2019, pp. 578–579. ISBN: 9781450362436. DOI: 10.1145/3298689.3346951.

[166] Yu Zheng and Xing Xie. “Learning Travel Recommendations from User-generated GPS Traces.” In: *ACM Transactions on Intelligent Systems and Technology* 2.1 (Jan. 2011), pp. 1–29. ISSN: 2157-6904. DOI: 10.1145/1889681.1889683.

[167] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. “Mining Interesting Locations and Travel Sequences from GPS Trajectories.” In: *18th International World Wide Web Conference*. WWW’09. New York, NY, USA: ACM, Apr. 2009. DOI: 10.1145/1526709.1526816.

[168] Andreas H. Zins. “Exploring Travel Information Search Behavior beyond Common Frontiers.” In: *Information Technology & Tourism* 9.3 (Sept. 2007), pp. 149–164. DOI: 10.3727/109830507782167015.

Appendix

A Embedded Publications

A.1 Publication 1: "Mining Trips from Location-based Social Networks for Clustering Travelers and Destinations"

Linus W. Dietz et al. "Mining Trips from Location-based Social Networks for Clustering Travelers and Destinations." In: *Information Technology & Tourism* 22.1 (Mar. 2020), pp. 131–166. ISSN: 1098-3058. DOI: 10.1007/s40558-020-00170-6

Summary

In this work, we present a data-driven method to mine trips from location-based social networks to understand how tourists travel the world. The obtained insights can be relevant building blocks for destination recommender systems, i.e., automatic preference elicitation, defining travel regions, and general traveler behavior. The primary artifact of this paper is the `tripmining` library, which quantifies collected trips with several metrics to capture the underlying mobility and assess the quality of the data.

We showcase two applications that utilize the mined trips. The first is an approach for clustering travelers in two case studies, one of Twitter and another of Foursquare, where the pure mobility metrics are enriched with social aspects, i.e., what activities the users have done. Clustering 133,614 trips from Twitter, we obtain three distinct groups of travelers based on the pure mobility trace. In the Foursquare data set, which includes the type of venues the users have checked in, six clusters can be determined. The second application area is the spatial clustering of destinations around the world. These discovered regions are solely formed by the mobility patterns of the trips and are, thus, independent of administrative regions such as countries. We identify 942 regions as destinations that can be directly used as a hierarchical region model in a destination recommender system.

Author Contributions

LD was the main author of the manuscript and supervised the individual research projects of AS and RR. LD developed the tripmining library, initiated the data collection, and performed all statistical analyses. RR developed the initial version of the trip clustering approach under the supervision of LD and helped to adapt the method for the novel data set. AS helped with the data collection from Twitter and developed the method for region discovery including the visualization. WW supervised the overall project and co-edited the manuscript.

Information Technology & Tourism (2020) 22:131–166
https://doi.org/10.1007/s40558-020-00170-6

ORIGINAL RESEARCH

Mining trips from location-based social networks for clustering travelers and destinations

Linus W. Dietz[1] · Avradip Sen[1] · Rinita Roy[1] · Wolfgang Wörndl[1]

Received: 23 June 2019 / Revised: 3 January 2020 / Accepted: 21 January 2020 /
Published online: 29 January 2020

Abstract
It is important to learn the characteristics of travelers and touristic regions when trying to generate recommendations for destinations to users. In this work, we first present a data-driven method to mine trips from location-based social networks to understand how tourists travel the world. These trips are quantified using a number of metrics to capture the underlying mobility patterns. We then present two applications that utilize the mined trips. The first one is an approach for clustering travelers in two case studies, one of Twitter and another of Foursquare, where the pure mobility metrics are enriched with social aspects, i.e., the kinds of venues into which the users checked-in. Clustering 133,614 trips from Twitter, we obtain three distinct clusters. In the Foursquare data set, however, six clusters can be determined. The second application area is the spatial clustering of destinations around the world. These discovered regions are solely formed by the mobility patterns of the trips and are, thus, independent of administrative regions such as countries. We identify 942 regions as destinations that can be directly used as a region model of a destination recommender system. This paper is the extended version of the conference article “Characterisation of Traveller Types Using Check-in Data from Location-Based Social Networks” presented at the 26th Annual ENTER eTourism Conference held from January 19 to February 1, 2019 in Nicosia, Cyprus.

Keywords Mobility modeling · Cluster analysis · Spatial clustering · Recommender systems

1 Introduction

Analyzing the mobility of travelers reveals a lot of information about their behavior, preferences, and the destinations they visit. This is interesting for a number of different purposes. Municipalities can obtain information about the popularity of destinations

✉ Linus W. Dietz
linus.dietz@tum.de

[1] Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany

within their district to build infrastructure and provide services in an informed way. Destination marketers can learn more about the context of their prospective guests and make improved offers to attract more visitors. Tourist agencies or travel recommender systems can characterize their clients and suggest serendipitous, yet accurate destinations to visit. Finally, prospective travelers can benefit from useful recommendations when planning their trips.

Tourist mobility can be observed in different ways. Analyzing the number of accommodation bookings in a city, tracking ticket sales of flights or trains, or analyzing the congestion of highway connections only captures aggregate travel patterns of one destination or the connections between them. To provide insights into individual travel, we analyze the movement of individual travelers with data from location-based social networks (LBSNs). Our definition of LBSNs follows the one of Roick and Heuser (2013), which includes both social networks that allow the geotagging of contents, such as Twitter or Flickr, as well as geosocial networking sites, such as Foursquare.

The basic idea of our approach is to chronologically sort all of a user's geotagged content into a stream of check-ins and to segment it into periods of being at home and of travel. Consecutive check-ins outside of one's home are combined into a trip that can be characterized using different metrics (Dietz et al. 2018a). Such trips can be used to find the destinations visited together, to derive the durations of stays (Dietz and Wörndl 2019), to cluster them into discernible groups (Dietz et al. 2018b), or to discover larger travel regions (Sen and Dietz 2019).

In this paper, we refine and extend the aforementioned approaches. After reviewing the most relevant literature on the respective topics, we describe three major contributions in the following sections: First, we thoroughly describe the method for deriving trips from various LBSNs, such as Foursquare, Twitter and Flickr. We describe metrics to quantify the quality of trips and compare the mobility metrics of several data sets in Sect. 3. Second, we extend the approach of Dietz et al. (2018b) to cluster the mined trips by social aspects and a second case study on a Twitter data set in Sect. 4. Third, in Sect. 5, we present a novel approach to transform the mobility patterns manifested in trips into a network of tourist flows and use a community detection approach to discover tourist destinations that are defined by actual tourist mobility as opposed to political and administrative boundaries. Finally, we conclude our findings in Sect. 6.

The contributions of this paper can be used to improve the personalization of destination recommender systems. By observing a large number of travelers, we learn how real users travel and should be able to make more realistic recommendations. The spatial clustering of cities to larger travel regions can be directly used as a region model within a destination recommender system that was formerly dependent on political boundaries (Wörndl 2017; Dietz 2018).

2 Related work

The motivation of this work is to improve several aspects of tourist recommender systems. For making good travel recommendations, we need deeper insights as to how people travel, what types of travelers exist, and which regions one should travel to. This section discusses the literature on tourist recommender systems, the analysis of human mobility, characterizing travelers, and previous approaches to discovering and defining travel regions.

2.1 Tourist recommender systems

Individual tourism is a challenging domain for recommender systems due to the substantial complexities of planning an independent trip and the huge economic importance of the travel and tourism industry (Chaudhari and Thakkar 2019). Big commercial player such as Booking, Tripadvisor, and Skyscanner focus on recommending single items, such as hotels, restaurants, and flights. Nowadays, the academic community is more concerned with recommendations of various touristic items, such as attractions, tourist packages, and composite trips (Borràs et al. 2014). Since these items are often not as well defined as hotels or restaurants, collaborative filtering methods have proven to be less suited. In addition, most people usually travel less frequently than e.g., consuming music or movies, making collaborative recommendations even less reliable due to the cold start problem. Instead, content-based and knowledge-based recommendation techniques are often employed (Burke and Ramezani 2011), or in case it is possible, hybridization (Kbaier et al. 2017).

To facilitate the content-based paradigm, users and items need to be placed in the same feature space that allows for the calculation of a similarity metric. This is usually done using a common categorization, and the similarity measure determines the ranking of the recommended items. This categorization problem is nontrivial, since it requires reliable information about both the user and the candidate items. It can, however, overcome the cold start problem (Burke 2007), since, unlike in a collaborative filtering approach, one can design intelligent systems that efficiently and effectively capture the user’s preferences (Braunhofer et al. 2014). For example, it is possible to characterize users based on their past trips (Dietz and Weimert 2018), or to define a more elaborate mapping between classes of users and destinations (Sertkan et al. 2017). Also, the information about a user’s social network and previously visited places can be used (Bao et al. 2015; Tsai et al. 2019). Dietz (2018) proposed a data-driven destination recommender system that would suggest composite trips of destinations to users. This paper contributes to the outlined ideas by improving the personalization of the recommendations, especially the duration of stay at the respective destination depending on the traveler type. Furthermore, in Sect. 5, we propose an approach to constructing a hierarchical model of travel regions.

2.2 Human mobility analysis

The analysis of human mobility gives insights into various aspects of everyday life. Before the advent of online social networks on GPS-enabled devices, data sources like mobile phone communication records (González et al. 2008), Wi-Fi usage (Zhang et al. 2012), and raw GPS trajectories (Zheng et al. 2009) have been used to analyze individual human mobility. Given today’s availability of LBSN data that enriches a pure location trace with further information, such as user-posted content and the user's social network, much research has been done analyzing mobility using data from Twitter, Foursquare, and other platforms (Hess et al. 2015).

A prominent research objective is the predictability of human mobility. Song et al. found that individual mobility patterns follow reproducible scaling laws (Song et al. 2010a) and described the limits of the extent to which human mobility can be predicted (Song et al. 2010b). More recently, Ouyang et al. (2016) have analyzed mobility data to predict travel trajectories using a deep learning framework. Similar approaches to predicting the next visited place exist for tourists as well (Zheng et al. 2017) The correlation of locations with a social activity that can be studied with LBSN data promises interesting insights into social behavior. Cheng et al. (2011) found recurring daily and weekly patterns of activity and Wang et al. (2011) found a positive pairwise correlation between social connectedness, i.e., the strength of interactions, and mobility. Noulas et al. analyze activity patterns of Foursquare users, such as the spatial and temporal distances between two check-ins (Noulas et al. 2011). They discover place transitions that could well be used to predict or recommend the future locations of users. The general idea behind their approach is quite similar to ours, however, their motivation was to uncover recurring patterns of human mobility, thus the resulting metrics go in a different direction.

LBSN mobility data have been used to improve recommender systems (Bao et al. 2015). Zheng and Xie (2011) studied spatial co-occurrences that can also be used to identify similar users and generate implicit ratings for collaborative filtering algorithms. Bao et al. (2012) matched the travelers in a foreign city to local experts based on their respective home behaviors to improve the accuracy of a point of interest recommender. LBSN data has also been used to capture cross-border movement (Blanford et al. 2015). The authors demonstrate how the movement dynamics of people in a country can be analyzed, however, this study is not about tourists and is limited to one country, Kenya. Hsieh et al. (2012) used past LBSN data to recommend traveling paths, while Zheng et al. (2019) proposed heuristics to approximate the similarity of tourist trips. For this they present solutions to derive the popularity, the proper time of day to visit, the transit time between venues and the best order to visit the places. In contrast to our scenario, the routes contain single points of interest in urban areas and they leave determining durations of stay at one place to future work. Recently, Dietz et al. (2018a) have proposed a metric-based approach that extracts foreign trips from LBSN data. In this paper, they analyze tourist mobility patterns with the goal of investigating the popularity and co-occurrences of tourist destinations in composite trips.

2.3 Tourist roles

The characterization of tourists has been discussed in literature for decades with an increasing level of complexity. One of the first works was Cohen's four different social roles of tourists: the "organized mass tourist", "individual mass tourist", "explorer", and "drifter" (Cohen 1972). Pearce used fuzzy set theory to define 15 different travel roles (Pearce 1982), while McKercher used an approach motivated by cultural sciences to classify tourists based on the importance of cultural motives when deciding which destination to visit and the depth of cultural experience gathered by the tourist (McKercher 2002). Finally, Yiannakis and Gibson (1992) took a sociological perspective to observe which roles—they identify 17—are enacted by people when they travel; and associated these with different psychological needs.

With such diversity of tourist categorizations in the literature, the best grouping of tourist preferences and needs to improve destination recommendation is unclear. More importantly, none of the existing categorizations have been validated with observational data (Neidhardt et al. 2014), so it is unclear whether the categories apply to real travelers. To address this challenge, Neidhardt et al. developed the *Seven Factor Model* of tourist behavioral patterns (Neidhardt et al. 2014) based on the *Big Five Factor Model* (McCrae and John 1992) from psychology and a factor analysis of the 17 tourist roles proposed by Yiannakis and Gibson (1992); Gibson and Yiannakis (2002). With a destination recommender system in mind, they elicited user preferences through an image classification task, where the users are to pick the most appealing travel-related photos from a collection. The classification of these pictures along the *Seven Factors* has been previously determined using a questionnaire. Thus, the user's selection of images constitutes a personalized mixture of taste model, allowing for content-based recommendation of points of interests that were rated by experts along the *Seven Factors* in the design stage. Continuing this line of research, Sertkan et al. used unsupervised learning to cluster 561 tourist destinations from a rich commercial data set based on 18 motivational and 7 geographical attributes (Sertkan et al. 2017). Using an expert mapping of the *Seven Factors* to these destinations, they could distill associations between destination attributes and the *Seven Factor Model* that indicate travel behaviors. The *Seven Factor Model* relies heavily on expert knowledge, which is a drawback if this information is not available or costly to obtain. To overcome these limitations, Dietz et al. (2018b) propose trips mined from LBSNs to cluster trips into distinct groups using mobility metrics. To obtain good cluster quality, they perform a correlation analysis of the mobility features and identify four important features: the number of countries visited, the duration of travel, the radius of gyration, and the displacement from home. The resulting clusters are "Vacationers", "Explorers", "Voyagers", and "Globetrotters". We improve upon this research by also analyzing domestic trips and compare a pure mobility-based cluster analysis with social aspects, i.e., to which kinds of establishments the travelers have checked-in during their trip.

2.4 Touristic region discovery via community detection

Researchers have already attempted to define regions based on human mobility data for various purposes such as administrative region discovery (del Prado and Alatrista-Salas 2016), topical region discovery (Taniguchi et al. 2015), and political redistricting (Joshi et al. 2009). Closest to our region discovery approach is the work of Hawelka et al. (2014), who aim to find larger regions of mobility by combining several countries. We aim to find touristic regions that are smaller and potentially independent of countries.

There are various algorithms to perform community detection in networks, such as the Louvain method (Blondel et al. 2008), GDBSCAN (Orman et al. 2011), and Infomap (Rosvall et al. 2009). The complexity of these methods is $\mathcal{O}(n \log n)$. GDBSCAN is less flexible compared to the two others, since it requires to use the distance between spatial points to form clusters that are geographically contiguous. This is a limitation that the other two methods do not have, since the weights of the edges can be chosen at the analyst's will. In the end, we decided to use Infomap, since it has been reported that it outperforms the Louvain method in the quality of the communities (Fortunato and Hric 2016), and there is an up-to-date implementation available.[1] This implementation of Infomap can recursively apply the algorithm to the detected clusters to detect hierarchies of clusters. This mitigates the resolution limit problem, where the size of communities depends on the size of the graph. Thus, the Infomap implementation was our choice to be used without modification in the spatial clustering of tourist destinations of Sect. 5.

3 Trip mining

In this section, we explain our method to mine trips from various LBSNs, namely Foursquare, Twitter and Flickr. Geo-tagged posts in LBSNs provide an incomplete view of a user's mobility, since a user's location is only recorded when she decides to share it. However, given the prevalent use of LBSNs on mobile devices, users often leave a nearly continuous spatio-temporal trace behind them. For example, if a user tweets using a mobile device and decides to enable the "Tweet with location" feature, her location will be recorded with every sent tweet. Similarly, if a user checks in at Foursquare venues, her presence at the venue at a particular time is recorded.

These posts can be seen as a continuous stream of check-ins: a check-in is a tuple of the unique identifier of a user, a location, and a timestamp. The precision of a location's coordinates does not have to be exact, but can also be on the granularity of destinations, such as cities or small islands. Since the data set does not include additional metadata, such as user profiles, users' home countries must be solely determined from the check-in stream. Literature lists several strategies for that, such as *Plurality*, the *geometric median*, or *nDays* (Kariryaa et al. 2018). Segmenting users' check-in stream into trips by periods of travel before returning home can then be done; however, the derived trips need to be checked for data quality, as some users might check in rarely, thus, their true location might be concealed.

3.1 Data sets

Human mobility has been of great interest to the scientific community, as it explains a lot about people's habits; however, location data is inherently privacy-sensitive and

[1] https://www.mapequation.org/code.html.

Table 1 Characteristics of the data sets

Feature	Foursquare	Flickr	Twitter
Number of users	266,909	214,204	2,662,741
Number of check-ins	33,263,633	48,469,177	263,926,396
Observation period	2012-04-03–2013-09-17	2001-07-22[a]–2014-04-26	2011-05-16[a]–2019-04-28

[a] Left 0.1% quantile

anonymizing it for research purposes is challenging, since correlating trajectories with single data points introduces many de-anonymization opportunities (de Montjoye et al. 2013). For this reason, location-based social networks are usually quite restrictive towards querying user location and enforce more or less strict API limits. In this paper, we analyze three data sets: a Foursquare data set from 2012/13 (Yang et al. 2015), the YFCC100M Flickr data set (Thomee et al. 2016), and a self-crawled Twitter data set from 2018/19. Bao et al. (2015) list further data sets stemming from LBSNs.

The raw data sets of Foursquare and Flickr are available following the respective references. Furthermore, we published the mined trips from all three data sets with redacted user identifiers and dates to protect the users' privacy.[2] Table 1 shows an general overview of the data sets. In the case of Flickr and Twitter, we sorted the check-ins chronologically and discarded the first 0.1%, since they were very sparse and potentially wrong, such as 1970-01-01 (the Unix timestamp 0).

3.1.1 Foursquare

Yang has published a check-in data set[3] stemming from Foursquare (Yang et al. 2015). It contains check-in data spanning 18 months (April 2012–September 2013) and 266,909 users at 3,680,126 venues in 77 countries; however, the data set only contains check-ins from the 415 most popular cities on Foursquare and, therefore, does not include data from travelers seeking recreation in the countryside. Table 2 shows how the distribution of the travelers' origin is influenced by the original data collection. The data set is interesting, because it features many users not residing in the Western countries. The large number of users in countries like Turkey and Indonesia is in line with reports on the regional popularity of Foursquare.

3.1.2 Flickr

The YFCC100M data set is described as the "largest public multimedia collection ever released" (Thomee et al. 2016). It comprises 100 million media objects, less than half of them enriched with geotags. The images were uploaded to Flickr

[2] https://github.com/LinusDietz/JITT2020-Mining-Trips-Replication.

[3] https://sites.google.com/site/yangdingqi/home/foursquare-dataset.

Table 2 Distribution of travelers' home country Foursquare

Home country	Fraction of travelers in %
Turkey	15.15
Brazil	14.28
USA	11.14
Japan	10.16
Indonesia	8.31
Chile	5.59
Malaysia	5.20
Mexico	4.71
Russia	2.85
Thailand	1.95
Other	20.66

Table 3 Distribution of travelers' home country Flickr

Home country	Fraction of travelers in %
USA	27.49
Great Britain	8.00
Spain	5.38
France	4.64
Canada	4.03
Germany	3.83
Japan	3.63
Australia	3.56
Italy	2.82
Brazil	2.49
Other	34.14

between 2004 and 2014 and have been published under a Creative Commons license. We only analyze the metadata of geotagged images for our purpose, since we are not interested in the images themselves. The distribution of user origin, cf. Table 3, is more diverse than in the other two data sets, with most users coming from highly developed countries.

3.1.3 Twitter

Twitter has been frequently used to analyze the individual mobility of the platform's users. The reasons for this are that—in contrast to other social media platforms for private communication—the content on Twitter is mostly public. Twitter also offers APIs to query information about it's users, including the approximate location of

Table 4 Distribution of travelers’ home country Twitter

Home country	Fraction of travelers in %
USA	60.38
Great Britain	11.39
Japan	4.20
Canada	2.91
Brazil	2.23
Germany	1.96
Mexico	1.72
Netherlands	1.62
Australia	1.51
Spain	1.32
Other	10.77

their tweets if they have enabled sharing the geolocation of their tweets. By querying the timelines of these users, we can follow their movement patterns.

We have continuously collected timelines of Twitter users since mid-2018 to build up a database of 267,853 timelines. These users tweet in all regions of the world, and the individual check-ins are matched to 24,186 cities with over 15,000 inhabitants each using the GeoNames Gazetteer.[4] As can be seen in Table 4, most users come from the United States, where Twitter is highly popular.

3.2 Method details

The chronologically sorted list of the check-ins of each user is segmented into periods of being at home and periods of travel. To determine the home location of the user, we use the plurality strategy, i.e., choosing the city with the highest number of check-ins. While this is the simplest heuristic to compute, literature shows that its accuracy is on par with more sophisticated methods such as the geometric median (Kariryaa et al. 2018). It may, however, be susceptible to the effects of commuting and users who predominately use social media when traveling. To reduce such false classifications, we discard travelers whose check-ins at home are fewer than a predefined threshold, in our case 50%. This threshold is an aggressive reduction of the data set, discarding about 95% of the users whose home city is unclear to us. It is, however, necessary, since incorrect classification of the user’s home would have severe consequences on the forthcoming analyses. The 50% cut-off could be lowered for analyses that are not so much dependent on the correct classification of the home location or the home location can be retrieved using other channels, such as a field in the user profile.

[4] http://download.geonames.org/export/dump/readme.txt.

$$\overbrace{\text{Munich } d_0}^{\textit{Home}} \rightarrow_2 \underbrace{\overbrace{d_3\ \textit{Paris}\ d_{11}}^{\text{Block}}}_{\text{Trip 1}} \rightarrow_0 \overbrace{d_{12}\ \text{Munich}\ d_{100}}^{\textit{Home}} \rightarrow_0$$

$$\rightarrow_0 \underbrace{\overbrace{d_{101}\ \textit{Paris}\ d_{101}}^{\text{Block}} \rightarrow_0 \overbrace{d_{101}\ \text{New York City}\ d_{105}}^{\text{Block}} \rightarrow_3 \overbrace{d_{109}\ \text{Washington, D.C.}\ d_{118}}^{\text{Block}}}_{\text{Trip 2}} \rightarrow_1$$

$$\rightarrow_1 \overbrace{d_{120}\ \text{Munich}}^{\textit{Home}} \rightarrow \dots$$

Fig. 1 Example of a user's check-in stream with two trips

Figure 1 exemplifies a check-in stream of a user from Munich. We define a number of continuous check-ins at one location as a block, which can be while traveling or while at home. In our example, the user's check-in stream starts on day 0 (d_0) in Munich and is followed by a block of 9 days (d_3–d_{11}) in Paris. In this case, the block is terminated by a check-in in Munich on the next day (d_{12}). Since Munich is the user's home location, the first trip is considered completed. This trip, thus, only consists of one block.

Staying at home for 83 days, the user is then observed checking-in in Paris on d_{101} and a few hours later in New York City. Since she was located in Munich the day before, it seems quite probable that she traveled from Munich to New York City with a stopover in Paris. The check-in stream shows several check-ins in New York City until d_{105} and continues with check-ins in Washington, D.C. from d_{109}–d_{118}, before the trip is again terminated by the return to Munich on d_{120}. Thus, this trip has a duration of 18 days and consists of three blocks.

The design decision to include short stopovers in the main trip was made due to the fact that stopovers can be often extended for several days without an increase in the flight ticket price. In fact, some travelers actively choose a city as a destination for the stopover; thus making it part of the trip. Having said that, there are some other uncertainties. As previously described, we only know the location of the user if she decides to check-in, tweet, or take a photo with a GPS tag. During a block, this is not much of an issue, since we assume that sequential check-ins at the same location mean that the user has not moved. The transition time, i.e., the time between two blocks, is more important, as it determines the duration the user has been at a location. In Fig. 1, we denote the transition time with a $\rightarrow_t$, where t is the days between the last check-in of the first block and the first check-in of the subsequent block. Trip 2 starts with perfect information: We know where the user was on all days from d_{100} to d_{105}. What we do not know is where she was on d_{106}–d_{108}, because the next check-in was just on d_{109} in Washington, D.C. Regarding the temporal segmentation of the blocks and trips, we follow a conservative strategy, which means that the block is terminated with the last check-in without the transition time. We think this strategy is sound, because it does not involve any speculation about the traveler's location. For example, at the end of trip 2, it is not clear when the traveler flew back from the United States to Munich. All we know that she was in Washington D.C. on d_{118} and in Munich on d_{120}. Since we do not have any evidence of the user's location, we do not add d_{119} to any of these blocks. The drawback of this is that the sum of the durations of a trip's blocks can be shorter than the duration of the trip.

3.3 Trip quality assurance

Using the aforementioned trip heuristic, one would potentially get a lot of trips comprising of only a single check-in. To filter out typical business trips, we only analyze trips with a minimum duration of 7 days. The maximum duration of the trips is set to 365 days, since longer durations are not considered a "visit" anymore in the recommendations for tourism statistics by the United Nations Department of Economic and Social Affairs (2010). Furthermore, we require the user to display relatively steady check-in behavior during travel. Thus, this section is about metrics that ensure a minimum quality of the check-in behavior.

The check-in frequency shown in Eq. (1) is not robust against a multitude of check-ins on 1 day, which makes it unsuitable for assessing the reliability of the check-in stream. In this regard, the better measure is the check-in density (Eq. 2), as it captures the fraction of days with a check-in during a trip. Thus, it captures how steady the check-in stream is, which is more important than having several check-ins at the same location on 1 day. We exclude trips that fall under the minimum check-in density of 0.2, which means that the user must have checked-in at least once in 5 days on average.

$$\text{Check-in frequency} = \frac{\text{check-ins}}{\text{days}} \tag{1}$$

The minimum value of check-in density should be chosen depending on the use case. For the purpose of analyzing global mobility patterns, we analyzed the consequences of enforcing a minimal check-in density. Figure 2 depicts the cumulative density function of the check-in densities of the trips. Since the curve is smooth and without an obvious "elbow", we set the threshold at 20%, which discards 32.88% of the trips. Recalling our initial goal with this heuristic, we reduced the mean transition time from 9.80 to 3.39 days while still keeping 67.12% of the trips.

$$\text{Check-in density} = \frac{\text{days with check-in}}{\text{days}} \tag{2}$$

3.4 Mobility metrics

Using this data-driven method, we obtain trips from each data set which we summarize in Table 5. The number of trips is the highest in the Twitter data set; the least amount of trips come from Flickr. The ratio of the number of foreign trips to domestic ones is about 1:19. While the metrics from the previous section were all about the quality of the users' check-in stream, the following metrics capture the mobility patterns of the users. We visualize the distribution of all metrics using the empirical cumulative distribution function (ECDF).

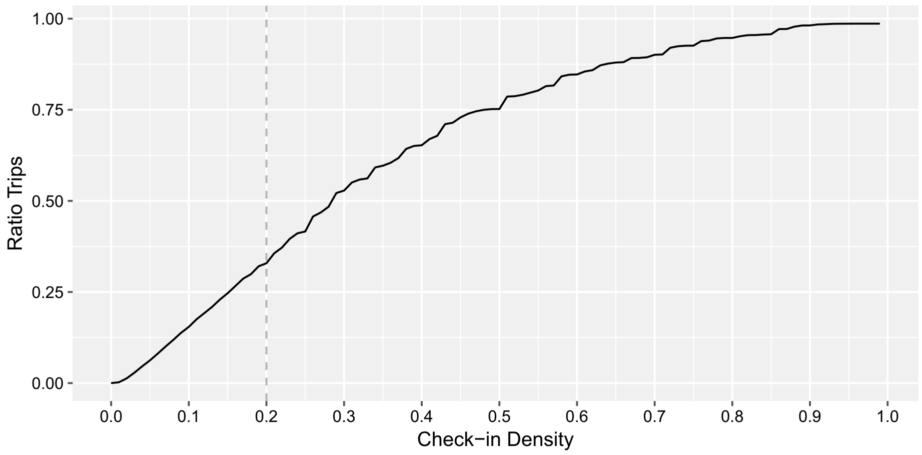


Fig. 2 Empirical distribution function of the check-in densities in the Foursquare data set

3.4.1 Trip duration

The trip duration is the number of days between the first and the last check-in of the trip. Figure 3 shows the cumulative distribution function of the trip durations. One can see the sharp increase in the curves of all three data sets. Overall, 90% of all trips are shorter than 30 days and 62% are shorter than 2 weeks. Flickr has an overall high mean duration of 31.59 days, followed by Twitter with 18.45 and Foursquare. When looking at the median, the data sets are quite similar with a value of 9 Foursquare, 11 for Twitter, and 12 in the case of Flickr. The main reason for the higher mean value of Flickr is that it has more very long trips in comparison to the other data sets. This is an indication that Flickr is used in a different way than the other LBSNs. It might also be that trips are not segmented correctly, possibly due to photographers mostly taking pictures when on travel as opposed to being at home.

3.4.2 Locations, blocks, and countries visited

We also analyzed the number of distinct locations, blocks, and countries within a trip. The number of locations is naturally lower than the number of blocks, since one location can be visited in several noncontiguous blocks of a trip. Figure 4 also reveals that in all data sets most trips span very few countries.

3.4.3 Check-in distance and radius of gyration

Check-in distance measures the mean geographic distance between two consecutive check-ins. This metric is heavily influenced by the check-in frequency. Thus, we prefer to use the radius of gyration for measuring how far the users traveled within a trip. To do so, we follow the definition of González et al. (2008). In simple terms, the radius of gyration measures the mean distance between the mean location of the trip to all other check-ins. It is, thus, more robust against skewed distributions of

Table 5 Trip statistics

Feature	Metric	Flickr	Foursquare	Twitter
Number	Trips	1254	20,317	133,614
Number	Travelers	1254	10,508	23,178
Number	Check-ins	96,111	101,759	2,665,987
Duration	Mean	31.59	11.70	18.45
Duration	SD	54.14	8.14	26.65
Duration	Max	362	222	364
Checkins	Mean	76.64	5.01	19.95
Checkins	SD	138.81	5.20	37.67
Checkins	Max	1063	355	991
Locations	Mean	22.39	3.16	5.27
Locations	SD	58.96	1.92	5.84
Locations	Max	942	25	284
Blocks	Mean	1.83	2.37	1.75
Blocks	SD	3.53	2.08	2.35
Blocks	Max	80	95	165
Countries	Mean	1.43	1.06	1.40
Countries	SD	1.33	0.28	0.93
Countries	Max	22	9	32
Checkin density	Mean	0.41	0.34	0.50
Checkin density	SD	0.22	0.13	0.23
Checkin density	Max	1.00	1.00	1.00
Radius of gyration	Mean	424.46	121.70	703.35
Radius of gyration	SD	1258.09	603.44	1435.82
Radius of gyration	Max	12,735.39	12,871.49	15,834.70
Displacement	Mean	295.19	186.58	1349.55
Displacement	SD	894.25	921.24	2484.86
Displacement	Max	13,183.50	16,279.58	19,430.03

The first three rows showcase the amount of data pruning in comparison to Table 1

check-ins than the check-in distance. In Fig. 5, one can see that the Twitter trips have the largest radius of gyration followed by the Flickr trips and the Foursquare trips.

3.4.4 Displacement

Displacement measures the distance between the user's home location and the mean position of the places visited during the trips. In our data, a similar trend as in the radius of gyration emerges: the Twitter users travel farther than the Foursquare and the Flickr users; however, Twitter shows a clearly slower increase of the distribution function than in the radius of gyration. The reason for this big difference is unclear. Possibly, the socio-economic background of some Twitter users is a different one than of the Foursquare users allowing them to make more intercontinental trips.

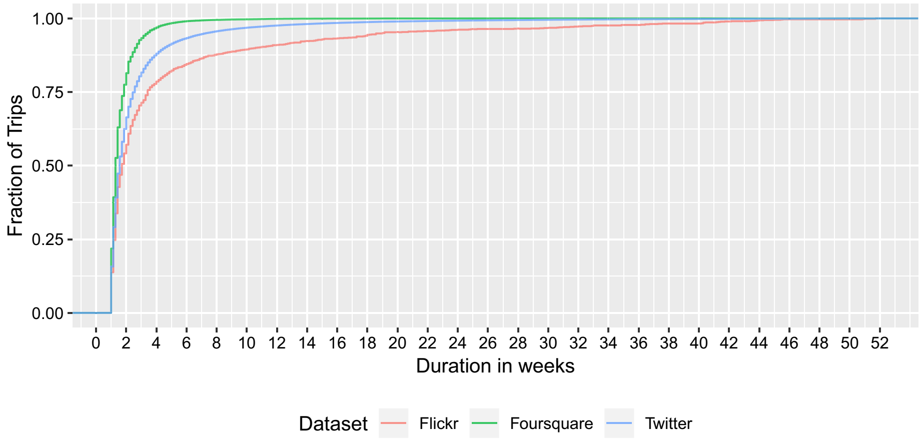


Fig. 3 ECDF of the trip duration

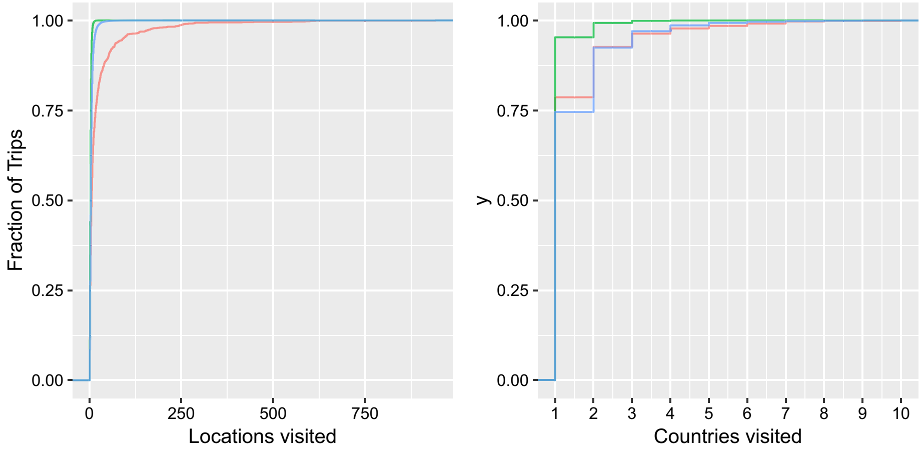


Fig. 4 ECDF of the number of visited locations and countries

3.4.5 Venue information

Finally, we looked at the types of venues checked into according to Foursquare's categorization.[5] Naturally, this information is only available for the Foursquare data set. Out of the ten top-level Foursquare categories, we took a subset of four of the most relevant categories to characterize a trip: Food, which comprises of restaurants and cafés, Nightlife, which are mostly bars and clubs, Arts and Entertainment, which also encompasses all kinds of cultural sites, and Outdoors and Recreation, which are parks and other sports-related sites. As can be seen in Fig. 6, Food is the most

[5] https://developer.foursquare.com/docs/api/venues/categories.

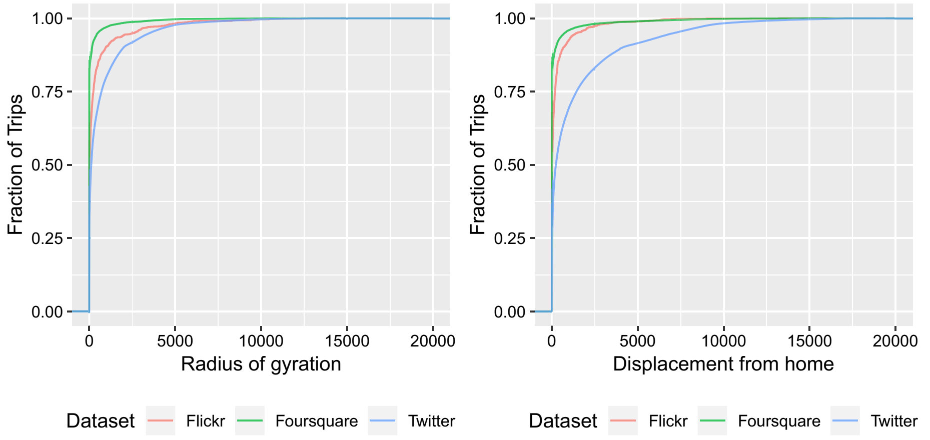


Fig. 5 Left: ECDF of the radius of gyration. Right: ECDF of the mean displacement

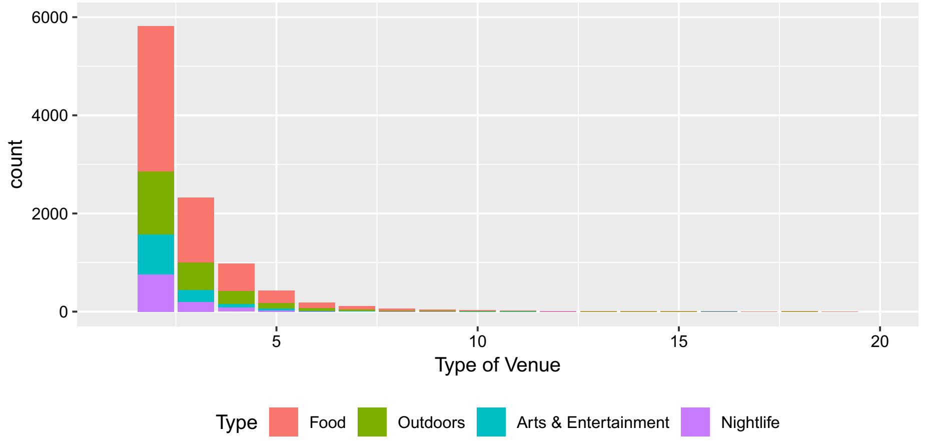


Fig. 6 Types of venues

common venue type, followed by Outdoors, Nightlife, and Arts and Entertainment check-ins. Furthermore, each trip typically has few check-ins that fall under these categories.

3.5 Summary

The proposed approach makes it possible to mine trips from the check-in stream of LBSN users. We have derived a number of metrics for the trips and have distinguished two types of metrics: first, metrics that capture the quality of the data, and second, metrics that capture the underlying mobility of the travelers. Making use of the former ones, we could determine which trips are plausible and sound, whereas the latter ones enable us to do further analyses in this domain.

The algorithms described in this section are implemented in a Python module, published under the permissive MIT License on Github.[6] It provides functionality for parsing the aforementioned data sets and can be extended for parsers of other data sets in about 30–50 lines of Python code.

We learned that not all LBSNs are equally good for such analysis, since users of a geosocial network like Foursquare comprise a different population than Twitter. As a consequence, Twitter users travel further and more often compared to users of Foursquare. Flickr is not well suited for this kind of analysis, since we observe that few people post geotagged pictures of their home. This makes it hard to determine their home locations with the information we have at hand and, thus, making this data set not well-suited for our further applications. Naturally, this sampling of trips based on the respective LBSN limits the generalizability of the findings. The proposed trip mining approach is tailored to LBSN data in large quantities. Since most of the data is not travel-related per se, the heuristics to filter the check-in information for touristic trips will inevitably throw away most of the data points. Furthermore, the data is not suitable for all analyses. For example, short holiday trips of 3–4 days are hard to distinguish from business trips of the same duration. For these reasons, we opted to only analyze long trips of a duration of at least 7 days.

The mined trips can serve as starting points for various improvements to recommender systems. First of all, they show the relative popularity of cities throughout the year. This can be used to increase the diversity of recommendations and, thus, avoid peak season visits for travelers who are sensitive to mass tourism. Furthermore, it shows patterns of destinations that are often visited together. In the composite destination recommendation scenario (Herzog et al. 2019; Dietz 2018), this can provide information on which cities should be combined. Finally, the trip data gives cues on how many destinations should be visited within a trip of a given length and also how long one should stay at each destination (Dietz and Wörndl 2019).

In the next section, we perform a cluster analysis to identify different kinds of trips. This can be useful for distinguishing traveler types in the preference elicitation phase of a travel recommender system. Another application for which we use the mined trips is touristic region discovery, which we describe in Sect. 5.

4 Clustering of traveler types

The first application of the mined trips for use in touristic information systems is a cluster analysis. Cluster analysis is the task of finding groups of data objects, where each group comprises similar objects, whereas the groups themselves are dissimilar to each other. This technique can uncover a structure within unlabeled data and is therefore categorized as unsupervised machine learning (Jain and Dubes 1988).

By revealing different kinds of tourist trips, we can offer insights about the general characteristics of different types of travelers. While this is an analytic result on its own, it can be directly used as part of user modeling within a recommender

[6] https://github.com/LinusDietz/tripmining.

system. But what kind of travelers are there? To answer this, we analyze trips from two data sources: Twitter and Foursquare. In the former, we have a pure mobility trajectory with a check-in granularity of cities, whereas in the latter, the check-ins are attributed to specific venues within a city with further information about the type of the venue.

4.1 Method

In both analyses, we follow the method introduced by Dietz et al. (2018b). The trips are characterized by several features derived from the check-in stream; however, not all metrics are useful for the analysis. Since the goal is to capture the underlying phenomena of the users’ travel behavior, we only use the metrics from Sect. 3.4 that capture the users’ mobility instead of the quality of data.

Among the remaining metrics, we perform a correlation analysis and remove redundant features, i.e., those whose correlation to another feature is very high. The threshold for this was set at a Pearson correlation coefficient > 0.75. The reason for this exclusion is that highly correlated features will not improve the segregation in the clustering algorithm.

Having decided upon the metrics, we normalize all features using min–max normalization and then run the K-means clustering algorithm with a different number of expected clusters. We evaluate the quality of the determined clusters in terms of the within-cluster sums of squares and the average silhouette (Rousseeuw 1987). The silhouette width measures how well a data object fits into its labeled cluster as opposed to all other clusters. Therefore, it is a robust and easy to interpret method that gives a broad overview of the overall solution quality, as well as information about each data object.

4.2 Case study 1: Twitter trips

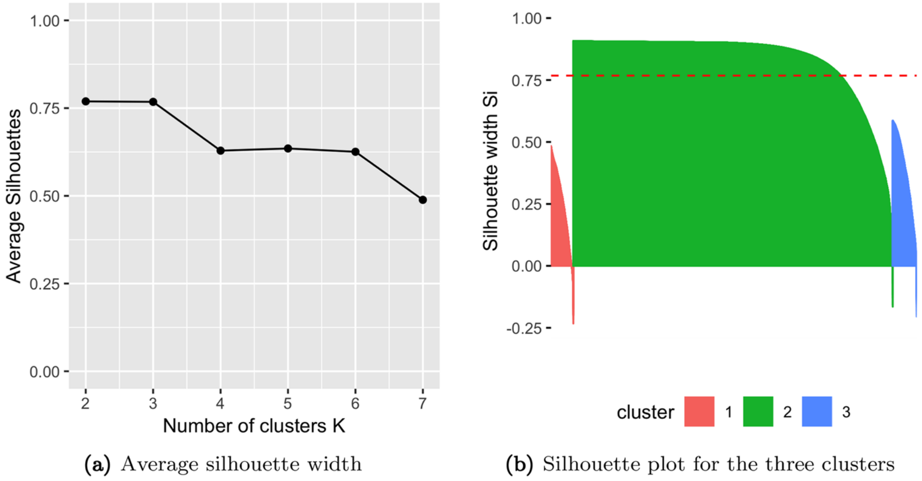
As already mentioned, the trips from Twitter are the most numerous. Due to memory limitations, we drew a random sample of 40,000 trips to run the clustering using K-means clustering. Since this is almost half of the trips, the sample is representative for the overall number of trips.

4.2.1 Features

As already mentioned, this case study is about the pure mobility of the travelers. After the correlation analysis, we could retain the four metrics for capturing the mobility: the duration of the trip, the number of locations, the number of blocks, the radius of gyration, and the displacement from home.

4.2.2 Results

Analyzing the results of the clustering from $K = 2$ to $K = 7$ clusters, there is always one dominant cluster and several smaller ones. According to the silhouette



(a) Average silhouette width (b) Silhouette plot for the three clusters

Fig. 7 Twitter: choosing the best number of clusters

width in Fig. 7a, the solution with two or three clusters has a higher quality, before it decreases to a lower plateau for $K \geq 4$. Thus, we chose the result of $K = 3$ as our final result.

Since we do not discriminate between international and domestic trips, unsurprisingly, most trips from the dominant cluster in green (cf. Fig. 7b) are domestic trips, with a low mean displacement of 565.57 km and only 1.22 countries on average. This result is potentially an outcome of most Twitter users residing in the USA. The two other clusters are smaller in number and more specialized. The Globetrotters travel further, visit the most countries, and display the highest radius of gyration. With 35 days duration, these trips are also the longest. Finally, the Distant Vacationers travel furthest from home, but are not so active during their travel. Their radius of gyration is only one third of the Globetrotters, despite visiting nearly as many distinct locations. Note that these names are only one way to name these clusters. We gave the clusters names to make it easier to refer to them and did our best to choose names that reflect the nature of the trips according to the clustering result (Table 6).

4.3 Case study 2: Foursquare trips

As opposed to the Twitter study and a previous analysis of Foursquare trips (Dietz et al. 2018b), we want to analyze what clusters are formed when taking the activities of the travelers into account.

4.3.1 Features

For this analysis, we use the mobility features of the trips enriched with the type of venues the travelers checked into on Foursquare. This results in the following

Table 6 Twitter: resulting clusters

	Domestic	Globetrotters	Distant vacationers
Ratio	87.1%	6.4%	6.5%
Silhouette width	0.83	0.25	0.38
Duration	19.65/53.45	34.99/96.03	25.05/59.18
Locations	5.13/6.6	9.73/12.35	8.18/7.92
Blocks	1.46/1.83	4.89/6.93	3.34/4.07
Countries	1.22/0.6	3.04/2.01	2.39/1.48
Radius of gyration	329.2/541.36	5172.25/2435.53	1677.74/1428.58
Displacement	565.57/887.83	5262.2/2323.11	8733.89/2834.18

Mean value/standard deviation

features: duration, countries visited, the displacement, the radius of gyration, and the number of check-ins in the categories Food, Nightlife, Arts and Entertainment, and Outdoors and Recreation. No features had to be removed after the correlation analysis.

4.3.2 Results

With this data set the results were more nuanced and a bigger number of clusters was found. The determination of the number of clusters using the silhouette width in Fig. 8a suggested six clusters. Analyzing the results summarized in Table 7 more closely, again a dominant cluster of “Short Domestic” trips arises with 76% of all trips residing in this group. These trips are on average the shortest, have the smallest radius of gyration, and are almost exclusively in the home country of the traveler, since the displacement is on average as small as 40 km. The other clusters are low in number and highly specialized.

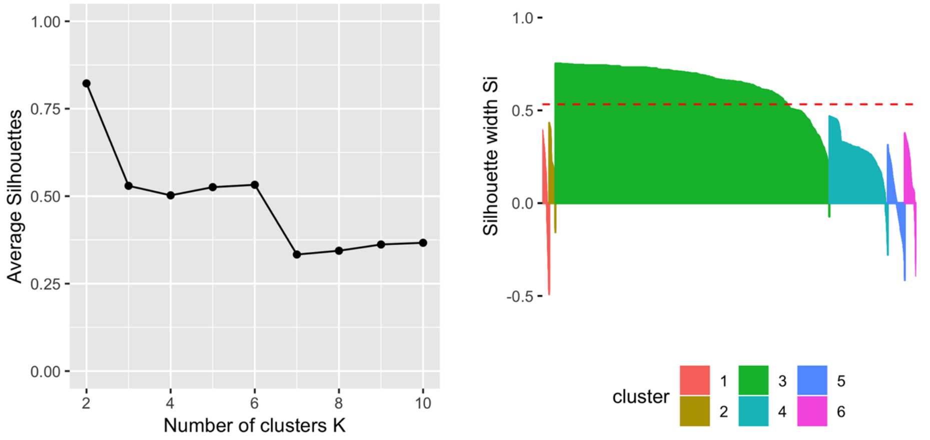
The “Party” trips are about 2-week long trips that visit around four cities in one country. They are distinguished by their high number of food check-ins and very high number of nightlife check-ins.

The “City” trips are quite similar to the domestic short trips, however, they visit more cities and these destinations are more distant to their home town.

The “Foreign” trips are about 2 week long trips to several cities located about 1700 km away from home. People travel quite extensively, as the radius of gyration of about 1300 km indicates.

The “World” trips are quite similar to the “Globetrotters” from Twitter. They visit the highest number of locations, travel the farthest, and also have the highest radius of gyration with nearly 4000 km.

Finally, there is the cluster of the “Long Domestic” trips that last about 6 weeks, which corresponds pretty well to the summer holidays of students at school or university. The small radius of gyration and the high number of outdoors and food



(a) Average silhouette width

(b) Silhouette plot for the six clusters

Fig. 8 Foursquare: choosing the best number of clusters

check-ins indicates that these trips might be monothematic vacations, e.g., at a beach resort during summer holidays.

4.4 Summary

This section described a method for finding tourist types using LBSN data. We presented two case studies of international and domestic trips stemming from Twitter and Foursquare. On Twitter, only three clusters emerged, whereas on Foursquare most trips resided in two clusters, with four more very specialized ones.

This method could be used to characterize prospective travelers from data about their past trips. Thus, it can be applied for preference elicitation and user modeling within recommender systems in tourism. Moreover, the analysis requires no user interaction, which is good for the user experience and is also computationally cheap; however, it requires access to the user's check-in history. This can be achieved through an app permission by which the user grants access to their timeline on an LBSN that they have been using, e.g., through a third-party Facebook or Twitter application. Obtaining the data in such a way, a classifier trained with this paper's approach can be used to classify the current user and, thus, be a foundation for providing personalized recommendations.

We have noticed that the cluster analysis results strongly depend on the input data. Developers of recommender systems should carefully evaluate only to include features that are useful for the preference elicitation and the recommendation outcome. Otherwise, the approach is at risk to overfit the data and reports outlier groups as happened in the Foursquare analysis. Finally, it seems to us that this kind of analysis is more suitable for analyzing international trips, as reported by Dietz et al. (2018b).

Table 7 Foursquare: resulting clusters

	Party	City	Foreign	World	Domestic long	Domestic short
Ratio	1.6%	14.7%	3.2%	1.2%	3.2%	76.2%
Silhouette	0.08	0.28	0.01	0.16	0.23	0.65
Duration	16.01/11.85	11.46/5.06	12.87/6.72	17.72/14.89	40.76/22.41	10.3/3.87
Locations	4.36/2.81	3.7/1.8	4.89/2.49	6.48/3.88	6.07/3.52	2.78/1.47
Blocks	3.24/2.8	2.53/1.8	4.14/2.26	5.71/3.67	5.5/6.06	2.07/1.45
Countries	1.06/0.24	1.02/0.14	1.85/0.54	2.39/1.05	1.03/0.17	1.01/0.1
Radius of gyration	107.65 /289.05	47.01 /138.24	1304.51 /963.4	3987.41 /2725.93	75.12 /216.55	27.85 /100.1
Displacement	192.69 /596.07	65.08 /204.52	1692.77 /1184.4	7224.18 /2678.58	95.05 /232.25	39.5 /146.52
Food check-ins	1.92/6.64	0.97/1.6	1.55/2.1	1.94/2.34	3.22/5.07	0.94/1.24
Arts check-ins	0.48/0.94	0.33/0.67	0.35/0.79	0.8/1.31	0.9/2.15	0.28/0.66
Outdoors check-ins	0.72/1.99	0.45/0.94	0.54/0.96	0.87/1.47	2.28/3.54	0.46/0.92
Nightlife check-ins	3.67/1.13	1.23/0.42	0.21/0.49	0.49/1	0.26/0.56	0/0

Mean value/standard deviation

5 Region discovery

The mobility of travelers manifested in the trips can be used for further applications. In this section, we describe a methodology to obtain a map of the world's travel regions that is entirely based on tourist travel behavior instead of political regions. With this approach we aim to uncover implicit tourist regions that are independent of administrative boundaries, e.g., in areas where travel can occur irrespective of national borders, such as the Schengen Area of Europe.

To achieve this, we construct a graph of flows from the trips and use a community detection algorithm to cluster single destinations into coherent travel regions (Sen and Dietz 2019). We use the Twitter data set described in Sect. 3.1.3, as it is the largest, most widespread, and most recent.

5.1 Method

As the Infomap algorithm (Rosvall et al. 2009) uses a weighted graph for community detection, we convert the trips into a graph of flows. Transforming the tourist trips into a graph is relatively straightforward, however, there are several options for quantifying the weights between the nodes.

5.1.1 Community detection

Infomap is a graph community detection algorithm that is designed to discover the underlying structure of the nodes and edges (Rosvall et al. 2009). It can be applied to large directed or undirected graphs and can yield multi-level hierarchies for communities. The algorithm accounts for weights of the edges and, thus, seems to be quite suitable for our application to the weighted flows and distances of tourist movement between cities. The algorithm tries to optimize communities to have more flows within themselves than other communities by using a random walker that traverses the graph. Since it uses a probabilistic model to find communities, the algorithm runs ten times to reduce the probability of obtaining a local minimum. Infomaps picks the best solution according to it's internal quality measure, the description length (Rosvall et al. 2009). In our approach, a community corresponds to a set of cities that form a region. Since the results of Infomap are hierarchical, it will return a tree of regions and subregions, depending on a graph-theoretic termination criterion. Choosing the right granularity of regions depends on the use case.

5.1.2 Graph creation

We transform the trips into an undirected graph, where each city is a node. To form the edges, we try to map the *traveled-together* relation, i.e. that two cities have been visited within the same trip, as closely as possible. The flow between two cities is computed by summing up the co-occurrences of the two nodes in a clique formed by all cities in a trip, over all trips. For example, if a trip consisted of travel from

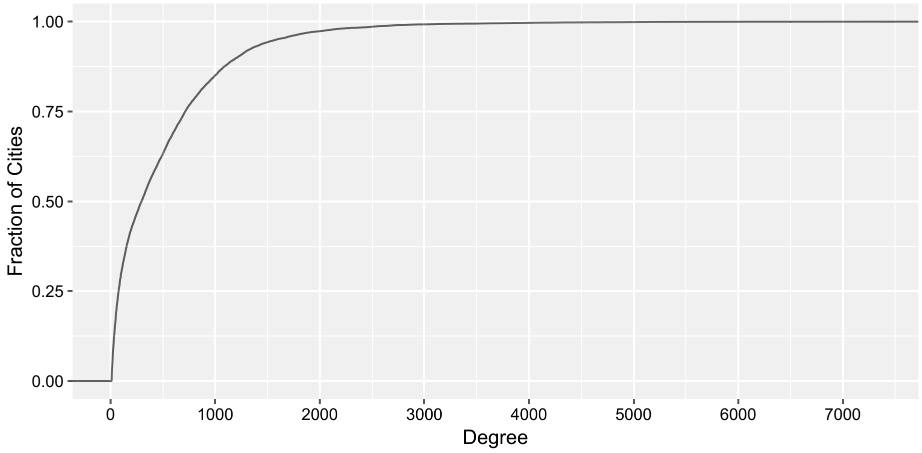


Fig. 9 ECDF of the node degrees

Munich to Berlin via Nuremberg, we would also count the flow from Munich to Berlin. The alternative of not adding this transitive connection to the weight is not appealing for us, since we think the *traveled-together* relation more accurately models the underlying mobility than one-to-one connections. In our example, this would mean that we lose the information that Munich and Berlin have been visited within the same trip. The final weight of each edge is the amount of flow divided by the Euclidean Distance between the two cities. Including the distance in the edge weight reduces the noise in the flow graph introduced by distant traffic hubs, such as airports. Transforming the mobility patterns into this graph-based representation enables us to run the Infomap community detection algorithm to see which cities form coherent clusters.

5.1.3 Graph description

The resulting graph consists of 14,558 nodes and 3,624,909 edges. The degree distribution depicted in Fig. 9 is long-tailed with very few high degree nodes and 87% of nodes having a degree of less than 1000. The graph density of 0.034 also indicates a sparse graph.

5.2 Results

The communities computed by the Infomap algorithm show four top-level regions that align well with existing continental boundaries. These are then further subdivided into a region hierarchy of up to five levels; however, for our purposes, level three and four are the interesting ones. On the second level, many regions are analogous to countries but there are a few interesting variations from this rule. The next level of regions tends to align mostly with travel destinations within federal states. The hierarchy and the discovered regions are discussed in detail in the following.

Table 8 Numerical description of the four top-level regions

Region name	Cities	2nd-level regions	3rd-level regions	4th-level regions
South America	1873	9	193	19
North and Central America	4193	17	254	145
Europe and West Africa	6591	14	381	116
Asia and Oceania	3201	13	114	196
World	15,858	53	942	476

Table 8 gives an overview of the hierarchy of the discovered regions. The South American region consists of the fewest cities, the European cluster has most. The number of second-level regions is small with the average number of cities in each region varying from 208 in South America to 470 in Europe. The third-level regions are clusters of about ten cities in South America, while the mean number of cities in the North American cluster and the European clusters is 17, and even 28 in Asia and Oceania. Most third-level clusters do not contain any sub-regions and most regions at lower-hierarchy levels are very small.

5.2.1 Level 1: continents

The four major regions found at the first level are loosely aligned with existing continental and cultural boundaries (cf. Fig. 10). The division between the Americas is a perfect cut between North and Central America and South America. Africa is under-represented in the data because it has only a few check-ins in Morocco, Algeria, Ghana, Nigeria, Kenya, and South Africa. These countries are merged in the European cluster with the exception of Kenya, which is in the Asian region. The European cluster is merged with all of Russia, Turkey, and the Arabian Peninsula. The Asian region comprises the Indian subcontinent, South East Asia, South Korea, Japan, Australia, and New Zealand.

On this level, the geography and the accessibility of the cities plays a dominant role. This explains the Arabian countries, which belong in the European cluster due to the major aviation hubs. The distance factor introduced to the edge weights seems to have a lower impact than the actual flows within the regions, since even the easternmost parts of Russia are clustered with Europe. Unfortunately, the lack of data from Africa and selected countries in Central Asia hinders the formation of clusters in these areas. We will discuss this limitation at the end of this section.

5.2.2 Level 2: countries

At the second level of the region hierarchy, we found that many regions align with national boundaries; however, there are exceptions observed in each of the top-level clusters.

In Europe a large second-level cluster is found spanning the countries of Germany, Austria, Switzerland, Hungary, the Czech Republic, Poland, and

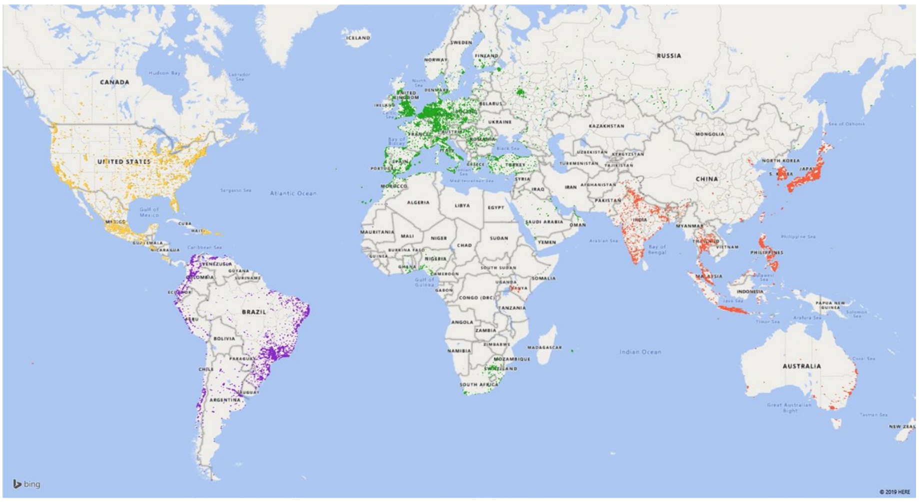


Fig. 10 The top-level regions

Romania (cf. Fig. 11). The Scandinavian countries are clustered together with Russia and the Baltic countries. Italy forms one region with Serbia, however, the unavailability of data from Croatia possibly influenced this result in an unpredictable manner. The Iberian countries are clustered with Morocco, which could be attributed to immigration patterns and the very cheap flight and ferry prices between these countries. Belgium and the Netherlands form another region, and also the British Isles are clustered together. France, Turkey and Greece, however, form regions identical to their national borders.

The second-level regions formed in North America in Fig. 12 mostly disregard national boundaries. Mexico is in one region with other Central American countries, while the USA and Canada are divided into fourteen clusters. The western Canadian states are merged together with Oregon and Washington, while California is split into two major clusters with the southern cluster expanding down to Tijuana and Mexicali in Mexico. Mexican cities on the borders of Arizona and Texas are also members of predominantly American clusters. Several other well-known regions, such as the New England states, Florida and the Great Lakes area form their own clusters.

In Asia (see Fig. 13), the Indian subcontinent and Pakistan are grouped together, which is surprising given the geopolitical context. Australia and New Zealand are clustered together, while the countries in South East and East Asia form their individual regions.

The second-level regions reveal that there are some countries that are traveled to exclusively, but other countries are more frequently traveled to together. In very large countries like Brazil and the USA, we observe a subdivision into multiple subregions at the second level. This is proof that the approach works well for domestic tourism regions.

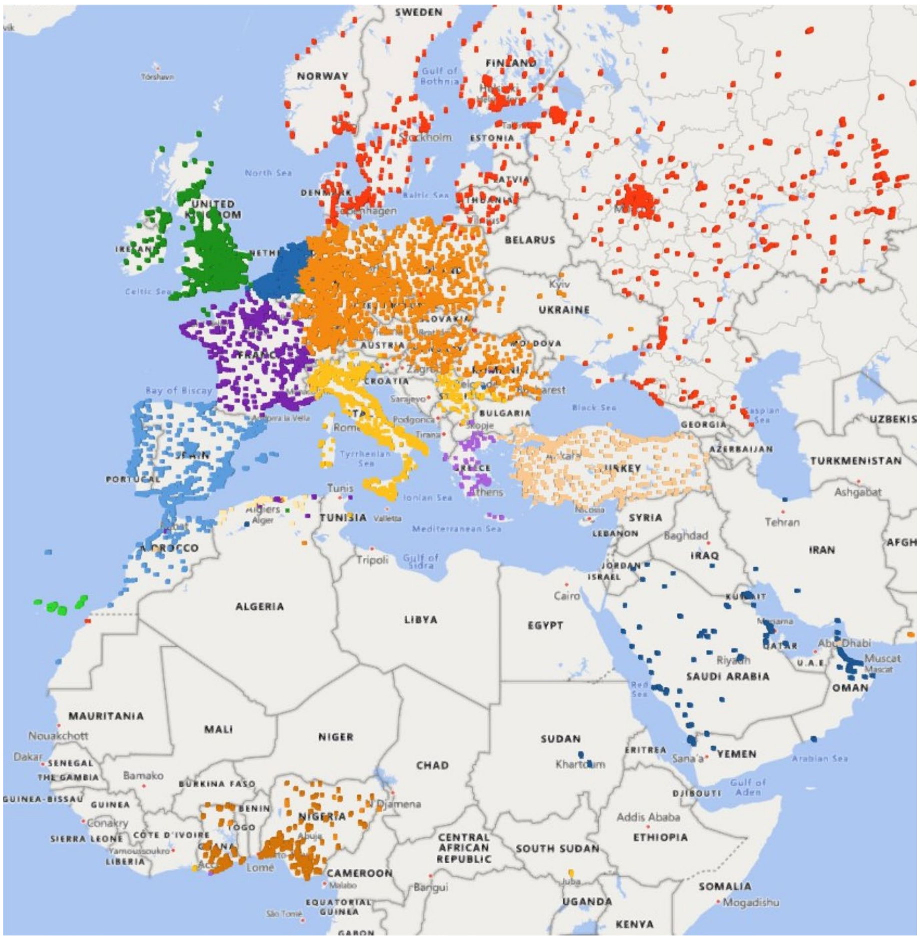


Fig. 11 The second-level community structure of Europe

5.2.3 Level 3: destinations

The regions formed at the third level of the hierarchy can be already seen as tourism destinations; however, the results show varying granularities in different parts of the world with some regions containing further subregions.

The third-level clusters of the big Central European cluster in Fig. 14 are varied in terms of the size and density of cities. The dense regions are typically very contiguous and are centered around a major city. For example, the region containing Munich is comparatively large and includes southern Bavaria. Large areas of the Czech Republic, Poland, and Hungary form homogeneous clusters with no further subregions.

Figure 15 shows that Belgium forms two regions at the third level, however, the Netherlands is divided into six regions that align well to the local divisions.

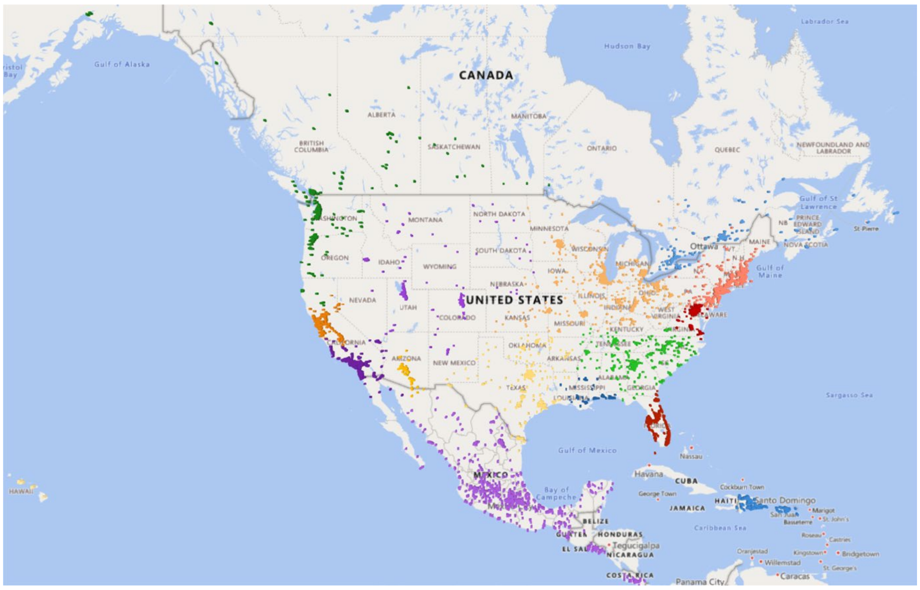


Fig. 12 The second-level communities of North America

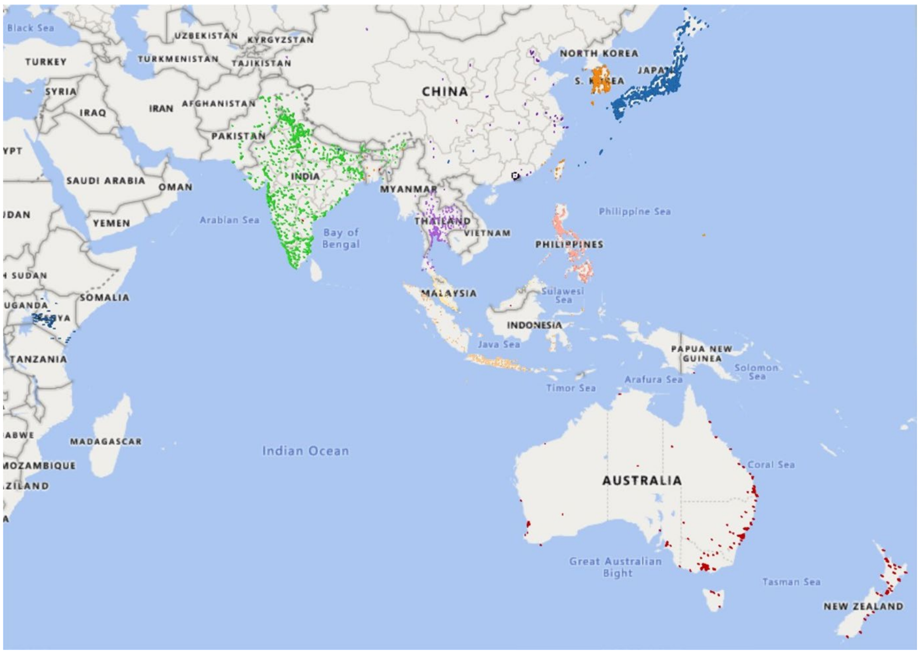


Fig. 13 The second-level community structure of South Asia

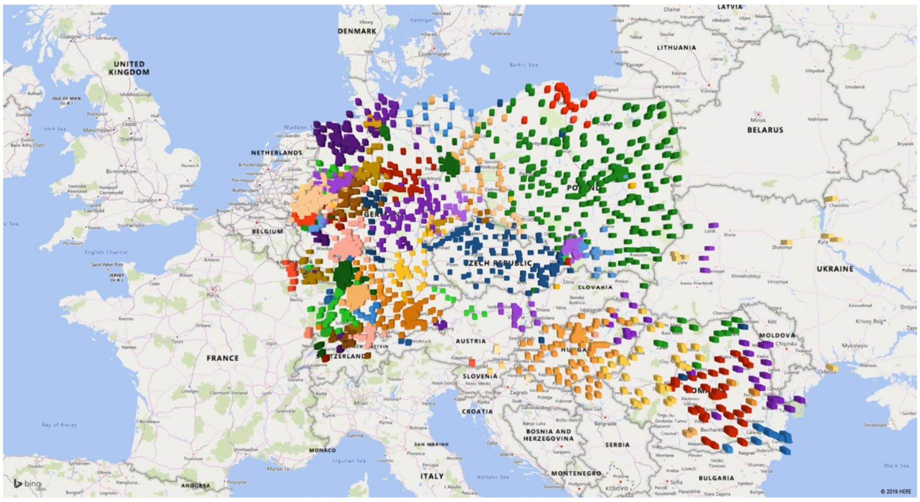


Fig. 14 The third-level community structure of the Central European cluster

Similar contiguous subdivisions are found in the British Isles, Spain, and Italy. The clustering of Morocco, cf. Fig. 16, with Gibraltar and Andalusia is interesting.

The third-level clusters in North America and South America show similar patterns to those in Europe with clusters being centered around cities, as can be seen in Fig. 17.

Pakistan and India are separated at this level with India forming four subregions (Fig. 18). In Thailand, one region is formed by places along the touristically very active coast, while the inland regions are divided into numerous smaller regions. The regions formed in Japan are similar to the political Japanese regions. Australia consists of three regions, one in Western Australia and two in the South East, while New Zealand forms it’s own region.

The third-level hierarchical result generally provides regions that can be seen as coherent tourist destinations. At this level, they become small enough to visit them exhaustively within few days and most do not contain further subregions.

5.2.4 Further levels

Some regions are further subdivided, which we discuss for the sake of completeness. In the third-level region of New York State, a fourth-level region with the Burroughs Manhattan, Bronx, Brooklyn, Queens, Staten Island, and Jersey City, which is not a part of New York City, is formed (cf. Fig. 19). Long Island contains two more regions, while four other regions surround New York City. This shows that if the Infomap algorithm obtains sufficient data, it is capable of discovering very fine-grained regions. This example of New York city is an artifact of the high-population density, the municipality structure, and the large amount of Twitter data in this area.

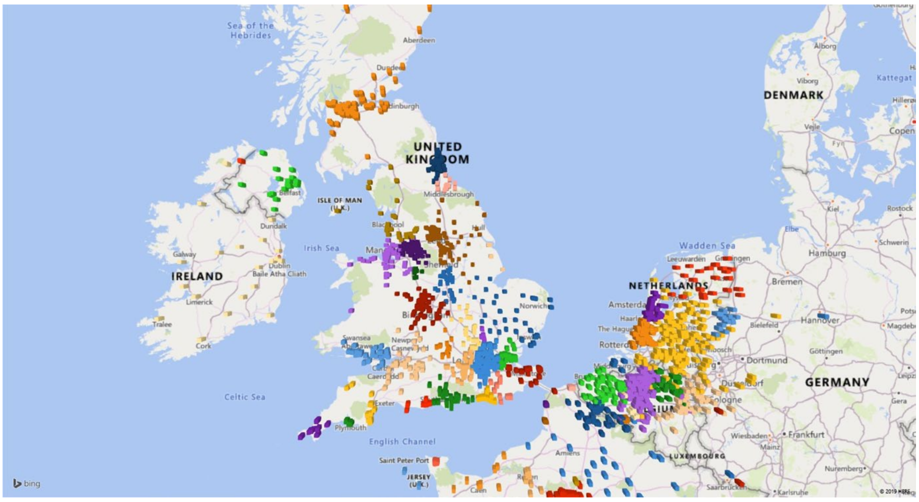


Fig. 15 The third-level community structure of the British and the Benelux cluster

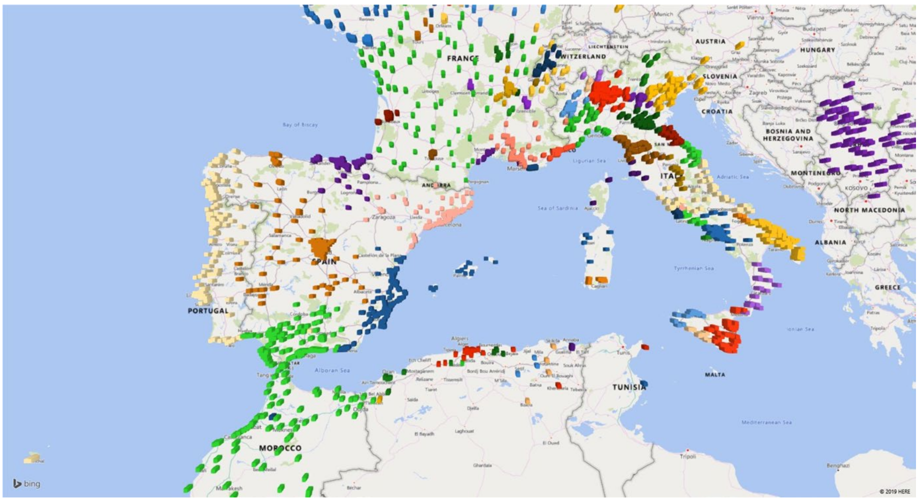


Fig. 16 The third-level community structure of the Iberian and Italian clusters

5.3 Summary

The first three levels of the cluster hierarchy roughly align to continents, countries, and travel destinations. At the second level of the hierarchy, we find that many countries form their own region, while larger and more populous countries, such as the USA and Brazil are subdivided at this level. India stands out, possibly due its lower per capita Twitter usage, and is only subdivided at the third level. Belgium and Netherlands as well as the Iberian countries forming common

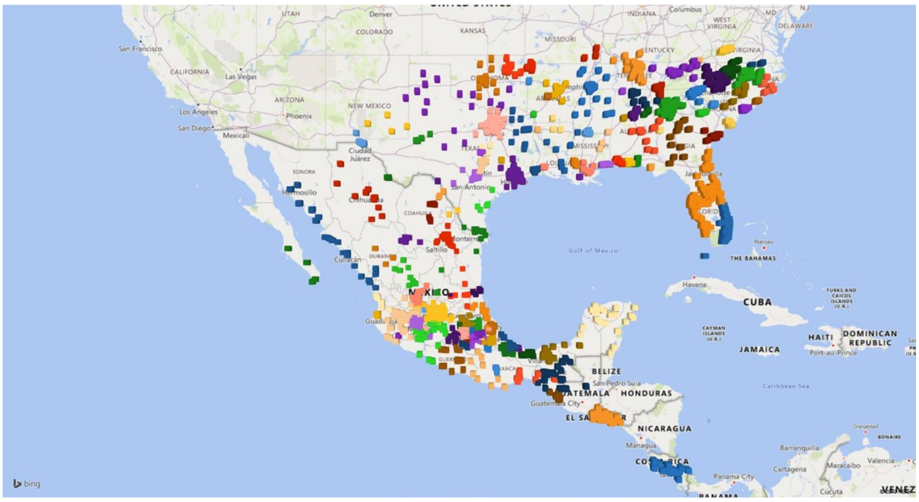


Fig. 17 The third-level community structure of south-east USA and Central America

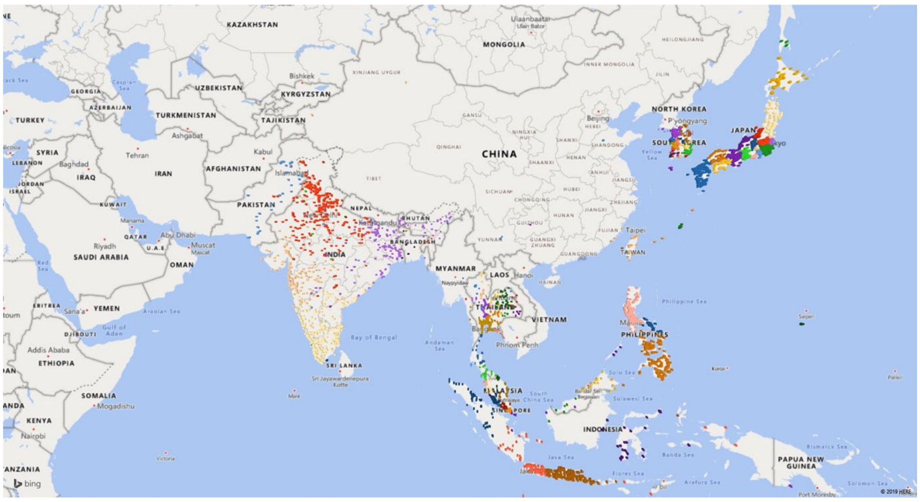


Fig. 18 The third-level community structure of South Asia

regions indicates a tendency for people to travel within countries that have close cultural ties. A similar grouping is formed by the German speaking countries of Austria, Switzerland, and Germany; however, the inclusion of Poland, Hungary, and Romania in the same cluster also underlines the high mobility within the European Union.

Sometimes, we would have expected different borders to be drawn, such as a clear separation at the outer border of the European Union. This is not the case in the countries around the Baltic Sea (cf. Fig. 11), where the Baltic States and

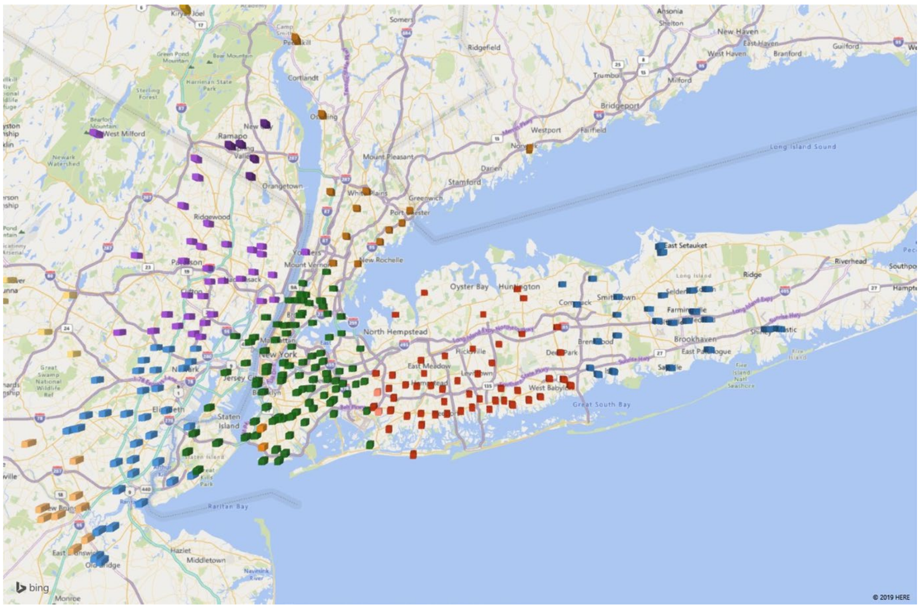


Fig. 19 The fourth-level community structure of New York

Russia are merged together on the second level of Europe. We attribute this to the high number of Russians living in these countries.

In some cases, also a fourth or fifth level exists in the hierarchy. The destination regions containing cities such as Los Angeles and London are found at the third level, while the third-level cluster containing New York consists of multiple fourth-level clusters, with the city of New York forming a cluster that is quite well aligned to the city burroughs. In India, the popular tourist states of Kerala and Rajasthan form fourth level clusters, while the rest of the country is decomposed to tourism destinations. These observations make a termination criterion for subdividing the regions an important problem. Additionally to this, a useful criterion would take into account whether the resulting areas fulfill a certain threshold for with regards to the area, the number of cities, and other metrics relevant to the purpose for the clustering. In our opinion, this cannot be decided with the current data, but requires a use-case-specific analysis of the regions; however, the third-level clusters are already very well defined and form understandable regions.

The stability of the algorithm’s output is quite high over several runs. We experimented a lot with Infomap and have not experienced noteworthy changes in the clustering given the same input parameters. This is mostly due to the fact that the algorithm does not recurse to deeper levels if the results become unstable due to insufficient data.

An important limitation of this approach is missing data. If the underlying data source is missing check-ins from a country for any reason, the algorithm does not have good means to counter this. In the case of China, where Twitter is a target of censorship (Bamman et al. 2012), independent clusters simply form around the large country. In the case of small countries with missing data such as Croatia or Belarus, the algorithm can ignore the missing data resulting in clusters that encompass the area, such as a sea.

Thus, this approach provides a fine-grained map of touristic travel regions in any information system concerning travel. Since the region model is hierarchical, its application scenarios are flexible and developers can pick the hierarchy that suits their needs best.

6 Conclusions

This paper presented three major contributions in the field of tourist recommender systems, mobility analysis and user modeling. The first is a metric-driven method to mine trips from various location-based social networks. We show how to extract domestic and international trips and ensure that the quality is sufficient for further analyses. By comparing several data sets, we find that the users display different mobility behavior on different platforms and that not all platforms are equally suited for this kind of analysis. For example, Flickr users typically have too few geotagged images, which results in only 1254 trips out of over 48 million check-ins.

Second, we present two case studies of cluster analyses of trips from Twitter and Foursquare. The purely mobility-based data set from Twitter revealed three clusters, while the Foursquare data that contained information about the type of venues was segmented into six clusters. This shows that the result is highly dependent on feature selection, which should be accounted for when using this method to classify users.

Finally, we presented an approach for the spatial clustering of touristic regions from Twitter trips. To the best of our knowledge, this is the first application of geolocated tweets to find travel regions with data spanning the whole world. The analysis of results reveals a hierarchy of regions, with tourist destinations residing on the third level. These results confirm that the use of volunteered geographic information to find traveler mobility patterns and define regions based on the patterns is a feasible approach.

The findings of this paper reveal much about how different user groups travel throughout the world. We have established a methodology that extracts travel trajectories from incomplete information sources. Naturally, not all LBSN sources are equally good for different use cases, but we have provided researchers and tourism analysts tools to evaluate this. Working with imperfect knowledge about the travelers' mobility has some limitations. Since we had to filter out many trips due to data quality issues, the results might be biased towards the behavior of travelers who continuously share their location on LBSNs. A generalization of the results should, thus, be done with care. Furthermore, the availability of spatio-temporal user data for independent research purposes is on the decline (Freelon 2018). For this reason, we had to work with two out-of date static data sets in the case of Foursquare and Flickr and had to put a lot of effort into building our own data set of Twitter trips using the official APIs. Other popular LBSNs, such as Facebook, Instagram, or Snapchat do not permit independent content extraction.

The cluster analyses of the trips provide tools to classify users in any tourism information system. This is useful, since knowing the type of traveler can be used to filter which items should be shown to the user. Performing the classification of a user can be done via the analysis of her past travel data, such as booking information or LBSN profiles, but also using a self-assessment of the traveler type. In the future, we plan to implement this in a global destination recommender system for composite trips, thereby extending previous approaches (Dietz 2018; Dietz et al. 2019). The results of the trip clustering approach showed a high dependence on choosing the right features for the given use case. For this reason, the resulting clusters should not necessarily be taken as a generally valid segmentation of traveler types; instead, the proposed method can be applied to determine the groups of users of one's own information system.

The discovered regions provide a hierarchical model of touristic regions. The advantage of this region model is that it is specific to the travel domain and is, thus, the preferable choice for visualizing regions in a travel recommender system over e.g., administrative boundaries. This resolves a problem, where previous systems had to make ad-hoc decisions on how to reasonably split large countries into smaller areas (Herzog and Wörndl 2014; Wörndl 2017). We think that this region model can help users to select their preferred travel destinations, especially in a composite trips scenario (Herzog et al. 2019) and to visualize various trends of global travel in more meaningful ways. While the output of the community detection algorithm itself were quite stable, the results might change with other data sources. Again, by using Twitter as the sole data source, people not active on this platform do not contribute to the region model. This is an important limitation, since this threatens the generalization of the results. In future, the results of different data sets should be compared systematically and also contrasted to official statistics about tourism movement.

Acknowledegments Open Access funding provided by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Bamman D, O'Connor B, Smith N (2012) Censorship and deletion practices in Chinese social media. First Monday. https://doi.org/10.5210/fm.v17i3.3943

Bao J, Zheng Y, Wilkie D, Mokbel M (2015) Recommendations in location-based social networks: a survey. GeoInformatica 19(3):525–565. https://doi.org/10.1007/s10707-014-0220-8

Bao J, Zheng Y, Mokbel MF (2012) Location-based and preference-aware recommendation using sparse geo-social networking data. In: 20th international conference on advances in geographic information systems, ACM, New York, NY, USA, SIGSPATIAL '12, pp 199–208. https://doi.org/10.1145/2424321.2424348

Blanford JI, Huang Z, Savelyev A, MacEachren AM (2015) Geo-located tweets. enhancing mobility maps and capturing cross-border movement. PLoS One 10(6):1–16. https://doi.org/10.1371/journal.pone.0129202

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 10:1–12. https://doi.org/10.1088/1742-5468/2008/10/P10008

Borràs J, Moreno A, Valls A (2014) Intelligent tourism recommender systems: a survey. Expert Syst Appl 41(16):7370–7389. https://doi.org/10.1016/j.eswa.2014.06.007

Braunhofer M, Elahi M, Ricci F (2014) Techniques for cold-starting context-aware mobile recommender systems for tourism. Intelligenza Artificiale 8(2):129–143. https://doi.org/10.3233/IA-140069

Burke RD (2007) Hybrid web recommender systems. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The adaptive web: methods and strategies of web personalization. Springer, Berlin, pp 377–408. https://doi.org/10.1007/978-3-540-72079-9_12

Burke RD, Ramezani M (2011) Recommender systems handbook, chap matching recommendation technologies and domains. Springer, Boston, pp 367–386. https://doi.org/10.1007/978-0-387-85820-3_11

Chaudhari K, Thakkar A (2019) A comprehensive survey on travel recommender systems. Arch Comput Methods Eng. https://doi.org/10.1007/s11831-019-09363-7

Cheng Z, Caverlee J, Lee K, Sui DZ (2011) Exploring millions of footprints in location sharing services. In: Fifth international conference on weblogs and social media, AAAI, Palo Alto, CA, USA, ICWSM '11, pp 81–88

Cohen E (1972) Towards a sociology of international tourism. Soc Res 39(1):164–182

de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. Sci Rep 3(1):1–5. https://doi.org/10.1038/srep01376

del Prado MN, Alatrista-Salas H (2016) Administrative regions discovery based on human mobility patterns and spatio-temporal clustering. In: 13th international conference on mobile ad hoc and sensor systems, IEEE, MASS'16, pp 65–74. https://doi.org/10.1109/mass.2016.019

Dietz LW (2018) Data-driven destination recommender systems. In: 26th conference on user modeling, adaptation and personalization, ACM, New York, NY, USA, UMAP '18, pp 257–260. https://doi.org/10.1145/3209219.3213591

Dietz LW, Weimert A (2018) Recommending crowdsourced trips on wOndary. In: RecSys workshop on recommenders in tourism, Vancouver, BC, Canada, RecTour'18, pp 13–17

Dietz LW, Wörndl W (2019) How long to stay where? On the amount of item consumption in travel recommendation. In: ACM RecSys 2019 late-breaking results, pp 31–35

Dietz LW, Herzog D, Wörndl W (2018a) Deriving tourist mobility patterns from check-in data. In: WSDM workshop on learning from user interactions, Los Angeles, CA, USA

Dietz LW, Roy R, Wörndl W (2018b) Characterisation of traveller types using check-in data from location-based social networks. In: Pesonen J, Neidhardt J (eds) Inf Commun Technol Tour. Springer, Cham, pp 15–26

Dietz LW, Myftija S, Wörndl W (2019) Designing a conversational travel recommender system based on data-driven destination characterization. In: ACM RecSys workshop on recommenders in tourism, pp 17–21

Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659(11):1–44. https://doi.org/10.1016/j.physrep.2016.09.002

Freelon D (2018) Computational research in the post-API age. Political Commun 35(4):665–668. https://doi.org/10.1080/10584609.2018.1477506

Gibson H, Yiannakis A (2002) Tourist roles: needs and the lifecourse. Ann Tour Res 29(2):358–383

González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature06958

Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260–271. https://doi.org/10.1080/15230406.2014.890072

Herzog D, Wörndl W (2014) A travel recommender system for combining multiple travel regions to a composite trip. CBRecSys@RecSys. Foster City, CA, USA, pp 42–48

Herzog D, Dietz LW, Wörndl W (2019) Tourist trip recommendations—foundations, state of the art and challenges. In: Augstein M, Herder E, Wolfgang W (eds) Personalized human–computer interaction. de Gruyter Oldenbourg, Berlin, pp 159–182

Hess A, Hummel KA, Gansterer WN, Haring G (2015) Data-driven human mobility modeling. ACM Comput Surv 48(3):1–39. https://doi.org/10.1145/2840722

Hsieh HP, Li CT, Lin SD (2012) Exploiting large-scale check-in data to recommend time-sensitive routes. In: ACM SIGKDD international workshop on urban computing, ACM, New York, NY, USA, UrbComp '12, pp 55–62. https://doi.org/10.1145/2346496.2346506

Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Upper Saddle River

Joshi D, Soh LK, Samal A (2009) Redistricting using heuristic-based polygonal clustering. In: Ninth IEEE international conference on data mining, IEEE, pp 830–835. https://doi.org/10.1109/ICDM.2009.126

Kariryaa A, Johnson I, Schöning J, Hecht B (2018) Defining and predicting the localness of volunteered geographic information using ground truth data. In: Conference on human factors in computing system, ACM, CHI'18. https://doi.org/10.1145/3173574.3173839

Kbaier MEBH, Masri H, Krichen S (2017) A personalized hybrid tourism recommender system. In: 2017 IEEE/ACS 14th international conference on computer systems and applications (AICCSA), pp 244–250. https://doi.org/10.1109/AICCSA.2017.12

McCrae RR, John OP (1992) An introduction to the five-factor model and its applications. Personality 60(2):175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McKercher B (2002) Towards a classification of cultural tourists. Int J Tour Res 4(1):29–38. https://doi.org/10.1002/jtr.346

Neidhardt J, Schuster R, Seyfang L, Werthner H (2014) Eliciting the users' unknown preferences. In: 8th ACM conference on recommender systems, ACM, New York, NY, USA, RecSys '14, pp 309–312. https://doi.org/10.1145/2645710.2645767

Noulas A, Scellato S, Mascolo C, Pontil M (2011) An empirical study of geographic user activity patterns in Foursquare. In: Fifth international conference on weblogs and social media, AAAI, Palo Alto, CA, USA, ICWSM '11, pp 570–573

Orman GK, Labatut V, Cherifi H (2011) On accuracy of community structure discovery algorithms. J Converg Inf Technol 6(11):283–292. https://doi.org/10.4156/jcit.vol6.issue11.32

Ouyang X, Zhang C, Zhou P, Jiang H (2016) Deepspace: an online deep learning framework for mobile big data to understand human mobility patterns. CoRR abs/1610.07009

Pearce PL (1982) The social psychology of tourist behavior. In: International series in experimental social psychology, vol, 3. Pergamon Press

Roick O, Heuser S (2013) Location based social networks—definition, current state of the art and research agenda. Trans GIS 5(17):763–784. https://doi.org/10.1111/tgis.12032

Rosvall M, Axelsson D, Bergstrom CT (2009) The map equation. Eur Phys J Spec Top 178(1):13–23. https://doi.org/10.1140/epjst/e2010-01179-1

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput Appl Math 20(1987):53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sen A, Dietz LW (2019) Identifying travel regions using location-based social network check-in data. Front Big Data. https://doi.org/10.3389/fdata.2019.00012

Sertkan M, Neidhardt J, Werthner H (2017) Mapping of tourism destinations to travel behavioural patterns. In: Stangl B, Pesonen J (eds) Information and communication technologies in tourism. Springer International Publishing, Cham, pp 422–434. https://doi.org/10.1007/978-3-319-72923-7_32

Song C, Koren T, Wang P, Barabási AL (2010a) Modelling the scaling properties of human mobility. Nat Phys 6(10):818–823. https://doi.org/10.1038/nphys1760

Song C, Qu Z, Blumm N, Barabási AL (2010b) Limits of predictability in human mobility. Science 327(5968):1018–1021. https://doi.org/10.1126/science.1177170

Taniguchi Y, Monzen D, Ariestien LS, Ikeda D (2015) Discover overlapping topical regions by geosemantic clustering of tweets. In: 29th international conference on advanced information networking and applications workshops, IEEE, pp 552–557. https://doi.org/10.1109/waina.2015.85

Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ (2016) YFCC100M: the new data in multimedia research. Commun ACM 59(2):64–73. https://doi.org/10.1145/2812802

Tsai CY, Paniagua G, Chen YJ, Lo CC, Yao L (2019) Personalized tour recommender through geotagged photo mining and LSTM neural networks. MATEC Web Conf. https://doi.org/10.1051/mateconf/201929201003

United Nations Department of Economic and Social Affairs (2010) International recommendations for tourism statistics 2008. https://unstats.un.org/unsd/tradekb/Knowledgebase/50551/IRTS-2008

Wang D, Pedreschi D, Song C, Giannotti F, Barabási AL (2011) Human mobility, social ties, and link prediction. In: 17th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD'11, pp 1100–1108. https://doi.org/10.1145/2020408.2020581

Wörndl W (2017) A web-based application for recommending travel regions. In: Adjunct publication of the 25th conference on user modeling, adaptation and personalization, ACM, New York, NY, USA, UMAP '17, pp 105–106. https://doi.org/10.1145/3099023.3099031

Yang D, Zhang D, Chen L, Qu B (2015) NationTelescope: monitoring and visualizing large-scale collective behavior in LBSNs. J Netw Comput Appl 55:170–180. https://doi.org/10.1016/j.jnca.2015.05.010

Yiannakis A, Gibson H (1992) Roles tourists play. Ann Tour Res 19(2):287–303. https://doi.org/10.1016/0160-7383(92)90082-z

Zhang Y, Wang L, Zhang YQ, Li X (2012) Towards a temporal network analysis of interactive WiFi users. Europhys Lett. https://doi.org/10.1209/0295-5075/98/68002

Zheng Y, Xie X (2011) Learning travel recommendations from user-generated GPS traces. ACM Trans Intell Syst Technol 2(1):1–29. https://doi.org/10.1145/1889681.1889683

Zheng W, Huang X, Li Y (2017) Understanding the tourist mobility using GPS: where is the next place? Tour Manag 59:267–280. https://doi.org/10.1016/j.tourman.2016.08.009

Zheng Y, Zhang L, Xie X, Ma WY (2009) Mining interesting locations and travel sequences from GPS trajectories. In: 18th international world wide web conference, ACM, New York, NY, USA, WWW'09. https://doi.org/10.1145/1526709.1526816

Zheng W, Zhou R, Zhang Z, Zhong Y, Wang S, Wei Z, Ji H (2019) Understanding the tourist mobility using GPS: how similar are the tourists? Tour Manag 71:54–66. https://doi.org/10.1016/j.tourman.2018.09.019

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.2 Publication 2: "Travelers vs. Locals: The Effect of Cluster Analysis in Point-of-Interest Recommendation"

Pablo Sánchez and Linus W. Dietz. "Travelers vs. Locals: The Effect of Cluster Analysis in Point-of-Interest Recommendation." In: *30th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP'22. New York, NY, USA: ACM, July 2022, pp. 132–142. DOI: 10.1145/3503252.3531320

Summary

The involvement of geographic information differentiates point-of-interest recommendation from traditional product recommendation. This geographic influence is usually manifested in the effect of users tending toward visiting nearby locations, but further mobility patterns can be used to model different groups of users. In this study, we characterize the check-in behavior of local and traveling users in a global Foursquare check-in data set. Based on the features that capture the mobility and preferences of the users, we obtain representative groups of travelers and locals through an independent cluster analysis. Interestingly, for locals, the mobility features analyzed in this work seem to aggravate the cluster quality, whereas these signals are fundamental in defining the traveler clusters. To measure the effect of such a cluster analysis when categorizing users, we compare the performance of a set of recommendation algorithms, first on all users and then on each user group separately in terms of ranking accuracy, novelty, and diversity. Our results on the Foursquare data set of 139,270 users in five cities show that locals, despite being the most numerous groups of users, tend to obtain lower values than the travelers in terms of ranking accuracy while they also seem to receive more novel and diverse POI recommendations. For travelers, we observe the advantages of popularity-based recommendation algorithms in terms of ranking accuracy by recommending venues related to transportation and large commercial establishments. However, there are considerable differences in the respective traveler groups, especially between predominantly domestic and international travelers. Due to the considerable influence of mobility on the recommendations, this article underlines the importance of analyzing user groups differently when making and evaluating personalized point-of-interest recommendations.

Author Contributions

PS and LD declare that they have equally contributed to the project. PS was responsible for the recommendation evaluation framework and the statistical analyses of the results. LD was responsible for the characterization and clustering of the traveler and locals. They jointly edited the paper.

Travelers vs. Locals: The Effect of Cluster Analysis in Point-of-Interest Recommendation

Pablo Sánchez
pablo.sanchezp@uam.es
Information Retrieval Group
Universidad Autónoma de Madrid
Madrid, Spain

Linus W. Dietz
linus.dietz@tum.de
Department of Informatics
Technical University of Munich
Garching, Germany

ABSTRACT

The involvement of geographic information differentiates point-of-interest recommendation from traditional product recommendation. This geographic influence is usually manifested in the effect of users tending toward visiting nearby locations, but further mobility patterns can be used to model different groups of users. In this study, we characterize the check-in behavior of local and traveling users in a global Foursquare check-in data set. Based on the features that capture the mobility and preferences of the users, we obtain representative groups of travelers and locals through an independent cluster analysis. Interestingly, for locals, the mobility features analyzed in this work seems to aggravate the cluster quality, whereas these signals are fundamental in defining the traveler clusters. To measure the effect of such a cluster analysis when categorizing users, we compare the performance of a set of recommendation algorithms, first on all users together, and then on each user group separately in terms of ranking accuracy, novelty, and diversity. Our results on the Foursquare data set of 139,270 users in five cities show that locals, despite being the most numerous groups of users, tend to obtain lower values than the travelers in terms of ranking accuracy while they also seem to receive more novel and diverse POI recommendations. For travelers, we observe the advantages of popularity-based recommendation algorithms in terms of ranking accuracy, by recommending venues related to transportation and large commercial establishments. However, there are huge differences in the respective travelers groups, especially between predominantly domestic and international travelers. Due to the large influence of mobility on the recommendations, this article underlines the importance of analyzing user groups differently when making and evaluating personalized point-of-interest recommendations.

CCS CONCEPTS

• **Information systems → Recommender systems**; **Retrieval effectiveness**.

KEYWORDS

Point-of-Interest recommendation, User Modeling, Human mobility analysis, Offline evaluation

ACM Reference Format:
Pablo Sánchez and Linus W. Dietz. 2022. Travelers vs. Locals: The Effect of Cluster Analysis in Point-of-Interest Recommendation. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22), July 4–7, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503252.3531320

1 INTRODUCTION

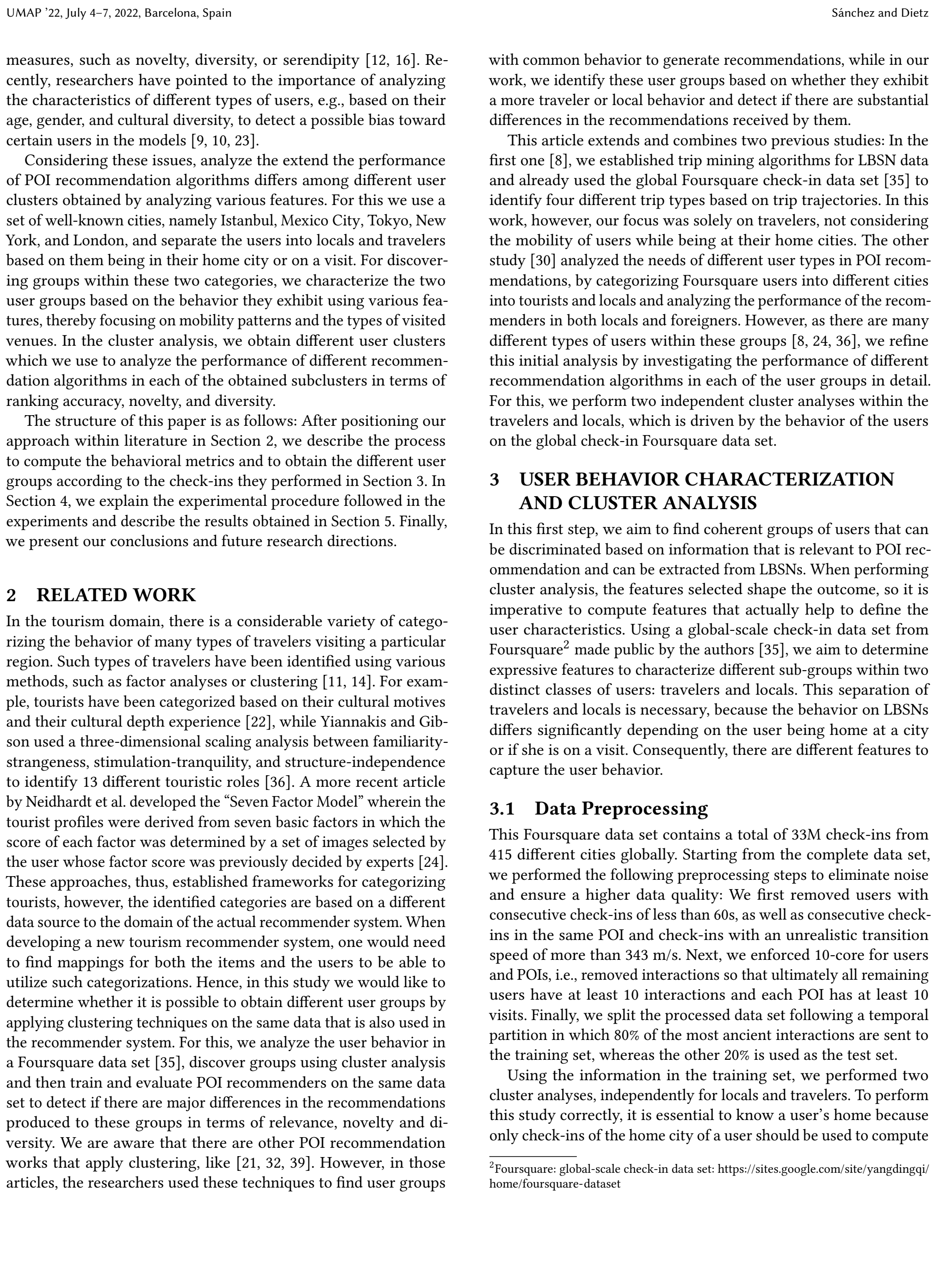
Recommender systems are prevalent in numerous areas including videos or movies (Netflix, Youtube), books (Goodreads), consumer products (Amazon), or social contact recommendations (Twitter, LinkedIn) [28]. In the travel and tourism domain, point-of-interest (POI) recommendation is an interesting challenge, where the items to be recommended are venues to be visited when the users arrive at a specific city or region [5, 33]. To perform POI recommendations, much of the data available to the scientific community stems from location-based social networks (LBSNs), such as Foursquare, Gowalla, or Yelp [3, 31]. LBSNs are so frequently used in research because the data usually comprises of several countries and provides additional information about social interactions between the users. Despite the richness and availability of LBSN data, POI recommendation has specific aspects that differ from the conventional recommendation of movies, books, or music that affect the recommenders' performance, including, the implicit information and repeated interactions, as users may check into at the same venue more than once; the relevance of external influences, such as social, temporal, sequential, and, most importantly, the geographical influence, since users tend to visit nearby locations [18, 19, 31]. Finally, the sparsity of the interaction data is typically more severe: For example, the global check-in Foursquare data set from [35] has a density of 0.0034%, making recommendations more difficult than the traditional scenario, such as the well-known Movielens25M data set[1] with a density of 0.2489%.

In addition to the abovementioned issues that affect the performance of the recommenders, we must also consider the different types of users that can be found in LBSNs. Traditionally, when measuring the recommendation quality in offline settings, all users are treated in the same way; hence, there is hardly a recommender systems study that does not report accuracy metrics for each algorithm such as Precision or nDCG averaged over all users, although the focus of evaluation has shifted from only accuracy to further

UMAP '22, July 4–7, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9207-5/22/07…$15.00
https://doi.org/10.1145/3503252.3531320

[1] Movielens25M data set: https://grouplens.org/datasets/movielens/



measures, such as novelty, diversity, or serendipity [12, 16]. Recently, researchers have pointed to the importance of analyzing the characteristics of different types of users, e.g., based on their age, gender, and cultural diversity, to detect a possible bias toward certain users in the models [9, 10, 23].

Considering these issues, analyze the extend the performance of POI recommendation algorithms differs among different user clusters obtained by analyzing various features. For this we use a set of well-known cities, namely Istanbul, Mexico City, Tokyo, New York, and London, and separate the users into locals and travelers based on them being in their home city or on a visit. For discovering groups within these two categories, we characterize the two user groups based on the behavior they exhibit using various features, thereby focusing on mobility patterns and the types of visited venues. In the cluster analysis, we obtain different user clusters which we use to analyze the performance of different recommendation algorithms in each of the obtained subclusters in terms of ranking accuracy, novelty, and diversity.

The structure of this paper is as follows: After positioning our approach within literature in Section 2, we describe the process to compute the behavioral metrics and to obtain the different user groups according to the check-ins they performed in Section 3. In Section 4, we explain the experimental procedure followed in the experiments and describe the results obtained in Section 5. Finally, we present our conclusions and future research directions.

2 RELATED WORK

In the tourism domain, there is a considerable variety of categorizing the behavior of many types of travelers visiting a particular region. Such types of travelers have been identified using various methods, such as factor analyses or clustering [11, 14]. For example, tourists have been categorized based on their cultural motives and their cultural depth experience [22], while Yiannakis and Gibson used a three-dimensional scaling analysis between familiarity-strangeness, stimulation-tranquility, and structure-independence to identify 13 different touristic roles [36]. A more recent article by Neidhardt et al. developed the “Seven Factor Model” wherein the tourist profiles were derived from seven basic factors in which the score of each factor was determined by a set of images selected by the user whose factor score was previously decided by experts [24]. These approaches, thus, established frameworks for categorizing tourists, however, the identified categories are based on a different data source to the domain of the actual recommender system. When developing a new tourism recommender system, one would need to find mappings for both the items and the users to be able to utilize such categorizations. Hence, in this study we would like to determine whether it is possible to obtain different user groups by applying clustering techniques on the same data that is also used in the recommender system. For this, we analyze the user behavior in a Foursquare data set [35], discover groups using cluster analysis and then train and evaluate POI recommenders on the same data set to detect if there are major differences in the recommendations produced to these groups in terms of relevance, novelty and diversity. We are aware that there are other POI recommendation works that apply clustering, like [21, 32, 39]. However, in those articles, the researchers used these techniques to find user groups with common behavior to generate recommendations, while in our work, we identify these user groups based on whether they exhibit a more traveler or local behavior and detect if there are substantial differences in the recommendations received by them.

This article extends and combines two previous studies: In the first one [8], we established trip mining algorithms for LBSN data and already used the global Foursquare check-in data set [35] to identify four different trip types based on trip trajectories. In this work, however, our focus was solely on travelers, not considering the mobility of users while being at their home cities. The other study [30] analyzed the needs of different user types in POI recommendations, by categorizing Foursquare users into different cities into tourists and locals and analyzing the performance of the recommenders in both locals and foreigners. However, as there are many different types of users within these groups [8, 24, 36], we refine this initial analysis by investigating the performance of different recommendation algorithms in each of the user groups in detail. For this, we perform two independent cluster analyses within the travelers and locals, which is driven by the behavior of the users on the global check-in Foursquare data set.

3 USER BEHAVIOR CHARACTERIZATION AND CLUSTER ANALYSIS

In this first step, we aim to find coherent groups of users that can be discriminated based on information that is relevant to POI recommendation and can be extracted from LBSNs. When performing cluster analysis, the features selected shape the outcome, so it is imperative to compute features that actually help to define the user characteristics. Using a global-scale check-in data set from Foursquare[2] made public by the authors [35], we aim to determine expressive features to characterize different sub-groups within two distinct classes of users: travelers and locals. This separation of travelers and locals is necessary, because the behavior on LBSNs differs significantly depending on the user being home at a city or if she is on a visit. Consequently, there are different features to capture the user behavior.

3.1 Data Preprocessing

This Foursquare data set contains a total of 33M check-ins from 415 different cities globally. Starting from the complete data set, we performed the following preprocessing steps to eliminate noise and ensure a higher data quality: We first removed users with consecutive check-ins of less than 60s, as well as consecutive check-ins in the same POI and check-ins with an unrealistic transition speed of more than 343 m/s. Next, we enforced 10-core for users and POIs, i.e., removed interactions so that ultimately all remaining users have at least 10 interactions and each POI has at least 10 visits. Finally, we split the processed data set following a temporal partition in which 80% of the most ancient interactions are sent to the training set, whereas the other 20% is used as the test set.

Using the information in the training set, we performed two cluster analyses, independently for locals and travelers. To perform this study correctly, it is essential to know a user’s home because only check-ins of the home city of a user should be used to compute

[2] Foursquare: global-scale check-in data set: https://sites.google.com/site/yangdingqi/home/foursquare-dataset

their behavior as a local; likewise, a user's travel behavior should solely be characterized using check-ins outside of the home city. For determining a user's home in the context of LBSN check-in data, several methods exist [17]; however, taking the city where most check-ins are done consistently produces highly accurate results when used along with a threshold. As such, we determined exactly one city for each user as home city using a threshold of at least 50% of check-ins needed to be performed in the most frequent city. This step excludes another 8,548 (6.20%) users with an unclear home from the training data, resulting in 129,294 valid users in the training set.

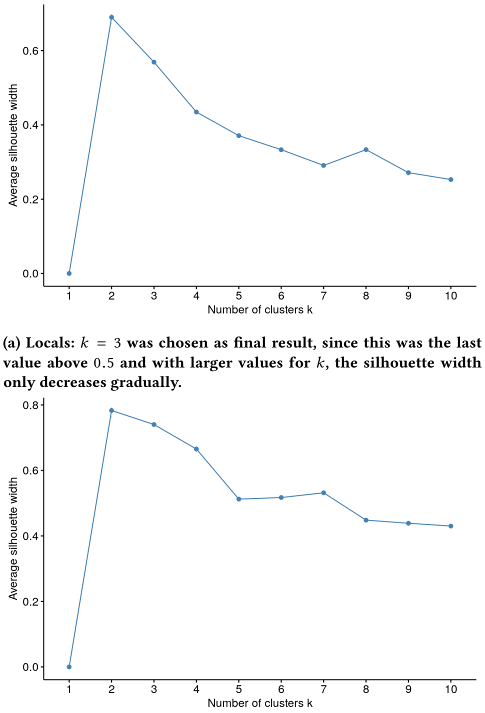
3.2 Local Behavior Cluster Analysis

To discover distinct groups of user activity in their home town, we exclusively analyzed check-ins they have performed in their home cities and computed various features including mobility metrics, such as the radius of gyration, the mean distance from the city center, and the mean distance between consecutive check-ins. Further features describe the activity of the users, e.g., the mean time between check-ins, the activity period, the number of check-ins, and the number of unique POIs visited. Finally, we also count check-ins in relevant categories separately, such as visiting POIs labeled with "Arts & Entertainment," "Outdoors & Recreation," "Food," "Nightlife Spot," and "Shops & Services."

First, we analyze correlations between features and eliminate those that have a high correlation > 0.7, as they are redundant. Likewise, we eliminate features that are orthogonal to all other ones identified by very low correlations with other attributes $[-0.1; 0.1]$. These features essentially treated as noise by the clustering algorithm and, thus, decrease the quality of the discovered groups. Concretely, this step resulted in the elimination of the following metrics: mean check-ins per day, total number of check-ins, and the number of check-ins in "Colleges & Universities."

Using the k-means algorithm, we systematically analyzed the outcome of the algorithm using the Euclidean Distance and min-max normalized features. Observing the quality of the resulting clusters using different values for k, we observed that the quality of the segmentation quality to be very low, despite having performed the relevant steps of the prior correlation analysis. Experimenting with different feature combinations, the silhouette width ranged in the area of 0.3 for 3–4 clusters and further dropped with a higher k. However, when dropping the mobility features (radius of gyration, mean check-in distance, and mean distance to city center), we obtained clearly better results, and finally choose the optimal configuration of a silhouette width of 0.57 for $k = 3$. We plot the silhouette width against k in Figure 1a and refer to the details of the final result in Table 1.

There are three clusters, two which respectively make up about a quarter of the users and one larger one, containing the remaining 46.6% of the locals. We interpret the fact that the mobility features, such as the radius of gyration and the distance to the city center prevented the algorithm from finding an acceptable segmentation of the locals, as a clear indication that these features are unsuitable for distinguishing different resident groups in the data set at hand. This may be due to several reasons: residents might be more active in their respective districts making it hard to characterize their behavior with metrics in relation to the entire city. In addition, commuting introduces noise, which is difficult to eliminate given the volatile usage of LBSNs during leisure and work time. Finally, the mobility metrics to characterize residents of five different cities might need more careful deliberation: cultural and geographic circumstances could be too different to find universal clusters across all cities. This means that the clustering result of the locals is mostly influenced by the user activity level.



(a) Locals: $k = 3$ was chosen as final result, since this was the last value above 0.5 and with larger values for k, the silhouette width only decreases gradually.

(b) Travelers: We chose $k = 4$, since with higher k, the silhouette width plateaus.

Figure 1: Determining the number of clusters using the silhouette width.

3.3 Traveler Behavior Cluster Analysis

Similar to the locals, we analyzed the behavior of the users when traveling outside of their respective home cities. The processing was performed using the `tripmining` library[3], which segments the user's check-ins into periods of being at home and in other cities [8]. Consecutive periods abroad are regarded as trips, provided

[3] `tripmining` library: https://github.com/LinusDietz/tripmining

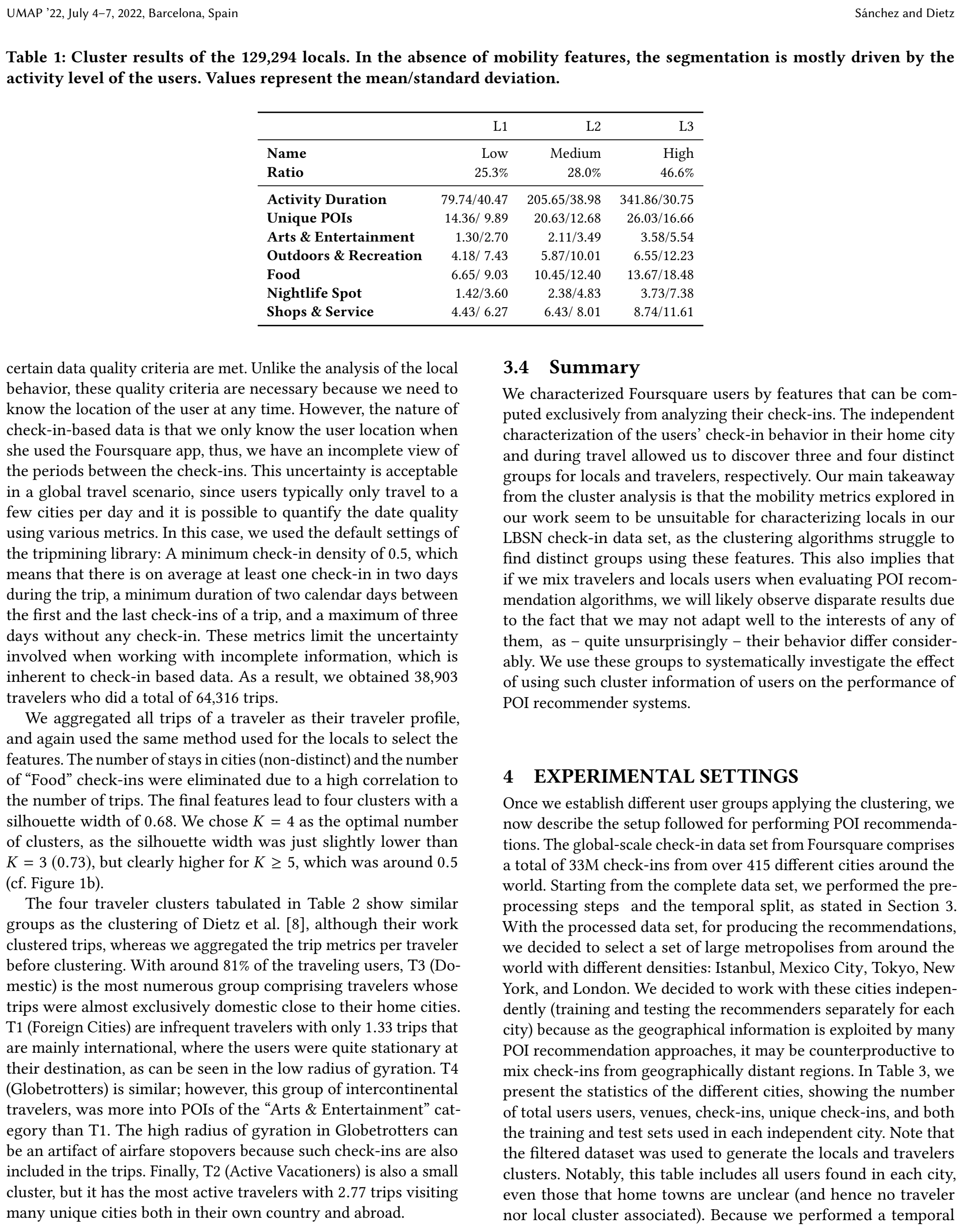


Table 1: Cluster results of the 129,294 locals. In the absence of mobility features, the segmentation is mostly driven by the activity level of the users. Values represent the mean/standard deviation.

	L1	L2	L3
Name	Low	Medium	High
Ratio	25.3%	28.0%	46.6%
Activity Duration	79.74/40.47	205.65/38.98	341.86/30.75
Unique POIs	14.36/ 9.89	20.63/12.68	26.03/16.66
Arts & Entertainment	1.30/2.70	2.11/3.49	3.58/5.54
Outdoors & Recreation	4.18/ 7.43	5.87/10.01	6.55/12.23
Food	6.65/ 9.03	10.45/12.40	13.67/18.48
Nightlife Spot	1.42/3.60	2.38/4.83	3.73/7.38
Shops & Service	4.43/ 6.27	6.43/ 8.01	8.74/11.61

certain data quality criteria are met. Unlike the analysis of the local behavior, these quality criteria are necessary because we need to know the location of the user at any time. However, the nature of check-in-based data is that we only know the user location when she used the Foursquare app, thus, we have an incomplete view of the periods between the check-ins. This uncertainty is acceptable in a global travel scenario, since users typically only travel to a few cities per day and it is possible to quantify the date quality using various metrics. In this case, we used the default settings of the tripmining library: A minimum check-in density of 0.5, which means that there is on average at least one check-in in two days during the trip, a minimum duration of two calendar days between the first and the last check-ins of a trip, and a maximum of three days without any check-in. These metrics limit the uncertainty involved when working with incomplete information, which is inherent to check-in based data. As a result, we obtained 38,903 travelers who did a total of 64,316 trips.

We aggregated all trips of a traveler as their traveler profile, and again used the same method used for the locals to select the features. The number of stays in cities (non-distinct) and the number of "Food" check-ins were eliminated due to a high correlation to the number of trips. The final features lead to four clusters with a silhouette width of 0.68. We chose $K = 4$ as the optimal number of clusters, as the silhouette width was just slightly lower than $K = 3$ (0.73), but clearly higher for $K \geq 5$, which was around 0.5 (cf. Figure 1b).

The four traveler clusters tabulated in Table 2 show similar groups as the clustering of Dietz et al. [8], although their work clustered trips, whereas we aggregated the trip metrics per traveler before clustering. With around 81% of the traveling users, T3 (Domestic) is the most numerous group comprising travelers whose trips were almost exclusively domestic close to their home cities. T1 (Foreign Cities) are infrequent travelers with only 1.33 trips that are mainly international, where the users were quite stationary at their destination, as can be seen in the low radius of gyration. T4 (Globetrotters) is similar; however, this group of intercontinental travelers, was more into POIs of the "Arts & Entertainment" category than T1. The high radius of gyration in Globetrotters can be an artifact of airfare stopovers because such check-ins are also included in the trips. Finally, T2 (Active Vacationers) is also a small cluster, but it has the most active travelers with 2.77 trips visiting many unique cities both in their own country and abroad.

3.4 Summary

We characterized Foursquare users by features that can be computed exclusively from analyzing their check-ins. The independent characterization of the users' check-in behavior in their home city and during travel allowed us to discover three and four distinct groups for locals and travelers, respectively. Our main takeaway from the cluster analysis is that the mobility metrics explored in our work seem to be unsuitable for characterizing locals in our LBSN check-in data set, as the clustering algorithms struggle to find distinct groups using these features. This also implies that if we mix travelers and locals users when evaluating POI recommendation algorithms, we will likely observe disparate results due to the fact that we may not adapt well to the interests of any of them, as – quite unsurprisingly – their behavior differ considerably. We use these groups to systematically investigate the effect of using such cluster information of users on the performance of POI recommender systems.

4 EXPERIMENTAL SETTINGS

Once we establish different user groups applying the clustering, we now describe the setup followed for performing POI recommendations. The global-scale check-in data set from Foursquare comprises a total of 33M check-ins from over 415 different cities around the world. Starting from the complete data set, we performed the preprocessing steps and the temporal split, as stated in Section 3. With the processed data set, for producing the recommendations, we decided to select a set of large metropolises from around the world with different densities: Istanbul, Mexico City, Tokyo, New York, and London. We decided to work with these cities independently (training and testing the recommenders separately for each city) because as the geographical information is exploited by many POI recommendation approaches, it may be counterproductive to mix check-ins from geographically distant regions. In Table 3, we present the statistics of the different cities, showing the number of total users users, venues, check-ins, unique check-ins, and both the training and test sets used in each independent city. Note that the filtered dataset was used to generate the locals and travelers clusters. Notably, this table includes all users found in each city, even those that home towns are unclear (and hence no traveler nor local cluster associated). Because we performed a temporal

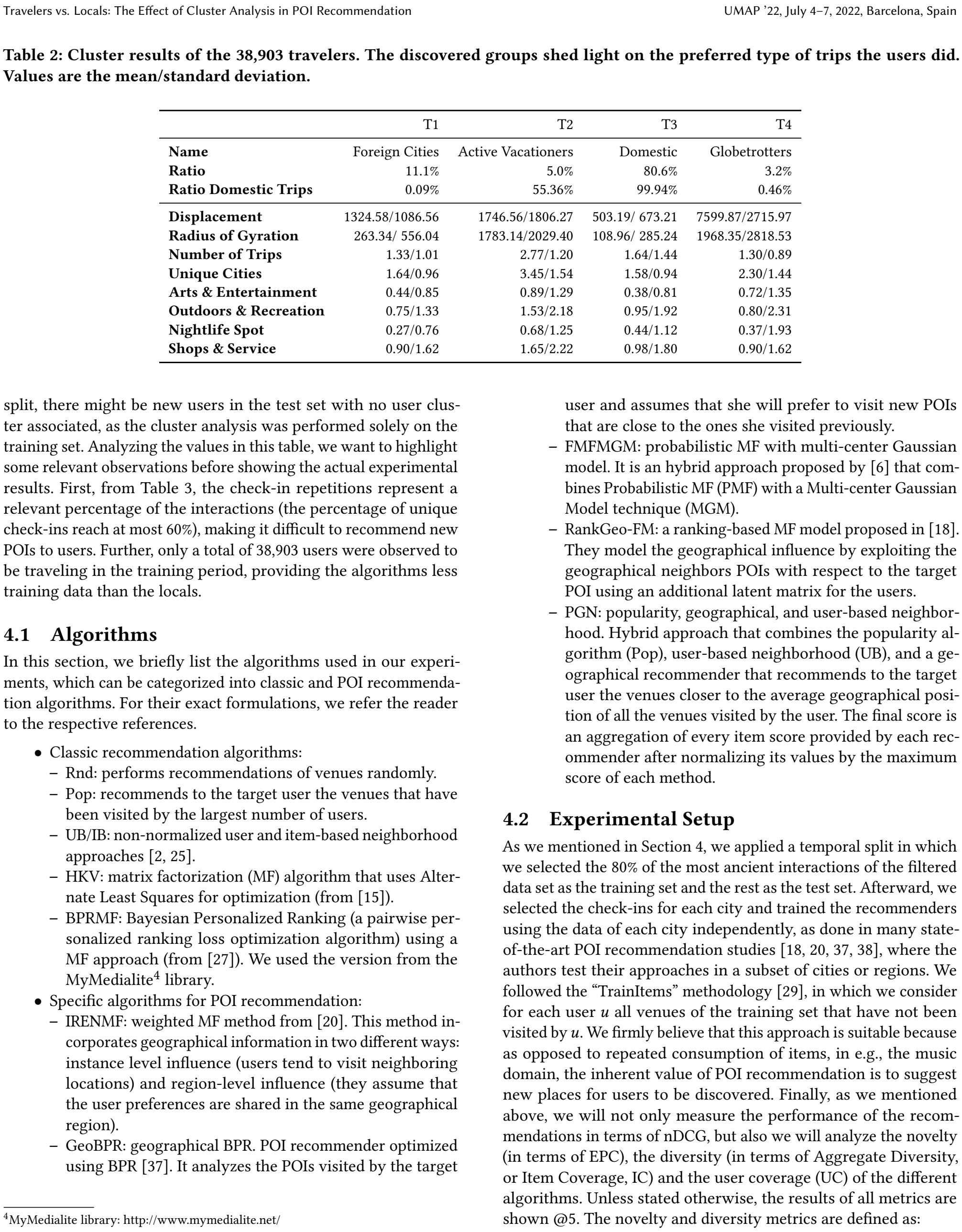


Table 2: Cluster results of the 38,903 travelers. The discovered groups shed light on the preferred type of trips the users did. Values are the mean/standard deviation.

	T1	T2	T3	T4
Name	Foreign Cities	Active Vacationers	Domestic	Globetrotters
Ratio	11.1%	5.0%	80.6%	3.2%
Ratio Domestic Trips	0.09%	55.36%	99.94%	0.46%
Displacement	1324.58/1086.56	1746.56/1806.27	503.19/ 673.21	7599.87/2715.97
Radius of Gyration	263.34/ 556.04	1783.14/2029.40	108.96/ 285.24	1968.35/2818.53
Number of Trips	1.33/1.01	2.77/1.20	1.64/1.44	1.30/0.89
Unique Cities	1.64/0.96	3.45/1.54	1.58/0.94	2.30/1.44
Arts & Entertainment	0.44/0.85	0.89/1.29	0.38/0.81	0.72/1.35
Outdoors & Recreation	0.75/1.33	1.53/2.18	0.95/1.92	0.80/2.31
Nightlife Spot	0.27/0.76	0.68/1.25	0.44/1.12	0.37/1.93
Shops & Service	0.90/1.62	1.65/2.22	0.98/1.80	0.90/1.62

split, there might be new users in the test set with no user cluster associated, as the cluster analysis was performed solely on the training set. Analyzing the values in this table, we want to highlight some relevant observations before showing the actual experimental results. First, from Table 3, the check-in repetitions represent a relevant percentage of the interactions (the percentage of unique check-ins reach at most 60%), making it difficult to recommend new POIs to users. Further, only a total of 38,903 users were observed to be traveling in the training period, providing the algorithms less training data than the locals.

4.1 Algorithms

In this section, we briefly list the algorithms used in our experiments, which can be categorized into classic and POI recommendation algorithms. For their exact formulations, we refer the reader to the respective references.

- Classic recommendation algorithms:
  - Rnd: performs recommendations of venues randomly.
  - Pop: recommends to the target user the venues that have been visited by the largest number of users.
  - UB/IB: non-normalized user and item-based neighborhood approaches [2, 25].
  - HKV: matrix factorization (MF) algorithm that uses Alternate Least Squares for optimization (from [15]).
  - BPRMF: Bayesian Personalized Ranking (a pairwise personalized ranking loss optimization algorithm) using a MF approach (from [27]). We used the version from the MyMedialite[4] library.
- Specific algorithms for POI recommendation:
  - IRENMF: weighted MF method from [20]. This method incorporates geographical information in two different ways: instance level influence (users tend to visit neighboring locations) and region-level influence (they assume that the user preferences are shared in the same geographical region).
  - GeoBPR: geographical BPR. POI recommender optimized using BPR [37]. It analyzes the POIs visited by the target user and assumes that she will prefer to visit new POIs that are close to the ones she visited previously.
  - FMFMGM: probabilistic MF with multi-center Gaussian model. It is an hybrid approach proposed by [6] that combines Probabilistic MF (PMF) with a Multi-center Gaussian Model technique (MGM).
  - RankGeo-FM: a ranking-based MF model proposed in [18]. They model the geographical influence by exploiting the geographical neighbors POIs with respect to the target POI using an additional latent matrix for the users.
  - PGN: popularity, geographical, and user-based neighborhood. Hybrid approach that combines the popularity algorithm (Pop), user-based neighborhood (UB), and a geographical recommender that recommends to the target user the venues closer to the average geographical position of all the venues visited by the user. The final score is an aggregation of every item score provided by each recommender after normalizing its values by the maximum score of each method.

4.2 Experimental Setup

As we mentioned in Section 4, we applied a temporal split in which we selected the 80% of the most ancient interactions of the filtered data set as the training set and the rest as the test set. Afterward, we selected the check-ins for each city and trained the recommenders using the data of each city independently, as done in many state-of-the-art POI recommendation studies [18, 20, 37, 38], where the authors test their approaches in a subset of cities or regions. We followed the "TrainItems" methodology [29], in which we consider for each user u all venues of the training set that have not been visited by u. We firmly believe that this approach is suitable because as opposed to repeated consumption of items, in e.g., the music domain, the inherent value of POI recommendation is to suggest new places for users to be discovered. Finally, as we mentioned above, we will not only measure the performance of the recommendations in terms of nDCG, but also we will analyze the novelty (in terms of EPC), the diversity (in terms of Aggregate Diversity, or Item Coverage, IC) and the user coverage (UC) of the different algorithms. Unless stated otherwise, the results of all metrics are shown @5. The novelty and diversity metrics are defined as:

[4] MyMedialite library: http://www.mymedialite.net/

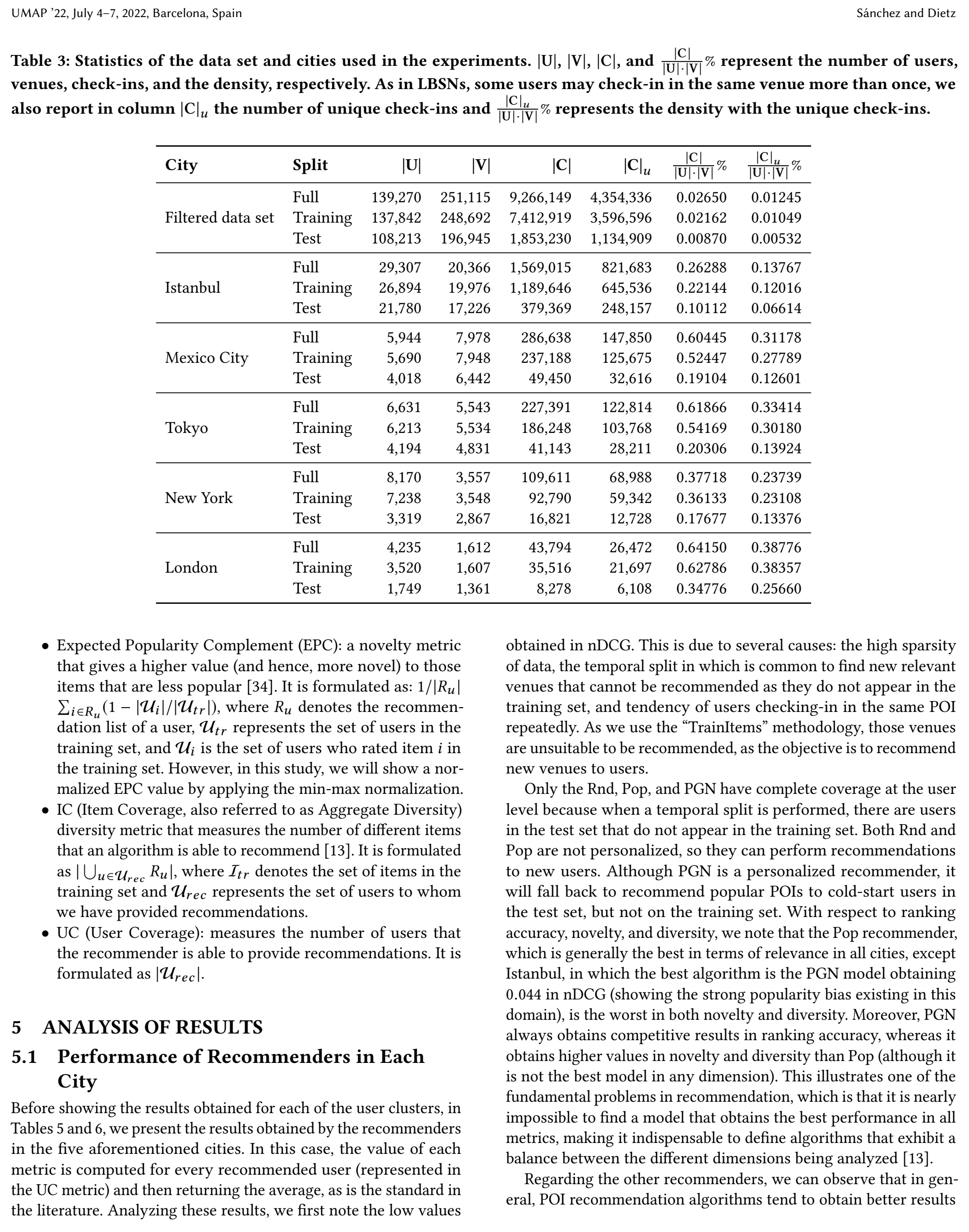


Table 3: Statistics of the data set and cities used in the experiments. |U|, |V|, |C|, and $\frac{|C|}{|U|\cdot|V|}\%$ represent the number of users, venues, check-ins, and the density, respectively. As in LBSNs, some users may check-in in the same venue more than once, we also report in column $|C|_u$ the number of unique check-ins and $\frac{|C|_u}{|U|\cdot|V|}\%$ represents the density with the unique check-ins.

City	Split	|U|	|V|	|C|	$\|C\|_u$	$\frac{\|C\|}{\|U\|\cdot\|V\|}\%$	$\frac{\|C\|_u}{\|U\|\cdot\|V\|}\%$
Filtered data set	Full	139,270	251,115	9,266,149	4,354,336	0.02650	0.01245
	Training	137,842	248,692	7,412,919	3,596,596	0.02162	0.01049
	Test	108,213	196,945	1,853,230	1,134,909	0.00870	0.00532
Istanbul	Full	29,307	20,366	1,569,015	821,683	0.26288	0.13767
	Training	26,894	19,976	1,189,646	645,536	0.22144	0.12016
	Test	21,780	17,226	379,369	248,157	0.10112	0.06614
Mexico City	Full	5,944	7,978	286,638	147,850	0.60445	0.31178
	Training	5,690	7,948	237,188	125,675	0.52447	0.27789
	Test	4,018	6,442	49,450	32,616	0.19104	0.12601
Tokyo	Full	6,631	5,543	227,391	122,814	0.61866	0.33414
	Training	6,213	5,534	186,248	103,768	0.54169	0.30180
	Test	4,194	4,831	41,143	28,211	0.20306	0.13924
New York	Full	8,170	3,557	109,611	68,988	0.37718	0.23739
	Training	7,238	3,548	92,790	59,342	0.36133	0.23108
	Test	3,319	2,867	16,821	12,728	0.17677	0.13376
London	Full	4,235	1,612	43,794	26,472	0.64150	0.38776
	Training	3,520	1,607	35,516	21,697	0.62786	0.38357
	Test	1,749	1,361	8,278	6,108	0.34776	0.25660

- Expected Popularity Complement (EPC): a novelty metric that gives a higher value (and hence, more novel) to those items that are less popular [34]. It is formulated as: $1/|R_u| \sum_{i\in R_u}(1 - |\mathcal{U}_i|/|\mathcal{U}_{tr}|)$, where R_u denotes the recommendation list of a user, $\mathcal{U}_{tr}$ represents the set of users in the training set, and $\mathcal{U}_i$ is the set of users who rated item i in the training set. However, in this study, we will show a normalized EPC value by applying the min-max normalization.
- IC (Item Coverage, also referred to as Aggregate Diversity) diversity metric that measures the number of different items that an algorithm is able to recommend [13]. It is formulated as $|\bigcup_{u\in\mathcal{U}_{rec}} R_u|$, where $\mathcal{I}_{tr}$ denotes the set of items in the training set and $\mathcal{U}_{rec}$ represents the set of users to whom we have provided recommendations.
- UC (User Coverage): measures the number of users that the recommender is able to provide recommendations. It is formulated as $|\mathcal{U}_{rec}|$.

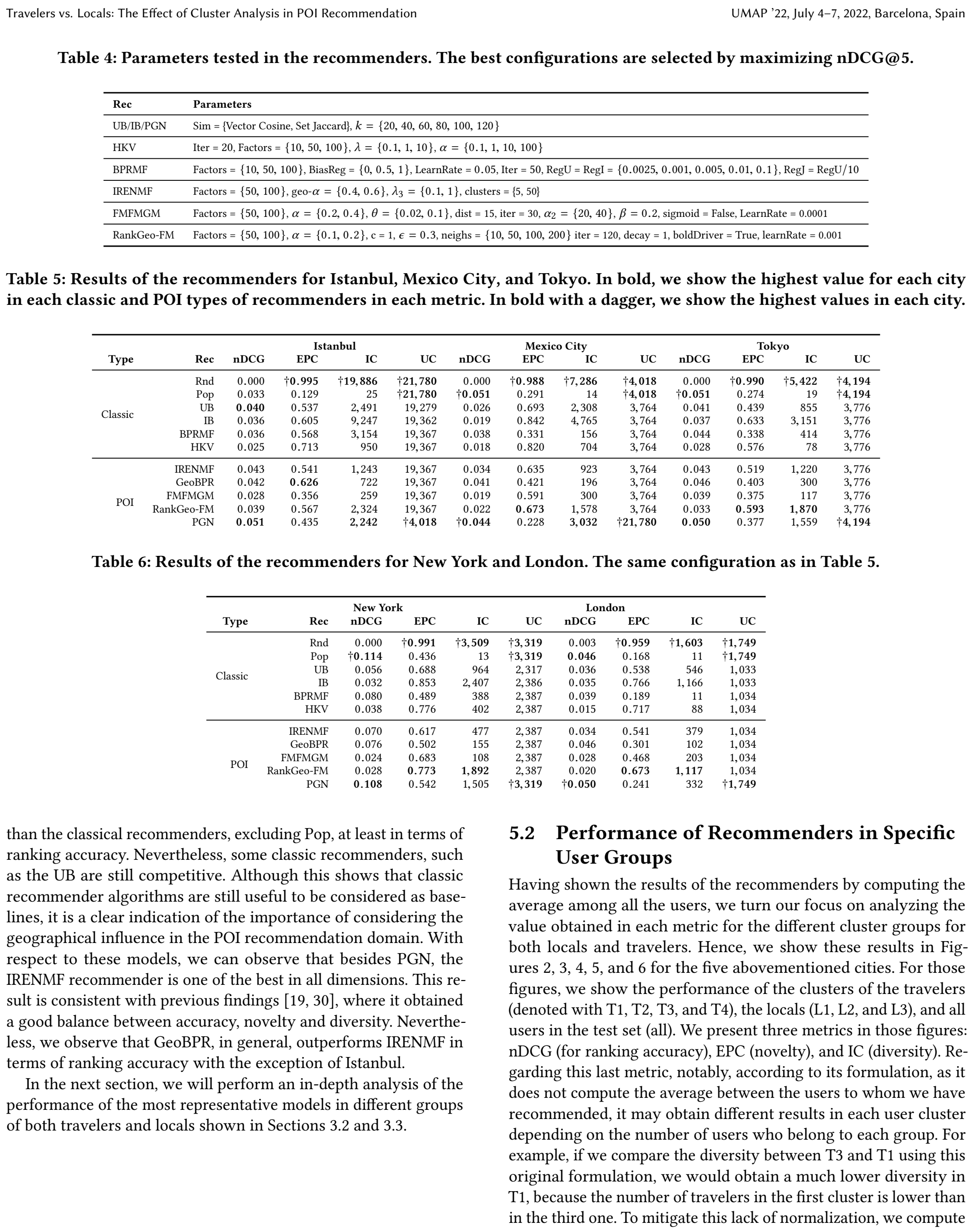
5 ANALYSIS OF RESULTS

5.1 Performance of Recommenders in Each City

Before showing the results obtained for each of the user clusters, in Tables 5 and 6, we present the results obtained by the recommenders in the five aforementioned cities. In this case, the value of each metric is computed for every recommended user (represented in the UC metric) and then returning the average, as is the standard in the literature. Analyzing these results, we first note the low values obtained in nDCG. This is due to several causes: the high sparsity of data, the temporal split in which is common to find new relevant venues that cannot be recommended as they do not appear in the training set, and tendency of users checking-in in the same POI repeatedly. As we use the "TrainItems" methodology, those venues are unsuitable to be recommended, as the objective is to recommend new venues to users.

Only the Rnd, Pop, and PGN have complete coverage at the user level because when a temporal split is performed, there are users in the test set that do not appear in the training set. Both Rnd and Pop are not personalized, so they can perform recommendations to new users. Although PGN is a personalized recommender, it will fall back to recommend popular POIs to cold-start users in the test set, but not on the training set. With respect to ranking accuracy, novelty, and diversity, we note that the Pop recommender, which is generally the best in terms of relevance in all cities, except Istanbul, in which the best algorithm is the PGN model obtaining 0.044 in nDCG (showing the strong popularity bias existing in this domain), is the worst in both novelty and diversity. Moreover, PGN always obtains competitive results in ranking accuracy, whereas it obtains higher values in novelty and diversity than Pop (although it is not the best model in any dimension). This illustrates one of the fundamental problems in recommendation, which is that it is nearly impossible to find a model that obtains the best performance in all metrics, making it indispensable to define algorithms that exhibit a balance between the different dimensions being analyzed [13].

Regarding the other recommenders, we can observe that in general, POI recommendation algorithms tend to obtain better results



Travelers vs. Locals: The Effect of Cluster Analysis in POI Recommendation UMAP '22, July 4–7, 2022, Barcelona, Spain

Table 4: Parameters tested in the recommenders. The best configurations are selected by maximizing nDCG@5.

Rec	Parameters
UB/IB/PGN	Sim = {Vector Cosine, Set Jaccard}, $k = \{20, 40, 60, 80, 100, 120\}$
HKV	Iter = 20, Factors = $\{10, 50, 100\}$, $\lambda = \{0.1, 1, 10\}$, $\alpha = \{0.1, 1, 10, 100\}$
BPRMF	Factors = $\{10, 50, 100\}$, BiasReg = $\{0, 0.5, 1\}$, LearnRate = 0.05, Iter = 50, RegU = RegI = $\{0.0025, 0.001, 0.005, 0.01, 0.1\}$, RegJ = RegU/10
IRENMF	Factors = $\{50, 100\}$, geo-$\alpha = \{0.4, 0.6\}$, $\lambda_3 = \{0.1, 1\}$, clusters = {5, 50}
FMFMGM	Factors = $\{50, 100\}$, $\alpha = \{0.2, 0.4\}$, $\theta = \{0.02, 0.1\}$, dist = 15, iter = 30, $\alpha_2 = \{20, 40\}$, $\beta = 0.2$, sigmoid = False, LearnRate = 0.0001
RankGeo-FM	Factors = $\{50, 100\}$, $\alpha = \{0.1, 0.2\}$, c = 1, $\epsilon = 0.3$, neighs = $\{10, 50, 100, 200\}$ iter = 120, decay = 1, boldDriver = True, learnRate = 0.001

Table 5: Results of the recommenders for Istanbul, Mexico City, and Tokyo. In bold, we show the highest value for each city in each classic and POI types of recommenders in each metric. In bold with a dagger, we show the highest values in each city.

		Istanbul				Mexico City				Tokyo			
Type	Rec	nDCG	EPC	IC	UC	nDCG	EPC	IC	UC	nDCG	EPC	IC	UC
Classic	Rnd	0.000	**†0.995**	**†19,886**	**†21,780**	0.000	**†0.988**	**†7,286**	**†4,018**	0.000	**†0.990**	**†5,422**	**†4,194**
	Pop	0.033	0.129	25	**†21,780**	**†0.051**	0.291	14	**†4,018**	**†0.051**	0.274	19	**†4,194**
	UB	**0.040**	0.537	2,491	19,279	0.026	0.693	2,308	3,764	0.041	0.439	855	3,776
	IB	0.036	0.605	9,247	19,362	0.019	0.842	4,765	3,764	0.037	0.633	3,151	3,776
	BPRMF	0.036	0.568	3,154	19,367	0.038	0.331	156	3,764	0.044	0.338	414	3,776
	HKV	0.025	0.713	950	19,367	0.018	0.820	704	3,764	0.028	0.576	78	3,776
POI	IRENMF	0.043	0.541	1,243	19,367	0.034	0.635	923	3,764	0.043	0.519	1,220	3,776
	GeoBPR	0.042	**0.626**	722	19,367	0.041	0.421	196	3,764	0.046	0.403	300	3,776
	FMFMGM	0.028	0.356	259	19,367	0.019	0.591	300	3,764	0.039	0.375	117	3,776
	RankGeo-FM	0.039	0.567	2,324	19,367	0.022	**0.673**	1,578	3,764	0.033	**0.593**	**1,870**	3,776
	PGN	**0.051**	0.435	**2,242**	**†4,018**	**†0.044**	0.228	**3,032**	**†21,780**	**0.050**	0.377	1,559	**†4,194**

Table 6: Results of the recommenders for New York and London. The same configuration as in Table 5.

		New York				London			
Type	Rec	nDCG	EPC	IC	UC	nDCG	EPC	IC	UC
Classic	Rnd	0.000	**†0.991**	**†3,509**	**†3,319**	0.003	**†0.959**	**†1,603**	**†1,749**
	Pop	**†0.114**	0.436	13	**†3,319**	**0.046**	0.168	11	**†1,749**
	UB	0.056	0.688	964	2,317	0.036	0.538	546	1,033
	IB	0.032	0.853	2,407	2,386	0.035	0.766	1,166	1,033
	BPRMF	0.080	0.489	388	2,387	0.039	0.189	11	1,034
	HKV	0.038	0.776	402	2,387	0.015	0.717	88	1,034
POI	IRENMF	0.070	0.617	477	2,387	0.034	0.541	379	1,034
	GeoBPR	0.076	0.502	155	2,387	0.046	0.301	102	1,034
	FMFMGM	0.024	0.683	108	2,387	0.028	0.468	203	1,034
	RankGeo-FM	0.028	**0.773**	**1,892**	2,387	0.020	**0.673**	**1,117**	1,034
	PGN	**0.108**	0.542	1,505	**†3,319**	**†0.050**	0.241	332	**†1,749**

than the classical recommenders, excluding Pop, at least in terms of ranking accuracy. Nevertheless, some classic recommenders, such as the UB are still competitive. Although this shows that classic recommender algorithms are still useful to be considered as baselines, it is a clear indication of the importance of considering the geographical influence in the POI recommendation domain. With respect to these models, we can observe that besides PGN, the IRENMF recommender is one of the best in all dimensions. This result is consistent with previous findings [19, 30], where it obtained a good balance between accuracy, novelty and diversity. Nevertheless, we observe that GeoBPR, in general, outperforms IRENMF in terms of ranking accuracy with the exception of Istanbul.

In the next section, we will perform an in-depth analysis of the performance of the most representative models in different groups of both travelers and locals shown in Sections 3.2 and 3.3.

5.2 Performance of Recommenders in Specific User Groups

Having shown the results of the recommenders by computing the average among all the users, we turn our focus on analyzing the value obtained in each metric for the different cluster groups for both locals and travelers. Hence, we show these results in Figures 2, 3, 4, 5, and 6 for the five abovementioned cities. For those figures, we show the performance of the clusters of the travelers (denoted with T1, T2, T3, and T4), the locals (L1, L2, and L3), and all users in the test set (all). We present three metrics in those figures: nDCG (for ranking accuracy), EPC (novelty), and IC (diversity). Regarding this last metric, notably, according to its formulation, as it does not compute the average between the users to whom we have recommended, it may obtain different results in each user cluster depending on the number of users who belong to each group. For example, if we compare the diversity between T3 and T1 using this original formulation, we would obtain a much lower diversity in T1, because the number of travelers in the first cluster is lower than in the third one. To mitigate this lack of normalization, we compute

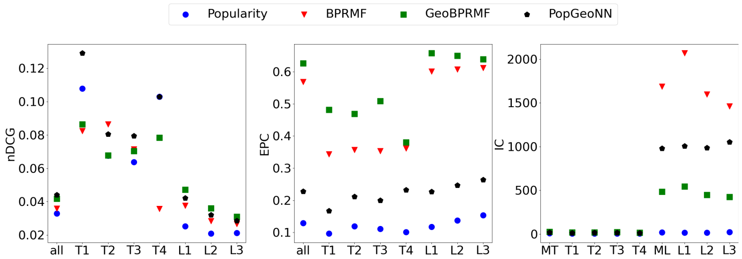


Figure 2: Results for Istanbul.

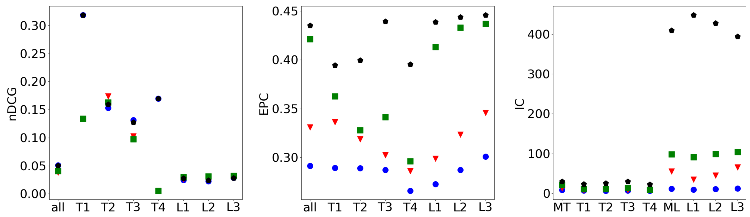


Figure 3: Results for Mexico City.

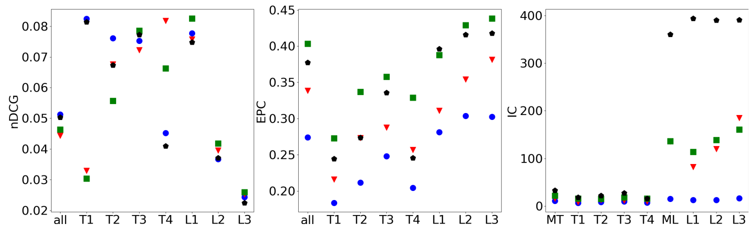


Figure 4: Results for Tokyo.

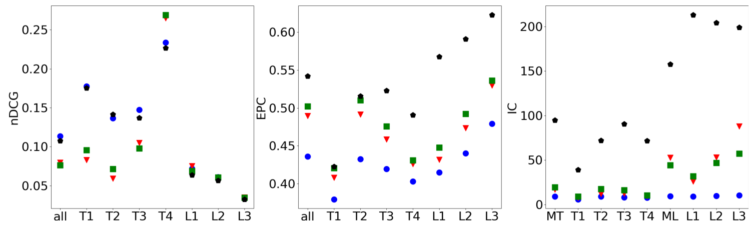


Figure 5: Results for New York City.

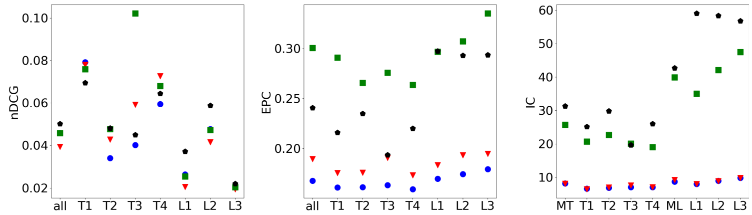


Figure 6: Results for London.

this metric performing different subsamples. Hence, we selected for each major group (travelers and locals) the cluster with the smallest number of users and then computed the value of the IC for this number of random users. The final value is the mean after repeating the sampling 1, 000 times, thereby making the values comparable. This is why instead of the "all" label, we use two additional ones when representing the IC metric, "MT" (mean of travelers) and "ML" (mean of locals), which would be used for selecting a subsample of all users with the size of the lowest traveler and local group, respectively. We repeat this process 1, 000 times, and then the result shown is the average of the 1, 000 runs. Finally, due to the large number of recommenders, we decided to show for each city the performance of Pop and BPRMF from the classical recommenders and GeoBPR and PGN from the POI recommenders, as they are the algorithms which generally obtain the best nDCG results.

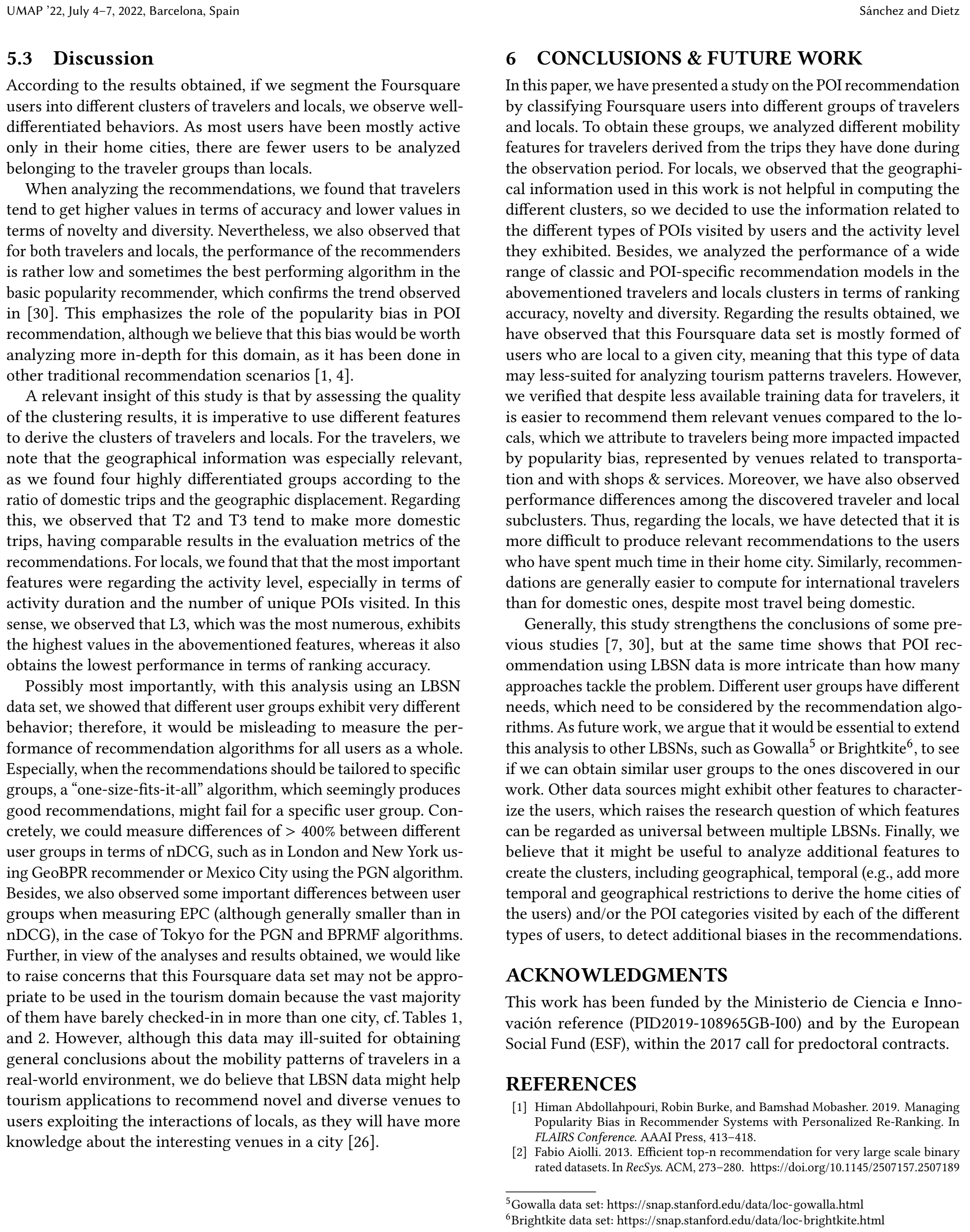
Analyzing the figures, we can observe interesting effects. First of all, travelers generally obtain higher values than locals in terms of accuracy in most cities (e.g., in Istanbul and in Mexico City all traveler groups obtain higher values in terms of nDCG than any local cluster), despite being the group with less users (e.g., 7% in the case of Mexico City and 24% of New York users in the test set are travelers). However, notably, travelers generally have a slightly lower novelty than locals, indicating that they tend to receive recommendations of more popular POIs. This makes sense, because when a tourist visits a city, she is more likely to visit the most popular venues than if she is a local. Besides, we analyzed the top-5 most popular venues in each city and we observed that most of them belong to transport and commerce categories. For example, both airports (e.g., Kennedy, Atatürk and Benito Juárez, in New York, Istanbul, and Mexico City respectively) and train stations (Euston Railway Station in New York and Akihabara Station in Tokyo) are the venues that have received most visits in the training sets. In addition, shopping malls and commerce districts like Harrods (London), Perisur (Mexico City), and Times Square (New York) are also in the top-5 most popular venues.

Moreover, we can observe how locals tend to receive more diverse recommendations. This again may be because locals are commonly sightseeing extensively within their home city. Furthermore, locals are more likely to have visited numerous different POIs in the training set (including the most popular ones), which are then unavailable for recommendation in the test set. By contrast, most travelers will visit a city for the first time during the evaluation period; thus, it is more probable that they visit one of the recommended popular POIs, which will result in a decrease in novelty and diversity. Finally, as there are far fewer travelers than locals, it is normal that despite having computed the IC metric using the subsamples, we obtain much lower results for travelers than for locals, making a direct comparison between them impossible.

In general, the big picture of these results tends to support the findings of [30], although we performed a different data preprocessing, splitting methodology, and also a different analysis and characterization of travelers and locals. Hence, in addition to the analysis performed for travelers and locals, it is also interesting to study the behavior of the models among the different types of travelers and locals, i.e., all clusters shown in Tables 1 and 2.

First, regarding the travelers, there is no a common behavior in the different cities. For example, T4 is the group that obtains the highest values in nDCG for New York in all recommenders, whereas, in other cities, such as Istanbul and Mexico City, there are some models which obtain very low values for these users. Regarding novelty and diversity, T1 obtains the worst results in the cities of Tokyo and New York, whereas, in London it is one of the best group in both aspects. Despite these discrepancies among the travelers, we also perceive common behavior, such as T2 and T3 generally obtaining similar results. This may be explained by the features shown in Table 2, where we can observe that these two groups have the highest ratio of domestic trips, whereas T1 and T4 tend to make more abroad travels, visiting more popular POIs as we can observe in the performance in both nDCG and EPC metrics. In fact, except for Mexico City, in the rest of the cities the Pop algorithm achieve higher values in EPC for both T2 and T3.

Regarding the locals, in all cities, except Mexico City, L3 is the cluster that obtains the lowest levels in nDCG, comparing it with all locals and travelers groups whereas, in general, it also obtains higher levels of novelty than the other clusters. Besides, for L3, all models have similar nDCG performance, thus, exhibiting much fewer variations in this group than in any other group. From Table 1, besides L3 being the most numerous, it is also the cluster that, in general, contains the most active users (represented by the "Unique POIs" feature). Hence, it is more probable to recommend these users less popular venues given the probability that they have visited more venues before than the other two groups with a lower activity level, making more difficult to recommend to them both novel and relevant venues.



5.3 Discussion

According to the results obtained, if we segment the Foursquare users into different clusters of travelers and locals, we observe well-differentiated behaviors. As most users have been mostly active only in their home cities, there are fewer users to be analyzed belonging to the traveler groups than locals.

When analyzing the recommendations, we found that travelers tend to get higher values in terms of accuracy and lower values in terms of novelty and diversity. Nevertheless, we also observed that for both travelers and locals, the performance of the recommenders is rather low and sometimes the best performing algorithm in the basic popularity recommender, which confirms the trend observed in [30]. This emphasizes the role of the popularity bias in POI recommendation, although we believe that this bias would be worth analyzing more in-depth for this domain, as it has been done in other traditional recommendation scenarios [1, 4].

A relevant insight of this study is that by assessing the quality of the clustering results, it is imperative to use different features to derive the clusters of travelers and locals. For the travelers, we note that the geographical information was especially relevant, as we found four highly differentiated groups according to the ratio of domestic trips and the geographic displacement. Regarding this, we observed that T2 and T3 tend to make more domestic trips, having comparable results in the evaluation metrics of the recommendations. For locals, we found that that the most important features were regarding the activity level, especially in terms of activity duration and the number of unique POIs visited. In this sense, we observed that L3, which was the most numerous, exhibits the highest values in the abovementioned features, whereas it also obtains the lowest performance in terms of ranking accuracy.

Possibly most importantly, with this analysis using an LBSN data set, we showed that different user groups exhibit very different behavior; therefore, it would be misleading to measure the performance of recommendation algorithms for all users as a whole. Especially, when the recommendations should be tailored to specific groups, a "one-size-fits-it-all" algorithm, which seemingly produces good recommendations, might fail for a specific user group. Concretely, we could measure differences of > 400% between different user groups in terms of nDCG, such as in London and New York using GeoBPR recommender or Mexico City using the PGN algorithm. Besides, we also observed some important differences between user groups when measuring EPC (although generally smaller than in nDCG), in the case of Tokyo for the PGN and BPRMF algorithms. Further, in view of the analyses and results obtained, we would like to raise concerns that this Foursquare data set may not be appropriate to be used in the tourism domain because the vast majority of them have barely checked-in in more than one city, cf. Tables 1, and 2. However, although this data may ill-suited for obtaining general conclusions about the mobility patterns of travelers in a real-world environment, we do believe that LBSN data might help tourism applications to recommend novel and diverse venues to users exploiting the interactions of locals, as they will have more knowledge about the interesting venues in a city [26].

6 CONCLUSIONS & FUTURE WORK

In this paper, we have presented a study on the POI recommendation by classifying Foursquare users into different groups of travelers and locals. To obtain these groups, we analyzed different mobility features for travelers derived from the trips they have done during the observation period. For locals, we observed that the geographical information used in this work is not helpful in computing the different clusters, so we decided to use the information related to the different types of POIs visited by users and the activity level they exhibited. Besides, we analyzed the performance of a wide range of classic and POI-specific recommendation models in the abovementioned travelers and locals clusters in terms of ranking accuracy, novelty and diversity. Regarding the results obtained, we have observed that this Foursquare data set is mostly formed of users who are local to a given city, meaning that this type of data may less-suited for analyzing tourism patterns travelers. However, we verified that despite less available training data for travelers, it is easier to recommend them relevant venues compared to the locals, which we attribute to travelers being more impacted impacted by popularity bias, represented by venues related to transportation and with shops & services. Moreover, we have also observed performance differences among the discovered traveler and local subclusters. Thus, regarding the locals, we have detected that it is more difficult to produce relevant recommendations to the users who have spent much time in their home city. Similarly, recommendations are generally easier to compute for international travelers than for domestic ones, despite most travel being domestic.

Generally, this study strengthens the conclusions of some previous studies [7, 30], but at the same time shows that POI recommendation using LBSN data is more intricate than how many approaches tackle the problem. Different user groups have different needs, which need to be considered by the recommendation algorithms. As future work, we argue that it would be essential to extend this analysis to other LBSNs, such as Gowalla[5] or Brightkite[6], to see if we can obtain similar user groups to the ones discovered in our work. Other data sources might exhibit other features to characterize the users, which raises the research question of which features can be regarded as universal between multiple LBSNs. Finally, we believe that it might be useful to analyze additional features to create the clusters, including geographical, temporal (e.g., add more temporal and geographical restrictions to derive the home cities of the users) and/or the POI categories visited by each of the different types of users, to detect additional biases in the recommendations.

ACKNOWLEDGMENTS

This work has been funded by the Ministerio de Ciencia e Innovación reference (PID2019-108965GB-I00) and by the European Social Fund (ESF), within the 2017 call for predoctoral contracts.

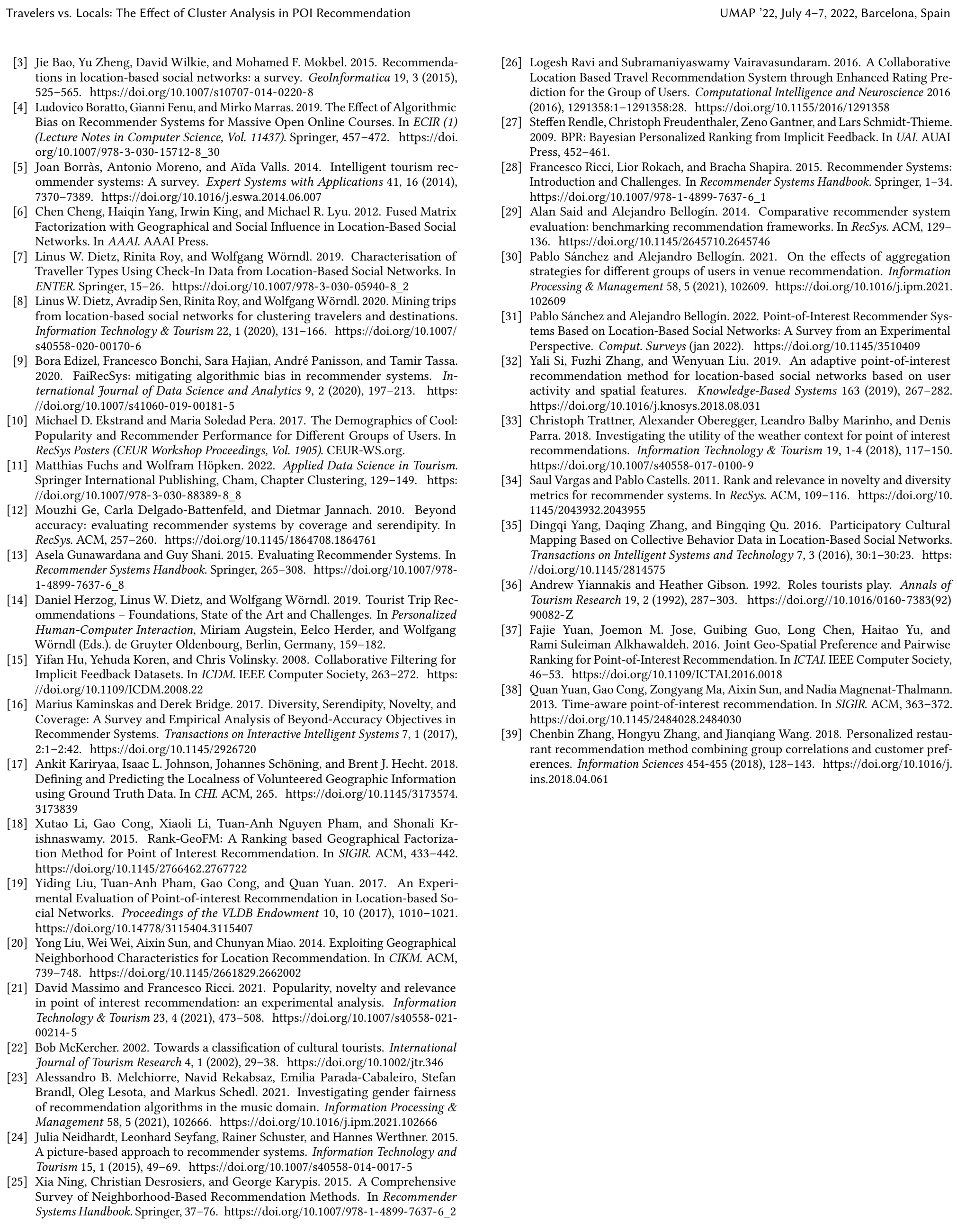
REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-Ranking. In *FLAIRS Conference*. AAAI Press, 413–418.

[2] Fabio Aiolli. 2013. Efficient top-n recommendation for very large scale binary rated datasets. In *RecSys*. ACM, 273–280. https://doi.org/10.1145/2507157.2507189

[5] Gowalla data set: https://snap.stanford.edu/data/loc-gowalla.html

[6] Brightkite data set: https://snap.stanford.edu/data/loc-brightkite.html



[3] Jie Bao, Yu Zheng, David Wilkie, and Mohamed F. Mokbel. 2015. Recommendations in location-based social networks: a survey. *GeoInformatica* 19, 3 (2015), 525–565. https://doi.org/10.1007/s10707-014-0220-8

[4] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2019. The Effect of Algorithmic Bias on Recommender Systems for Massive Open Online Courses. In *ECIR (1) (Lecture Notes in Computer Science, Vol. 11437)*. Springer, 457–472. https://doi.org/10.1007/978-3-030-15712-8_30

[5] Joan Borràs, Antonio Moreno, and Aïda Valls. 2014. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications* 41, 16 (2014), 7370–7389. https://doi.org/10.1016/j.eswa.2014.06.007

[6] Chen Cheng, Haiqin Yang, Irwin King, and Michael R. Lyu. 2012. Fused Matrix Factorization with Geographical and Social Influence in Location-Based Social Networks. In *AAAI*. AAAI Press.

[7] Linus W. Dietz, Rinita Roy, and Wolfgang Wörndl. 2019. Characterisation of Traveller Types Using Check-In Data from Location-Based Social Networks. In *ENTER*. Springer, 15–26. https://doi.org/10.1007/978-3-030-05940-8_2

[8] Linus W. Dietz, Avradip Sen, Rinita Roy, and Wolfgang Wörndl. 2020. Mining trips from location-based social networks for clustering travelers and destinations. *Information Technology & Tourism* 22, 1 (2020), 131–166. https://doi.org/10.1007/s40558-020-00170-6

[9] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. 2020. FaiRecSys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics* 9, 2 (2020), 197–213. https://doi.org/10.1007/s41060-019-00181-5

[10] Michael D. Ekstrand and Maria Soledad Pera. 2017. The Demographics of Cool: Popularity and Recommender Performance for Different Groups of Users. In *RecSys Posters (CEUR Workshop Proceedings, Vol. 1905)*. CEUR-WS.org.

[11] Matthias Fuchs and Wolfram Höpken. 2022. *Applied Data Science in Tourism*. Springer International Publishing, Cham, Chapter Clustering, 129–149. https://doi.org/10.1007/978-3-030-88389-8_8

[12] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *RecSys*. ACM, 257–260. https://doi.org/10.1145/1864708.1864761

[13] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook*. Springer, 265–308. https://doi.org/10.1007/978-1-4899-7637-6_8

[14] Daniel Herzog, Linus W. Dietz, and Wolfgang Wörndl. 2019. Tourist Trip Recommendations – Foundations, State of the Art and Challenges. In *Personalized Human-Computer Interaction*, Miriam Augstein, Eelco Herder, and Wolfgang Wörndl (Eds.). de Gruyter Oldenbourg, Berlin, Germany, 159–182.

[15] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *ICDM*. IEEE Computer Society, 263–272. https://doi.org/10.1109/ICDM.2008.22

[16] Marius Kaminskas and Derek Bridge. 2017. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *Transactions on Interactive Intelligent Systems* 7, 1 (2017), 2:1–2:42. https://doi.org/10.1145/2926720

[17] Ankit Kariryaa, Isaac L. Johnson, Johannes Schöning, and Brent J. Hecht. 2018. Defining and Predicting the Localness of Volunteered Geographic Information using Ground Truth Data. In *CHI*. ACM, 265. https://doi.org/10.1145/3173574.3173839

[18] Xutao Li, Gao Cong, Xiaoli Li, Tuan-Anh Nguyen Pham, and Shonali Krishnaswamy. 2015. Rank-GeoFM: A Ranking based Geographical Factorization Method for Point of Interest Recommendation. In *SIGIR*. ACM, 433–442. https://doi.org/10.1145/2766462.2767722

[19] Yiding Liu, Tuan-Anh Pham, Gao Cong, and Quan Yuan. 2017. An Experimental Evaluation of Point-of-interest Recommendation in Location-based Social Networks. *Proceedings of the VLDB Endowment* 10, 10 (2017), 1010–1021. https://doi.org/10.14778/3115404.3115407

[20] Yong Liu, Wei Wei, Aixin Sun, and Chunyan Miao. 2014. Exploiting Geographical Neighborhood Characteristics for Location Recommendation. In *CIKM*. ACM, 739–748. https://doi.org/10.1145/2661829.2662002

[21] David Massimo and Francesco Ricci. 2021. Popularity, novelty and relevance in point of interest recommendation: an experimental analysis. *Information Technology & Tourism* 23, 4 (2021), 473–508. https://doi.org/10.1007/s40558-021-00214-5

[22] Bob McKercher. 2002. Towards a classification of cultural tourists. *International Journal of Tourism Research* 4, 1 (2002), 29–38. https://doi.org/10.1002/jtr.346

[23] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management* 58, 5 (2021), 102666. https://doi.org/10.1016/j.ipm.2021.102666

[24] Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. 2015. A picture-based approach to recommender systems. *Information Technology and Tourism* 15, 1 (2015), 49–69. https://doi.org/10.1007/s40558-014-0017-5

[25] Xia Ning, Christian Desrosiers, and George Karypis. 2015. A Comprehensive Survey of Neighborhood-Based Recommendation Methods. In *Recommender Systems Handbook*. Springer, 37–76. https://doi.org/10.1007/978-1-4899-7637-6_2

[26] Logesh Ravi and Subramaniyaswamy Vairavasundaram. 2016. A Collaborative Location Based Travel Recommendation System through Enhanced Rating Prediction for the Group of Users. *Computational Intelligence and Neuroscience* 2016 (2016), 1291358:1–1291358:28. https://doi.org/10.1155/2016/1291358

[27] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. AUAI Press, 452–461.

[28] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender Systems: Introduction and Challenges. In *Recommender Systems Handbook*. Springer, 1–34. https://doi.org/10.1007/978-1-4899-7637-6_1

[29] Alan Said and Alejandro Bellogín. 2014. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *RecSys*. ACM, 129–136. https://doi.org/10.1145/2645710.2645746

[30] Pablo Sánchez and Alejandro Bellogín. 2021. On the effects of aggregation strategies for different groups of users in venue recommendation. *Information Processing & Management* 58, 5 (2021), 102609. https://doi.org/10.1016/j.ipm.2021.102609

[31] Pablo Sánchez and Alejandro Bellogín. 2022. Point-of-Interest Recommender Systems Based on Location-Based Social Networks: A Survey from an Experimental Perspective. *Comput. Surveys* (jan 2022). https://doi.org/10.1145/3510409

[32] Yali Si, Fuzhi Zhang, and Wenyuan Liu. 2019. An adaptive point-of-interest recommendation method for location-based social networks based on user activity and spatial features. *Knowledge-Based Systems* 163 (2019), 267–282. https://doi.org/10.1016/j.knosys.2018.08.031

[33] Christoph Trattner, Alexander Oberegger, Leandro Balby Marinho, and Denis Parra. 2018. Investigating the utility of the weather context for point of interest recommendations. *Information Technology & Tourism* 19, 1-4 (2018), 117–150. https://doi.org/10.1007/s40558-017-0100-9

[34] Saul Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys*. ACM, 109–116. https://doi.org/10.1145/2043932.2043955

[35] Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory Cultural Mapping Based on Collective Behavior Data in Location-Based Social Networks. *Transactions on Intelligent Systems and Technology* 7, 3 (2016), 30:1–30:23. https://doi.org/10.1145/2814575

[36] Andrew Yiannakis and Heather Gibson. 1992. Roles tourists play. *Annals of Tourism Research* 19, 2 (1992), 287–303. https://doi.org//10.1016/0160-7383(92)90082-Z

[37] Fajie Yuan, Joemon M. Jose, Guibing Guo, Long Chen, Haitao Yu, and Rami Suleiman Alkhawaldeh. 2016. Joint Geo-Spatial Preference and Pairwise Ranking for Point-of-Interest Recommendation. In *ICTAI*. IEEE Computer Society, 46–53. https://doi.org/10.1109/ICTAI.2016.0018

[38] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat-Thalmann. 2013. Time-aware point-of-interest recommendation. In *SIGIR*. ACM, 363–372. https://doi.org/10.1145/2484028.2484030

[39] Chenbin Zhang, Hongyu Zhang, and Jianqiang Wang. 2018. Personalized restaurant recommendation method combining group correlations and customer preferences. *Information Sciences* 454-455 (2018), 128–143. https://doi.org/10.1016/j.ins.2018.04.061

A.3 Publication 3: "How Long to Stay Where? On the Amount of Item Consumption in Travel Recommendation"

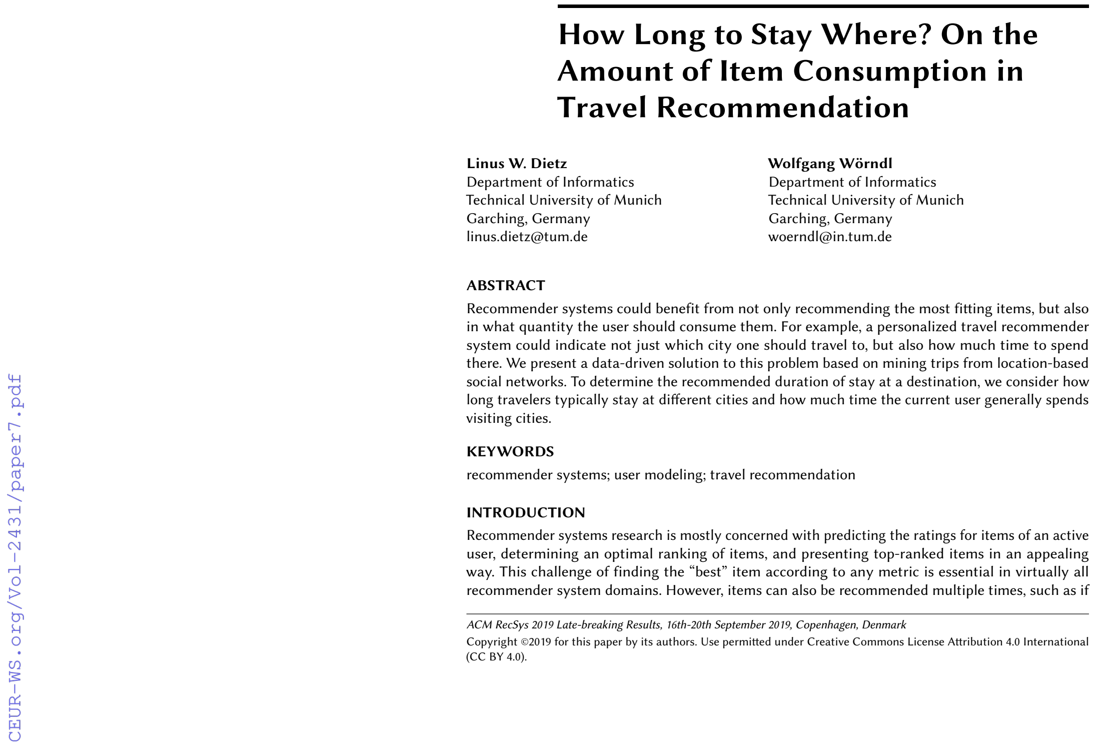
Linus W. Dietz and Wolfgang Wörndl. "How Long to Stay Where? On the Amount of Item Consumption in Travel Recommendation." In: *ACM RecSys 2019 Late-breaking Results*. Sept. 2019, pp. 31–35

Summary

Recommender systems could benefit from not only recommending the most fitting items but also in what quantity the user should consume them. In this paper, we tackle the problem of recommending the personalized duration of stay at a destination. We present a data-driven solution to this problem based on mining trips from location-based social networks. To determine the recommended duration of stay at a destination, we use a statistical approach based on how long travelers typically stay in different cities and how much time the current user generally spends visiting cities. The method can serve as an extension of personalized travel planning systems by not just recommending which city one should travel to but also how much time to spend there.

Author Contributions

LD was the main author of the manuscript, developed the methods, and did all statistical analyses. WW contributed the literature survey and co-edited the paper.



How Long to Stay Where? On the Amount of Item Consumption in Travel Recommendation

Linus W. Dietz
Department of Informatics
Technical University of Munich
Garching, Germany
linus.dietz@tum.de

Wolfgang Wörndl
Department of Informatics
Technical University of Munich
Garching, Germany
woerndl@in.tum.de

ABSTRACT

Recommender systems could benefit from not only recommending the most fitting items, but also in what quantity the user should consume them. For example, a personalized travel recommender system could indicate not just which city one should travel to, but also how much time to spend there. We present a data-driven solution to this problem based on mining trips from location-based social networks. To determine the recommended duration of stay at a destination, we consider how long travelers typically stay at different cities and how much time the current user generally spends visiting cities.

KEYWORDS

recommender systems; user modeling; travel recommendation

INTRODUCTION

Recommender systems research is mostly concerned with predicting the ratings for items of an active user, determining an optimal ranking of items, and presenting top-ranked items in an appealing way. This challenge of finding the “best” item according to any metric is essential in virtually all recommender system domains. However, items can also be recommended multiple times, such as if

ACM RecSys 2019 Late-breaking Results, 16th-20th September 2019, Copenhagen, Denmark

favorite artists appear repeatedly in a long music playlist. In this case, the assumption is that an item should be watched, listened to or experienced as a whole, not only parts of it.

In some scenarios, it is important to decide not just which items should be recommend, but how much of an item should be consumed. For example, a destination recommender should not only recommend where to go, but also the optimal duration of one's stay at each location. The duration can vary depending on the relevance of the item and other domain-related factors, such as the type of traveler [4]. Furthermore, items may be recommended multiple times within a travel package [8]. For example, a recommendation regarding the perfect day at an amusement park might call for riding on the same attraction multiple times.

In this paper, we examine the problem of determining the personalized amount of recommended item consumption in recommender systems, since previous approaches have not solved this problem convincingly. We then present a way to derive the duration of stays in the domain of destination recommendation. Finally, we discuss the generalizability of our method and draw conclusions.

RELATED WORK

There is very limited related work with regard to determining the amount of item consumption in recommender systems for travel and tourism.

Melià-Seguí et al. have investigated the typical duration of stays for tourists visiting different point-of-interest categories using a Foursquare data set; for example, they considered the average amount of time that users spent in restaurants [7]. Google Maps also presents information on how much time visitors spent at selected venues in its search results. However, this information represents only the duration of visits to individual locations or categories of locations; it cannot be used directly to construct recommendations on how long to stay in a city or travel region.

There are several approaches to recommending travel packages, such as the Tourist-Area-Season Topic (TAST) Model [6]. The idea underlying this model is to analyze features of travel packages with regard to their item and user representations, which can then be utilized in a recommender system. In this and similar work, features such as seasonality and item prices are often taken into account, but the duration of stay is either fixed and predetermined, or not considered at all. A related problem is to combine several destinations in a single composite trip. Since travelers' time availability and budget are usually constrained, this recommendation problem can be modeled as a knapsack problem with a scoring function that balances the benefit and cost of items within the package. Herzog and Wörndl have presented an approach to scoring travel regions based on user preferences and then combining them into a longer trip [5]. The score of a region is gradually decreased on a weekly basis, so different regions with lower initial scores may be added to the knapsack. However, this adjustment of the duration of stay is very coarse and not adapted to item or user characteristics in more detail.

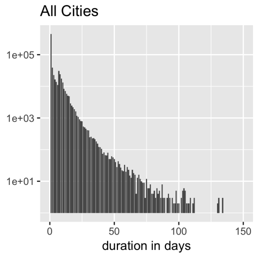


Figure 1: Distribution of the durations of blocks of trips in all 3,938 cities

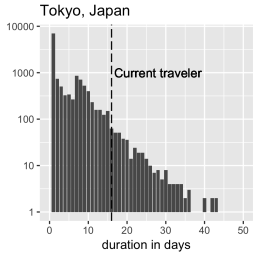


Figure 2: Distribution of the durations of blocks of trips in Tokyo, Japan

When recommending a sequence of travel-related items, such as an itinerary for a city visit, the problem of how much time to spend at individual locations arises as well. For example, De Choudhury et al. have analyzed Flickr photo streams to reconstruct paths of tourists in a city [1]. This information is useful for creating an interesting itinerary once a tourist has already selected a city to visit, but it does not tackle the problem of where to go on a trip or for how long. The determination of the duration of stay is an open research problem [2], which could be resolved through mobility analysis of traveler data [3]. To the best of our knowledge, no existing approach adequately addresses the problem raised here.

DERIVING THE DURATION OF STAY IN A DESTINATION RECOMMENDER SYSTEM

Having analyzed the related work, we will now sketch our ideas as to how to resolve the problem in the domain of destination recommendation. Our proposed solution addresses the question of making personalized recommendations regarding the duration of a tourist's visit to a city by considering two factors: the typical time that all tourists spend in that city and the particular user's average length of stay at a given destination. Initially, we need to know the distribution of the durations of people's stay at a destination, since there can be substantial differences between destinations as to how long one needs to explore it. For example, a smaller city can be covered within a day or two, whereas a major metropolis might require more time. The second aspect is the pace at which the particular traveler visits destinations. Some tourists want to immerse themselves deeply into a culture and therefore stay at each location for a longer time, whereas others want to visit as many different places as possible during their holidays. To quantify these behaviors, we need a database of previous trips to establish a distribution of how long people stay at a specific destination, such as a country or a city.

We employ our previously proposed approach to mine trips from a data set stemming from Foursquare [3], a location-based social network (LBSN), where people can check in at venues all over the world. However, the analysis presented in that paper is at a country-level granularity, whereas we look at the duration of stays at the city level. Using a Foursquare data set of 33,278,683 check-ins by 266,909 users [9], we mine 223,688 domestic and 10,963 international trips, requiring a minimum duration of seven days to mitigate the confounding effect of short business trips. These trips are further segmented into blocks, which are consecutive check-ins at the same municipality with over 15,000 inhabitants. The trips have a mean value of 2.944 blocks, resulting in a total of 690,897 blocks for further analysis. The bar plot in Figure 1 shows the distribution of the durations of all blocks, regardless of the city. The logarithmized counts show a bimodal distribution, with most blocks being one day long and another peak at seven days. This second peak can be attributed our decision to set the minimum duration of the whole trip at seven days.

The next step is to determine the pace at which our particular user typically travels, i.e., the distribution of the duration the individual traveler's past blocks. To obtain this information, we can

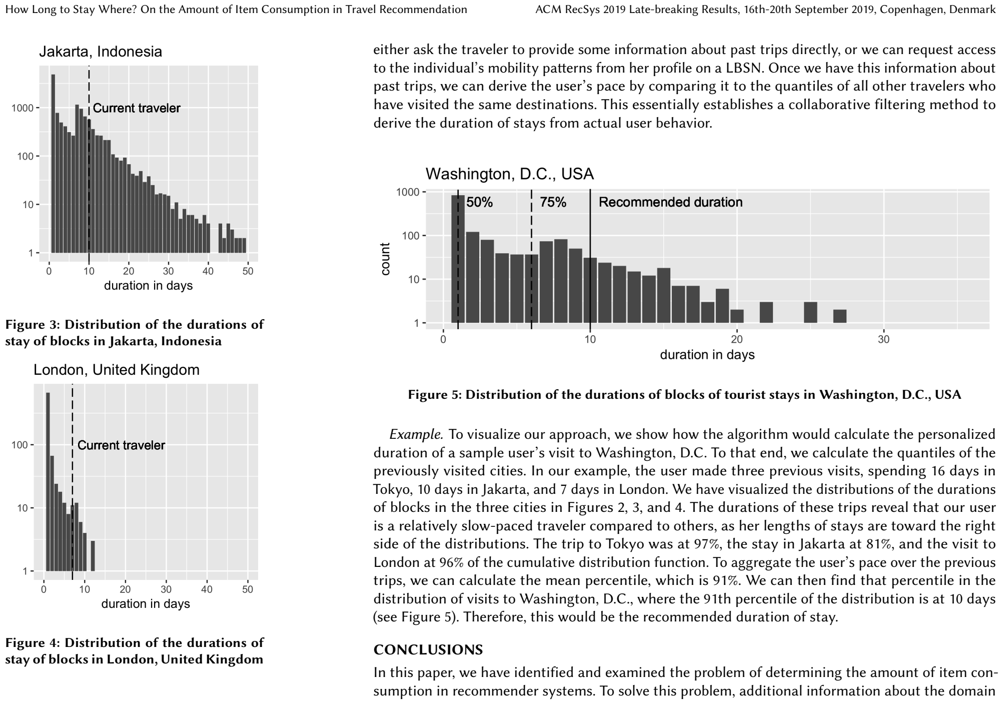


Figure 3: Distribution of the durations of stay of blocks in Jakarta, Indonesia

Figure 4: Distribution of the durations of stay of blocks in London, United Kingdom

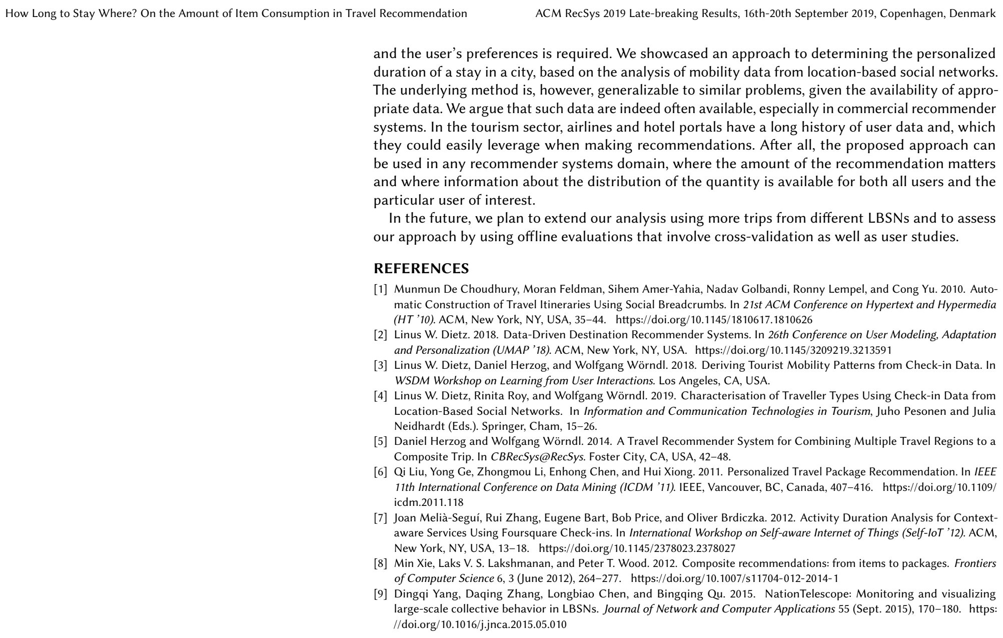
either ask the traveler to provide some information about past trips directly, or we can request access to the individual’s mobility patterns from her profile on a LBSN. Once we have this information about past trips, we can derive the user’s pace by comparing it to the quantiles of all other travelers who have visited the same destinations. This essentially establishes a collaborative filtering method to derive the duration of stays from actual user behavior.

Figure 5: Distribution of the durations of blocks of tourist stays in Washington, D.C., USA

Example. To visualize our approach, we show how the algorithm would calculate the personalized duration of a sample user’s visit to Washington, D.C. To that end, we calculate the quantiles of the previously visited cities. In our example, the user made three previous visits, spending 16 days in Tokyo, 10 days in Jakarta, and 7 days in London. We have visualized the distributions of the durations of blocks in the three cities in Figures 2, 3, and 4. The durations of these trips reveal that our user is a relatively slow-paced traveler compared to others, as her lengths of stays are toward the right side of the distributions. The trip to Tokyo was at 97%, the stay in Jakarta at 81%, and the visit to London at 96% of the cumulative distribution function. To aggregate the user’s pace over the previous trips, we can calculate the mean percentile, which is 91%. We can then find that percentile in the distribution of visits to Washington, D.C., where the 91th percentile of the distribution is at 10 days (see Figure 5). Therefore, this would be the recommended duration of stay.

CONCLUSIONS

In this paper, we have identified and examined the problem of determining the amount of item consumption in recommender systems. To solve this problem, additional information about the domain



and the user's preferences is required. We showcased an approach to determining the personalized duration of a stay in a city, based on the analysis of mobility data from location-based social networks. The underlying method is, however, generalizable to similar problems, given the availability of appropriate data. We argue that such data are indeed often available, especially in commercial recommender systems. In the tourism sector, airlines and hotel portals have a long history of user data and, which they could easily leverage when making recommendations. After all, the proposed approach can be used in any recommender systems domain, where the amount of the recommendation matters and where information about the distribution of the quantity is available for both all users and the particular user of interest.

In the future, we plan to extend our analysis using more trips from different LBSNs and to assess our approach by using offline evaluations that involve cross-validation as well as user studies.

REFERENCES

[1] Munmun De Choudhury, Moran Feldman, Sihem Amer-Yahia, Nadav Golbandi, Ronny Lempel, and Cong Yu. 2010. Automatic Construction of Travel Itineraries Using Social Breadcrumbs. In *21st ACM Conference on Hypertext and Hypermedia (HT '10)*. ACM, New York, NY, USA, 35–44. https://doi.org/10.1145/1810617.1810626

[2] Linus W. Dietz. 2018. Data-Driven Destination Recommender Systems. In *26th Conference on User Modeling, Adaptation and Personalization (UMAP '18)*. ACM, New York, NY, USA. https://doi.org/10.1145/3209219.3213591

[3] Linus W. Dietz, Daniel Herzog, and Wolfgang Wörndl. 2018. Deriving Tourist Mobility Patterns from Check-in Data. In *WSDM Workshop on Learning from User Interactions*. Los Angeles, CA, USA.

[4] Linus W. Dietz, Rinita Roy, and Wolfgang Wörndl. 2019. Characterisation of Traveller Types Using Check-in Data from Location-Based Social Networks. In *Information and Communication Technologies in Tourism*, Juho Pesonen and Julia Neidhardt (Eds.). Springer, Cham, 15–26.

[5] Daniel Herzog and Wolfgang Wörndl. 2014. A Travel Recommender System for Combining Multiple Travel Regions to a Composite Trip. In *CBRecSys@RecSys*. Foster City, CA, USA, 42–48.

[6] Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. 2011. Personalized Travel Package Recommendation. In *IEEE 11th International Conference on Data Mining (ICDM '11)*. IEEE, Vancouver, BC, Canada, 407–416. https://doi.org/10.1109/icdm.2011.118

[7] Joan Melià-Seguí, Rui Zhang, Eugene Bart, Bob Price, and Oliver Brdiczka. 2012. Activity Duration Analysis for Context-aware Services Using Foursquare Check-ins. In *International Workshop on Self-aware Internet of Things (Self-IoT '12)*. ACM, New York, NY, USA, 13–18. https://doi.org/10.1145/2378023.2378027

[8] Min Xie, Laks V. S. Lakshmanan, and Peter T. Wood. 2012. Composite recommendations: from items to packages. *Frontiers of Computer Science* 6, 3 (June 2012), 264–277. https://doi.org/10.1007/s11704-012-2014-1

[9] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. 2015. NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. *Journal of Network and Computer Applications* 55 (Sept. 2015), 170–180. https://doi.org/10.1016/j.jnca.2015.05.010

A.4 Publication 4: "A Comparative Study of Data-driven Models for Travel Destination Characterization"

Linus W. Dietz et al. "A Comparative Study of Data-driven Models for Travel Destination Characterization." In: *Frontiers in Big Data* 5 (Apr. 2022). ISSN: 2624-909X. DOI: 10.3389/fdata.2022.829939

Summary

Characterizing items for content-based recommender systems is challenging in complex domains such as travel and tourism. In the case of destination recommendation, no feature set can be readily used as a similarity ground truth, which makes it hard to evaluate the quality of destination characterization approaches. Furthermore, the process should scale well for many items, be cost-efficient, and, most importantly, correct. To evaluate which data sources are most suitable, we investigate 18 characterization methods that fall into the following categories: venue data, textual data, and factual data. We make these data models comparable using rank agreement metrics and reveal which data sources capture similar underlying concepts. To support choosing more suitable data models, we capture the desired concept using an expert survey and evaluate our characterization methods toward it. We find that the textual models to characterize cities perform best overall, with data models based on factual and venue data being less competitive. However, we show that data models with explicit features can be optimized by learning weights for their features.

Author Contributions

LD has compiled the data, developed the metrics, conducted the analyses, and was the main author of the manuscript. MS contributed various data sources and co-wrote the manuscript. SM contributed two data models and was the main developer of the expert survey instrument. ST contributed two data models. JN recruited domain experts and together with WW supervised the project. All authors read and approved the final manuscript.

frontiers | Frontiers in Big Data

ORIGINAL RESEARCH
published: 07 April 2022
doi: 10.3389/fdata.2022.829939

A Comparative Study of Data-Driven Models for Travel Destination Characterization

Linus W. Dietz[1], Mete Sertkan[2], Saadi Myftija[1], Sameera Thimbiri Palage[1], Julia Neidhardt[2] and Wolfgang Wörndl[1]*

[1] Department of Informatics, Technical University of Munich, Garching, Germany, [2] Research Unit E-Commerce, TU Wien, Vienna, Austria

Characterizing items for content-based recommender systems is a challenging task in complex domains such as travel and tourism. In the case of destination recommendation, no feature set can be readily used as a similarity ground truth, which makes it hard to evaluate the quality of destination characterization approaches. Furthermore, the process should scale well for many items, be cost-efficient, and most importantly correct. To evaluate which data sources are most suitable, we investigate 18 characterization methods that fall into three categories: venue data, textual data, and factual data. We make these data models comparable using rank agreement metrics and reveal which data sources capture similar underlying concepts. To support choosing more suitable data models, we capture a desired concept using an expert survey and evaluate our characterization methods toward it. We find that the textual models to characterize cities perform best overall, with data models based on factual and venue data being less competitive. However, we show that data models with explicit features can be optimized by learning weights for their features.

Keywords: destination characterization, rank agreement metrics, expert evaluation, data mining, recommender systems, content-based filtering

OPEN ACCESS

Edited by:
Alejandro Bellogín,
Autonomous University of Madrid,
Spain

Reviewed by:
Mehdi Elahi,
University of Bergen, Norway
Christine Bauer,
Utrecht University, Netherlands

****Correspondence:***
Linus W. Dietz
linus.dietz@tum.de

Specialty section:
This article was submitted to
Recommender Systems,
a section of the journal
Frontiers in Big Data

***Received:** 06 December 2021*
***Accepted:** 21 February 2022*
***Published:** 07 April 2022*

Citation:
Dietz LW, Sertkan M, Myftija S,
Thimbiri Palage S, Neidhardt J and
Wörndl W (2022) A Comparative
Study of Data-Driven Models for Travel
Destination Characterization.
Front. Big Data 5:829939.
doi: 10.3389/fdata.2022.829939

1. INTRODUCTION

The performance of data-driven systems is inherently determined by the underlying quality of data, which is becoming increasingly hard to judge in the current era of big data. When deciding on which features to use in the data model of an information retrieval or content-based recommender system, there are often several options to choose from. Out of the many options how to model a domain, how can one determine which instantiation of the available data is the *best?* The data-driven characterization of real-world items should capture each entity as closely as possible with respect to the user task supported by the system. As a principle, similar items in the physical world should also be similar in the information space, despite the loss of information and granularity. Thus, authors of content-based filtering algorithms (Pazzani and Billsus, 2007) should evaluate whether their data model matches the user goals, since a divergence might cause confusion and inevitably decrease the trust in and satisfaction with the system. Sometimes, the mapping to the physical world is obvious. When recommending a computer configuration, the feature values, such as the available memory or number of USB outlets have clear meaning and can be easily interpreted by the users and algorithms (Zhang and Pu, 2006). In other domains, however, a ground truth for items similarity is hard to capture, which is a fundamental problem (Yao and Harper, 2018): What are the movies most similar to “Fight Club”? Which cities are most similar to Munich? We as humans might have an intuition about such similarity concepts, but it is hard to develop recommendation algorithms that emulate such, possibly latent, concepts. Evaluating this is an under-researched challenge in the area of content-based recommendation, especially in the travel and tourism domain, where items such as destinations or travel packs are often not as clearly defined as consumer products. Approaches that rely on a history of explicit or implicit user interactions, such as collaborative filtering or bandit algorithms do not face this problem since there is a clear connection between the item and the user’s rating (Su and Khoshgoftaar, 2009). Still, in cold-start situations, i.e., when too little interaction data is available for employing such approaches, a common strategy used is hybridization, which again requires using content-based algorithms to compute the initial recommendations (Çano and Morisio, 2017).

To make these considerations concrete, take a destination recommender system as an example. The CityRec system (Dietz et al., 2022), allows users to refine their travel preferences based on six features that were obtained and derived from various sources such as Foursquare and open data portals: “Nightlife,” “Food,” “Arts and Entertainment,” “Outdoor and Recreation,” “Average Temperature,” and “Cost.” However, when developing this system, we faced the issue of determining which data set and features are the most accurate and useful for prospective travelers to reason about destinations. Given that recommendations are computed using the cities’ features and the users base their decisions on them, inaccuracies in the data model negatively impact the trust in the system. To the best of our knowledge, this two-fold challenge of choosing accurate data sources to quantify specific aspects travel destinations, as well as choosing which features to incorporate in a content-based recommender system has not been analyzed in a systematic way (Yao and Harper, 2018). This motivated us to develop this toolbox of methods to compare data sources with each other and also with respect to what is important in the domain of such a recommender system.

To make our contributions generalizable for different data models in various domains, we rely on rank agreement methods (Kendall, 1970), which operate on ranked lists based on the similarity measure of the recommender system. The proposed methods quantify correlations between conceptually diverse characterization methods to enable informed decision making with respect to which one to employ. For example, if it turns out that two characterization methods are highly correlated, i.e., both capture the same underlying concepts using different features, one could go ahead and exchange one for another without introducing disruptive changes in the resulting recommendations.

Furthermore, we propose a method to assess the quality of data models with respect to a *desired* concept. We argue that a destination recommender system should use a data model that results in recommendations that emulate the destination experience as closely as possible. To achieve this, it is imperative to assess which available data source and feature set approximate the concept of touristic experience best. However, such a *“gold standard”* is readily not available and typically can only be elicited for a small subset of the recommendation items. We elicit the concept of touristic experience using an expert study and propose methods to assess the quality of the characterization methods with such incomplete information.

To showcase the utility of our approach, we exercise the methods within the particularly challenging domain of content-based destination recommendation (Le Falher et al., 2015; Liu et al., 2018), where recommendations are solely computed based on the items’ characteristics as opposed to rating or interaction data in collaborative filtering approaches. For this, we introduce 18 destination characterization methods for 140 cities, which we have collected from literature or constructed ourselves. Using well-established rank correlation methods (Kendall, 1970), we compute their pairwise similarities, thereby revealing families of similar data sources. To evaluate the data sources with respect to how tourists experience a destination, we conduct an expert study to elicit this latent concept. Using variants of established top-k rank agreement methods, we are able to assess the quality of the data sources by their similarity to the expert opinions.

The main contributions of this work are as follows:

- We propose a method to assess the similarities and the quality of data models characterizing items in content-based recommender systems.
- We introduce, instantiate, and compare 18 different destination characterization methods using the proposed methodology.
- We conduct a survey among travel experts to establish a similarity baseline of the different destination characterization methods. Using this expert-elicited concept, we assess and optimize the data sources with respect to this concept.

While we use destination characterization as our running example, our methods are not specific to this domain, since the proposed methods operate on ranked lists of any kind.

The structure of the paper is the following: after discussing the prior work in Section 2, we provide a description of the data for destination characterization. In Section 4, we introduce our methodology of how we made the data sources comparable using rank agreement metrics. The expert study in Section 5 shows how we elicited our desired concept. In Section 6, we present the analysis of which data sources capture similar concepts and which approximates the concept of touristic experience best. Finally, we conclude our findings and point out future work in Section 7.

2. RELATED WORK

In recommender systems research, most algorithms traditionally use the collaborative filtering paradigm, i.e., interpreting user ratings (or similar explicit feedback) of items. However, in cold-start recommendation situations, where such interaction data does not exist to sufficient degree, content-based algorithms (Lops et al., 2011) play a role to be able to generate

Abbreviations: LBSN, location-based social network; OSM, OpenStreetMap.

meaningful and personalized recommendations to the users. This research is motivated by practical challenges of content-based information retrieval systems, especially personalized recommender systems.

Concerning the similar item recommendation problem, Yao and Harper have shown that content-based algorithms have outperformed ratings- and clickstream-based ones with respect to the perceived similarity of items and the overall quality of the recommendations in the movie domain (Yao and Harper, 2018). Our work goes one step further: We provide a framework that allows to compare different data models of the same items with respect to the similarity according to a desired concept and additionally provide means to optimize the feature weights to approximate the concept even better.

We use the travel and tourism recommendation domain as a running example; however, other domains face similar challenges. In their survey, Borràs et al. (2014) identify four different tasks that tourism recommender systems have to cope with: recommending travel destinations or travel packages (Liu et al., 2011), suggesting attractions (Massimo and Ricci, 2018; Sánchez and Bellogín, 2022), planning trips (Gavalas et al., 2014; Dietz and Weimert, 2018), and accounting for social aspects (Gretzel, 2011). We aim to contribute to the feature engineering challenges of the first task: recommending travel destinations. We focus on the characterization of destinations, which is the task to establish the underlying data model for destination recommender systems. Herein, “destination” refers to cities. Key challenges in recommending cities are the intangibility of the recommended item, high consumption costs, and high emotional involvement (Werthner and Ricci, 2004).

Burke and Ramezani (2011) suggest the content-based recommendation paradigm as one of the appropriate ones for the tourism domain. Content-based recommenders need a domain model and an appropriate distance measure to enable effective matchmaking between user preferences and items to generate recommendations without details rating or interaction data. For this reason, they can be successful in situations where the interaction with the recommender is very rare and short-term and the user model can be derived from alternative information sources. Domain models in recommenders have been constructed using various data sets (Dietz et al., 2019) derived through analyses and user studies (Neidhardt et al., 2014, 2015) or realized through ontologies (Moreno et al., 2013; Grün et al., 2017). In this work, we compare different data-driven destination characterization methods, which project destinations onto the respective search spaces using different types of data, cf. Section 3.

Naturally, the question of which features are useful to characterize a destination for efficient retrieval in an information system arises. Dietz (2018) mentions challenges of characterizing destinations: the destination boundaries must be clearly defined, the data needs to be kept up-to date, and the features should be relevant with respect to the recommendation goal. Analyzing LBSN data to characterize cities and their districts has been an active topic in previous years (Silva et al., 2019). It has been shown that such data is quite useful to unveil characteristics of certain districts within a city (Le Falher et al., 2015). McKenzie and Adams (2017) suggested the use of Kernel density estimation models of check-ins to identify thematic areas within a city.

A related line of research is concerned with capturing and visualizing intangible concepts in urban areas: Quercia et al. (2015) used LBSN data and Google Street View imagery[1] to determine intangible concepts such as the smell, the soundscape (Aiello et al., 2016), and general happiness (Quercia et al., 2014) on a street granularity. Analogously, street imagery can also be reliably used to measure distributions of income, education, unemployment, housing, living environment, health and crime as Suel et al. (2019) have demonstrated. Finally, it has been shown that it possible to automatically distinguish cities based on their architectural elements learned from street imagery (Doersch et al., 2015). Using features derived from such approaches could also be used to compute similarities of cities and their districts. Obtaining Google Street View images is feasible on a small and medium scale, however, the costs to do such an analysis on a global scale prevented us from experimenting with this data source.

In the area of content-based characterization of destinations for use in recommender systems there are so far few approaches. Sertkan et al. (2017) characterized a huge data set of 16,950 destinations based on 26 motivational ratings and 12 geographical attributes within the Seven Factor Model of tourism motifs. They proposed a cluster analysis and regression analysis to map the destinations to the vector space of the Seven Factor Model (Neidhardt et al., 2015). The framework was recently also used by Grossmann et al. (2019) to elicit preferences of prospective tourists using picture of destinations While modeling the user’s interests using travel-related pictures has been shown to be possible, obtaining a representative set of images of global destinations in an automated fashion is an open research problem (Sertkan et al., 2020a,b). The development of the CityRec recommender system (Dietz et al., 2019, 2022) partly motivated investigating different data models for recommender system: CityRec uses a domain model based on Foursquare venue categories and further information such as a cost index or climate data collected from web APIs (Dietz et al., 2019). This data model is used both to elicit user preferences via conversational refinement, i.e., turn-based adjustment of the preferences in a dialogue with the system and to compute the recommendations in a content-based way.

It is striking that researchers invest a lot of energy into capturing signals from various online sources to approximate complex, intangible concepts. To the best of our knowledge, these data models are rarely systematically verified as to what extend they approximate the recommendation domain. In this paper, we propose a collection of several methods to provide researchers with tools to evaluate this.

3. DATA SOURCES

To characterize destinations, we used various online data sources showcased in **Table 1**. Our selection criteria for the data sources

[1] https://www.google.com/streetview/

were that they should have touristic relevancy, i.e., prospective travelers should be able to use them to familiarize themselves with a destination, or that they are already part of travel-related information systems, as is the case with the data set from Foursquare and the data sources in the factual category.

To obtain a balanced collection of destinations on all continents that would reflect the cultural differences of the travel destinations, we initially gathered an extensive list of prominent cities such as capital cities or relevant travel destinations. Unfortunately, not all cities could be characterized with all methods. Several destinations were not included in Nomad List and OpenStreetMap did often not have proper city boundaries for several cities in Asia and Africa. Our proposed approaches to make the destination characterization methods comparable, however, require complete data for each city. Thus, our data set for this study comprises a set of 140 cities, which are those that could be characterized with all data sources. Looking at **Figure 1**, the distribution of destinations on the planet is missing cities in Central Asia and Africa that had to be excluded due to this requirement, otherwise the distribution would roughly correspond to the world’s population density. The full list is available in the replication pack.

Throughout this paper, we work with ranked lists of the cities most similar to one city. Such lists are based on one data source and start with the base city followed by the 139 other cities. This means that for each data source there are 140 such ranked lists. In the following, we describe the data sources and how we computed the similarity metrics between the cities.

3.1. Venue Data

The first type of data we used to characterize destinations is venue-based data. Intuitively, the variety of venues one can visit at a destination might reflect the experience of a traveler. The underlying assumption is that a destination can be characterized using the distribution of all its touristic venues. The following characterization methods rely on the assumption that the larger the variety of, e.g., restaurants or cultural sites of a city is, the better the score should be in these categories. This also means that we do not aim to assess the quality of the venues, since most venues do not come with quality indications such as ratings.

3.1.1. Foursquare

Foursquare is a LBSN that offers a rich, well-structured taxonomy of venue categories and also allows reasonably generous API rate limits to crawl data from it. Using the “search venues” endpoint[2], we were able to obtain a collection of each city’s venues using a recursive algorithm that exhaustively queried all Foursquare venues specified within a bounding box. Using this method, we collected 2,468,736 venues in 140 cities that had at least 5,000 venues each.

To create the specific set of lists, we needed to establish an association between the cities and the venue types. Foursquare provides a well-defined venue category hierarchy[3], which allows us to map every venue to a top-level category, e.g., Science Museum → Museum → Arts & Entertainment. We use the tourism-related subset of these categories to create a feature set that enables us to characterize the cities, shown in **Figure 2**. These features can be conceptualized as a multi-dimensional vector space, however, to perform reasonable comparisons the data must be normalized to make large and small cities comparable. By normalizing the number of venues in each category using the total venue count of the city, we obtain the percentage of each category in the city’s category distribution. This approach relies on the assumption that a larger number of venues in a certain category improves the touristic experience while visiting it. A simple example helps in demonstrating this: The cities in our data set have a certain distribution of venue categories; if the number of venues labeled with “Arts & Entertainment” in a city is on the high end of that distribution, it can be assumed that it likely offers a larger number of opportunities and should, thus, get a higher score in this category. **Figure 2** shows the category distributions of a few cities of different continents and sizes that we have chosen as illustrative examples. Note that, unlike in the data model, this visualization is not normalized with respect to the number of venues. Examining **Figure 2**, one can see that many cities have a somewhat similar distribution of venue categories, where “Food” and “Shops & Service” dominate in general. To eliminate this effect, we apply min-max scaling to the calculated percentages. This way we obtain the final city scores for each of the features, which take values in $[0, 1]$.

Using this method, we constructed two data models from Foursquare. The first on the four top-level categories – “Arts & Entertainment,” “Food,” “Outdoors & Recreation,” and “Nightlife” – and another one using the 337 second-level categories as aggregation target.

3.1.2. OpenStreetMap

With OSM, we used a similar approach as with Foursquare. To obtain all map features, we set up our own OSM server and developed a querying client to obtain the entities within the city relations. The map entities are again hierarchically categorized on three levels[4]. The 27 top-level categories are subdivided into several subcategories which finally contain 1,032 types of map features. For example, the “amenity” category has several subcategories, such as “Healthcare,” “Transportation,” and “Entertainment, Arts and Culture.” These subcategories again contain numerous entities, uniquely identified by the full path, for example “amenity:entertainment/arts” and “culture:cinema.” As opposed to the Foursquare characterization, we also had exact city boundaries, so we could compute the area of the destination.

Leveraging this hierarchy, we again built a top-level model, which collapses the map entities to tourism-related categories: “tourism,” “leisure,” “historic,” “natural,” and “sport” as well as the venue count and the area. The second-level model comprises entities of 14 tourism-related subcategories as well as the venue count and area.

[2] https://developer.foursquare.com/docs/api/venues/search

[3] https://developer.foursquare.com/docs/resources/categories

[4] https://wiki.openstreetmap.org/wiki/Map_Features

TABLE 1 | Overview of the data sources for characterizing cities.

Type	Name	Category	Data objects	Number of objects	Acronym
Venue	Foursquare	LBSN	Venues	2,468,736	FSQ
Data	OpenStreetMap	Collaborative map	Map entities	3,106,856	OSM
Textual	Wikipedia	Collaborative encyclopedia	Documents	1,150,719 words	WP
	Wikitravel	Travel-related Wiki	Documents	984,777 words	WT
	Google travel	Travel information	Documents	56,499 words	GT
Factual	Webologen	Travel information provider	City features	49 tourism facts/city	TF
	Nomad list	Collaborative travel information	City features	8 features / city	Nomadlist
	Seven factor model	Scientific characterization	Derived factors	7 factors / city	7FM-2018
	Geographic location	Geographic location	Latitude, longitude	1 coordinate pair / city	GEO

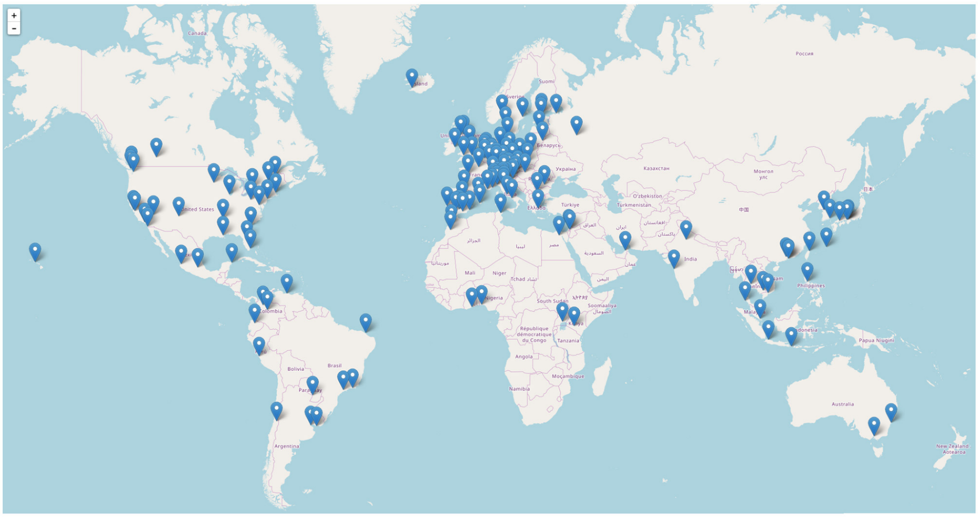


FIGURE 1 | Geographic distribution of the characterized cities. Map data © OpenStreetMap contributors, see https://www.openstreetmap.org/copyright.

3.2. Textual Data

Using document similarity assessments such as the Jaccard Distance, Word2Vec embeddings and a transformers-based approach on texts describing a destination, we are able to compute pairwise similarities between the cities. As textual basis, we used three online resources: Wikipedia, Wikitravel, and Google Travel.

3.2.1. Wikipedia

We used the articles of the English Wikipedia[5] about the 140 cities to compute the similarity between cities. The mean length of the articles about our destinations was 8,219 words.

3.2.2. Wikitravel

This collaborative travel guide[6] provides useful information about touristic destinations. It is free of charge and offers detailed information about possible activities, recommended restaurants, and general advice for traveling. Some cities have sub-pages about their districts, however, we have only used the main articles to maintain comparability. The mean length of the articles was 7,034 words.

3.2.3. Google Travel

Another popular platform for learning about travel destinations and planning trips is Google Travel[7]. Based on actual traveler visits and local insights, the platform provides a list of most

[5] https://en.wikipedia.org

[6] https://wikitravel.org

[7] https://www.google.com/travel

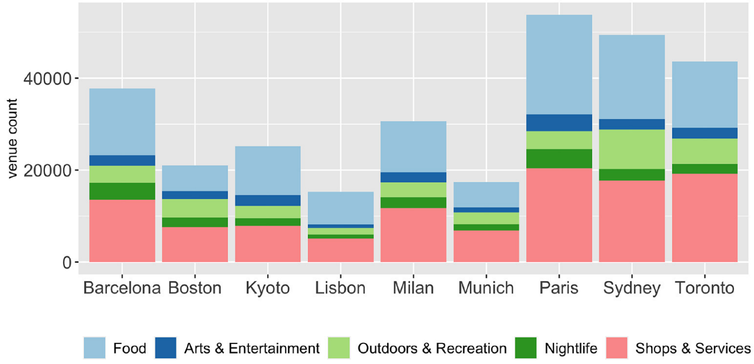


FIGURE 2 | Venue category distribution for a subset of cities.

iconic attractions. In this work, we used the short description of each attraction at a destination. For example, the description of *Schönbrunn Palace* in Vienna is *"Baroque palace with opulent interiors."* Since there are often myriads of attractions in a city, we concatenated these descriptions into one document to obtain the overall description of the considered city. This resulted in one document per city with a mean length of 404 words.

3.2.4. Text Processing and Similarity Measures

For the raw text of the three text sources, we used the same pre-processing steps. After the HTML tags were removed, the text was put in lower case and stripped of all special characters, such as line breaks and punctuation marks. Then, the terms were tokenized using a standard word tokenizer and the stop words eliminated to reduce noise.

For the Jaccard models, the document term matrix was computed based on the cleaned text and then the similarity between the cities was computed using the Jaccard Distance.

We use pre-trained Word2Vec and BERT-based models as zero-shot encoders to embed the documents. In case of the Word2Vec-based models, we aggregated the pre-trained word embeddings using mean-pooling to obtain the document embedding and used the cosine similarity to compute the similarity. We utilize the open-source library *spaCy*[8] and in particular the english-core-web-large model, which outputs 300-dimensional vectors and is trained on *OntoNotes 5* (Weischedel et al., 2011), *ClearNLP Constituent-to-Dependency Conversion* (Choi et al., 2015), *WordNet 3.0* (Miller, 1995), and *GloVe Common Crawl* (Pennington et al., 2014). This means, that there was no need to fine-tuning the models; the default hyperparameters of Spacy could be re-used.

The BERT-based sentence encoder (Yang et al., 2021) we employed was also pre-trained on Wikipedia and Common Crawl[9] to encode the documents, thus, again making further fine-tuning of (hyper-) parameters obsolete. We use the cosine similarity to rank the cities. Thus, we obtained nine textual ranked lists: three data sources × three similarity measures.

3.3. Factual Data

The third category is factual data with a focus on travel and tourism. This group comprises data sources that had readily available facts about destinations, such as rated features or geo-social features relevant for travelers. However, this does not imply that the quality of the data is beyond scrutiny.

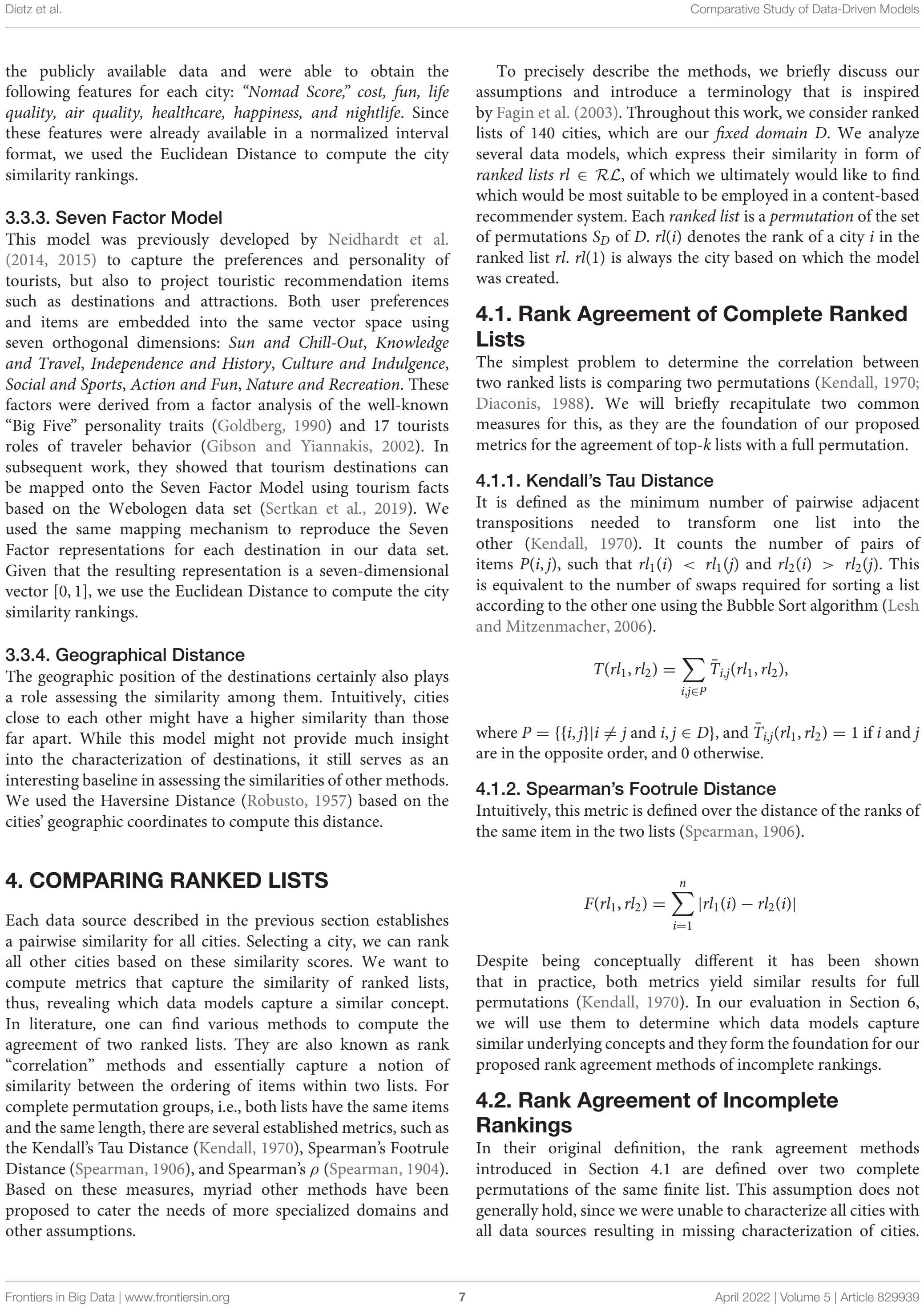
3.3.1. Webologen Tourism Facts

The former German eTourism start-up Webologen compiled a data set of 30,000 cities, which are described by 22 geographical attributes and 27 "motivational" ratings. The geographical attributes have binary values indicating the presence or absence of various geographical attributes: *sea, mountain, lake, island, etc.* The motivational ratings were assessed using proprietary methods at Webologen, taking into account infrastructure, climate, marketing, and economic data. With a score between 0 and 1, the motivational ratings, such as *nightlife, wellness, shopping, nature and landscape*, measure the quality of those touristic aspects at a destination. The higher the value, the better this aspect is for a traveler. Given this multitude of features, this data set provides a very detailed image of a tourism destination. Since there are multiple types of data (i.e., binary and interval scale), we use the Gower Distance (Gower, 1971) to compute the similarity for the city rankings.

3.3.2. Nomad List

As opposed to Webologen's approach, Nomadlist[10] employed a mixture of own data modeling and crowdsourcing to characterize cities for their suitability for digital nomadism. Built as a specialized platform for this community, it offers rich information about the cities in its database. We crawled

[8] https://spacy.io
[9] https://commoncrawl.org
[10] https://nomadlist.com



the publicly available data and were able to obtain the following features for each city: *"Nomad Score," cost, fun, life quality, air quality, healthcare, happiness, and nightlife*. Since these features were already available in a normalized interval format, we used the Euclidean Distance to compute the city similarity rankings.

3.3.3. Seven Factor Model

This model was previously developed by Neidhardt et al. (2014, 2015) to capture the preferences and personality of tourists, but also to project touristic recommendation items such as destinations and attractions. Both user preferences and items are embedded into the same vector space using seven orthogonal dimensions: *Sun and Chill-Out*, *Knowledge and Travel*, *Independence and History*, *Culture and Indulgence*, *Social and Sports*, *Action and Fun*, *Nature and Recreation*. These factors were derived from a factor analysis of the well-known "Big Five" personality traits (Goldberg, 1990) and 17 tourists roles of traveler behavior (Gibson and Yiannakis, 2002). In subsequent work, they showed that tourism destinations can be mapped onto the Seven Factor Model using tourism facts based on the Webologen data set (Sertkan et al., 2019). We used the same mapping mechanism to reproduce the Seven Factor representations for each destination in our data set. Given that the resulting representation is a seven-dimensional vector $[0, 1]$, we use the Euclidean Distance to compute the city similarity rankings.

3.3.4. Geographical Distance

The geographic position of the destinations certainly also plays a role assessing the similarity among them. Intuitively, cities close to each other might have a higher similarity than those far apart. While this model might not provide much insight into the characterization of destinations, it still serves as an interesting baseline in assessing the similarities of other methods. We used the Haversine Distance (Robusto, 1957) based on the cities' geographic coordinates to compute this distance.

4. COMPARING RANKED LISTS

Each data source described in the previous section establishes a pairwise similarity for all cities. Selecting a city, we can rank all other cities based on these similarity scores. We want to compute metrics that capture the similarity of ranked lists, thus, revealing which data models capture a similar concept. In literature, one can find various methods to compute the agreement of two ranked lists. They are also known as rank "correlation" methods and essentially capture a notion of similarity between the ordering of items within two lists. For complete permutation groups, i.e., both lists have the same items and the same length, there are several established metrics, such as the Kendall's Tau Distance (Kendall, 1970), Spearman's Footrule Distance (Spearman, 1906), and Spearman's ρ (Spearman, 1904). Based on these measures, myriad other methods have been proposed to cater the needs of more specialized domains and other assumptions.

To precisely describe the methods, we briefly discuss our assumptions and introduce a terminology that is inspired by Fagin et al. (2003). Throughout this work, we consider ranked lists of 140 cities, which are our *fixed domain D*. We analyze several data models, which express their similarity in form of *ranked lists* $rl \in \mathcal{RL}$, of which we ultimately would like to find which would be most suitable to be employed in a content-based recommender system. Each *ranked list* is a *permutation* of the set of permutations S_D of D. $rl(i)$ denotes the rank of a city i in the ranked list rl. $rl(1)$ is always the city based on which the model was created.

4.1. Rank Agreement of Complete Ranked Lists

The simplest problem to determine the correlation between two ranked lists is comparing two permutations (Kendall, 1970; Diaconis, 1988). We will briefly recapitulate two common measures for this, as they are the foundation of our proposed metrics for the agreement of top-k lists with a full permutation.

4.1.1. Kendall's Tau Distance

It is defined as the minimum number of pairwise adjacent transpositions needed to transform one list into the other (Kendall, 1970). It counts the number of pairs of items $P(i,j)$, such that $rl_1(i) < rl_1(j)$ and $rl_2(i) > rl_2(j)$. This is equivalent to the number of swaps required for sorting a list according to the other one using the Bubble Sort algorithm (Lesh and Mitzenmacher, 2006).

$$T(rl_1, rl_2) = \sum_{i,j \in P} \bar{T}_{i,j}(rl_1, rl_2),$$

where $P = \{\{i,j\} | i \neq j \text{ and } i,j \in D\}$, and $\bar{T}_{i,j}(rl_1, rl_2) = 1$ if i and j are in the opposite order, and 0 otherwise.

4.1.2. Spearman's Footrule Distance

Intuitively, this metric is defined over the distance of the ranks of the same item in the two lists (Spearman, 1906).

$$F(rl_1, rl_2) = \sum_{i=1}^{n} |rl_1(i) - rl_2(i)|$$

Despite being conceptually different it has been shown that in practice, both metrics yield similar results for full permutations (Kendall, 1970). In our evaluation in Section 6, we will use them to determine which data models capture similar underlying concepts and they form the foundation for our proposed rank agreement methods of incomplete rankings.

4.2. Rank Agreement of Incomplete Rankings

In their original definition, the rank agreement methods introduced in Section 4.1 are defined over two complete permutations of the same finite list. This assumption does not generally hold, since we were unable to characterize all cities with all data sources resulting in missing characterization of cities.

We sidestepped this problem by considering only a subset of destinations that could be characterized with all methods.

4.2.1. Problem Formulation

The outcome of the expert study (cf. Section 5) is a collection of top-k lists. Each top-k list contains the $k \geq 10$ most similar cities to the city the expert characterized. To find out which data model is the most similar to the experts' opinions, we need to modify the rank agreement methods to cope with this scenario. Concretely, this means that we need to compute the rank agreement between an expert's top-k ranking τ, where $10 \leq k \leq 139$ and a complete permutation of length 140. To the best of our knowledge, we are first to systematically analyze this special case.

4.2.2. Approaches in Literature

In literature, similar problems have been tackled in the area of biostatistics and information retrieval. Critchlow was first to establish a theoretical basis for such rankings (Critchlow, 1985), assuming a fixed domain of items D. One of the most comprehensive papers on the rank agreement of top-k lists is the one of Fagin et al. (2003). Unlike Critchlow and us, they did not assume a fixed domain of items and, thus, proposed very general distance measures for top-k lists that are not directly useful to our scenario. The authors also proved that in the general case, the measures for top-k lists reside in the same equivalence class and showcased further applications of these measures in the context of the rank aggregation problem (Dwork et al., 2001; Lin and Ding, 2008).

An important property of Kendall's Tau and Spearman's Footrule is that all ranks are treated equal, i.e., they do not take the potentially non-uniform relevancy of top-ranked or bottom ranked into account. In many domains, the assumption of uniform relevancy does not hold, thus, several other measures have been proposed. Iman and Conover (1987) proposed a concordance measure that prioritizes rank agreements at the top of the rankings, while Shieh proposed a weighted variant of Kendall's Tau, where the analyst can prioritize either low-ranked or high-ranked items (Shieh, 1998). The Average Precision (AP) correlation is another important measure in information retrieval that more heavily penalizes differences of top-ranked items compared to Kendall's Tau (Yilmaz et al., 2008).

In our domain at hand, the issues motivating the aforementioned papers are not present. Since our top-k lists are very short, we do not need to come up with additional weights based on the position within the list. Furthermore, given the underlying data sources and similarity measures used, the probability of tied ranks in the lists is very low so that this case can be neglected as well (Urbano and Marrero, 2017).

The alternative to dealing with the rank agreement problem of a top-k list and a permutation would be to disregard the inherent order of the ranked list and view it as a set. This would open the door to interpret each element of the list as an independent query, on which traditional information retrieval metrics can be computed such as Precision or the Reciprocal Rank. By repeating this process for each item, one could assess the quality of the expert's selection just as it is frequently done with search engines or recommender systems resulting in metrics, such as the Precision (Precision@K) or the Mean Reciprocal Rank (MRR). Instead, we aim to retain the ranking information by the experts and discuss various methods to compute metrics that operate on the ranked list semantics.

4.2.3. Proposed Methods

Under the assumptions that 1) we have a fixed domain of items, and 2) only the relative ranking in the ranked lists matters, i.e., the concrete values of the agreement are not of importance, there is the option to randomly fill the missing items of a top-k list τ with the remaining items $\{D - \tau\}$ (Ekstrøm et al., 2018). This essentially constructs two permutations, which can then be assessed with the standard metrics from Section 4.1. By repeating this process a large number of times, the effect of the random items at the tail of the list is eliminated and, finally, the ranking is computed based on the mean value of all iterations. This simple idea would be a permissible option for our scenario; however, it requires much overhead computation and does not provide concrete values for the rank agreement, since only the top k items contribute to the signal, while the remaining ones are pure noise.

Thus, an analytic solution for this problem would be preferable. In our scenario, we can always assume that we have a fixed domain of items, since each data model will be able to produce similarity scores between all items. Thus, our problem is similar to the one Fagin et al. resolve in their approach, i.e., comparing one top-k list τ of length k with a ranked list rl (Fagin et al., 2003, Section 3.1), however, due to the fixed domain assumption, we only need to discriminate three cases. This results in a simpler problem without any room for uncertainty that might arise from having items that are in one top-k list, but not in the other.

- Case 1: $i \in \tau$, and $rl(i) \leq k$ (the item is in the top-k list and the rank of the item in the permutation is at most k)
- Case 2: $i \in \tau$, but $rl(i) > k$ (the item is in the top-k list but the rank of the item in the permutation is greater than k)
- Case 3: $i \notin \tau$ (the item is not in the top-k list)

Using this insight, we propose variants to Kendall's Tau and Spearman's Footrule distance for top-k lists.

- **Modified Spearman's Footrule Distance**
  If a city is in the top-k list (Case 1 & 2), we can compute the distance between $\tau(i)$ and $rl(i)$ as before, since all information is still available. In Case 3, we do not add any penalty, since we have no information about which penalty should be applied. Thus, $F'(\tau, rl)$ is simply the footrule distance between all elements of τ and the corresponding elements in rl.

$$F'(\tau, rl) = \sum_{i=1}^{k} |\tau(i) - rl(i)|$$

  Fagin et al. (2003) discuss another variant, $F^{(l)}$, the footrule distance with location parameter l, where they set $l = k + 1$. This is not applicable in our scenario, since we have a fixed domain and have already applied a penalty for each element τ.

- **Modified Kendall's Tau Distance**
  For a modified Kendall's Tau Distance, we again count the number of discordant pairs between τ and rl. This situation is similar to the modified footrule distance, as only the penalties from the elements of τ are applied.

  $$T'(\tau, rl) = \sum_{i,j \in P} \bar{T}_{i,j}(\tau, rl),$$

  where $P = \{\{i,j\} | i \neq j \text{ and } i,j \in D\}$, and $\bar{T}_{i,j}(\tau, rl) = 1$ if i and j are in the opposite order, and 0 otherwise.

Table 2 at the end of the following section exemplifies how we use these derived rank agreement metrics adapted to our scenario to compare top-k lists with complete permutations in Section 6.2.

5. ELICITING A DESIRED CONCEPT THROUGH EXPERT OPINIONS

Now, we want to find out which data source is best suited for the domain of destination recommendation. To do so, we have developed a web-based expert study to capture a very specific concept we are interested in: *"similar experience when visiting cities as a tourist."* To make this latent concept explicit, we asked experts from the travel and tourism domain to give their opinion on this matter, by selecting the most similar destinations to a given city.

In a pilot study, we realized that even for experts, the task to rank the k most similar destinations from a list of 140 cities in world be challenging. Therefore, to obtain a sufficient number of characterizations per city, we restricted the characterization of our expert study to 50 prominent cities. Naturally, we would have preferred to perform a characterization of all cities in the data set, but given that the experts' time was limited, we focused on the 50 cities, which we expected our experts to be most familiar with.

5.1. Expert Survey Instrument

We now describe the user interface of the online survey application and elaborate on the design choices that influenced the system.

5.1.1. Landing Page

The experts were contacted via email and, when they followed the link, they were presented with a landing page, which contained general instructions and the contact data of the authors. They were allowed to choose either from the 50 cities to be surveyed, or in the case of local experts a predefined city they should characterize.

5.1.2. City Similarity Ranking Task

After selecting the city to be characterized, the experts were presented with their task, as shown in **Figure 3**. First, they were asked to provide their familiarity with the given city on a five-point named Likert Scale. Then the concrete task followed, which was to be completed by ranking the cities using three columns. The left "Most Similar Cities" column was initially empty. The middle column contained a precomputed candidate list of 30 cities, that were the most similar to the base city according to the aggregation of all methods. The decision to introduce this column – as opposed to a two-column solution – was not taken lightly. It was necessary, though, since going through an unordered list of 139 items is not practical for human experts, as it would have taken a long time depleting their concentration. For this reason, we added this shortlist in a randomized order to ease the task for the experts without introducing bias in favor of a specific data source. We chose 30 as length of this shortlist, since this is three times longer than the minimum of 10 cities that needed to be dragged to the left result column. These precautions prevent biasing the results toward a specific data model. Finally, the right column contained all remaining 109 cities in alphabetical order, to provide the experts the possibility to incorporate them into their ranking.

When the experts finished dragging at least 10 cities to the left column and indicated their familiarity with the base city, a prominent "Submit" button became available. The minimum number of 10 cities was chosen to give the partial rank agreement methods sufficient information to compute meaningful results and to limit the time it takes for the experts to complete a city ranking. It also corresponds to the reality of information retrieval or recommender systems, where only few highly relevant items are of importance. When clicking the button, the results were not yet finally submitted; instead, a modal pop-up window appeared, where the users were asked to adjust the ranking of their current shortlist: *"Please adjust the order of the cities in this list before submitting."* We decided to introduce this additional step, because when we observed test subjects in our pilot runs, it became apparent that some users simply dragged the cities into the left column without taking much care of the internal ordering of the left column. By explicitly reminding the user to revise this result, we aimed to improve the ranking, as otherwise the left column might have had set semantics instead of a ranked list semantics.

5.2. Sampling of Tourism Experts

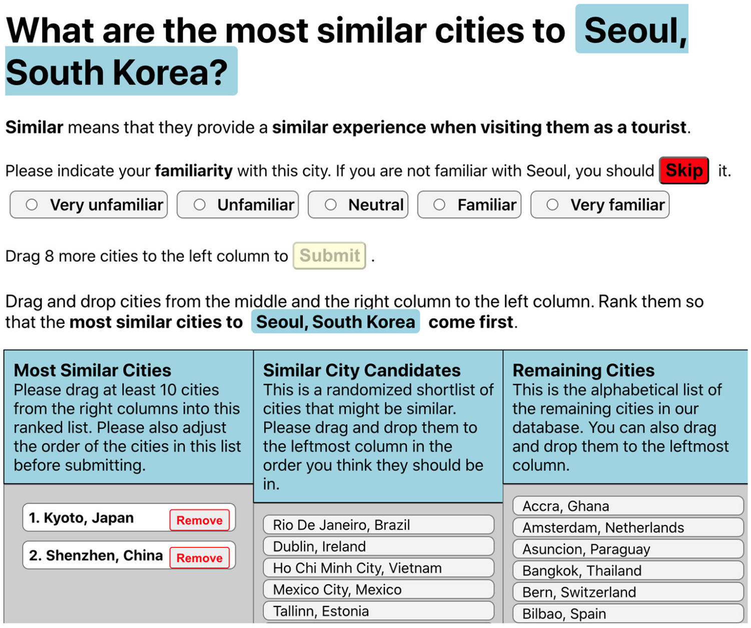
To obtain a high-quality ranking data set, we reached out to experts having relevant experience with global tourism in three ways: first, we distributed leaflets to tourism experts and researchers at the ENTER eTourism Conference held in January 2020 in Guildford, United Kingdom. Second, we directly contacted representatives and researchers on the tourism boards of the 50 cities and their respective regions. Our reasoning was that these experts in the local tourism boards know best whom they compete with, and we hope this helped to establish higher diversity of where the participants of our study originated from. This group did receive a special link to the survey, which forced them to first complete the ranking of their local city, before having the chance to rank other cities as well. Finally, we also shared the user study with the TRINET Tourism Research Information Network[11] mailing list of *accredited* members of the international tourism research and education community.

We are confident that this rigorous sampling method ensured that both the quality and the quantity of the responses are very high, despite being a web-based study conducted during the Covid-19 Pandemic.

[11] https://tim.hawaii.edu/about-values-vision-mission-accreditation/trinet/

TABLE 2 | Three expert opinions on the city of Munich are contrasted with the WP-jaccard ranked lists. The ranking of Expert 1 is closer to the ranked list than the two others.

	WP-jaccard	Expert 1	Expert 2	Expert 3
1	Vienna	Salzburg	Vienna	Frankfurt
2	Dusseldorf	Vienna	Milan	Brussels
3	Leipzig	Cologne	Dusseldorf	Heidelberg
4	Berlin	Graz	Paris	Budapest
5	Frankfurt	Milan	Boston	Hamburg
6	Heidelberg	Edinburgh	Luxembourg	Barcelona
7	Cologne	Dusseldorf	Berlin	Vienna
8	Nuremberg	Hamburg	Cologne	Prague
9	Salzburg	Amsterdam	Vancouver	Berlin
10	Copenhagen	Brussels	Dubai	Rome
	$F'(\tau, m)\,\bar{x} = 233.67$	146	292	263
	$T'(\tau, m)\,\bar{x} = 16.33$	14	15	20



What are the most similar cities to Seoul, South Korea?

Similar means that they provide a **similar experience when visiting them as a tourist**.

Please indicate your **familiarity** with this city. If you are not familiar with Seoul, you should **Skip** it.

○ Very unfamiliar ○ Unfamiliar ○ Neutral ○ Familiar ○ Very familiar

Drag 8 more cities to the left column to Submit .

Drag and drop cities from the middle and the right column to the left column. Rank them so that the **most similar cities to Seoul, South Korea come first**.

Most Similar Cities
Please drag at least 10 cities from the right columns into this ranked list. Please also adjust the order of the cities in this list before submitting.

1. Kyoto, Japan Remove
2. Shenzhen, China Remove

Similar City Candidates
This is a randomized shortlist of cities that might be similar. Please drag and drop them to the leftmost column in the order you think they should be in.

Rio De Janeiro, Brazil
Dublin, Ireland
Ho Chi Minh City, Vietnam
Mexico City, Mexico
Tallinn, Estonia

Remaining Cities
This is the alphabetical list of the remaining cities in our database. You can also drag and drop them to the leftmost column.

Accra, Ghana
Amsterdam, Netherlands
Asuncion, Paraguay
Bangkok, Thailand
Bern, Switzerland
Bilbao, Spain

FIGURE 3 | User interface of the expert survey.

5.3. Data Preparation and Cleaning

In total, we received 164 destination rankings from the survey. Since it was a web study, we took the following precautions to protect the data quality against potential low-effort submissions: We excluded responses that were completed in shorter time than 1 min and all those who did not adjust the internal ordering within the results column at all. Furthermore, we removed responses where the experts indicated their familiarity with the city on the “Very unfamiliar” or “Unfamiliar” levels. Looking at the number of completed rankings by destination, we have 28 cities with at least two submissions. Excluding the rankings of cities that were only ranked once, the results in Section 6.2

are based on the final 88 rankings of 28 cities coming from 37 different IP addresses. The characterizations done by the experts are available in the supplementary material.

The median time the experts needed to rank one city was 3m 14s and the number of re-rankings in the left results column had a median value of 4. Most submissions (74) comprised the minimum number of 10 most similar cities; eleven characterized 11–12, and the remaining three rankings were of length 13–14. The top five most characterized cities were London, UK; New York City, NY, USA; Miami, FL, USA; Barcelona, Spain; and Nice, France.

5.4. Example

To concertize the approach, we show the experts' rankings for the city of Munich in **Table 2**. The first column shows the first 10 cities of the Wikipedia-jaccard list. To obtain the score of a data model with respect to the expert's opinion, we compute the two modified rank agreement metrics between the ranked list and the experts' partial rankings. The overall score is the mean value of the rank agreement metric of all experts and all cities. The lower part of the table shows individual values and the aggregation: In this example, the opinion of Expert 1 is quite close to the ranked list according to both metrics. According to the modified Kendall's Tau, Expert 2's ranking is closer than Expert 3, however the modified Footrule distance is lower for Expert 3. This is due to the potentially exotic choices of Expert 2 to include Dubai (rank 48 in the ranked list), Vancouver (rank 54), and Boston (rank 84), which are heavily penalized in Footrule distance.

The final score of a data model according to one of the metrics is computed by the mean value of all expert rankings over all cities.

5.5. Expert Ranking Behavior

To provide some insights into the expert opinions, we first tabulate the number of cities that came from the 30 destination shortlist against the cities that were in the right column of **Figure 3**. Overall, the expert rankings comprised 80% of cities from the shortlist, whereas they still included 20% from the arguably more arduous longer list of 109 alphabetically sorted items. We see this as a confirmation that the recruited experts were serious about their task and did not only follow the ranking provided by the shortlist. Nevertheless, the shortlist might still have influenced the reviewers in a way that we cannot quantify using this study design.

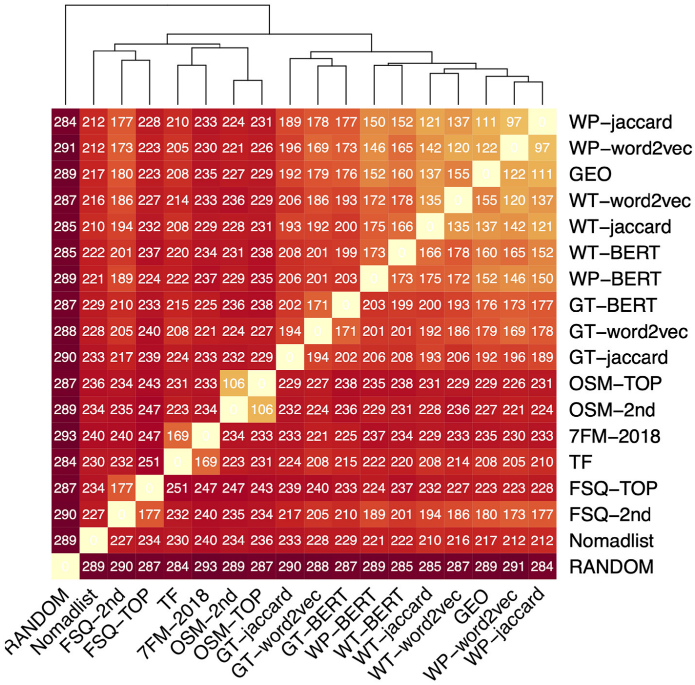
In the right column of **Table 3**, we quantify the level of agreement among the experts. Since this ranking task is different from traditional rating data, where the agreement could be quantified using metrics such as Fleiss' kappa (Fleiss, 1971), we use a set-theoretic measure to quantify the agreement of the experts for each city. We compute the agreement as the pairwise size of the intersection over the union of two annotators. The reported number is the mean value over all pairs to make results of cities with a different number of annotators comparable. The agreement ranges between 11% in the case of Brussels, Mumbai, and Osaka, while it reaches up to 54% in the case of San Diego. On average, the experts' lists had an overlap of about 25%, which we consider as quite good, given that they chose at most 14 out of 139 other cities. Agreeing on about one-fourth of the most similar destinations both shows that there is clear common ground among the experts, but also that an intangible concept such as the touristic experience cannot be determined in a purely objective way. Finally, it should be noted that our proposed rank agreement metrics deal well with potentially diverging opinions about a concept.

TABLE 3 | Expert annotators behavior: amount of cities selected from the shortlist vs. the full alphabetical list and percentages of the same cities selected.

City name	Alphabetical list %	Shortlist %	Expert agreement %
Amsterdam	20.00	80.00	17.65
Bangkok	30.43	69.57	27.78
Barcelona	40.00	60.00	12.57
Berlin	0.00	100.00	36.51
Brussels	55.00	45.00	11.11
Chicago	30.00	70.00	17.65
Copenhagen	22.73	77.27	15.79
Hamburg	23.33	76.67	27.78
Hong Kong	0.00	100.00	40.00
London	8.06	91.94	21.92
Madrid	18.00	82.00	29.50
Miami	48.00	52.00	26.59
Moscow	10.00	90.00	25.00
Mumbai	20.00	80.00	11.11
Munich	23.33	76.67	17.92
New York City	13.33	86.67	24.39
Nice	28.85	71.15	14.41
Osaka	10.00	90.00	11.11
Oslo	0.00	100.00	37.50
Paris	9.38	90.62	31.02
Rome	20.00	80.00	21.69
Saint Petersburg	43.33	56.67	17.92
San Diego	35.00	65.00	53.85
Seville	3.12	96.88	46.98
Singapore	11.63	88.37	30.72
Stockholm	12.50	87.50	32.83
Vancouver	10.00	90.00	26.32
Vienna	18.92	81.08	37.16
Overall	20.18	79.82	25.88

6. RESULTS

We evaluate our work in three ways: first, an exploratory approach using pairwise comparisons of the ranked lists to capture commonalities between them. Second, the comparison of each individual data model against the top-k lists that encode the expert-elicited concept, and, finally, the results of black-box optimization of selected data sources against the expert-elicited concept.



	RANDOM	Nomadlist	FSQ–2nd	FSQ–TOP	TF	7FM–2018	OSM–2nd	OSM–TOP	GT–jaccard	GT–word2vec	GT–BERT	WP–BERT	WT–BERT	WT–jaccard	WT–word2vec	GEO	WP–word2vec	WP–jaccard
WP–jaccard	284	212	177	228	210	233	224	231	189	178	177	150	152	121	137	111	97	0
WP–word2vec	291	212	173	223	205	230	221	226	196	169	173	146	165	142	120	122	0	97
GEO	289	217	180	223	208	235	227	229	192	179	176	152	160	137	155	0	122	111
WT–word2vec	287	216	186	227	214	233	236	229	206	186	193	172	178	135	0	155	120	137
WT–jaccard	285	210	194	232	208	229	228	231	193	192	200	175	166	0	135	137	142	121
WT–BERT	285	222	201	237	220	234	231	238	208	201	199	173	0	166	178	160	165	152
WP–BERT	289	221	189	224	222	237	229	235	206	201	203	0	173	175	172	152	146	150
GT–BERT	287	229	210	233	215	225	236	238	202	171	0	203	199	200	193	176	173	177
GT–word2vec	288	228	205	240	208	221	224	227	194	0	171	201	201	192	186	179	169	178
GT–jaccard	290	233	217	239	224	233	232	229	0	194	202	206	208	193	206	192	196	189
OSM–TOP	287	236	234	243	231	233	106	0	229	227	238	235	238	231	229	229	226	231
OSM–2nd	289	234	235	247	223	234	0	106	232	224	236	229	231	228	236	227	221	224
7FM–2018	293	240	240	247	169	0	234	233	233	221	225	237	234	229	233	235	230	233
TF	284	230	232	251	0	169	223	231	224	208	215	222	220	208	214	208	205	210
FSQ–TOP	287	234	177	0	251	247	247	243	239	240	233	224	237	232	227	223	223	228
FSQ–2nd	290	227	0	177	232	240	235	234	217	205	210	189	201	194	186	180	173	177
Nomadlist	289	0	227	234	230	240	234	236	233	228	229	221	222	210	216	217	212	212
RANDOM	0	289	290	287	284	293	289	287	290	288	287	289	285	285	287	289	291	284

FIGURE 4 | Crosswise-analysis of all data models using Kendall's Tau method. The entries are sorted using hierarchical clustering; the dendrogram reveals families of data sources. The colors are scaled according to Kendall's Tau with the bright yellow corresponding to 0 (the diagonal) and dark red representing no correlation (Random).

6.1. Assessing the Similarity of Data Sources

To reveal correlations, we compare our data sources for each city against each other using the rank agreement metrics for full permutations. In this particular case, we chose Kendall's Tau, but we could have likewise chosen Spearman's Footrule distance, which gives a similar picture. The heatmap in **Figure 4** visualizes the mean pairwise distances among all ranked lists derived from the data models. The sort order was adjusted using hierarchical clustering using the Euclidean Distance, which is also the basis of the dendrogram on the top. The values in the cells are the Kendall's Tau distance, rounded to integers.

We now describe the resulting clusters. RANDOM is clearly separated from all other data models, since it has no correlation to any of them. The first group is the family of textual data sources together with GEO. The respective ranked lists (Jaccard, Word2vec, and BERT) are based on the same data source and closest to each other. Within this family, one can also see that those based on Google Travel are a bit further away from the remaining ones. We attribute the very close grouping of GEO with the ones from Wikipedia and Wikitravel to the amount of geographic information that is encoded within the articles describing the cities. Nomadlist seems to be unrelated to any other data source in particular, but unlike RANDOM still has a low correlation to all other data models. The remaining three clusters are the ones of Foursquare, OSM, and the ranked lists based on the Webologen data (TF). The high agreement between TF and 7FM-2018 is interesting, because it shows that the tourism facts are still manifested in the Seven Factor Model of the destinations. This analysis is a very compact representation of the similar concepts behind the respective data sources and their instantiation. Thus, we want to outline some further observations:

It seems that the choice of document similarity, i.e., Jaccard distance vs. cosine similarity based on word vectors, is more important in the Google Travel documents than in the Wikitravel or Wikipedia, which can be attributed to the topic, but also to the length of documents. Google Travel descriptions are around 400 words compared to 7,000 in the case of Wikitravel and 8,200 in Wikipedia.

When we compare the similarities between the top-level aggregation of OSM and Foursquare to their second-level variants, the distances are quite small between the two OSM aggregations, however, relatively large between the Foursquare aggregations. Revisiting the data, this can be explained with the very huge branching factor of the Foursquare's category tree, where the four top-level features are expanded into 337 second-level features. In the case of OSM, the four top-level categories are only expanded to 14 second-level features. This makes the Foursquare data models more dissimilar to each other than the ones from OSM.

This analysis was interesting to get a broad overview of the data sources and their commonalities. The hierarchical clustering grouped the data models in well-comprehensible families; however, the pairwise comparisons also revealed that some models that one could have expected to be quite similar, such as the top-level and second-level aggregation of Foursquare, are indeed not that similar. The benefit of this analysis is that an analyst can quickly recognize whether data models capture similar concepts to make a decision if they can be interchanged in case of them being highly correlated.

6.2. Comparison With the Touristic Experience

Finally, we get to answer which characterization method would be most suited to use within a content-based information retrieval system such as a destination recommender. Having elicited the concept of *"similar experience when visiting cities as a tourist"* with the expert study, we can now compare the partial rankings of the experts with our characterization methods. We use the two proposed methods from Section 4 that compare a full permutation with a top-*k* list. Furthermore, we also tabulate the Mean Reciprocal Rank (MRR), and Precision to comparison baselines. Precision@1 was very near to 0 for all characterization methods. Note that the MRR and Precision do not capture the internal rankings provided by the experts. To compute them, we treated the rankings provided by the experts as a set and aggregated the metrics over all cities included in the lists, treating each element as an individual query.

Note that the expert study is only one way to determine such a latent concept. In other domains, there are potentially different ways to elicit a baseline, but we argue that it is commonplace that a latent concept is only partially observable with respect to the set of rated items and the list of most similar items per item.

Generally, the results in **Table 4** confirm the picture that was already painted in **Figure 4**: the versions that have shown to be similar there also rank similarly in the comparison to the expert ranking. The textual data models derived from Wikipedia, Wikitravel and Google Travel, as well as the geographic location performed best, followed by the 2nd-level aggregation of Foursquare, and the factual ones. OSM and the Foursquare top-level categories conclude the ranking with the random model unsurprisingly performing worst.

The general stability of the ranking among the rank agreement metrics is high. This should not come as a surprise, since the metrics do capture the same concept; thus, we can confirm the findings of Fagin et al. (2003) that distance measures within the same equivalence class behave similarly. The absolute values of the data sources and the random baseline are quite close in some metrics, which we attribute to the small signal-to-noise ratio in the data: the rankings have only been computed on the basis of 10 – 14 items out of 140. Comparing the results to the MRR and Precision baselines, the overall trends are also similar. We again attribute this to the low signal-to-noise ratio in the evaluation of top-k lists, however, one can already see that, for example, the geographic distance becomes less successful when the ordering of the experts' lists is not taken into account.

The fact that the most successful data models according to the rank agreement with the expert study stem from freely available textual descriptions of the destinations, as well as the geographic location, is an interesting finding. The good result of the geographic location can be explained using the intuition that nearby destinations are often within a similar culture and climate and, thus, also have a similar experience when visiting them according to our expert rankers. The articles in the Wikipedia and Wikitravel also do a good job of emulating the expert-elicited concept. Many travelers already use such sources to inform themselves about potential destinations and we attribute the consistently higher ranking of the Wikipedia over Wikitravel to the different target audiences. As a travel guide, Wikitravel is more oriented toward travelers already at the destination seeking practical travel information such as restaurant suggestions, while the Wikipedia offers a more comprehensive overview of the culture, history, and attractions of a city.

We now can also see that the differences between the two document similarities, cosine similarity based on word vectors and the Jaccard Distance, do matter with respect to the baseline. For the shorter Google Travel documents, the word embeddings outperformed their counterpart, whereas the Jaccard Distance was slightly better for the longer Wikipedia and Wikitravel texts. We attribute the lesser performance of the BERT transformer encoder architecture due to the fact that the touristic information is mostly encoded within the terms, thus, using full contextual embeddings does not benefit the performance of the characterization.

When looking at the expressiveness of the data sources, we see no connection between the amount of information that is explicitly encoded within the features of a data model and its performance. This suggests that more information is not needed to build a successful data model, but features that are of high quality with respect to the target concept. The highly successful geographic distance only consists of two floating numbers $[-180; 180]$, but of course, implicitly encodes much relevant information for travelers such as the culture and climate of a city. As we will see in the next section, this analysis can be used to improve the performance of some data models by dropping features that are not useful toward the target domain.

Why were the factual and venue-based destination characterization methods, of which some are already employed in destination recommender systems (Sertkan et al., 2019; Myftija and Dietz, 2020) outperformed? The reason lies within the very specific concept that we elicited using the expert survey. The factual and venue-based data models could not have been optimized toward the concept of "touristic experience" based on the insights from the survey, since when constructing them,

TABLE 4 | Ranking of the different data sources using the modified rank agreement methods for top-*k* lists as well as MRR and Precision.

Spearman's FR top-*k*		Kendall's Tau top-*k*		Mean Reciprocal Rank		Precision@5		Precision@10	
WP-jaccard	297.011	GEO	18.284	WP-jaccard	0.101	WP-word2vec	0.186	WP-jaccard	0.304
GEO	304.091	WP-jaccard	18.750	WP-word2vec	0.101	WT-word2vec	0.182	GEO	0.302
WP-word2vec	318.080	WP-word2vec	19.068	WT-word2vec	0.100	WP-jaccard	0.178	WT-jaccard	0.297
WT-word2vec	322.489	WT-jaccard	19.227	FSQ-2nd	0.094	GEO	0.162	WT-word2vec	0.297
WT-jaccard	330.057	WP-BERT	19.568	WP-BERT	0.093	WT-jaccard	0.161	WP-word2vec	0.291
FSQ-2nd	330.420	GT-word2vec	19.852	GEO	0.093	FSQ-2nd	0.154	WP-BERT	0.279
WP-BERT	343.307	WT-word2vec	20.011	WT-jaccard	0.093	WP-BERT	0.147	FSQ-2nd	0.264
GT-word2vec	346.955	FSQ-2nd	20.625	GT-word2vec	0.088	GT-word2vec	0.139	GT-word2vec	0.263
TF	395.841	WT-BERT	20.818	WT-BERT	0.081	WT-BERT	0.139	WT-BERT	0.243
GT-BERT	396.375	TF	21.409	TF	0.075	TF	0.113	TF	0.231
WT-BERT	402.159	GT-BERT	21.477	GT-BERT	0.074	GT-BERT	0.103	GT-BERT	0.202
GT-jaccard	408.943	GT-jaccard	21.864	GT-jaccard	0.067	OSM-2nd	0.099	OSM-2nd	0.195
7FM-2018	457.909	OSM-2nd	22.375	7FM-2018	0.065	Nomadlist	0.092	GT-jaccard	0.187
Nomadlist	461.830	Nomadlist	22.420	OSM-2nd	0.065	OSM-TOP	0.090	7FM-2018	0.187
FSQ-TOP	506.500	FSQ-TOP	22.864	Nomadlist	0.063	GT-jaccard	0.087	OSM-TOP	0.180
OSM-TOP	516.114	7FM-2018	22.966	OSM-TOP	0.063	7FM-2018	0.086	Nomadlist	0.169
OSM-2nd	521.273	OSM-TOP	23.045	FSQ-TOP	0.054	FSQ-TOP	0.060	FSQ-TOP	0.122
RANDOM	649.398	RANDOM	23.341	RANDOM	0.039	RANDOM	0.033	RANDOM	0.073

the respective authors had no instantiation of the concept available or potentially decided to optimize toward a different concept. For example, the Nomad List characterization is aimed at digital nomads instead of tourists. Thus, it would have been be surprising if it was in a front runner position, as digital nomads have different information needs than a typical tourist does. For the same reason, the OSM performed quite poorly. Instead of the touristic experience, they simply captured the similarity of the distribution of the different map entities. On the contrary, the textual data sources are there to learn about the characteristics of a city, so it is not surprising that they encode the most useful information for travelers. This means that the proposed methods are able to discriminate between similar and somewhat orthogonal concepts and do so by quantifying the distance. Since some features of Nomad List, OSM, and Foursquare were not aimed at encoding the same concept that we have elicited in the expert study, it is just natural that these data sources perform underwhelmingly in our initial comparison. Our methods reveal the degree to which the expert-elicited concept is not (well) encoded within the features, but since the characterizations are somewhat related to traveling, they are not orthogonal.

To summarize, the proposed rank agreement metrics for top-*k* lists have been successfully employed in determining the quality of the data sources with respect to the expert-elicited concept. They produce comparable rankings as established information retrieval metrics, such as MRR and Precision. The advantage is that rank agreement metrics operate on ranked lists instead on sets, making them conceptually more fitting than MRR, and Precision or similar metrics.

6.3. Optimization of Data Models

With this tooling established, there is now potential to refine existing data models based on tourism facts and the venue distributions by learning the importance of the respective features or even constructing a composite data model with features from different data sources. By assigning different weights to the features based on their importance in computing similarity metrics, rich models with several features can be fine-tuned toward the expert-elicited concept. This is useful, since standard similarity metrics in content-based recommendation, such as the Euclidean Distance give same weight to all features. In practice, however, not all features equally contribute to the expert-elicited concept of the touristic experience. By decreasing the weights of less-relevant features, the similarity metric can be improved to emulate the expert concept even better.

TABLE 5 | Optimization toward the Expert Opinion using Spearman's Footrule top-*k*.

Model	Unoptimized	Optimized	Improvement
Nomadlist	461.83	426.27	7.70%
FSQ-TOP	506.50	503.33	0.63%
FSQ-2nd	330.42	312.47	5.43%
OSM-TOP	516.11	508.38	1.50%
OSM-2nd	521.27	490.33	5.94%

Given the combinatorial explosion of the search space for weights, we have used black-box learning, namely Simulated Annealing (Kirkpatrick et al., 1983) for tuning the weights [0,1] of the data sources with explicit features. The proprietary TF and 7FM-2018 sources were only provided to us as rankings, thus, we could not optimize those.

The optimization tabulated in **Table 5** works better with more features, as can be seen with Nomadlist, FSQ-2nd, and OSM-2nd.

We attribute the small relative changes in FSQ-TOP and OSM-TOP to the fact that they capture slightly orthogonal concepts to the expert-elicited baseline and due to their smaller number of features, they are harder to optimize toward this concept. However, as discussed before, this domain has a high signal-to-noise ratio, making these small relative improvements relevant in the overall comparison. Concretely, the optimized version of FSQ-2nd would be the third most competitive data model in **Table 4**.

What is more, even with minor contributions to the overall performance, the learned weights for the features gives further insight into their importance. Features with a very low weight could be dropped, while the feature selection of a potential combined data model of several data sources should be guided by the learned weights.

To exemplify the insights from this analysis, we discuss the learned weights of the Nomadslist data: The features "cost," "life quality," "air quality," and "happiness" got relatively high values ranging from 0.58 to 0.75, while the other features, "nomad score," "fun," "healthcare," and "nightlife" were reduced to low weights ranging between 0.2 and 0.35. Such low values indicating that they are not in line with the elicited concept of the expert study. In FSQ-TOP, which is employed in the CityRec system (Dietz et al., 2019), "food" gets a very low weight, which is an indication that it could be dropped from the recommendation algorithm.

The results of this in-depth analysis of the weights are certainly quite specialized with respect to the target concept and the intricacies of the respective data sources. For this reason, we do not further elaborate on the other optimized models but refer the reader to the full results of this optimization tabulated in the reproducibility material. The method, however, is again generalizable for any domain, where the data source's features are known and a baseline exists in the form of ranked lists.

7. CONCLUSIONS

We presented a comprehensive overview of data-driven methods to characterize cities at scale using online data. Motivated by the question of model choice in destination recommender systems, we proposed methods to make such data models of destinations comparable against each other as well as against a – potentially latent – concept, that the recommender system should emulate when computing content-based recommendations. To derive this concept, we conducted an expert study that provided us with partial rankings and provided us opportunity to further optimize the data models that are based on explicit features. The decision of eliciting this baseline using experts instead of large-scale crowd-sourcing was done due to the difficulty of the task. Since this is the first study analyzing latent concepts encoded in features of content-based recommender systems, we decided to elicit a high-quality data set with less noise, than having a large-scale data set that is less to be trusted.

In a first step, we were able to unveil commonalities of data sources, through which it became apparent that, for example, articles about destinations on Wikipedia and Wikitravel encode much geographic information. The second contribution are methods to compare top-*k* lists with permutations in our specific scenario. We used these to show that, according to the expert opinion, the touristic experience was best approximated using the textual similarities from Wikipedia, Wikitravel, and the geographic location. This means that when simply retrieving the most similar destinations according to the touristic experience, one can choose one of the top-ranked entries from **Table 4**. Finally, we were able to show that it is possible to optimize the distance metric of a content-based recommender system toward a desired concept.

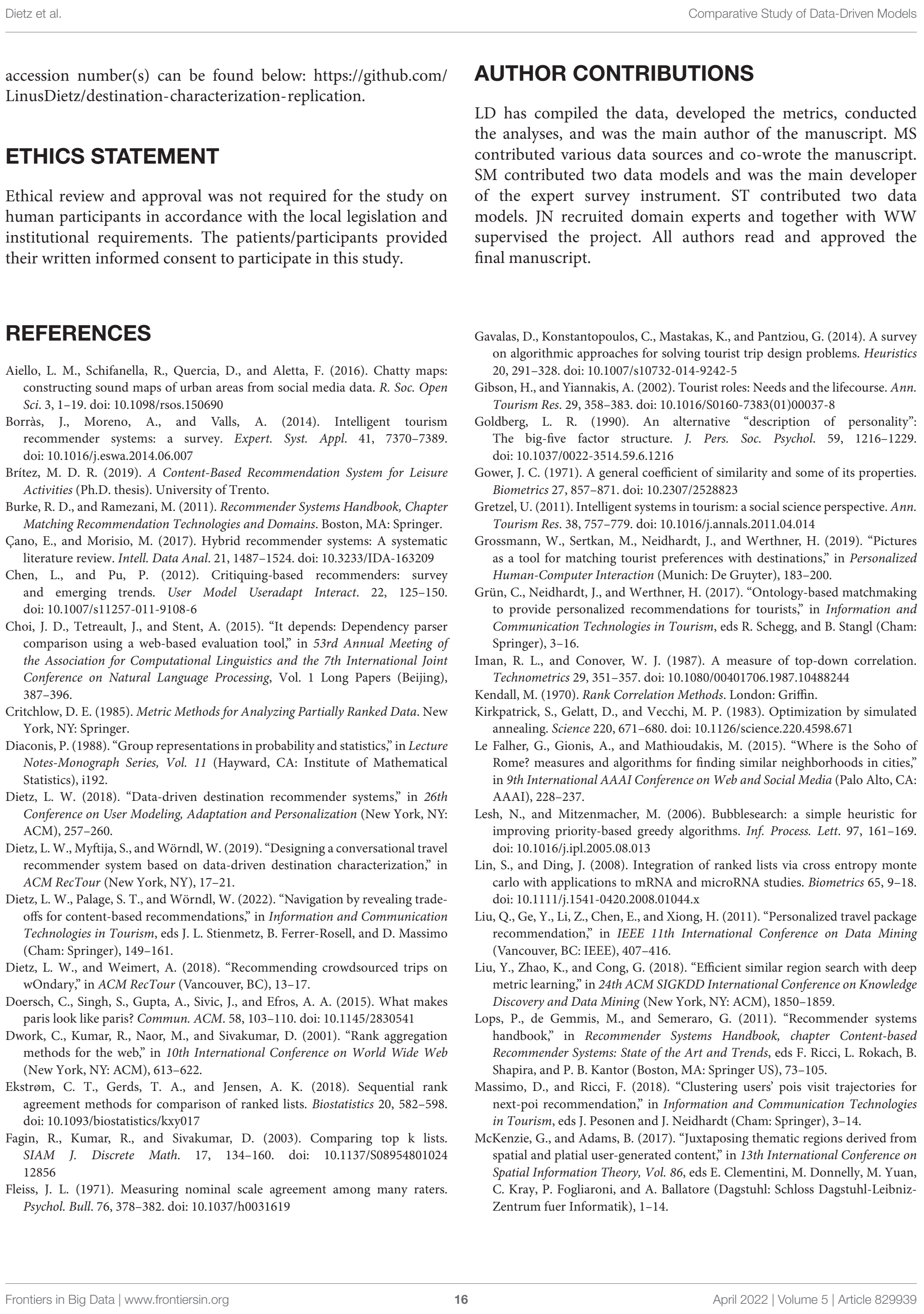
From a recommender systems research perspective, the results show that existing destination recommender systems do not necessarily use data models that capture the concept of similar touristic experience very well. This might be intentional, if the system's purpose is to capture a different concept, or possibly due to the previous lack of a concrete instantiation of the concept. A limitation of the top-ranked textual or geographic characterizations is that they do not come with specific features the user can interact with. This is a drawback, since it means that they cannot directly place the user's preferences and the items in a common vector space to perform content-based recommendation as frequently done in travel recommender systems (Burke and Ramezani, 2011). Furthermore, common recommendation techniques such as critiquing (Chen and Pu, 2012), i.e., giving a system feedback about the features of a suggested item, are only possible if the items are characterized with a fixed number of features.

Our work has provided the community with adequate tools to optimize feature-based data models toward a desired concept such as the similar touristic experience. The methodological contribution, is, however, not limited to recommender systems in the tourism domain, but can be applied in other domains similarly as the proposed metrics operate on ranked lists. Latent similarity concepts are prevalent in many domains such as music (Yoshii et al., 2006) or leisure activities (Brítez, 2019); generally anywhere, where the accuracy of the information retrieval system depends on the embeddings of items in a search space.

A logical continuation of this work would be to investigate the potential to construct better, potentially combined data models. This research can help to improve all kinds of data-driven characterizations of travel destinations as it provides direct feedback about the data quality with respect to the touristic experience. While this time we used an expert study, we also plan to apply these methods in other domains in a large-scale crowd-sourcing setting. Finally, it would be worthwhile to perform an analysis of the effect of improved data model quality with respect to further evaluation metrics such as accuracy.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and



accession number(s) can be found below: https://github.com/LinusDietz/destination-characterization-replication.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LD has compiled the data, developed the metrics, conducted the analyses, and was the main author of the manuscript. MS contributed various data sources and co-wrote the manuscript. SM contributed two data models and was the main developer of the expert survey instrument. ST contributed two data models. JN recruited domain experts and together with WW supervised the project. All authors read and approved the final manuscript.

REFERENCES

Aiello, L. M., Schifanella, R., Quercia, D., and Aletta, F. (2016). Chatty maps: constructing sound maps of urban areas from social media data. *R. Soc. Open Sci.* 3, 1–19. doi: 10.1098/rsos.150690

Borràs, J., Moreno, A., and Valls, A. (2014). Intelligent tourism recommender systems: a survey. *Expert. Syst. Appl.* 41, 7370–7389. doi: 10.1016/j.eswa.2014.06.007

Brítez, M. D. R. (2019). *A Content-Based Recommendation System for Leisure Activities* (Ph.D. thesis). University of Trento.

Burke, R. D., and Ramezani, M. (2011). *Recommender Systems Handbook, Chapter Matching Recommendation Technologies and Domains*. Boston, MA: Springer.

Çano, E., and Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intell. Data Anal.* 21, 1487–1524. doi: 10.3233/IDA-163209

Chen, L., and Pu, P. (2012). Critiquing-based recommenders: survey and emerging trends. *User Model Useradapt Interact.* 22, 125–150. doi: 10.1007/s11257-011-9108-6

Choi, J. D., Tetreault, J., and Stent, A. (2015). “It depends: Dependency parser comparison using a web-based evaluation tool,” in *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 1 Long Papers (Beijing), 387–396.

Critchlow, D. E. (1985). *Metric Methods for Analyzing Partially Ranked Data*. New York, NY: Springer.

Diaconis, P. (1988). “Group representations in probability and statistics,” in *Lecture Notes-Monograph Series, Vol. 11* (Hayward, CA: Institute of Mathematical Statistics), i192.

Dietz, L. W. (2018). “Data-driven destination recommender systems,” in *26th Conference on User Modeling, Adaptation and Personalization* (New York, NY: ACM), 257–260.

Dietz, L. W., Myftija, S., and Wörndl, W. (2019). “Designing a conversational travel recommender system based on data-driven destination characterization,” in *ACM RecTour* (New York, NY), 17–21.

Dietz, L. W., Palage, S. T., and Wörndl, W. (2022). “Navigation by revealing trade-offs for content-based recommendations,” in *Information and Communication Technologies in Tourism*, eds J. L. Stienmetz, B. Ferrer-Rosell, and D. Massimo (Cham: Springer), 149–161.

Dietz, L. W., and Weimert, A. (2018). “Recommending crowdsourced trips on wOndary,” in *ACM RecTour* (Vancouver, BC), 13–17.

Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. A. (2015). What makes paris look like paris? *Commun. ACM.* 58, 103–110. doi: 10.1145/2830541

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). “Rank aggregation methods for the web,” in *10th International Conference on World Wide Web* (New York, NY: ACM), 613–622.

Ekstrøm, C. T., Gerds, T. A., and Jensen, A. K. (2018). Sequential rank agreement methods for comparison of ranked lists. *Biostatistics* 20, 582–598. doi: 10.1093/biostatistics/kxy017

Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. *SIAM J. Discrete Math.* 17, 134–160. doi: 10.1137/S0895480102412856

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382. doi: 10.1037/h0031619

Gavalas, D., Konstantopoulos, C., Mastakas, K., and Pantziou, G. (2014). A survey on algorithmic approaches for solving tourist trip design problems. *Heuristics* 20, 291–328. doi: 10.1007/s10732-014-9242-5

Gibson, H., and Yiannakis, A. (2002). Tourist roles: Needs and the lifecourse. *Ann. Tourism Res.* 29, 358–383. doi: 10.1016/S0160-7383(01)00037-8

Goldberg, L. R. (1990). An alternative “description of personality”: The big-five factor structure. *J. Pers. Soc. Psychol.* 59, 1216–1229. doi: 10.1037/0022-3514.59.6.1216

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27, 857–871. doi: 10.2307/2528823

Gretzel, U. (2011). Intelligent systems in tourism: a social science perspective. *Ann. Tourism Res.* 38, 757–779. doi: 10.1016/j.annals.2011.04.014

Grossmann, W., Sertkan, M., Neidhardt, J., and Werthner, H. (2019). “Pictures as a tool for matching tourist preferences with destinations,” in *Personalized Human-Computer Interaction* (Munich: De Gruyter), 183–200.

Grün, C., Neidhardt, J., and Werthner, H. (2017). “Ontology-based matchmaking to provide personalized recommendations for tourists,” in *Information and Communication Technologies in Tourism*, eds R. Schegg, and B. Stangl (Cham: Springer), 3–16.

Iman, R. L., and Conover, W. J. (1987). A measure of top-down correlation. *Technometrics* 29, 351–357. doi: 10.1080/00401706.1987.10488244

Kendall, M. (1970). *Rank Correlation Methods*. London: Griffin.

Kirkpatrick, S., Gelatt, D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680. doi: 10.1126/science.220.4598.671

Le Falher, G., Gionis, A., and Mathioudakis, M. (2015). “Where is the Soho of Rome? measures and algorithms for finding similar neighborhoods in cities,” in *9th International AAAI Conference on Web and Social Media* (Palo Alto, CA: AAAI), 228–237.

Lesh, N., and Mitzenmacher, M. (2006). Bubblesearch: a simple heuristic for improving priority-based greedy algorithms. *Inf. Process. Lett.* 97, 161–169. doi: 10.1016/j.ipl.2005.08.013

Lin, S., and Ding, J. (2008). Integration of ranked lists via cross entropy monte carlo with applications to mRNA and microRNA studies. *Biometrics* 65, 9–18. doi: 10.1111/j.1541-0420.2008.01044.x

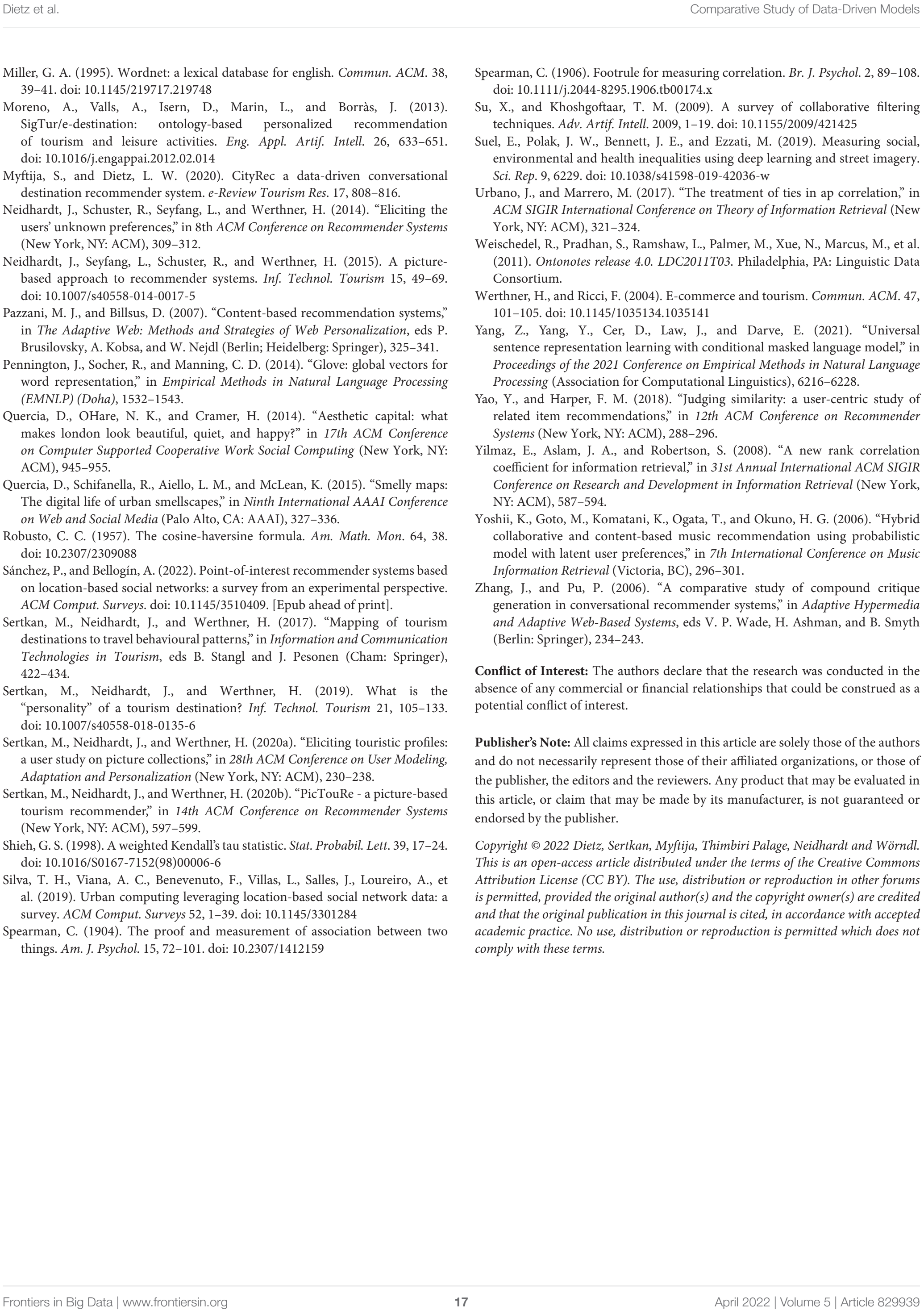
Liu, Q., Ge, Y., Li, Z., Chen, E., and Xiong, H. (2011). “Personalized travel package recommendation,” in *IEEE 11th International Conference on Data Mining* (Vancouver, BC: IEEE), 407–416.

Liu, Y., Zhao, K., and Cong, G. (2018). “Efficient similar region search with deep metric learning,” in *24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 1850–1859.

Lops, P., de Gemmis, M., and Semeraro, G. (2011). “Recommender systems handbook,” in *Recommender Systems Handbook, chapter Content-based Recommender Systems: State of the Art and Trends*, eds F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor (Boston, MA: Springer US), 73–105.

Massimo, D., and Ricci, F. (2018). “Clustering users' pois visit trajectories for next-poi recommendation,” in *Information and Communication Technologies in Tourism*, eds J. Pesonen and J. Neidhardt (Cham: Springer), 3–14.

McKenzie, G., and Adams, B. (2017). “Juxtaposing thematic regions derived from spatial and platial user-generated content,” in *13th International Conference on Spatial Information Theory, Vol. 86*, eds E. Clementini, M. Donnelly, M. Yuan, C. Kray, P. Fogliaroni, and A. Ballatore (Dagstuhl: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik), 1–14.



Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM*. 38, 39–41. doi: 10.1145/219717.219748

Moreno, A., Valls, A., Isern, D., Marin, L., and Borràs, J. (2013). SigTur/e-destination: ontology-based personalized recommendation of tourism and leisure activities. *Eng. Appl. Artif. Intell*. 26, 633–651. doi: 10.1016/j.engappai.2012.02.014

Myftija, S., and Dietz, L. W. (2020). CityRec a data-driven conversational destination recommender system. *e-Review Tourism Res*. 17, 808–816.

Neidhardt, J., Schuster, R., Seyfang, L., and Werthner, H. (2014). "Eliciting the users' unknown preferences," in 8th *ACM Conference on Recommender Systems* (New York, NY: ACM), 309–312.

Neidhardt, J., Seyfang, L., Schuster, R., and Werthner, H. (2015). A picture-based approach to recommender systems. *Inf. Technol. Tourism* 15, 49–69. doi: 10.1007/s40558-014-0017-5

Pazzani, M. J., and Billsus, D. (2007). "Content-based recommendation systems," in *The Adaptive Web: Methods and Strategies of Web Personalization*, eds P. Brusilovsky, A. Kobsa, and W. Nejdl (Berlin; Heidelberg: Springer), 325–341.

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP) (Doha)*, 1532–1543.

Quercia, D., OHare, N. K., and Cramer, H. (2014). "Aesthetic capital: what makes london look beautiful, quiet, and happy?" in *17th ACM Conference on Computer Supported Cooperative Work Social Computing* (New York, NY: ACM), 945–955.

Quercia, D., Schifanella, R., Aiello, L. M., and McLean, K. (2015). "Smelly maps: The digital life of urban smellscapes," in *Ninth International AAAI Conference on Web and Social Media* (Palo Alto, CA: AAAI), 327–336.

Robusto, C. C. (1957). The cosine-haversine formula. *Am. Math. Mon*. 64, 38. doi: 10.2307/2309088

Sánchez, P., and Bellogín, A. (2022). Point-of-interest recommender systems based on location-based social networks: a survey from an experimental perspective. *ACM Comput. Surveys*. doi: 10.1145/3510409. [Epub ahead of print].

Sertkan, M., Neidhardt, J., and Werthner, H. (2017). "Mapping of tourism destinations to travel behavioural patterns," in *Information and Communication Technologies in Tourism*, eds B. Stangl and J. Pesonen (Cham: Springer), 422–434.

Sertkan, M., Neidhardt, J., and Werthner, H. (2019). What is the "personality" of a tourism destination? *Inf. Technol. Tourism* 21, 105–133. doi: 10.1007/s40558-018-0135-6

Sertkan, M., Neidhardt, J., and Werthner, H. (2020a). "Eliciting touristic profiles: a user study on picture collections," in *28th ACM Conference on User Modeling, Adaptation and Personalization* (New York, NY: ACM), 230–238.

Sertkan, M., Neidhardt, J., and Werthner, H. (2020b). "PicTouRe - a picture-based tourism recommender," in *14th ACM Conference on Recommender Systems* (New York, NY: ACM), 597–599.

Shieh, G. S. (1998). A weighted Kendall's tau statistic. *Stat. Probabil. Lett*. 39, 17–24. doi: 10.1016/S0167-7152(98)00006-6

Silva, T. H., Viana, A. C., Benevenuto, F., Villas, L., Salles, J., Loureiro, A., et al. (2019). Urban computing leveraging location-based social network data: a survey. *ACM Comput. Surveys* 52, 1–39. doi: 10.1145/3301284

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol*. 15, 72–101. doi: 10.2307/1412159

Spearman, C. (1906). Footrule for measuring correlation. *Br. J. Psychol*. 2, 89–108. doi: 10.1111/j.2044-8295.1906.tb00174.x

Su, X., and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Adv. Artif. Intell*. 2009, 1–19. doi: 10.1155/2009/421425

Suel, E., Polak, J. W., Bennett, J. E., and Ezzati, M. (2019). Measuring social, environmental and health inequalities using deep learning and street imagery. *Sci. Rep*. 9, 6229. doi: 10.1038/s41598-019-42036-w

Urbano, J., and Marrero, M. (2017). "The treatment of ties in ap correlation," in *ACM SIGIR International Conference on Theory of Information Retrieval* (New York, NY: ACM), 321–324.

Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., et al. (2011). *Ontonotes release 4.0. LDC2011T03*. Philadelphia, PA: Linguistic Data Consortium.

Werthner, H., and Ricci, F. (2004). E-commerce and tourism. *Commun. ACM*. 47, 101–105. doi: 10.1145/1035134.1035141

Yang, Z., Yang, Y., Cer, D., Law, J., and Darve, E. (2021). "Universal sentence representation learning with conditional masked language model," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics), 6216–6228.

Yao, Y., and Harper, F. M. (2018). "Judging similarity: a user-centric study of related item recommendations," in *12th ACM Conference on Recommender Systems* (New York, NY: ACM), 288–296.

Yilmaz, E., Aslam, J. A., and Robertson, S. (2008). "A new rank correlation coefficient for information retrieval," in *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY: ACM), 587–594.

Yoshii, K., Goto, M., Komatani, K., Ogata, T., and Okuno, H. G. (2006). "Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences," in *7th International Conference on Music Information Retrieval* (Victoria, BC), 296–301.

Zhang, J., and Pu, P. (2006). "A comparative study of compound critique generation in conversational recommender systems," in *Adaptive Hypermedia and Adaptive Web-Based Systems*, eds V. P. Wade, H. Ashman, and B. Smyth (Berlin: Springer), 234–243.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

A.5 Publication 5: "Designing a Conversational Travel Recommender System Based on Data-driven Destination Characterization"

Linus W. Dietz, Saadi Myftija, and Wolfgang Wörndl. "Designing a Conversational Travel Recommender System Based on Data-driven Destination Characterization." In: *ACM RecSys Workshop on Recommenders in Tourism*. Sept. 2019, pp. 17–21

Summary

Recommending complex, intangible items in a domain with high consequences, such as destinations for traveling, requires additional care when deriving and confronting the users with recommendations. To address these challenges, we developed a first version of the CityRec system, a destination recommender system that makes two contributions. The first is a data-driven approach to characterize cities according to the availability of venues and travel-related features, such as the climate and travel costs. The second is a conversational recommender system using unit critiquing with 180 destinations around the globe based on the data-driven characterization, which provides prospective travelers with inspiration for and information about their next trip. An online user study with 104 participants revealed that the proposed system has a significantly higher perceived accuracy compared to the baseline approach, however, at the cost of ease of use.

Author Contributions

LD was the main author of the manuscript, designed the user study, and performed the statistical analyses. SM developed the initial prototype and conducted the user study under supervision of LD. WW co-edited the paper and supervised the project.

Designing a Conversational Travel Recommender System Based on Data-Driven Destination Characterization

Linus W. Dietz
Department of Informatics
Technical University of Munich
Garching, Germany
linus.dietz@tum.de

Saadi Myftija
Department of Informatics
Technical University of Munich
Garching, Germany
saadi.myftija@tum.de

Wolfgang Wörndl
Department of Informatics
Technical University of Munich
Garching, Germany
woerndl@in.tum.de

ABSTRACT

Recommending complex, intangible items in a domain with high consequences, such as destinations for traveling, requires additional care when deriving and confronting the users with recommendations. In order to address these challenges, we developed CityRec, a destination recommender that makes two contributions. The first is a data-driven approach to characterize cities according to the availability of venues and travel-related features, such as the climate and costs of travel. The second is a conversational recommender system with 180 destinations around the globe based on the data-driven characterization, which provides prospective travelers with inspiration for and information about their next trip. An online user study with 104 participants revealed that the proposed system has a significantly higher perceived accuracy compared to the baseline approach, however, at the cost of ease of use.

KEYWORDS

Tourism recommendation, Data mining, Cluster analysis, Conversational recommender systems

1 INTRODUCTION

In complex recommendation domains, such as the recommendation of tourist destinations, tweaking the algorithmic accuracy ad ultimo brings diminishing returns. It has been shown that the embedding of the algorithm in an adequate user interface is of similar importance [16]. Thus, in this paper, we present a data-driven conversational destination recommender system that has two contributions: it presents a novel, data-driven approach for characterizing destinations on user-understandable dimensions and shows how this characterization can be facilitated in a conversational recommender. This approach can be seen as an evolution of Burke's FindMe Approach [3] in the area of tourism. We thoroughly evaluated the system from the users' perspective to understand the effect of critiquing on the perceived accuracy of the recommendations and the satisfaction of the users from using the system.

After the literature review in the subsequent section, we will present the proposed method for characterizing destinations to realize content-based recommendations. Section 4, presents the the design and evaluation of the conversational recommender system that heavily relies on the previous characterization. We conclude our findings and point out future work in Section 5.

2 RELATED WORK

Tourism recommendation is inherently complex and has several facets. Borràs et al. enumerate four general functionalities of tourism recommender systems [2]: recommend travel destinations and tourist packs [17, 31], suggesting attractions [18], trip planners [10, 12], and social aspects [13]. In this paper, we focus on the first aspect and acknowledge that there are further definitions [1]. Herein, "destination" refers to cities. The challenge in recommending cities to a user at home arises from the intangibility of the items and the high emotional involvement [33]. It has been shown that leisure travel has a positive effect on an individual's happiness; however, it does not impact the overall life satisfaction, which has been attributed to poor tourism products [23]. An alternative conclusion could be that travelers visit the wrong places. This gives rise to researching improved destination recommender systems that can efficiently and effectively capture the user's preferences to overcome the cold start problem [5]. Given the characteristics of this domain, Burke and Ramezani suggested either the content-based [27] or the knowledge-based [3] paradigm [7].

In traditional information retrieval or static content-based recommendation, continuously querying for relevant items does not necessarily lead to better results [4]. Instead, a directed exploration of the search space using a conversational method is more promising [8, 11]. Burke et al. proposed and evaluated the FindMe approach [6], which allows the critiquing of single items so that the user can refine the recommendations iteratively until she is satisfied with the result. More advanced approaches on this topic are those of McCarthy et al., who propose a method to generate compound critiques [19], and McGinty and Smyth, who use the adaptive selection strategy to ensure diverse, yet fitting recommendations over the course of several critiquing cycles [21]. Recently, Xie et al. showed that incorporating the user experience into a critiquing system can improve the performance and recommendations at a reduced effort by the user [35]. In this study, we present a recommender system leveraging the potentials of the interplay between data science and user interface design. The items are characterized by a multidimensional space of features, which are intuitively understandable by the user and can then be critiqued in any direction. To overcome the problem of skeptical users hesitating to reveal their complete preferences [29] and the observation that users find it difficult to assess their exact preferences until when they are dealing with the actual set of offered options [26], the proposed method uses a mixture of explicit preference elicitation methods.

Using the content-based recommendation paradigm, one has to choose a domain model and distance metric to compute the most fitting items for the user. Such models can be realized through ontologies as done in SigTur [22] or in a the work of Grün et al. [14]. The latter is an example of ontologies being used to refine user profiles by enriching the generic preferences of a tourist through more specific interests. More often, items are simply characterized

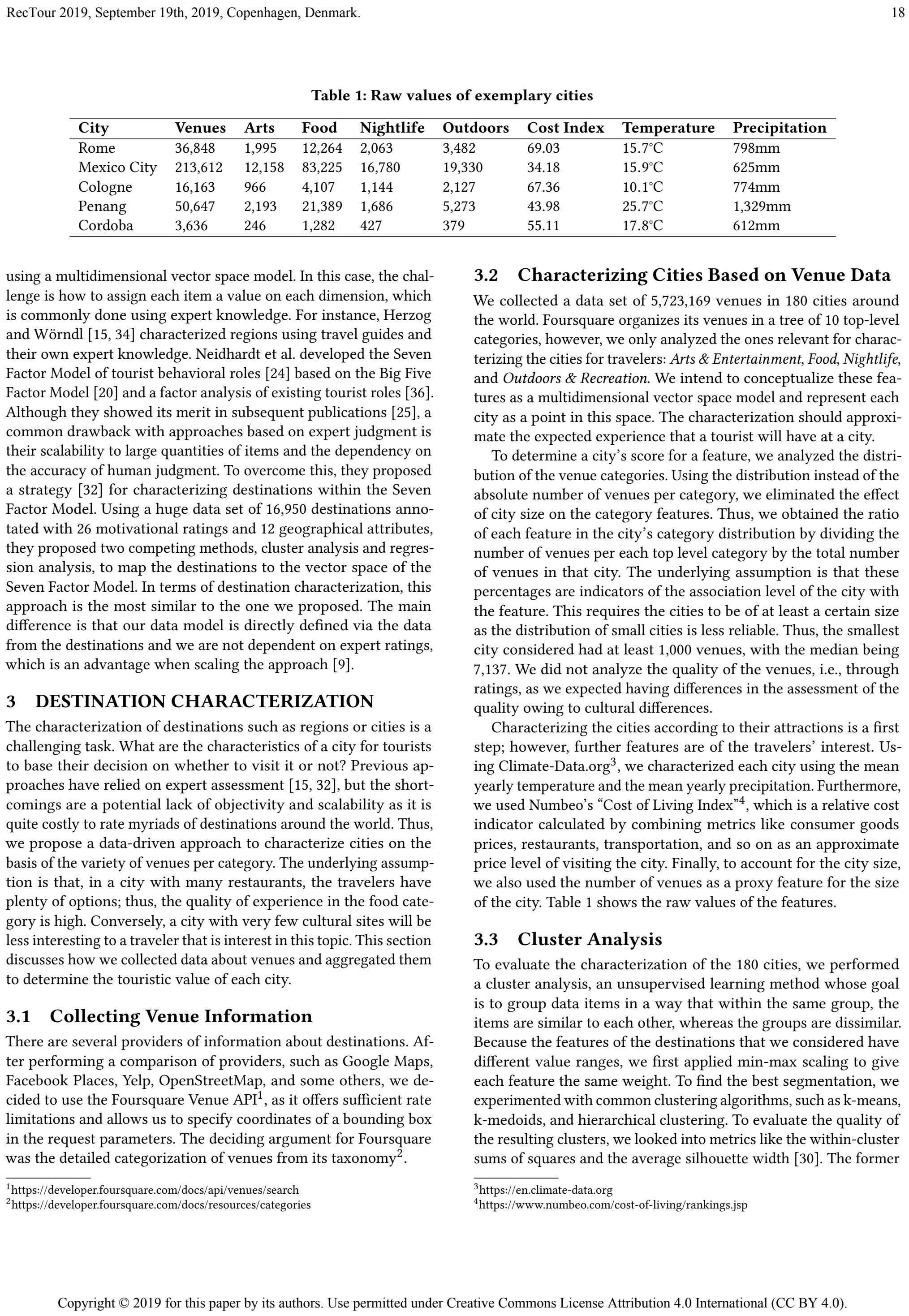


Table 1: Raw values of exemplary cities

City	Venues	Arts	Food	Nightlife	Outdoors	Cost Index	Temperature	Precipitation
Rome	36,848	1,995	12,264	2,063	3,482	69.03	15.7°C	798mm
Mexico City	213,612	12,158	83,225	16,780	19,330	34.18	15.9°C	625mm
Cologne	16,163	966	4,107	1,144	2,127	67.36	10.1°C	774mm
Penang	50,647	2,193	21,389	1,686	5,273	43.98	25.7°C	1,329mm
Cordoba	3,636	246	1,282	427	379	55.11	17.8°C	612mm

using a multidimensional vector space model. In this case, the challenge is how to assign each item a value on each dimension, which is commonly done using expert knowledge. For instance, Herzog and Wörndl [15, 34] characterized regions using travel guides and their own expert knowledge. Neidhardt et al. developed the Seven Factor Model of tourist behavioral roles [24] based on the Big Five Factor Model [20] and a factor analysis of existing tourist roles [36]. Although they showed its merit in subsequent publications [25], a common drawback with approaches based on expert judgment is their scalability to large quantities of items and the dependency on the accuracy of human judgment. To overcome this, they proposed a strategy [32] for characterizing destinations within the Seven Factor Model. Using a huge data set of 16,950 destinations annotated with 26 motivational ratings and 12 geographical attributes, they proposed two competing methods, cluster analysis and regression analysis, to map the destinations to the vector space of the Seven Factor Model. In terms of destination characterization, this approach is the most similar to the one we proposed. The main difference is that our data model is directly defined via the data from the destinations and we are not dependent on expert ratings, which is an advantage when scaling the approach [9].

3 DESTINATION CHARACTERIZATION

The characterization of destinations such as regions or cities is a challenging task. What are the characteristics of a city for tourists to base their decision on whether to visit it or not? Previous approaches have relied on expert assessment [15, 32], but the shortcomings are a potential lack of objectivity and scalability as it is quite costly to rate myriads of destinations around the world. Thus, we propose a data-driven approach to characterize cities on the basis of the variety of venues per category. The underlying assumption is that, in a city with many restaurants, the travelers have plenty of options; thus, the quality of experience in the food category is high. Conversely, a city with very few cultural sites will be less interesting to a traveler that is interest in this topic. This section discusses how we collected data about venues and aggregated them to determine the touristic value of each city.

3.1 Collecting Venue Information

There are several providers of information about destinations. After performing a comparison of providers, such as Google Maps, Facebook Places, Yelp, OpenStreetMap, and some others, we decided to use the Foursquare Venue API[1], as it offers sufficient rate limitations and allows us to specify coordinates of a bounding box in the request parameters. The deciding argument for Foursquare was the detailed categorization of venues from its taxonomy[2].

3.2 Characterizing Cities Based on Venue Data

We collected a data set of 5,723,169 venues in 180 cities around the world. Foursquare organizes its venues in a tree of 10 top-level categories, however, we only analyzed the ones relevant for characterizing the cities for travelers: *Arts & Entertainment*, *Food*, *Nightlife*, and *Outdoors & Recreation*. We intend to conceptualize these features as a multidimensional vector space model and represent each city as a point in this space. The characterization should approximate the expected experience that a tourist will have at a city.

To determine a city's score for a feature, we analyzed the distribution of the venue categories. Using the distribution instead of the absolute number of venues per category, we eliminated the effect of city size on the category features. Thus, we obtained the ratio of each feature in the city's category distribution by dividing the number of venues per each top level category by the total number of venues in that city. The underlying assumption is that these percentages are indicators of the association level of the city with the feature. This requires the cities to be of at least a certain size as the distribution of small cities is less reliable. Thus, the smallest city considered had at least 1,000 venues, with the median being 7,137. We did not analyze the quality of the venues, i.e., through ratings, as we expected having differences in the assessment of the quality owing to cultural differences.

Characterizing the cities according to their attractions is a first step; however, further features are of the travelers' interest. Using Climate-Data.org[3], we characterized each city using the mean yearly temperature and the mean yearly precipitation. Furthermore, we used Numbeo's “Cost of Living Index”[4], which is a relative cost indicator calculated by combining metrics like consumer goods prices, restaurants, transportation, and so on as an approximate price level of visiting the city. Finally, to account for the city size, we also used the number of venues as a proxy feature for the size of the city. Table 1 shows the raw values of the features.

3.3 Cluster Analysis

To evaluate the characterization of the 180 cities, we performed a cluster analysis, an unsupervised learning method whose goal is to group data items in a way that within the same group, the items are similar to each other, whereas the groups are dissimilar. Because the features of the destinations that we considered have different value ranges, we first applied min-max scaling to give each feature the same weight. To find the best segmentation, we experimented with common clustering algorithms, such as k-means, k-medoids, and hierarchical clustering. To evaluate the quality of the resulting clusters, we looked into metrics like the within-cluster sums of squares and the average silhouette width [30]. The former

[1] https://developer.foursquare.com/docs/api/venues/search

[2] https://developer.foursquare.com/docs/resources/categories

[3] https://en.climate-data.org

[4] https://www.numbeo.com/cost-of-living/rankings.jsp

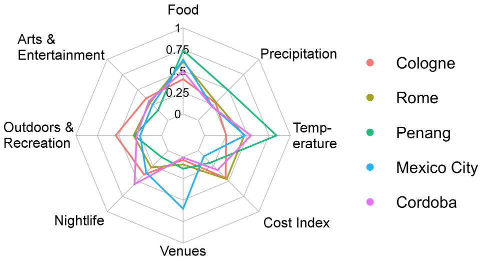


Figure 1: Normalized values of selected destinations

is a measure of the variability of the instances within each cluster, whereas the latter is a measure of how well the instances fit into their assigned cluster, as opposed to all the other clusters.

Using a systematic approach, we obtained the best results using hierarchical clustering and five clusters. The clusters named after the city closest to the centroid are "Cologne, Germany," with 74 Central European and North American cities; "Rome, Italy" with 35 cities in the Mediterranean and Oceania; "Penang, Malaysia" with 48 destinations residing mostly in Asia; "Mexico City, Mexico" with five metropoleis all around the world; and "Cordoba, Spain," with 18 small and relatively warm cities in different continents. Figure 1 shows the normalized values of the five characteristic cities.

4 A DATA-DRIVEN CONVERSATIONAL DESTINATION RECOMMENDER SYSTEM

Having characterized the destinations on eight dimensions, we facilitate it in a content-based critiquing recommender system. CityRec is implemented as a web application using NodeJS[5] and ReactJS[6] in the frontend. The codebase comprises about 3,500 lines of code and is available on Github[7]. A demo can be viewed at http://cityrec.cm.in.tum.de.

4.1 User Interaction with CityRec

The recommender system has three steps: (1) initial preference elicitation, shown in Figure 2 (a); (2) refinement through critiquing, shown in Figure 2 (b); and (3) a results page. In Step (1), we obtain the initial scores for the user profile by asking the user to select the destinations that best reflect her preferences from a set of 12 cities. We then construct an initial user model by averaging the feature values of the selected cities. This initial seed of 12 destinations is not random, but a diverse representation of the data set. We fill in the first nine slots by selecting two cities from each of the five previously established destination clusters (one in the case of the small "Mexico City" cluster). The remaining three slots are randomly selected cities to account for the size differences of the clusters. Using this approach, we can generate numerous, diverse, but equivalent shortlists because each cluster is represented. From these 12 cities, the users may choose three to five that best reflect their preferences. If a user does not recognize many cities, she can request another set of cities. Furthermore, a tooltip encourages the user to select cities that she finds generally interesting, including those she has already visited. This ensures that the system has enough data to work with for generating the initial user profile but avoids cases where users select many displayed cities, which end up in generic profiles with averaged-out feature values. The result of this step is an initial profile of the user that resides in the same vector space as the items.

In Step (2), we display a set of four initial destinations, computed using the Euclidean Distance. To give the users more control over their preference profile, we ask them to provide feedback on the initial recommendations by critiquing the cities' features one after another on a five-point Likert Scale: *"much lower" – "lower" — "just right" — "higher" — "much higher."* As can be seen in Figure 2 (b), the user now has more information about the cities, which establishes transparency and enables her to more informed decisions compared to in the first step. Using this feedback, we statically update the user profile scores by -0.2, -0.1, 0, 0.1, or 0.2 to attain a more refined preference model for the user.

Finally, in the last step, Step (3), the user is presented with a results page that shows a ranked list of the top five recommendations and their attributes, which can be explored. This page also contains the questionnaire for the evaluation.

4.2 Experimental Setup

The independent variable of the experiment is the version of the recommender system. Because we wanted to investigate the potential advantages and drawbacks of using critiquing in this domain, we created a baseline system in addition to the previously described critiquing-based recommender. The only difference in the baseline system was that the critiquing step, Step (2), is entirely skipped; that is, the outcome of the initial preference elicitation of Step (1) is the final result and is displayed in the same way as in Step (3).

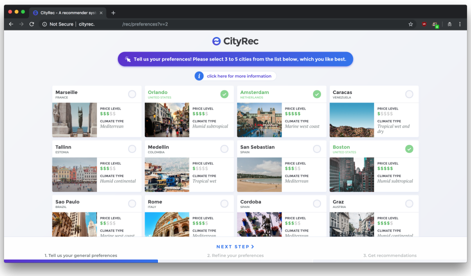
The dependent variables are the usage metrics, such as the choices made at each step, the time taken to specify the preferences, and the number of clicks. Furthermore, we asked the user to fill out a subset of the ResQue Questionnaire, a validated, user-centric evaluation framework for recommender systems [28].

(Q1) The travel destinations recommended to me by CityRec matched my interests
(Q2) The recommender system helped me discover new travel destinations
(Q3) I understood why the travel destinations were recommended to me
(Q4) I found it easy to tell the system what my preferences are
(Q5) I found it easy to modify my taste profile in this recommender system
(Q6) The layout and labels of the recommender interface are adequate
(Q7) Overall, I am satisfied with this recommender system
(Q8) I would use this recommender system again, when looking for travel destinations

4.3 Results

A total of 104 individuals participated in the online survey from December 2018 to March 2019. Participants (44% females, 56% males)

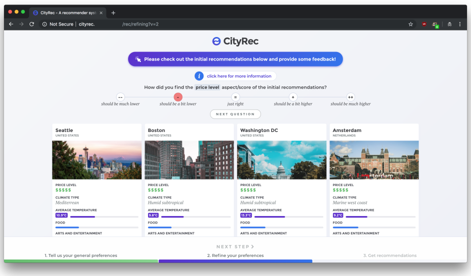
[5] https://nodejs.org/en/
[6] https://reactjs.org/
[7] https://github.com/divino5/cityrec-prototype


CityRec - A recommender syst
Not Secure | cityrec. /rec/preferences?v=2
CityRec
Marseille FRANCE PRICE LEVEL $$$$$ CLIMATE TYPE Mediterrean
Orlando UNITED STATES PRICE LEVEL $$$$$ CLIMATE TYPE Humid subtropical
Amsterdam NETHERLANDS PRICE LEVEL $$$$$ CLIMATE TYPE Marine west coast
Caracas VENEZUELA PRICE LEVEL $$$$$ CLIMATE TYPE Tropical wet and dry
Tallinn ESTONIA PRICE LEVEL $$$$$ CLIMATE TYPE Humid continental
Medellin COLOMBIA PRICE LEVEL $$$$$ CLIMATE TYPE Tropical wet
San Sebastian SPAIN PRICE LEVEL $$$$$ CLIMATE TYPE Mediterrean
Boston UNITED STATES PRICE LEVEL $$$$$ CLIMATE TYPE Humid subtropical
Sao Paulo BRAZIL PRICE LEVEL $$$$$ CLIMATE TYPE Marine west coast
Rome ITALY PRICE LEVEL $$$$$ CLIMATE TYPE Mediterrean
Cordoba SPAIN PRICE LEVEL $$$$$ CLIMATE TYPE Mediterrean
Graz AUSTRIA PRICE LEVEL $$$$$ CLIMATE TYPE Humid continental
NEXT STEP
1. Tell us your general preferences
2. Refine your preferences
3. Get recommendations

Tell us your preferences! Please select 3 to 5 cities from the list below, which you like best.

click here for more information

Figure 2 (a): Selection of favorable cities, Step (1)


CityRec - A recommender syst
Not Secure | cityrec. /rec/refining?v=2
CityRec
-- should be much lower
- should be a bit lower
= just right
+ should be a bit higher
++ should be much higher
NEXT QUESTION
Seattle UNITED STATES PRICE LEVEL $$$$$ CLIMATE TYPE Mediterrean AVERAGE TEMPERATURE 10.9°C FOOD ARTS AND ENTERTAINMENT
Boston UNITED STATES PRICE LEVEL $$$$$ CLIMATE TYPE Humid subtropical AVERAGE TEMPERATURE 9.8°C FOOD ARTS AND ENTERTAINMENT
Washington DC UNITED STATES PRICE LEVEL $$$$$ CLIMATE TYPE Humid subtropical AVERAGE TEMPERATURE 13.3°C FOOD ARTS AND ENTERTAINMENT
Amsterdam NETHERLANDS PRICE LEVEL $$$$$ CLIMATE TYPE Marine west coast AVERAGE TEMPERATURE 9.2°C FOOD ARTS AND ENTERTAINMENT
NEXT STEP
1. Tell us your general preferences
2. Refine your preferences
3. Get recommendations

Please check out the initial recommendations below and provide some feedback!

click here for more information

How did you find the price level aspect/score of the initial recommendations?

Figure 2 (b): Critiquing of initial recommendations, Step (2)

were recruited by sharing the user study on social media and among groups of friends and colleagues. The self-reported ages were 0–20 (7%), 21–30 (69%,) 31–40 (9%), and 41–50 (5%). Random assignment of the systems was performed after a landing page and had almost equal (51% versus 49%) completion of the survey.

Table 2: Differences between the two systems

Variable	Basel.	Critiqu.	p	W	Sig.
(Q1) Interest match	3.58	**3.88**	**0.043**	645	$*$
(Q2) Novelty	3.44	3.75	0.118	705	ns
(Q3) Understanding	3.46	3.77	0.073	673.5	ns
(Q4) Tell prefs.	3.73	3.90	0.328	775	ns
(Q5) Modify profile	3.24	3.48	0.17	723.5	ns
(Q6) Interface	**4.15**	3.62	**0.009**	1,044	$**$
(Q7) Satisfaction	3.66	**3.92**	**0.037**	649	$*$
(Q8) Future use	3.49	3.67	0.166	724	ns
Time to results	60.92s	**184.07s**	**<0.001**		$***$
Clicks	6.32	**21.35**	**<0.001**		$***$
PCC Food	-0.11	-0.01	0.341		ns
PCC Arts	0.05	0.38	0.066		ns
PCC Outdoors	0.02	**0.45**	**0.024**		$*$
PCC Nightlife	0.2	**0.57**	**0.028**		$*$

Significance levels: $* \, p < 0.05$; $** \, p < 0.01$; $*** \, p < 0.001$

The upper part of Table 2 shows the differences in the mean values and the significance tests of the dependent variables. The mean values of the ordinal answers to the questionnaire (Q1–Q8) are for viewing purposes only; the test statistic was calculated using the Wilcoxon rank sum test with continuity correction for independent populations. The null hypotheses were that the medians of variables of the two groups are equal. In three cases, (Q1), (Q6), and (Q7), we could refute the null hypothesis, which provides interesting insights into the users’ assessment of the system.

In the survey, we also asked the participants to rate their personal importance of tourism-related aspects. Thus, we could compute the Pearson Correlation Coefficient (PCC) between the actual profile from the system and the self-assessment from the survey. The lower part of Table 2 shows these correlations per system and the result of the one-sided Fisher’s r-to-Z test for independent samples.

4.4 Discussion

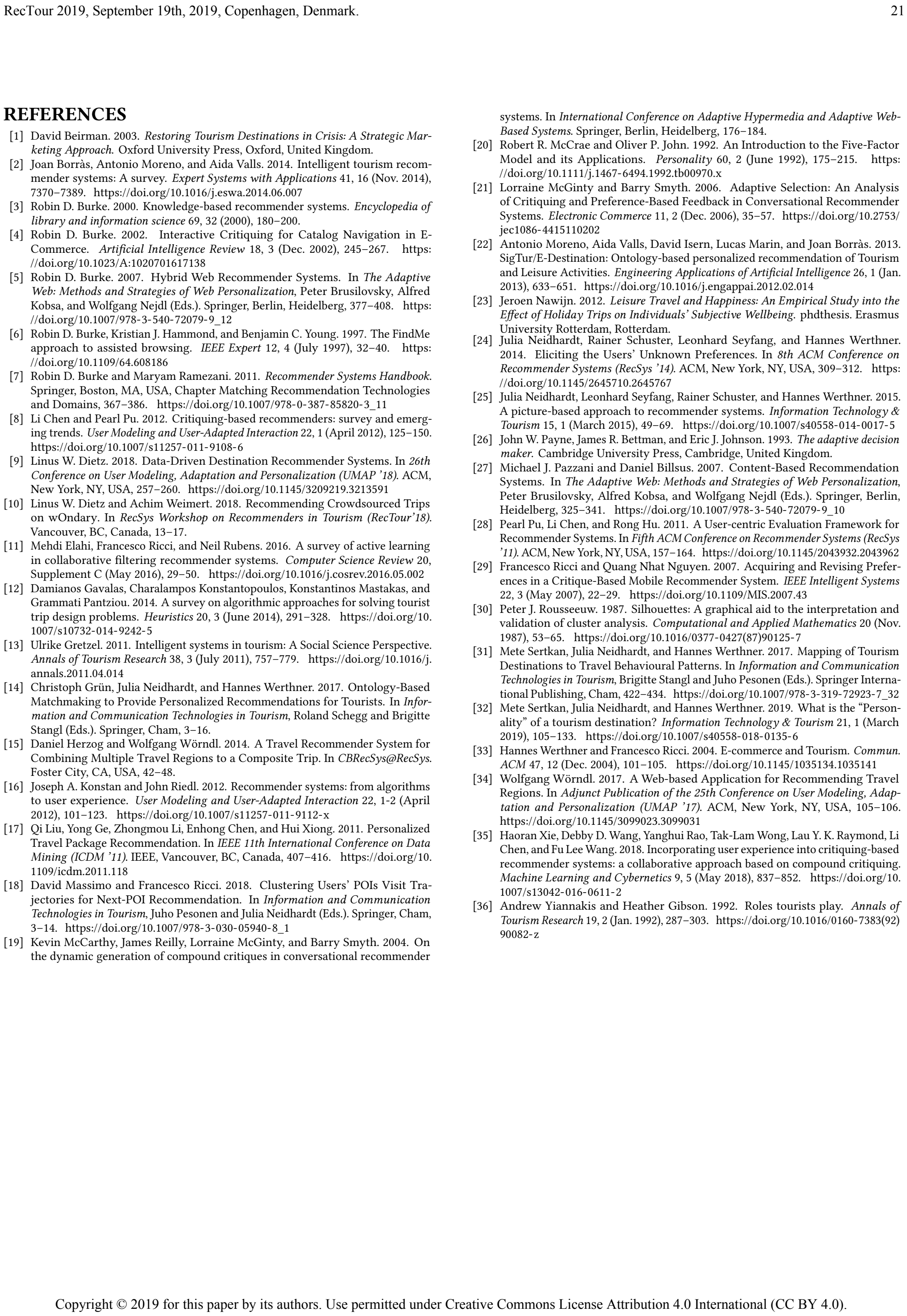
The significant difference in (Q1) shows that the perceived recommendation accuracy is higher, when using the proposed critiquing recommender system, however, at the cost of worse interface adequacy (Q6). This is attributable to the overhead of the critiquing step, Step (2), as it takes triple the time to complete the first two steps and more than triple the number of clicks. Interestingly, the users value higher accuracy more than the adequacy of the interface and the effort as can be seen in the significantly higher user satisfaction (Q7) and the similar levels of potential future use (Q8).

Furthermore, we observed that the user profiles of the critiquing system are significantly higher correlated with the self-assessment in the case of Outdoors & Recreation and Nightlife. This is further evidence that the critiquing recommender version performs better in capturing the preferences of the user. In conclusion, the critiquing version should be preferred as it provides better recommendations from the users’ perspective.

5 CONCLUSIONS

In this paper, we proposed an approach for tackling the problem of recommending complex items in the domain of travel recommendation. We characterized destinations around the globe in a user-understandable way and directly used this characterization in an online recommender system. From the evaluation experiments conducted, we discovered an interesting trade-off between the perceived recommendation accuracy and the perceived adequacy of the user interface; however, the users seemed to favor better recommendations over less effort to obtain them.

Because CityRec’s source code has been released, it can also serve as a foundation for the community to investigate conversational recommender systems based on data-driven item characterization. The destination characterization showed decent results; however, it would be worthwhile to investigate further useful features of destinations that can be derived from other data sources. In this study, we found that, despite higher perceived accuracy (Q1), the interface adequacy (Q6) was rated lower in the critiquing system. Thus, we regard this study as a first step that is to be extended with a more sophisticated preference elicitation approach using active learning. Furthermore, the behavior of the algorithm, with respect to the diversity of the recommendations, should be analyzed as well.



REFERENCES

[1] David Beirman. 2003. *Restoring Tourism Destinations in Crisis: A Strategic Marketing Approach.* Oxford University Press, Oxford, United Kingdom.

[2] Joan Borràs, Antonio Moreno, and Aida Valls. 2014. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications* 41, 16 (Nov. 2014), 7370–7389. https://doi.org/10.1016/j.eswa.2014.06.007

[3] Robin D. Burke. 2000. Knowledge-based recommender systems. *Encyclopedia of library and information science* 69, 32 (2000), 180–200.

[4] Robin D. Burke. 2002. Interactive Critiquing for Catalog Navigation in E-Commerce. *Artificial Intelligence Review* 18, 3 (Dec. 2002), 245–267. https://doi.org/10.1023/A:1020701617138

[5] Robin D. Burke. 2007. Hybrid Web Recommender Systems. In *The Adaptive Web: Methods and Strategies of Web Personalization*, Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Springer, Berlin, Heidelberg, 377–408. https://doi.org/10.1007/978-3-540-72079-9_12

[6] Robin D. Burke, Kristian J. Hammond, and Benjamin C. Young. 1997. The FindMe approach to assisted browsing. *IEEE Expert* 12, 4 (July 1997), 32–40. https://doi.org/10.1109/64.608186

[7] Robin D. Burke and Maryam Ramezani. 2011. *Recommender Systems Handbook*. Springer, Boston, MA, USA, Chapter Matching Recommendation Technologies and Domains, 367–386. https://doi.org/10.1007/978-0-387-85820-3_11

[8] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22, 1 (April 2012), 125–150. https://doi.org/10.1007/s11257-011-9108-6

[9] Linus W. Dietz. 2018. Data-Driven Destination Recommender Systems. In *26th Conference on User Modeling, Adaptation and Personalization (UMAP '18).* ACM, New York, NY, USA, 257–260. https://doi.org/10.1145/3209219.3213591

[10] Linus W. Dietz and Achim Weimert. 2018. Recommending Crowdsourced Trips on wOndary. In *RecSys Workshop on Recommenders in Tourism (RecTour'18).* Vancouver, BC, Canada, 13–17.

[11] Mehdi Elahi, Francesco Ricci, and Neil Rubens. 2016. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review* 20, Supplement C (May 2016), 29–50. https://doi.org/10.1016/j.cosrev.2016.05.002

[12] Damianos Gavalas, Charalampos Konstantopoulos, Konstantinos Mastakas, and Grammati Pantziou. 2014. A survey on algorithmic approaches for solving tourist trip design problems. *Heuristics* 20, 3 (June 2014), 291–328. https://doi.org/10.1007/s10732-014-9242-5

[13] Ulrike Gretzel. 2011. Intelligent systems in tourism: A Social Science Perspective. *Annals of Tourism Research* 38, 3 (July 2011), 757–779. https://doi.org/10.1016/j.annals.2011.04.014

[14] Christoph Grün, Julia Neidhardt, and Hannes Werthner. 2017. Ontology-Based Matchmaking to Provide Personalized Recommendations for Tourists. In *Information and Communication Technologies in Tourism*, Roland Schegg and Brigitte Stangl (Eds.). Springer, Cham, 3–16.

[15] Daniel Herzog and Wolfgang Wörndl. 2014. A Travel Recommender System for Combining Multiple Travel Regions to a Composite Trip. In *CBRecSys@RecSys.* Foster City, CA, USA, 42–48.

[16] Joseph A. Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (April 2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x

[17] Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. 2011. Personalized Travel Package Recommendation. In *IEEE 11th International Conference on Data Mining (ICDM '11).* IEEE, Vancouver, BC, Canada, 407–416. https://doi.org/10.1109/icdm.2011.118

[18] David Massimo and Francesco Ricci. 2018. Clustering Users' POIs Visit Trajectories for Next-POI Recommendation. In *Information and Communication Technologies in Tourism*, Juho Pesonen and Julia Neidhardt (Eds.). Springer, Cham, 3–14. https://doi.org/10.1007/978-3-030-05940-8_1

[19] Kevin McCarthy, James Reilly, Lorraine McGinty, and Barry Smyth. 2004. On the dynamic generation of compound critiques in conversational recommender systems. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems.* Springer, Berlin, Heidelberg, 176–184.

[20] Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and its Applications. *Personality* 60, 2 (June 1992), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

[21] Lorraine McGinty and Barry Smyth. 2006. Adaptive Selection: An Analysis of Critiquing and Preference-Based Feedback in Conversational Recommender Systems. *Electronic Commerce* 11, 2 (Dec. 2006), 35–57. https://doi.org/10.2753/jec1086-4415110202

[22] Antonio Moreno, Aida Valls, David Isern, Lucas Marin, and Joan Borràs. 2013. SigTur/E-Destination: Ontology-based personalized recommendation of Tourism and Leisure Activities. *Engineering Applications of Artificial Intelligence* 26, 1 (Jan. 2013), 633–651. https://doi.org/10.1016/j.engappai.2012.02.014

[23] Jeroen Nawijn. 2012. *Leisure Travel and Happiness: An Empirical Study into the Effect of Holiday Trips on Individuals' Subjective Wellbeing.* phdthesis. Erasmus University Rotterdam, Rotterdam.

[24] Julia Neidhardt, Rainer Schuster, Leonhard Seyfang, and Hannes Werthner. 2014. Eliciting the Users' Unknown Preferences. In *8th ACM Conference on Recommender Systems (RecSys '14).* ACM, New York, NY, USA, 309–312. https://doi.org/10.1145/2645710.2645767

[25] Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. 2015. A picture-based approach to recommender systems. *Information Technology & Tourism* 15, 1 (March 2015), 49–69. https://doi.org/10.1007/s40558-014-0017-5

[26] John W. Payne, James R. Bettman, and Eric J. Johnson. 1993. *The adaptive decision maker.* Cambridge University Press, Cambridge, United Kingdom.

[27] Michael J. Pazzani and Daniel Billsus. 2007. Content-Based Recommendation Systems. In *The Adaptive Web: Methods and Strategies of Web Personalization*, Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Springer, Berlin, Heidelberg, 325–341. https://doi.org/10.1007/978-3-540-72079-9_10

[28] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Fifth ACM Conference on Recommender Systems (RecSys '11).* ACM, New York, NY, USA, 157–164. https://doi.org/10.1145/2043932.2043962

[29] Francesco Ricci and Quang Nhat Nguyen. 2007. Acquiring and Revising Preferences in a Critique-Based Mobile Recommender System. *IEEE Intelligent Systems* 22, 3 (May 2007), 22–29. https://doi.org/10.1109/MIS.2007.43

[30] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics* 20 (Nov. 1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[31] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. 2017. Mapping of Tourism Destinations to Travel Behavioural Patterns. In *Information and Communication Technologies in Tourism*, Brigitte Stangl and Juho Pesonen (Eds.). Springer International Publishing, Cham, 422–434. https://doi.org/10.1007/978-3-319-72923-7_32

[32] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. 2019. What is the "Personality" of a tourism destination? *Information Technology & Tourism* 21, 1 (March 2019), 105–133. https://doi.org/10.1007/s40558-018-0135-6

[33] Hannes Werthner and Francesco Ricci. 2004. E-commerce and Tourism. *Commun. ACM* 47, 12 (Dec. 2004), 101–105. https://doi.org/10.1145/1035134.1035141

[34] Wolfgang Wörndl. 2017. A Web-based Application for Recommending Travel Regions. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17).* ACM, New York, NY, USA, 105–106. https://doi.org/10.1145/3099023.3099031

[35] Haoran Xie, Debby D. Wang, Yanghui Rao, Tak-Lam Wong, Lau Y. K. Raymond, Li Chen, and Fu Lee Wang. 2018. Incorporating user experience into critiquing-based recommender systems: a collaborative approach based on compound critiquing. *Machine Learning and Cybernetics* 9, 5 (May 2018), 837–852. https://doi.org/10.1007/s13042-016-0611-2

[36] Andrew Yiannakis and Heather Gibson. 1992. Roles tourists play. *Annals of Tourism Research* 19, 2 (Jan. 1992), 287–303. https://doi.org/10.1016/0160-7383(92)90082-z

A.6 Publication 6: “Navigation by Revealing Trade-offs for Content-based Recommendations”

Linus W. Dietz, Sameera Thimbiri Palage, and Wolfgang Wörndl. “Navigation by Revealing Trade-offs for Content-based Recommendations.” In: *Information and Communication Technologies in Tourism*. Ed. by Jason L. Stienmetz, Berta Ferrer-Rosell, and David Massimo. Cham: Springer, Jan. 2022, pp. 149–161. ISBN: 978-3-030-94751-4. DOI: 10.1007/978-3-030-94751-4_14

Summary

Tourism is a complex domain for recommender systems because of the high cost of recommending an unsuitable item and the absence of ratings to learn user preferences. Conversational recommender systems have been introduced to provide users with an opportunity to give feedback on items in a turn-based dialog until a final recommendation is accepted. In a scenario such as recommending a city to visit, conversational content-based recommendation may be well-suited since users often struggle to specify their preferences without concrete examples. However, critiquing item features comes with challenges. Users might request item characteristics during recommendation that do not exist in reality, for example, demanding very high item quality for a very low price. To tackle this problem, we present a novel conversational user interface that focuses on revealing the trade-offs of choosing one item over another. The recommendations are driven by a utility function that assesses the user’s preference toward item features while learning the importance of the features to the user. This enables the system to guide the recommendation through the search space faster and accurately over prolonged interaction. We evaluated the system in an online study with 600 participants and found that our proposed paradigm leads to improved perceived accuracy and fewer conversational cycles compared to unit critiquing.

Author Contributions

LD was the main author of the manuscript and designed the algorithms and the user interface together with STP. Furthermore, he conducted user study and performed all statistical analyes. STP implemented the initial prototype after jointly designing the algorithms and user interface with LD. WW supervised the project and co-edited the manuscript.

Navigation by Revealing Trade-offs for Content-Based Recommendations

Linus W. Dietz(✉), Sameera Thimbiri Palage, and Wolfgang Wörndl

Department of Informatics, Technical University of Munich, Munich, Germany
linus.dietz@tum.de

Abstract. Conversational recommender systems have been introduced to provide users the opportunity to give feedback on items in a turn-based dialog until a final recommendation is accepted. Tourism is a complex domain for recommender systems because of high cost of recommending a wrong item and often relatively few ratings to learn user preferences. In a scenario such as recommending a city to visit, conversational content-based recommendation may be advantageous, since users often struggle to specify their preferences without concrete examples. However, critiquing item features comes with challenges. Users might request item characteristics during recommendation that do not exist in reality, for example demanding very high item quality for a very low price. To tackle this problem, we present a novel conversational user interface which focuses on revealing the trade-offs of choosing one item over another. The recommendations are driven by a utility function that assesses the user's preference toward item features while learning the importance of the features to the user. This enables the system to guide the recommendation through the search space faster and accurately over prolonged interaction. We evaluated the system in an online study with 600 participants and find that our proposed paradigm leads to improved perceived accuracy and fewer conversational cycles compared to unit critiquing.

1 Introduction

Nowadays, the algorithmic side of (RSs) research has reached an impressive maturity, such that it has become virtually impossible to tell which algorithms are objectively the best [1]. However, this improvement primarily applies to traditional RSs domains, such as e-commerce, movies, and to some extent music. For recommendations in complex domains, such as tourism, the algorithmic advances of the earlier decades are of lesser value. This is because there are insufficient ratings available, the items are not so well defined in terms of their scope, and it has also been shown that users demonstrate different decision making behavior compared to purchasing physical products [2]. These challenges necessitate employing sophisticated preference elicitation strategies, and instead of collaborative filtering algorithms, recommendations are often computed with a content-based or knowledge-based paradigm. Given that traveling is a relatively rare, emotional, and high-stakes decision making scenario, RSs should provide users

J. L. Stienmetz et al. (Eds.): ENTER 2022, *Information and Communication Technologies in Tourism 2022*, pp. 149–161, 2022.
https://doi.org/10.1007/978-3-030-94751-4_14

with the opportunity to familiarize themselves with the items in the domain and refine their initial preferences, since users often struggle to declare their true preferences [3]. For instance, recommending which city to travel, is a very good fit for the conversational, content-based recommendation paradigm, since there are no ratings available, despite the existence of several data sets [4, Table 2].

Conversational RSs allow a directed search through the item space using some kind of dialog between the system and user [5]. Early approaches, such as FindMe [6], allow users to critique certain aspects of suggested items, whereas more sophisticated approaches allow for compound critiques [7]. Based on the observation that critiques with concrete examples can be useful [8], we are astonished that not much attention has been paid to informing users about the trade-offs involved in their critiquing choices. For example, many users would love to do a dream vacation to a buzzing city with outstanding cultural attractions, great food, a buzzy nightlife scene, favorable climate, at an affordable price tag. In reality, the combination of such features might be an empty set, thus, requiring compromising between conflicting preferences.

In this paper, we present a novel concept to navigate the item space that we call *"Navigation by Revealing Trade-offs."* The motivation for this combination of a novel user interface and a corresponding recommendation algorithm stems from the observation that conversational RSs tend to neglect informing their users about the trade-off involved in their critiquing choices.

After surveying the related work in Sect. 2, we present the user interface in Sect. 3, and describe the recommendation algorithms in Sect. 4. We choose the destination recommendation domain, as there are suitable data sets available and it inherently requires to make trade-offs between certain aspects of the trip. The experimental setup of a large-scale user study with 600 participants is described in Sect. 5 and we present the results in Sect. 6. Finally, we conclude our findings and point out future work in Sect. 7.

2 Related Work

In this work, our application domain is recommending cities for tourist destinations. As opposed to the recommendation of hotels or point of interests [9], cities as items have no meaningful ratings, thus, the user profile and items need to be matched based on elicited preferences and features of the items. To improve the user modeling, Neidhardt et al. [10] proposed a factor analysis for tourist roles and personality traits to reveal seven tourist behavioral patterns. The authors used a set of travel-related pictures, which were assigned to each of the seven factors by experts. Since the destinations were also characterized in the feature space of the Seven Factor model [11], they could perform content-based filtering for destination recommendation. Herzog and Wörndl [12] proposed another travel RS where travel plans of multiple destinations satisfy user constraints such as budget and duration. The user modeling was done by binary indications of interest, i.e., check boxes, and the items were characterized using expert opinions and literature. Such expert-driven models are quite costly, thus, automated approaches are preferable to scale the item characterization. Prior approaches using mainly location-based social network (LBSN) data have been successfully employed in point-of interest recommendation [13] or to characterize cities [14]. The previously proposed city characterization approach [14] is based on the distribution of its venues, where a higher amount of venue relative to the city size leads to a higher scoring. The corresponding user study also suggested that unit critiquing is a fruitful approach in the destination recommendation domain. In this work, we re-use the prototype[1] and domain model of CityRec [14] to build a conversational RS.

Critiquing is a popular approach of eliciting and refining user preferences in a conversational manner. It is usually associated with content-based filtering, although there are some research incorporating collaborative approaches [15] or even unstructured item descriptions [16]. One of the early systems, FindMe [6] introduced the concept of unit critiquing that can be seen as the start of conversational exploration of the search space in RSs research. The static unit critiquing was quite successful in several domains [6, 17], but there is opportunity to perform a smarter exploration of the item space [18]. For example, McCarthy et al. [7] proposed dynamic critiquing, to show how compound critiques can be generated dynamically, cycle-by-cycle by mining the feature patterns of the remaining products.

The evolution of dynamic compound critiques is the multi-attribute utility theory (MAUT) [19], which introduced a utility function to rank a list of multi-attribute products. Once the user selects a critique, the corresponding product is set as the current preference product in the user model and a new set of critiques is generated using a utility function. The MAUT was successfully evaluated against dynamic critiquing [7] thereby reducing the number of critiquing cycles. Chen et al. extended the MAUT-based approach and called it "preference-based organization interfaces" [20]. In their approach, the authors organized all potential critiques in a trade-off vector showing whether the features were compromised or improved in comparison to the current recommendation. That enabled them to determine useful compound critiques and successfully evaluate it using a computer configuration data set. However, we feel that such an approach is more suited for products with clear specifications, since in tourism, relative differences between the features values of items are of higher importance.

One major issue with critiquing is the divergence of the intended direction of exploration. McGinty et al. [21] studied selection strategies for recommending items in critiquing. Their Adaptive Selection approach resulted in a reduction in critiquing cycles and they could prove that their critiquing-based approaches would converge faster than preference-based approaches. Another important insight of their work was that the user should not lose the progress, i.e., the previous recommendation should be included in the upcoming cycle.

Based on these observations, we introduce a paradigm to navigate the search space that we call *"Navigation by Revealing Trade-offs."* We propose a user interface element that visualizes the trade-offs involved in choosing one item instead of

[1] https://github.com/myftija/cityrec-prototype.

another in a less technical way than the preference-based organization interfaces by Chen and Pu [20]. Distinctively, our proposed interface gives the user an indication of the search space, i.e., where the current item’s feature are located within the whole feature space, which was not given in the dynamic and compound critiquing approaches [7,22]. Furthermore, we used a utility function that determines the proposed items, aimed to resolve the *“wishful-thinking problem”* of users requesting item characteristics from the RS that do not exist in reality.

3 A User Interface Concept for Revealing Trade-Offs

3.1 Domain Model

The pure content-based paradigm requires each item to be characterized along the same features to compute recommendations. In our case, we used an available data set of already characterized 180 cities all over the world [14]. This dataset comes with a score for each city in the categories of “Food”, “Nightlife”, “Arts & entertainment”, “Outdoor and recreation”, “Cost of living”, “Shops and services”, “Average temperature”, “Average precipitation”, and “Venue count”. The domain of traveling successfully motivates our approach, since these features are natural in competition, i.e., a larger city with abundant cultural scene usually has higher cost of living, or, conversely, the nightlife options might be limited in small cities.

3.2 User Interaction

The user interaction through a web browser[2] goes through three major steps:

Step (1): An initial user *Preference Elicitation Page*, where the system learns general user preferences,

Step (2): *Conversational Refining* of the recommendations, where the user can refine preferences and learn about the trade-offs in choosing an alternative destination, and

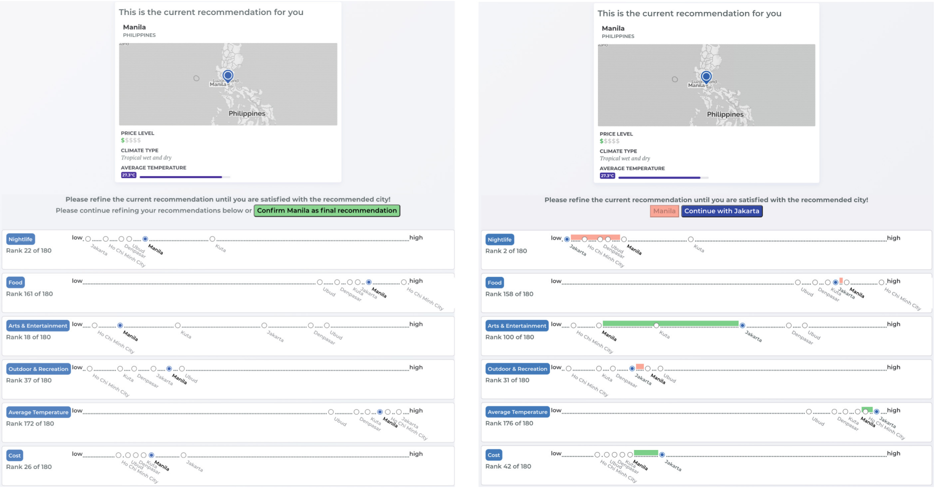
Step (3): *Final Recommendation Page*, where the user is shown the result.

The contribution of this paper focuses on Step (2), the *Conversational Refining*. However, this key step must be seen in the context of the whole interaction design, which we now present step-by-step.

Initial Preference Elicitation. Before the user can start refining, an initial item needs to be determined. Ideally, the system would already have an established user profile based, e.g., through previous interactions. As we have no prior information about the user, we used a previously proposed approach to present the user with an initial seed of destinations where the user can select 3–5 [14]. This seed comprises of randomly selected candidates of various clusters. By this,

[2] The system is available under http://conversational-cityrec.cm.in.tum.de.

the diversity of the sample is warranted, as the user is presented with a representative set of items to choose from. Also, this method is quite fitting for the domain and the initial set of selected cities can directly serve as input for the utility functions of Step (2). We do not aim to evaluate this method from literature [14], as we used it in the same way in all experimental conditions.



This is the current recommendation for you

Manila
PHILIPPINES

PRICE LEVEL
$$$$$

CLIMATE TYPE
Tropical wet and dry

AVERAGE TEMPERATURE
27.3°C

Please refine the current recommendation until you are satisfied with the recommended city!

Please continue refining your recommendations below or Confirm Manila as final recommendation

Nightlife Rank 22 of 180

Food Rank 161 of 180

Arts & Entertainment Rank 18 of 180

Outdoor & Recreation Rank 37 of 180

Average Temperature Rank 172 of 180

Cost Rank 26 of 180

This is the current recommendation for you

Manila
PHILIPPINES

PRICE LEVEL
$$$$$

CLIMATE TYPE
Tropical wet and dry

AVERAGE TEMPERATURE
27.3°C

Please refine the current recommendation until you are satisfied with the recommended city!

Manila Continue with Jakarta

Nightlife Rank 2 of 180

Food Rank 158 of 180

Arts & Entertainment Rank 100 of 180

Outdoor & Recreation Rank 31 of 180

Average Temperature Rank 176 of 180

Cost Rank 42 of 180

(a) Preference Refining Page. This is shown to the user at the beginning of each conversational cycle. The current city is marked bold, and five alternatives are displayed on the spectrum of each feature.

(b) Trade-off Visualization. The green and red shades indicate the trade-off involved should the user choose Jakarta instead of Manila. The user explores various alternatives before continuing.

Fig. 1. User interface of navigation by revealing trade-offs. (Color figure online)

Navigation by Revealing Trade-Offs. Figure 1 shows the interface element for our conversational *"Navigation by Revealing Trade-offs"* approach. At the top of the page, the currently recommended city is shown; below is the novel user interface. This component shows the current city along with five other cities recommended based on the utility function. For each feature the five candidate items are shown in an ordered list from low to high depending on the score. Users can select an item to see the feature value differences in all feature spaces compared to the currently recommended city. An increase in feature value is indicated using a green shade, a decrease is shown in red.

If the user is satisfied with the current recommendation, the user can choose not to continue with refining, but to confirm the current recommendation. In this case, the user is forwarded to the final recommendation page.

Final Recommendation Page. This page shows the final recommendation to the user along with a survey to measure the performance of the recommendation approaches. The final recommended city is shown with details such as the city name, country and feature values.

Baseline System. To evaluate our proposed approach, we used a modified version of the "CityRec" destination RS [14], where one could critique features of several destinations to refine them by buttons indicating "much lower", "lower", "just right", "higher", "much higher". As the source code of this system was readily available, we used it as foundation for our experiments. We re-used the system architecture and the front-end for the initial Preference Elicitation page Step (1), and the Final Recommendation page in Step (3). However, notable differences in the user interface are that we did not use photos of cities to avoid bias due to the selection of images. Furthermore, we re-worked the unit critiquing algorithm to make it more comparable with our system. The critiques can be selected below using the same labels and logic for the adjustment as in the original approach [14], although, it is possible to adjust all features at once and the user is not limited in the number of critiquing cycles, thus, can refine the items until she is satisfied with the recommendation.

4 Algorithms

Having described the user interface elements, this section presents the machinery that computes the recommendation and, therefore, directs the path the user takes through the search space. To enable reproducibility, the system and the study data set are available under an open-source license as a Dockerized software project on Github.[3]

4.1 Cold-Start User Modeling

Recall that in Step (1) of the system the initial input comprises a set of 3–5 items that are characterized along the aforementioned eight features. This already allows us to compute an initial user model by simply representing the user model as an eight dimensional vector with the mean feature values of the initial cities. Nevertheless, this method is quite simple and could be interchanged with any other strategy if more information about the user's preferences is available. Since this is not the case in our evaluation prototype, we used this simple method from literature.

4.2 Candidate Selection Strategy

The next step in the user interface requires finding candidates of which the user can choose one to progress the search for suitable recommendations. In typical content-based recommendation style, one could naively use any similarity metric, such as the Euclidean Distance on normalized feature values to compute some cities similar to the current user model. The top items can then be shown as alternatives to the user. One issue with this strategy is, that it does not consider the user preference variations during the refinement. Furthermore,

[3] https://github.com/LinusDietz/conversational-cityrec.

the convergence of the algorithm will be poor, since it presents the user with similar items to the current recommendation, thus, the user will not have the option to select a city with a significantly different feature value. Instead, we propose the "Variance Bi-distribution" utility function (Eq. 3) whose value is defined by two normal distributions per feature, each representing an increase of decrease of feature value. The two normal distributions are given as $\backsim N(\mu_1, \sigma^2)$ and $\backsim N(\mu_2, \sigma^2)$, where μ_1 and μ_2 define the position of the bell curves on the normalized value range of the feature, and σ defines the shape of the curve.

The distance between the currently selected reference item, ref_k and the respective bell curves are computed by adding or subtracting an offset computed in Eq. 1. This offset is the standard deviation of each feature value f of all previous items in the conversational history H by the number of previous conversational iterations n. The numerator of the offset needs to be moderated by a constant C_m, which we empirically determine for the dataset in Sect. 5.1. To summarize, the mean of the normal distribution is farther from the current user model if the variance of a feature is higher.

$$\mu_1 = ref_k - \frac{\sqrt{Var(f \in H)} \cdot C_m}{n} \quad \mu_2 = ref_k + \frac{\sqrt{Var(f \in H)} \cdot C_m}{n} \tag{1}$$

The second parameter of the normal distributions, σ, is computed in a similar way (cf. Eq. 2). This has the effect that with a higher variance, we obtain a flatter distribution and, thus, a lower impact of this feature on the utility score.

$$\sigma = \frac{\sqrt{Var(f \in H)} \cdot C_s}{n} \tag{2}$$

The intuition behind this is that if the user has a strong preference regarding a feature having a certain value and consistently picks cities with a high temperature, the system is quite certain of this user's preference toward temperature and, thus, should put high weight to this feature. Conversely, if a user has selected cities with another feature having both low and high values resulting in a high variance, it can be seen as a signal that the user has no specific preferences toward the feature as it is not of importance to the user. Thus, the impact of such a high-variance feature should be smaller than a low-variance feature. Over time, we increase this effect by dividing through the number of previous iterations n. This further helps the algorithm converge.

The maximum score of the two distribution functions for a given item feature is taken as the utility score of the respective feature. We then compute the overall utility of each item as the sum of all feature scores of the utility function.

$$\text{utility} = \sum_{f \in F} s(f) \tag{3}$$

Convergence Behavior. The effect of this utility function is that it balances exploration in the beginning and fine-tuning in later stages of the search. If a feature variance is high and the number of iterations small model adjusts μ_1 and μ_2 further away from the reference point, with a higher σ resulting in a flatter distribution of the feature's utility function. In this case, items far away in the feature space also would get higher utility scores, ensuring users are presented with cities more spread across the feature space. With a larger number of iterations, the user preferences for particular features are converging, i.e., the user will be presented with an increasingly narrower band of feature values to refine the preferences. As a result, μ_1 and μ_2 are closer to the feature value of the current recommended item, with a smaller σ, such that items with similar feature values have a substantially higher utility score than the cities with dissimilar feature values. However, if the variance of a feature is still high, the curve will stay quite flat giving this feature less weight, thus, recognizing that the user is rather indifferent toward this feature. This convergence behavior can be observed in Fig. 1. After some iterations, the algorithm determined that the user has a clear preference for high scores in the food and temperature aspects, and low scores in nightlife, outdoor & recreation, and cost. Thus, the refining candidates are quite close by each other, whereas they are spread along the spectrum in the arts & entertainment spectrum.

Elimination of Candidates. To further improve the convergence, we propose a variant that eliminates items whose feature values have been refined in a contrary way. The reasoning behind this elimination of candidates is that if a user refines a feature of an item, it becomes an explicit information that the value of the feature is unsatisfactory and should take only values toward the direction of the refinement. Thus, we can compute candidates just as before, however, items that have a lower (or higher) value than the original item ref_k are removed from the search space. For example, if the user refines the value of *Arts & Entertainment* of Manila in Fig. 1b in favor of Jakarta, the system will assume that all cities that have a lower value in *Arts & Entertainment* than Manila should be excluded from future suggestions.

5 User-Centric Evaluation

For the evaluation of the system, we chose a between-subject design to perform a large-scale online user study. First, we need to determine the constants of the Variance Bi-distribution Model for the current data set.

5.1 Instantiation for the Domain

During the development of the system, we noticed that using only the standard deviation divided by the number of iterations in Eq. 1 and 2, μ_1 and μ_2 would be too extreme, which will result in items recommended that are too far away from the current city. To moderate this effect, the constants C_m and C_s of Eqs. 1 and 2 were introduced for the Variance Bi-distribution Model. This step ensures an efficient navigation should be seen as an adjustment of the algorithmic properties to the data set at hand, as different domains can have different characteristics, i.e., a different number of items.

Determining Constants. The values of C_m and C_s can be determined in an offline setting using a simulation. This is because by systematically altering the values of C_m and C_s, we can see how quickly the algorithms converge from an initial setting after Step (1) to a desired item while making *consistent decisions*. In the context of the simulation, we define consistent decisions by choosing the item that is nearest to the target item using the distance metric of the RS. Thus, the simulator chooses candidates toward the target recommendation, just as a real user would, until that recommendation is part of the set of candidate items. For the cities, we used user interaction data to perform a realistic simulation [14]. The data set of 63 user sessions contained the initial city selections by the user and the final recommendation the user had selected. Having historic data for the simulation, we can now train the parameters using relevant scenarios, as opposed the randomized or exhaustive simulation strategies.

Result. Regarding parameters of the simulation, we varied C_m from 2 to 6, and C_s from 4 to 20, both in 0.5 intervals. For each these parameters' configuration, we recorded the session length of the 63 user sessions of the data set. The result of the simulation reveals a global optimum at $C_m = 3$ and $C_s = 8$.

5.2 Online User Study

We conducted the user study using the online experimentation platform Prolific.[4] We used a between-subject design and invited participants of the platform who had indicated "Traveling" as one of their hobbies. Only one independent variable was randomly assigned to the users, i.e., the critiquing system in Step (2). The three options[5] were the baseline unit critiquing system and the trade-offs UI using the Variance Bi-distribution Model without and with the elimination variant. As dependent variables, we used metrics about the user interaction and a subset of the ResQue Questionnaire (cf. Table 1), which is a validated, user-centric evaluation framework for RSs [23], where users indicate their agreement with each statement on a Five-point Likert Scale.

6 Results

The user study was conducted in December 2020 with 600 participants. Out of the 600 participants, we excluded 181 responses, which failed an attention check, showed very low interaction with the system, i.e., an interaction of less than 35 s, and did not use a desktop browser as instructed. This left us with 419 valid submissions (59.9% female, 39.1% male, 1% other) from 42 different countries. The users predominantly came from Europe, due to the time zone when the survey was initiated. The age distribution was 20.8% of below 21 year olds, 55.6% were 21–30, 13.1% were 31–40, 6.2% were 41–50, 3.1% were 51–60,

[4] https://prolific.co/.

[5] The variants can be tested under http://conversational-cityrec.cm.in.tum.de.

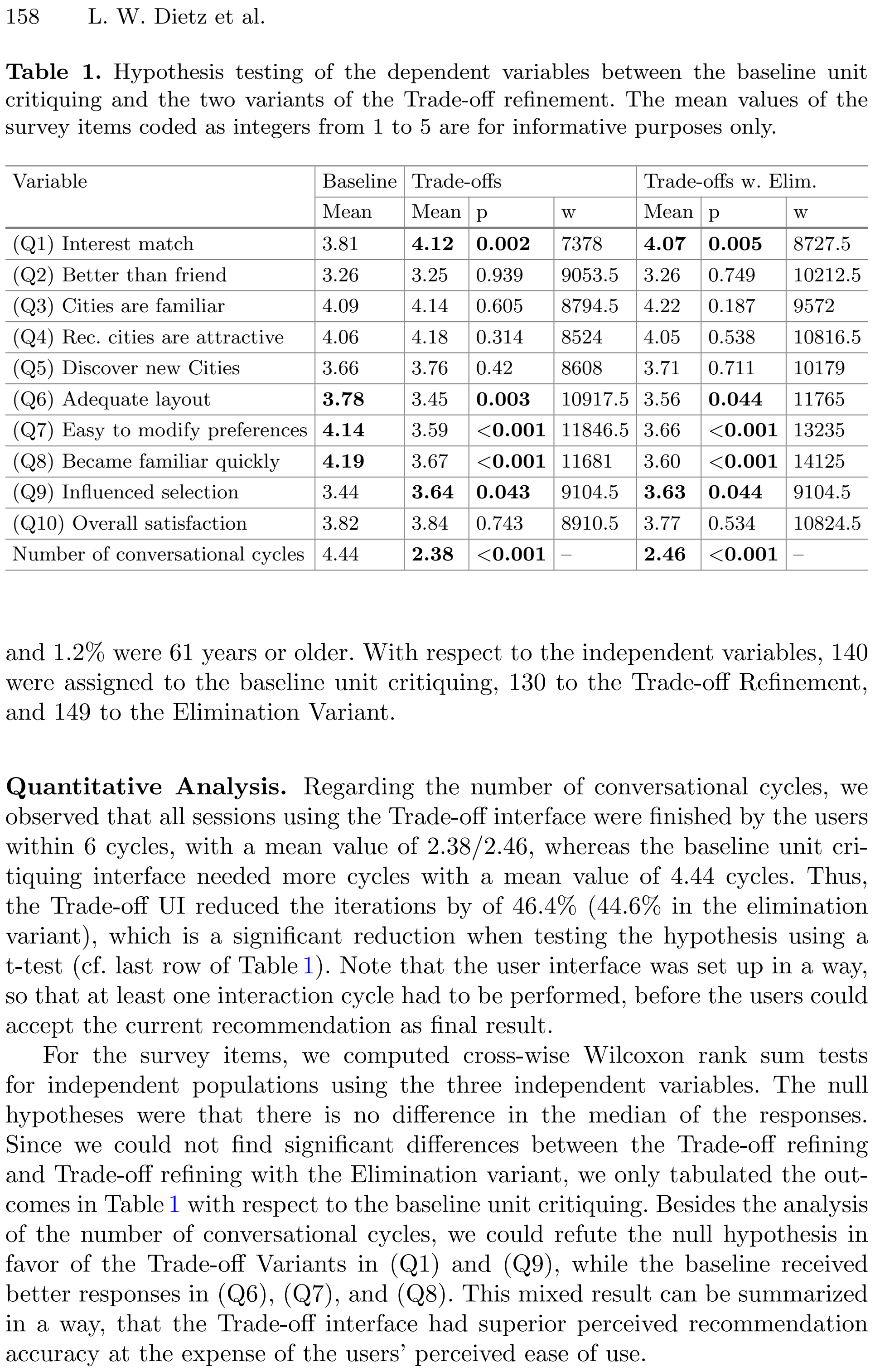


Table 1. Hypothesis testing of the dependent variables between the baseline unit critiquing and the two variants of the Trade-off refinement. The mean values of the survey items coded as integers from 1 to 5 are for informative purposes only.

Variable	Baseline	Trade-offs			Trade-offs w. Elim.		
	Mean	Mean	p	w	Mean	p	w
(Q1) Interest match	3.81	**4.12**	**0.002**	7378	**4.07**	**0.005**	8727.5
(Q2) Better than friend	3.26	3.25	0.939	9053.5	3.26	0.749	10212.5
(Q3) Cities are familiar	4.09	4.14	0.605	8794.5	4.22	0.187	9572
(Q4) Rec. cities are attractive	4.06	4.18	0.314	8524	4.05	0.538	10816.5
(Q5) Discover new Cities	3.66	3.76	0.42	8608	3.71	0.711	10179
(Q6) Adequate layout	**3.78**	3.45	**0.003**	10917.5	3.56	**0.044**	11765
(Q7) Easy to modify preferences	**4.14**	3.59	**<0.001**	11846.5	3.66	**<0.001**	13235
(Q8) Became familiar quickly	**4.19**	3.67	**<0.001**	11681	3.60	**<0.001**	14125
(Q9) Influenced selection	3.44	**3.64**	**0.043**	9104.5	**3.63**	**0.044**	9104.5
(Q10) Overall satisfaction	3.82	3.84	0.743	8910.5	3.77	0.534	10824.5
Number of conversational cycles	4.44	**2.38**	**<0.001**	–	**2.46**	**<0.001**	–

and 1.2% were 61 years or older. With respect to the independent variables, 140 were assigned to the baseline unit critiquing, 130 to the Trade-off Refinement, and 149 to the Elimination Variant.

Quantitative Analysis. Regarding the number of conversational cycles, we observed that all sessions using the Trade-off interface were finished by the users within 6 cycles, with a mean value of 2.38/2.46, whereas the baseline unit critiquing interface needed more cycles with a mean value of 4.44 cycles. Thus, the Trade-off UI reduced the iterations by of 46.4% (44.6% in the elimination variant), which is a significant reduction when testing the hypothesis using a t-test (cf. last row of Table 1). Note that the user interface was set up in a way, so that at least one interaction cycle had to be performed, before the users could accept the current recommendation as final result.

For the survey items, we computed cross-wise Wilcoxon rank sum tests for independent populations using the three independent variables. The null hypotheses were that there is no difference in the median of the responses. Since we could not find significant differences between the Trade-off refining and Trade-off refining with the Elimination variant, we only tabulated the outcomes in Table 1 with respect to the baseline unit critiquing. Besides the analysis of the number of conversational cycles, we could refute the null hypothesis in favor of the Trade-off Variants in (Q1) and (Q9), while the baseline received better responses in (Q6), (Q7), and (Q8). This mixed result can be summarized in a way, that the Trade-off interface had superior perceived recommendation accuracy at the expense of the users’ perceived ease of use.

Discussion. The superior perceived accuracy measured by (Q1) at about 45% fewer conversational cycles, underlines the merit of our proposed user interface. However, the subjects rated the usability-related metrics of the unit critiquing system higher (Q6–Q8). We suspect that this due to that unit critiquing has already been employed in various RSs, so it is quite possible that many users were already familiar with this concept. Dealing with a new refinement interface involving reasoning about trade-offs certainly involves more cognitive effort and, thus, might need more familiarization (Q8) than only one session. The study was designed in a way that users could only submit the survey once and we did not familiarize the users with the system before their session to avoid learning effects. The significant difference in (Q9) "This recommender system influenced my selection of cities." in favor of the Trade-off interface is likely an artifact of the comparative lengthy search in the unit critiquing, since both values are in the center of the Likert Scale. Interestingly, there were no significant differences in any dependent variables between the Trade-off refinement and its Elimination variant. We attribute this to the low number of conversational cycles that were needed to come up with a satisfactory result. In the given data set of 180 cities, the elimination of candidates was probably not necessary, as the utility function was able to recommend attractive items after two or three cycles. Nevertheless, we are confident that the concept of elimination of parts of the search space based on the users' choices could be useful and we plan to analyze the merit of the Elimination variant with larger item sets of over 1000 items.

7 Conclusions

The success of modern recommender systems depends on the seamless integration of algorithms and user interface elements. Given that existing critiquing systems have often neglected to explicitly inform users about the trade-offs of the critiquing actions, we developed the Navigation by Revealing Trade-offs system, which integrates a user interface concept with a utility function to compute refinement candidates. The evaluation shows that perceived accuracy is better than the unit critiquing baseline at similar reductions in the number of conversational cycles as other advanced critiquing approaches have demonstrated [19,21].

Based on this promising result, further analyses of this refinement paradigm should follow with larger item sets to analyze the merits of the Elimination variant. Since our study followed a between-subject design, we also can not answer whether the higher ratings for the interface adequacy are due to that unit critiquing being conceptually easier to understand or users are more familiar with such a long-established paradigm. Therefore, the usability and learnability should be investigated in a usability analysis in a controlled laboratory setting.

References

1. Dacrema MF, Cremonesi P, Jannach D (2019) Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In: 13th ACM conference on recommender systems, September 2019. ACM, New York, pp 101–109
2. Werthner H, Ricci F (2004) E-commerce and tourism. CACM 47(12):101–105
3. Sertkan M, Neidhardt J, Werthner H (2020) Eliciting touristic profiles: a user study on picture collections. In: 28th ACM conference on user modeling, adaptation and personalization. ACM, New York, pp 230–38
4. Jannach D, Manzoor A, Cai W, Chen L (2021) A survey on conversational recommender systems. ACM Comput Surv 54(5):1–36
5. Salamó M, Reilly J, McGinty L, Smyth B (2005) Knowledge discovery from user preferences in conversational recommendation. In: Jorge AM, Torgo L, Brazdil P, Camacho R, Gama J (eds) Knowledge discovery in databases. Springer, Heidelberg, pp 228–239. https://doi.org/10.1007/11564126_25
6. Burke RD, Hammond KJ, Young BC (1997) The FindMe approach to assisted browsing. IEEE Expert 12(4):32–40
7. McCarthy K, Reilly J, McGinty L, Smyth B (2004) On the dynamic generation of compound critiques in conversational recommender systems. In: De Bra PME, Nejdl W (eds) International conference on adaptive hypermedia and adaptive web-based systems. Springer, Heidelberg, pp 176–184. https://doi.org/10.1007/978-3-540-27780-4_21
8. Viappiani P, Faltings B, Pearl P (2006) Preference-based search using example-critiquing with suggestions. JAIR 27(1):465–503
9. David M, Ricci F (2018) Clustering users' POIs visit trajectories for next-POI recommendation. In: Pesonen J, Neidhardt J (eds) Information and communication technologies in tourism, December 2018. Springer, Cham, pp 3–14
10. Neidhardt J, Seyfang L, Schuster R, Werthner H (2015) A picture-based approach to recommender systems. JITT 15(1):49–69
11. Sertkan M, Neidhardt J, Werthner H (2019) What is the "personality" of a tourism destination? JITT 21(1):105–133
12. Herzog D, Wörndl W (2014) A travel recommender system for combining multiple travel regions to a composite trip. In: CBRecSys, pp 42–48
13. Zhiwen Yu, Huang X, Yang Z, Guo B (2015) Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints. IEEE Trans Hum-Mach Syst 46(1):151–158
14. Dietz LW, Myftija S, Wörndl W (2019) Designing a conversational travel recommender system based on data-driven destination characterization. In: ACM RecTour, September 2019, pp 17–21
15. Xie H et al (2018) Incorporating user experience into critiquing-based recommender systems: a collaborative approach based on compound critiquing. Mach Learn Cybern 9(5):837–852. https://doi.org/10.1007/s13042-016-0611-2
16. Rana A, Bridge D (2020) Navigation-by-preference: a new conversational recommender with preference-based feedback. In: 25th International conference on intelligent user interfaces, March 2020. ACM, New York, pp 155–165
17. Burke RD (2002) Interactive critiquing for catalog navigation in e-commerce. Artif Intell Rev 18(3):245–267
18. McGinty L, Reilly J (2011) On the evolution of critiquing recommenders. In: Ricci F, Rokach L, Shapira B, Kantor P (eds) RecSys handbook, October 2011. Springer, Boston, pp 419–453. https://doi.org/10.1007/978-0-387-85820-3_13

19. Zhang J, Pu P (2006) A comparative study of compound critique generation in conversational recommender systems. In: Wade VP, Ashman H, Smyth B (eds) Adaptive hypermedia and adaptive web-based systems. Springer, Heidelberg, pp 234–243. https://doi.org/10.1007/11768012_25

20. Chen L, Pu P (2007) Preference-based organization interfaces: aiding user critiques in recommender systems. In: Conati C, McCoy K, Paliouras G (eds) User modeling. Springer, Heidelberg, pp 77–86. https://doi.org/10.1007/978-3-540-73078-1_11

21. McGinty L, Smyth B (2006) Adaptive selection: an analysis of critiquing and preference-based feedback in conversational recommender systems. Electron Commer 11(2):35–57

22. Reilly J, Zhang J, McGinty L, Pu P, Smyth B (2007) Evaluating compound critiquing recommenders: a real-user study. In: 8th ACM conference on electronic commerce. ACM, pp 114–123

23. Pu P, Chen L, Hu R (2011) A user-centric evaluation framework for recommender systems. In: 5th ACM conference on recommender systems. ACM, New York, pp 157–164