Predicting mobility patterns based on profiles of social media users: tourists case study

Helen C. de Mattos Senefonte
UEL - Londrina
helen@uel.br

Thiago H. Silva
UTFPR - Curitiba
thiagoh@utfpr.edu.br

Ricardo Lüders
UTFPR - Curitiba
luders@utfpr.edu.br

Abstract

Studies based on traditional data sources like surveys, for instance, offer poor scalability. The experiments are limited, and the results are restricted to small regions (such as a city or a state). The use of location-based social network (LBSN) data can mitigate the scalability problem by enabling the study of social behavior in large populations. When explored with Data Mining and Machine Learning techniques, LBSN data can be used to provide predictions of relevant cultural and behavioral data from cities or countries around the world. The main goal of this work is to predict and explore user behavior from LBSNs in the context of tourists' mobility patterns. To achieve this goal, we propose PredicTour, which is an approach used to process LBSN users' check-ins and to predict mobility patterns of tourists with or without previous visiting records when visiting new countries. PredicTour is composed of three key blocks: mobility modeling, profile extraction, and tourist' mobility prediction. In the first block, sequences of check-ins in a time interval are associated with other user information to produce a new structure called "mobility descriptor". In the profile extraction, self-organizing maps and fuzzy C-means work together to group users according to their mobility descriptors. PredicTour then identifies tourist' profiles and estimates their mobility patterns in new countries. When comparing the performance of PredicTour with three well-known machine learning-based models, the results indicate that PredicTour outperforms the baseline approaches. Therefore, it is a good alternative for predicting and understanding international tourists' mobility, which has an economic impact on the tourism industry, particularly when services and logistics across international borders should be provided. The proposed approach can be used in different applications such as recommender systems for tourists, and decision-making support for urban planners interested in improving both the tourists' experience and attractiveness of venues through personalized services.

Keywords: location-based social network, machine learning, tourist mobility, cross-cultural analysis, social computing

1 Problem Description and Motivation

An increasing number of people living in cities imposes new challenges to segments essential for a good quality of life, such as mobility, social interaction, and health. Consequently, cities worldwide seek novel solutions to face these challenges [1].

There are many data sources available to help urban/social computing studies. For example, Location-based social networks (LBSN) are a type of social media that allows users to share data containing spatiotemporal information [9]. LBSNs, such as Foursquare, Waze, Twitter, and Instagram, provide a new range of possibilities for obtaining data on a large scale, especially with the massive increase in social media users. LBSNs have been explored in different applications and offer a cheap platform for studying human behavior in near real-time and planetary-scale [10].

The application explored in the present work relates to urban tourism. Planning for urban tourism is important in many aspects. According to the World Tourism Organization (WTO), the flux of tourists around the world generated revenues of more than one billion US dollars in 2019 [5], and created millions of direct and indirect jobs [5]. It is essential to understand the tourists' preferences to maintain the attractiveness of the tourist activity in a specific location, offering better and smarter services. For that, study tourist mobility is essential. The study of tourist mobility has been an under-explored aspect of the tourism industry, receiving scarce research funding [4]. Despite previous efforts, only some works have attempted to model the mobility patterns of tourists on large scale [3, 14]. LBSN data represents a more recent alternative for research. Understanding tourist mobility is critical for many applications in urban tourism planning. For instance, constructing specialized recommendation systems, suggesting attractive services and products, and improving transportation and attractions strategies.

Motivated by the previously mentioned aspects, this thesis addresses the following questions: (i) In terms of mobility, how do tourists behave in countries where they travel to? (ii) Do the origin and the destination have any influence over the mobility pattern? (iii) How can we model useful mobility characteristics of tourists? (iv) Is the tourist mobility information useful to improve existing applications or propose new ones?

2 Objectives

The main goals of the present work are to characterize, understand, and predict tourists' mobility from LBSN data. By exploring tourists' mobility patterns, tourist agencies and public administrations could provide better services like more

In: V Concurso de Teses e Dissertações (CTD 2023), Ribeirão Preto, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2023.

ISSN 2596-1683

engaging activities, logistic planning, and public transportation.

The specific objectives are: (i) to characterize tourists' mobility patterns in different countries; (ii) to model data from LBSN aiming to build mobility structures capable of representing non-trivial information; (iii) to predict urban users' issues based on the developed models, such as auto-identification of venue categories; (iv) to analyze the mobility pattern of specific groups aiming to extract the dependency of some aspects such as nationality, geo-location, and cultural differences between the users; (v) to consider additional features beyond the trivial origin and destination information in the users profiling process; and (vi) to predict tourists mobility behavior based on the profile information extracted by our methodology.

3 Main Contributions

The contributions of our work are summarized in the following points.

1. **Mobility Patterns Modeling and Characterization:** mobility patterns are modeled from LBSN and explored in different ways. We model the mobility of tourists by exploring structures proposed in the literature, and we also propose a new structure called mobility descriptor. These mobility structures are vastly used in all contributions of this thesis. In addition, we analyze and identify users' mobility patterns in different steps: (a) we present original results of the exploratory analysis from LBSN data, where we investigate some of our initial questions about country preferences on data sharing (b) we study the regional influences on tourists mobility to characterize users behavior by the origin and destination countries (that we called behavioral distance) (c) in a novel way, we explore the behavioral and cultural distances to construct a regression on the tourists behavior.
2. **Venue Classification:** we develop a new classification approach k-FN to identify the venue categories from unlabeled geo-located check-ins with noisy data. This new approach is inspired by k-NN, which is modified to deal with the particular characteristics of the problem. Simplicity and straightforward update processes are some of the proposed approach's main advantages compared to traditional training-based classification approaches.
3. **Tourists Mobility Prediction:** we propose a novel methodology called PredicTour to predict international tourists' mobility patterns using LBSNs. Users' check-in information underlies the construction of the PredicTour model that predicts the mobility of users when visiting new countries. Unlike the previous contributions, the mobility vector is expanded to include data other than the transition information. The structure called *mobility descriptor* considers tourist origin and destination countries and also tourist classification between returner and explorer classes beyond the mobility vector data. The PredicTour methodology consists of two main tasks: (a) tourists profile extraction; (b) prediction of mobility pattern of known tourists in unvisited countries. In the profile extraction task, self-organizing maps (SOM) and fuzzy c-means (FCM) models work jointly to group users according to their mobility descriptors. In the mobility prediction task, PredicTour identifies tourist profiles and estimates the mobility patterns of tourists visiting new countries. To comparatively evaluate the performance of PredicTour, we also use well-known machine learning-based models – Deep AutoEncoder, Multi-layer Perceptron, and Collaborative Filtering – as comparison approaches.

The following papers have been published as a result of the thesis contributions.

- Two conference papers concerning mobility patterns modeling and characterization. The mobility modeling and regional influences characterization are published at the Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - SBRC 2020 [13], and a linear regression model of tourists' behaviors exploring the behavioral and cultural distances are published at the International Conference on Social Informatics - SocInfo 2020 [6].
- One conference paper regarding venue classification, where a novel classification approach k-FN to identify the venue categories from unlabeled geo-located check-ins with noisy data using mobility patterns of users is published at the International Conference on Distributed Computing in Sensor Systems -DCOSS 2019 [2].
- One journal paper and one conference paper investigating the users' behavior, comparing the LBSN and traditional data results. The initial results from this investigation are published at the Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, SBRC 2021 [11]. Aiming to provide an extensive comparison between official surveys and social media data, we analyze the differences and similarities between both types of data, results published at the Online Social Networks and Media 2022 [12].
- One journal paper related to PredicTour contributions, published at the IEEE Access journal in 2022 [8].

4 Main Results

The results obtained by the experiments of our contributions are mentioned and discussed here.

4.1 Mobility Patterns Modeling and Characterization

In our first contribution, we presented the description of our dataset and the main observations obtained by the exploratory analysis. All observations presented in this chapter suggest significant differences concerning the behavior of users in certain places of the world. Users' origin, destination, and cultural characteristics reflect their behavior dynamic. Additionally, we described how data are modeled in the mobility context. We also presented a study comparing official census and LBSN data. According to the results, most of the countries have strong positive correlations. It represents how similar the country is concerning the others in both datasets. Finally, we investigated the mobility of tourists by country through the lens of regional influences. Linear regression was used to evaluate the relationship between behavioral and cultural distances. The results indicated an influence on tourists' mobility behavior associated with their country of origin. This fact opens up different opportunities for exploration in the proposition of new systems and improving existing ones.

4.2 Venue Classification

In this contribution, we investigated users' mobility patterns to classify venue subcategories from unlabeled check-ins with geographic location. We described the proposal of a new classification algorithm k-FN to deal with the particular characteristics of our problem. The outcomes using k-FN were satisfactory for all considered cities, reaching, in some cases, around 96% accuracy (in more realistic simulated scenarios). It suggests that users' mobility is relevant to be considered for better performance in classification problems with LBSN data.

4.3 PredicTour: Tourists Mobility Prediction

Some of the goals of this contribution were to understand what kind of intrinsic relationships can be identified when we group users, and what kind of pattern can be identified in each profile. We explore the mobility pattern of each profile by scrutinizing the characteristics of the mobility descriptor associated with each profile centroid. Clustering MDs of Foursquare users in different profiles (using our methodology) allows us to characterize each group according to specific patterns.

PredicTour uses profile information to predict mobility patterns. We consider five baseline approaches to compare the results with PredicTour. In this way, the overall RMSE performance is shown in Figure 1 for different difficulty levels. It compares the performance of PredicTour with all baselines, which consider as input an approximated mobility pattern of a tourist t and as output: Baseline 1) the mobility vector of a randomly chosen tourist with the same destination; Baseline 2) an average of mobility vectors among tourists with the same destination; Baseline 3) the output vector provided by collaborative filtering; Baseline 4) the output vector provided by a multi-layer perceptron; Baseline 5) the linearization of the matrix provided by a deep autoencoder applied to the input transition matrix.

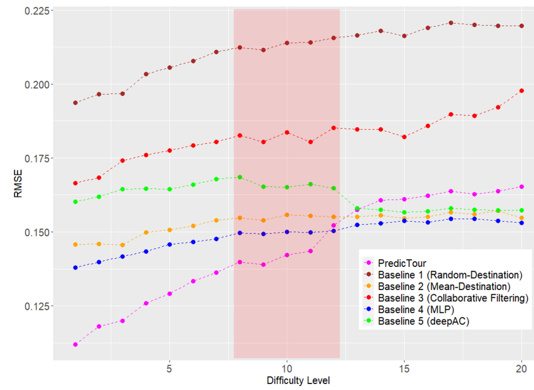

Figure 1. RMSE of PredicTour and baselines for different difficulty levels.

The results of RMSE show that PredicTour produces smaller errors for predicting tourists' mobility when they visit different countries, especially when the number of tourists with historical information is representative (the red region, where the proportion between non-historical and historical information is balanced).

Aiming to understand if the prediction errors are occurring for more relevant components of the mobility vectors, we also consider the nDCG metric. The results are significantly better for PredicTour in almost all cases. Therefore, the general prediction of tourist' behavior is satisfactory if we consider more relevant components (top 5 and top 10) that should be correctly predicted in practice.

Additionally, we presented a clustering sensibility study of the profiles' generation. PredicTour demonstrates consistency, with many users clustered in the same group when many experiments are performed.

As a final result of this contribution, in a non-trivial way, we also analyzed PredicTour's impact on the behavioral and cultural study. Outcomes from this investigation can state the power of PredicTour to generate synthetic data in the context of LBSN's mobility.

5 Discussion

Our first contribution proposed a mobility characterization where we presented a preliminary exploratory analysis and a regional influence characterization on tourists' mobility behavior. In this part of the work, we highlighted significant observations concerning the behavior of users in certain places of the world. We observed patterns in most countries, as cultural, social, and other habits tend to reflect the behavior of some specific groups. Tourists typically have behaviors influenced by their country of origin, mainly when large cultural distances from the locals are observed. We also presented a relevant contribution: a multidimensional comparison between the LBSN and WTO data, showing that most countries have high correlation values, which means that both datasets share similar information.

Another contribution of this thesis is the approach to classify venues of unlabeled check-ins with geographic locations. For example, this approach could aggregate missing information in the transition matrix. We proposed a new classification algorithm, k-FN, to deal with the particular characteristics of our problem. The results using k-FN were satisfactory for all considered cities, reaching, in some cases, around 96% accuracy (in realistic scenarios). It suggests that users' mobility is important to be considered for better performance in classification problems with LBSN data.

The final contribution is to extract and explore patterns available on LBSN data to improve the understanding of users' behavior and use this information to understand and predict international tourists' mobility patterns in different countries. In this contribution, the proposed approach PredicTour explored LBSN data to construct a *mobility descriptor* necessary to express non-trivial information regarding international tourists' mobility. PredicTour was then capable of predicting the mobility of tourists in unvisited countries based on the tourists' profiles extracted from Self Organizing Maps (SOMs) and Fuzzy C-means clustering.

The results showed evidence that PredicTour can extract important user profile characteristics, which can be further explored in a tourism context. We also evaluated the performance of PredicTour in different scenarios and against relevant baselines. The results showed that our approach could perform satisfactorily, providing smaller RMSE than the addressed baselines, especially for non-pessimistic scenarios. Furthermore, if we focus on the performance for the top 5 and top 10 features, the most important ones, PredicTour outperformed the baselines in virtually all test cases, except on the extreme ones expected to occur less in practice. Additionally, the results from clustering sensibility analysis showed that most users are clustered together in all cross-validation experiments, ratifying that the profile extraction task from PredicTour is consistent when clustering profiles of users by their mobility patterns. Surprisingly, the outcomes from investigating the PredicTour's impact revealed the power of our methodology to generate synthetic data, increasing the applicability range of the approach.

The PredicTour approach can be helpful for many applications in tourism planning. For instance, it can build recommender systems of new places for particular groups of international tourists. It can help in the suggestion of attractive services and products and improvement of transport and attractions strategies. The map of intrinsic relationships provided by SOM could also be helpful to show (in a visual tool for tourism planning, for example) tourists with similar behaviors, which could be grouped into the same activities. Although this study has focused on international tourism, we believe that PredicTour could also provide results for domestic tourism with minor adaptations for home and destination locations. Limitations and future works are presented in the Conclusion Chapter of the thesis [7].

Acknowledgment. This work was supported in part by the São Paulo Research Foundation (FAPESP) through the project GoodWeb (Grant 2018/23011-1), and SocialNet (Grant 2023/00148-0); and the National Council for Scientific and Technological Development - Brazil (CNPq) (Grant 310998/2020-4).

References

[1] Ayoub Arroub, Bassma Zahi, Essaid Sabir, and Mohamed Sadik. 2016. A literature review on Smart Cities: Paradigms, opportunities and open problems. *WINCOM. Fez, Morocco.* (2016), 180–186.

[2] H. C. de Mattos Senefonte, T. H. Silva, R. Lüders, and M. R. B. S. Delgado. 2019. Classifying Venue Categories of Unlabeled Check-ins Using Mobility Patterns. In *15th DCOSS*. 562–569.

[3] Ana P.G. Ferreira, Thiago H. Silva, and Antonio A.F. Loureiro. 2020. Uncovering spatiotemporal and semantic aspects of tourists mobility using social sensing. *Computer Communications* 160 (2020), 240 – 252.

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

[5] World Tourism Organization. 2021. *International Tourism Highlights, 2020 Edition.* World Tourism Organization, Madrid, Spain. https://doi.org/10.18111/9789284422456

[6] Helen Senefonte, Gabriel Frizzo, Myriam Delgado, Ricardo Lüders, Daniel Silver, and Thiago Silva. 2020. Regional Influences on Tourists Mobility Through the Lens of Social Sensing. In *Social Informatics*. Cham, 312–319.

[7] Helen C. de Mattos SENEFONTE. 2022. *Predicting mobility patterns based on profiles of social media users: tourists case study.* Ph. D. Dissertation. Graduate Program in Electrical and Computer Engineering of the Universidade Tecnológica Federal do Paraná - Curitiba.

[8] Helen C. Mattos Senefonte, Myriam Regattieri Delgado, Ricardo Lüders, and Thiago H. Silva. 2022. PredicTour: Predicting Mobility Patterns of Tourists Based on Social Media User's Profiles. *IEEE Access* 10 (2022), 9257–9270.

[9] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, and Antonio Alfredo Ferreira Loureiro. 2014. Large-scale study of city dynamics and urban social behavior using participatory sensing. *IEEE Wireless Communications* 21 (2014), 42–51.

[10] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Juliana F. S. Salles, and Antonio Alfredo Ferreira Loureiro. 2014. Revealing the City That We Cannot See. *ACM Trans. Internet Techn.* 14 (2014), 26:1–26:23.

[11] Lucas Skora, Helen Senefonte, Myriam Delgado, Ricardo Lüders, and Thiago Silva. 2021. Mobilidade de Turistas Internacionais: Uma Comparação entre Dados Oficiais e de LBSN. In *Anais do V CoUrb* (Uberlândia). SBC, 112–125.

[12] Lucas E.B. Skora, Helen C.M. Senefonte, Myriam Regattieri Delgado, Ricardo Lüders, and Thiago H. Silva. 2022. Comparing global tourism flows measured by official census and social sensing. *Online Social Networks and Media* 29 (2022), 100204.

[13] David Veiga, Gabriel Frizzo, Helen Senefonte, Ricardo Lüders, Myriam Delgado, and Thiago Silva. 2020. Influências Regionais na Mobilidade de Turistas e Residentes Usando Dados de Mídia Social. In *Anais do XXXVIII SBRC* (Rio de Janeiro). SBC, Porto Alegre, Brasil, 589–602.

[14] Huy Quan Vu, Gang Li, and Rob Law. 2020. Cross-Country Analysis of Tourist Activities Based on Venue-Referenced Social Media Data. *Journal of Travel Research* 59, 1 (2020), 90–106.