Hierarchical Clustering and Measure for Tourism Profiling

Sonia Djebali[0000−0003−2249−7727], Quentin Gabot, and Guillaume Guerard[0000−0002−6773−221X]

Léonard De Vinci, Research Center, 92 916 Paris La Défense, France
l_name.f_name@devinci.fr
All authors are corresponding authors.

Abstract. Social network analysis has become widespread in recent years, especially in digital tourism. Indeed, the vast amount of data that tourists produce during their travels represents an effective source for interpreting their behaviors (geographics, demographics, psychographics, movement patterns). Since the classic measures unfit to those kind of information, this article presents a new measure to determine tourist profiles thanks to the digital traces left on social networks. This measure is based on geographic, demographic and pattern's behaviors of the tourists as the context and the content of their trips. The approach is simulated and evaluated experimentally with a hierarchical clustering on the traces left by tourists on TripAdvisor in the French capital *Paris*. Clusters found correspond to tourism segment determined by the Tourism Office of Paris.

Keywords: Tourism Profiling · Machine learning · Distance Measure.

1 Introduction

The *World Tourism Organization* recorded 1.5 billion international tourist arrivals worldwide in 2019, an increase of 4% over 2018. Tourism is responsible for 10.3% of the world's gross domestic product and is considered one of the largest and fastest-growing industries. Tourism actors such as tourist offices, cultural and commercial services analyze the behavior of tourists to know their motivations as segments, to adapt to their demands, and thus to help them make decisions [7]. Profiles are typically determined by surveys and polls. However, the emergence of social networks, such as Facebook, Flickr, TripAdvisor, and Booking, has created a new paradigm for the study of tourism profiling.

In the literature, to create tourist profiles, tourist experiences are processed and common characteristics of tourists with similar experiences are captured to extract knowledge. Profiling is mostly performed as in the case of recommendation systems by finding similar people as filtering methods.Those methods are biased because the profiles are selected in advance. In order not to induce bias, we consider that profiling should be unsupervised.

In this paper, we propose a new measure called `Tourists Profile Measure (TPM)`, used by a hierarchical clustering, to determine tourist profiles considering geographic, demographic and behavioral information left by tourist on social medias. From the TPM measure, an hierarchical clustering algorithm determines groups of tourists' stays. They are examined to extract information, as a profile and perform various comparisons between them. This method can be applied to any dataset without the need for expertise.

The main contributions of our work can be summarized as follows:

- A summary of tourists' stays based on data shared via social networks.
- `TPM`, a new measure to qualify the proximity between two tourist stays.
- A knowledge extraction of profiles.

This article is organized as follows. In section 2, we present related work on tourism profiling. In section 3, we formalize and enrich our dataset. In section 4, we present our new measure to compare the tourist experience and to generate the tourist profile using the classification method presented in section 5. Our method is implemented and is the subject of a case study on a TripAdvisor dataset in section 6. We finish with a conclusion about the presented works.

2 Literature review

Our objective in this study is to establish tourism profiles that are not biased by this preliminary choice. We seek to create profiles using an unsupervised method to extract knowledge. To achieve this goal, we must address three major challenges. The first is how to define an experience in the context of tourism; the second is how to define tourist profiles and the third is how to extract knowledge from these profiles. The literature review presented below is structured along these three axes.

To define tourist's experiences. The initial challenge of profiling tourists is to identify the key characteristics of tourist experiences. In the literature, some studies consider the demographic data of the tourist as a characteristic to achieve a classification [10]. Other studies explore other features such as interests, order of visits, semantic analysis of comments, or photo location [11]. Some studies consider stays with their context *i.e.* season, duration, weather, etc. [8]. The objective of our study is to determine tourist profiles, so we need all the information about tourists, the context of their stays, and their interests.

Define profiles. Apart filtering methods and polls, most studies use machine learning approaches (supervised and unsupervised). Concerning supervised learning, many recent studies use polls and/or social network data to improve the profiling of tourists and enrich existing (already labeled) profiles [3]. Popular classification algorithms for profile enrichment include *K-Nearest Neighbors*, *Naive Bayes* and *Support Vector Machine* [4]. However, supervised learning methods have the same biases as filtering methods. In this case, an apriori choice of profiles on which to infer the rest of the data. About unsupervised learning, studies dealing with tourism recommendation systems consider a matrix composed of the set of tourism locations and implement methods such as *Latent Class Analysis* on it [5]. However, given the diversity of tourist places, it is often unlikely to find tourists with similar visiting experiences. Many other studies group tourists based on point-of-interest ratings to find tourist preferences [6]. However, the context of the stays or the social information of the tourists is often neglected.

Extracting Knowledge. Although unsupervised learning represents a popular and useful approach, it is more difficult to handle than supervised learning. One reason is the often opaque meaning or meaningless of the clusters discovered by unsupervised learning algorithms. It is a significant challenge to extract knowledge from them and analyze it against reality.

Many studies focus on a very precise piece of information deduced from tourists' stays and ignore essential elements such as the content of the stay (points of interest visited) or the context of the stay (duration and season). In the absence of a measure that can compare all of this information, the studies focus on either the content or the context. The main contribution of the paper is a new measure dedicated to the tourism profiling.

3 Touristic Data

We focus this on the study and analysis of tourist profiles based on the digital traces left by tourists on social networks. Digital traces refer to the digital data intentionally left by tourists on these networks. Data includes information about tourists, information about the places they visit, and their interactions.

Tourists' behaviors and decisions are influenced by a set of external parameters called contextual factors. They refer to the general background within which the tourist operates, like the season, weather conditions, length of the stay, social factors, etc. Contextual factors are not present during the extraction of digital traces. Therefore, we will enrich the data set.

Tourists make a series of stays consisting of visits to various places. A stay refers to a length of time beginning with the time the tourist leaves its usual place of residence and the time the tourist leave the destination area. Each stay is a chronological succession of places that the tourist has visited. To build this set of stays, we will rely on the comments left by tourists on the networks. The method was previously presented in a previous paper [1].

Contextual factors of a stay can be of two kinds, push factors and pull factors. Push factors cause tourists to go. These include natural motivations like the climate of the home country and institutionalized ones like school vacations. Pull factors attract tourists and relate to the destination area. They include the climate of the country visited, cultural events, or sports seasons. To study tourism profiling, we will focus on pull factors. We compute season and length of stay from the stay's building.

Determining the season of the tourist's country of origin is complex due to the lack of information of its departure.We will take into account only the season of the destination deduced from the dates of the beginning and the end of a stay and the country visited. The duration of the stay is equal to the date difference between the first comment of the stay and the last comment of the same stay.

Table 1: Ontology of places.

Category	**Subcategory**
Heritage	Monuments, Parks and Gardens, Urbanism (neighborhoods, bridges, cemeteries, streets)
Cultural Buildings	Art galleries and Museums, Holy sites and Places of worship, Historic buildings, Theaters and Auditorium
Food and Services	Shops, Restaurants and Bars, Gastronomy, Hotels
Entertainment	Music buildings (concerts, discotheques), Cinemas, Amusement park, Sports
Viewpoints	(no sub-categories)
Nature	Woods, Watering place (river, lake), Beaches and Mountains

To study tourism content, we will classify tourist places based on an ontology. In the literature, many studies propose ontologies to categorize tourist places [2, 9]. We compute a resume of tgese studies in Table 1. The first level will be composed of six key categories and the second level will be composed of several subcategories. Each place belongs to at least one category and one subcategory. Note that a place can belong to several categories and subcategories.

4 Tourism Profiling Measure

To use an unsupervised clustering algorithm, we propose a measure `Tourists Profile Measure (TPM)` that allows comparing stays. Our measure is used to compute the similarity between two stays by taking into account the context and the content of the stays. The `TPM` between two stays can be seen as the sum between the context distance and the content distance, both normalized. Given S_a and S_b two stays:

$$TPM(S_a, S_b) = \ distance_{context}(S_a, S_b) \ + distance_{content}(S_a, S_b) \qquad (1)$$

The context distance is defined as an addition of the duration distance and the season distance.Let S_a and S_b be two stays with ΔS_a and ΔS_b their respective duration, p represents the normal distribution on the duration, the distance of duration between these two stays is defined as follows:

$$distance_{context_{duration}}(S_a, S_b) = |p(\Delta S_a) - p(\Delta S_b)| \qquad (2)$$

We base our season distance on the seasonal calendar.Since the seasons are cyclical, we can represent them in a cyclic graph where the seasons are nodes.Let S_a and S_b be two stays with $Season_a$ and $Season_b$ their respective seasons:

$$distance_{context_{season}}(S_a, S_b) = \begin{cases} 0 \text{ if } Season_a \text{ and } Season_b \text{ are the same node} \\ 0.5 \text{ if } Season_a \text{ and } Season_b \text{ are adjacent} \\ 1 \text{ if } Season_a \text{ and } Season_b \text{ are distant nodes} \end{cases} \tag{3}$$

Content distance. We recall that a stay contains a set of visited places. Our ontology allows us to know the number of visited places for each category and subcategory. It is composed of six subvectors corresponding of the main categories counting the monument of each subsategory. To calculate the distance between two content vectors, we sum the distance *cosinus* of each sub-vector. The *cosinus* compare the distribution of two vectors, not their magnitude wich fit with a behaviours comparison.

$$distance_{content}(S_a, S_b) = \sum_{i=0}^{n} 1 - cosine(Vec_{a_i}, Vec_{b_i})$$

where Vec_{a_i} is the ith subvector of stay S_a.

$$cosine(Vec_a, Vec_b) = \frac{Vec_a.Vec_b}{\|Vec_a\|\|Vec_b\|} \tag{4}$$

Note that we are computing a distance, so we are inverting the bounds of the cosine.

5 Creating Profiles

The unsupervised algorithm will work on the stays independently of the tourists who made them, which means that stays made by the same tourist can be in different groups. As a result, it is necessary to re-inject the tourist's demographic information into each of his or her stays. We generate the tourist profiles using a machine learning method that will consist of:

- To construct the distance matrix by calculating the distance based on the text between the stays in pairs. This matrix is symmetric.
- To use an unsupervised clustering algorithm that will take the distance matrix as input and derive groups. We use AGNES [12], a hierarchical algorithm with a Ward linkage and Elbow method for the number of clusters.
- To inject the tourists' demographic data into the groups containing at least one of his stays.

Each cluster is then analyzed to extract the tourist profile. The summary of a cluster consists in calculating: 1) the statistics on the length of the stay: average and standard variation; 2) the statistics on the cluster: average and standard variation of the numerical traces by stay and by group size; 3) the distribution of seasons; 4) the distribution of nationalities; 5) the distribution of categories and sub-categories of the content of the stays. In addition to a summary for each cluster, an overall summary of the data set is constructed. Finally, these summaries are analyzed to extract interesting information about tourist behavior to create typical tourist profiles.

6 Result and Discussion

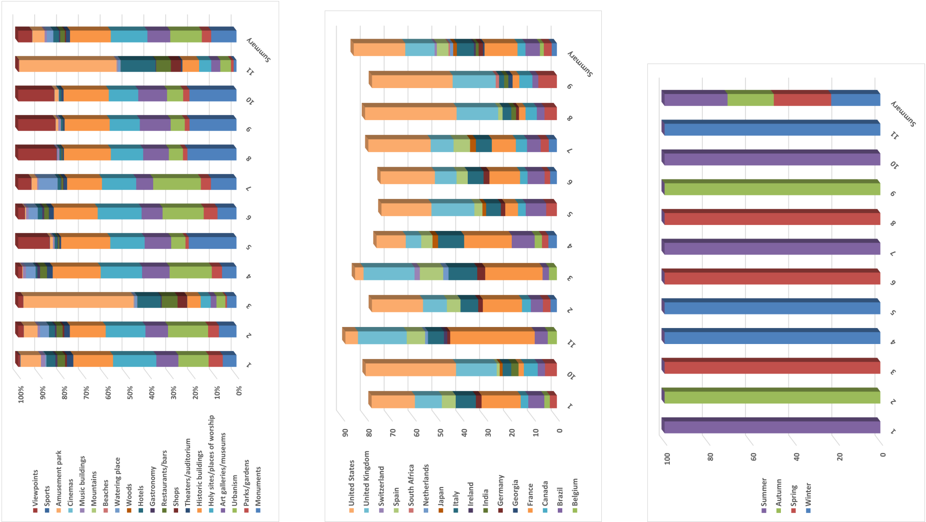
To validate our tourism profiling method, we will apply it to data from the social network *TripAdvisor* over a period from 2015 to 2018. For our case study, we have chosen the city of *Paris*, because it is one of the most attractive cities in the world, regularly ranking first among the most visited cities in the world.

Table 2: Statistics for each cluster.

Cluster	Duration		Places		Number of stays
	Mean	Standard deviation	Mean	Standard deviation	
Global	**1.92**	**2.139**	**4.935**	**2.923**	**100%**
1	1.3	0.747	3.736	2.021	5%
2	2.065	2.513	4.673	3.034	5.8%
3	1.624	1.615	3.621	2.039	7.6%
4	1.994	2.106	4.972	3.038	16.8%
5	1.332	0.761	5.435	2.353	6.3%
6	2.314	2.584	4.999	3.179	12.1%
7	4.191	3.458	6.987	3.681	4%
8	1.245	0.604	5.515	2.345	16.4%
9	1.269	0.664	5.827	2.434	4.4%
10	1.29	0.734	5.874	2.525	6.4%
11	1.565	1.422	3.532	2.093	15.2%

Our database is composed of $4,222,838$ comments distributed among $1,571,362$ tourists for a ratio of approximately 2.7 comments per tourist (with the date of the comment and the concerned monument). We compute the stays and we obtain a set of $150,306$ stays.The Elbow method returns a total of 11 clusters, we summarise them and the whole data set in Table 2.

We can notice cluster 7 represents the biggest average of duration of stays of 4.19 days with an average of visited places the most important in the event 6.98 (two more than the average) but with the lowest density per day, the cluster represents 4% of the total. We can observe in Figure 1a representing the percentage distribution of the visited subcategories in each cluster, the visits made in cluster 7 are very close to the global summary. We notice the presence of the 10 most represented nationalities in the global summary as seen in Figure 1b. From the Figure 1c, the entirety of the stays is realized during the Parisian summer period. We can conclude that cluster 7 contains tourists without any particular preferences on the visited places. These tourists tend to make/comment few vis-



(a) Distribution of categories.

(b) Distribution of nationalities.

(c) Distribution of seasons.

Fig. 1: Profiles summaries.

its to places during their long stay, which may imply a desire to take advantage of the summer sun and to enjoy the streets of Paris.

The analysis of cluster 7 is made without context, i.e. without comparisons to the results of sociological studies on tourist behaviors. For the remainder of the analysis, we will compare the profiles obtained with studies from tourist offices and sociological research on tourism. In the discussions, we will refer specifically to the public reports of the *Paris* regional tourism committee[1].

We notice, for example, that clusters 3 and 11 are mainly interested in *amusement parks* and the infrastructures that accompany them such as *hotels* and *restaurants*. Two nationalities are mainly present, *France* and *United Kingdom*. According to the Table and the Figure, both clusters come to *Paris* on average for one day and a half in winter (15.2% of all stays) and in spring (7.6%) to enjoy the amusement parks. This profile is confirmed by the reports from the Paris tourist office of French, and British tourists.

A similar observation can be made about clusters 5, 8, 9 and 10 (corresponding to the four seasons). The most represented categories of places are *Viewpoints* and *Monuments*. In terms of nationalities, countries from the anglosphere are the most present corresponding to the reports from the *Paris* tourist office.

Clusters 4 and 6 show a similar distribution of categories of places visited. In this case, an overwhelming proportion of places are related to the culture and urbanism of *Paris* for an average stay of two days. Nationalities far from *France* are more present showing the cultural appeal of *Paris* in the world. This

[1] https://pro.visitparisregion.com/chiffres-du-tourisme/profil-clientele-tourisme

tourism, having a particular attraction for indoor visits, is more dominant during the winter and spring seasons, with fewer outdoor attractions.

Clusters 1 and 2 represent a similar distribution of categories of places visited and nationalities with 5.0% and 5.8% of the total number of stays respectively with a majority of *parks/gardens*, *urbanism* and *amusement parks*. These clusters represent a summer tourism profile, privileging outdoor activities and summer attractions of *Paris* (fairs, amusement parks, music festivals).

The tourism profiles found by our method are very interesting in their accuracy with real-world data. Similar data set on the Hauts-de-France region and Nouvelle-Aquitaine region (popular region of France) have been studied in a similar way with equal relevant results.

7 Conclusion

In this article, we propose a method to discover tourist profiling. We have proposed a measure of distance based on both context and content data from tourist stays. We have shown that this measure highlights tourist profiles heretofore known in the literature, but with a finer knowledge. Our experiments demonstrate the validity of our results by comparing them to tourism management reports.Thus, the tourism industry can widely exploit our method in any geographical area without resorting to sociological studies of tourism, which are often complex to set up and must be spread over many years.

References

1. Ben Baccar, L., Djebali, S., Guérard, G.: Tourist's tour prediction by sequential data mining approach. In: International Conference on Advanced Data Mining and Applications. pp. 681–695. Springer (2019)
2. Borràs, J., Moreno, A., Valls, A.: Intelligent tourism recommender systems: A survey. Expert systems with applications **41**(16), 7370–7389 (2014)
3. Bu, N.T., Pan, S., Kong, H., Fu, X., Lin, B.: Profiling literary tourists: A motivational perspective. Journal of Destination Marketing & Management **22**, 100659 (2021)
4. Bzdok, D., Krzywinski, M., Altman, N.: Machine learning: supervised methods. Nature methods **15**(1), 5 (2018)
5. Jia, Z., Yang, Y., Gao, W., Chen, X.: User-based collaborative filtering for tourist attraction recommendations. In: 2015 IEEE international conference on computational intelligence & communication technology. pp. 22–25. IEEE (2015)
6. Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Li, X.: Efficient user profiling based intelligent travel recommender system for individual and group of users. Mobile Networks and Applications **24**(3), 1018–1033 (2019)
7. March, R., Woodside, A.G.: Tourism behaviortravelers' decisions and actions. No. G155. A1 M2655 2005, Ovid Technologies, Inc. (2005)
8. Massimo, D., Ricci, F.: Clustering users' pois visit trajectories for next-poi recommendation. In: Information and Communication Technologies in Tourism 2019, pp. 3–14. Springer (2019)

9. Moreno, A., Valls, A., Isern, D., Marin, L., Borràs, J.: Sigtur/e-destination: ontology-based personalized recommendation of tourism and leisure activities. Engineering applications of artificial intelligence **26**(1), 633–651 (2013)
10. Refanidis, I., Emmanouilidis, C., Sakellariou, I., Alexiadis, A., Koutsiamanis, R.A., Agnantis, K., Tasidou, A., Kokkoras, F., Efraimidis, P.S.: myvisitplanner gr: Personalized itinerary planning system for tourism. In: Hellenic Conference on Artificial Intelligence. pp. 615–629. Springer (2014)
11. Rodríguez, J., Semanjski, I., Gautama, S., Van de Weghe, N., Ochoa, D.: Unsupervised hierarchical clustering approach for tourism market segmentation based on crowdsourced mobile phone data. Sensors **18**(9), 2972 (2018)
12. Struyf, A., Hubert, M., Rousseeuw, P., et al.: Clustering in an object-oriented environment. Journal of Statistical Software **1**(4), 1–30 (1997)