Tourists Profiling by Interest Analysis

Sonia Djebali[1], Quentin Gabot[1,2], and Guillame Guérard[1]

[1] Léonard De Vinci, Research Center, 92 916 Paris La Défense, France
f_name.l_name@devinci.fr
[2] Research Student f_name.l_name@edu.devinci.fr

Abstract. With the recent digital revolution, analyzing of tourists' behaviors and research fields associated with it have changed profoundly. It is now easier to examine behaviors of tourists using digital traces they leave during their travels. The studies conducted on diverse aspects of tourism focus on quantitative aspects of digital traces to reach its conclusions. In this paper, we suggest a study focused on both qualitative and quantitative aspect of digital traces to understand the dynamics governing tourist behavior, especially those concerning attractions networks.

Keywords: Data Mining · Measure of Interest · Profiling· Big data · Social network · Tourism

1 Introduction

Nowadays, tourism industry is considered as one of the largest and fastest growing industries [9]. In 2019, *World Tourism Organisation* UNWTO[3] has recorded 1.5 billion international tourists, 4% more than the previous year.

With the recent booming of digital tools and mobile internet technology, alternative sources of data to understand tourism behaviors and to capture tourists' experiences have emerged. Users of social networks, tend to share openly and frequently photos, videos, reviews and recommendations of their travels and their experiences. Thus, when users share photos or reviews, geographical information is included. These geo-located data represent tourism and sociological views [7]. These social networks constitute an interesting observation field to analyze tourists' behavior evolution through time and space.

In social networks, graph models are widely used to represent an interaction between entities and their relations. A graph model allows to understand both local phenomena (i.e. node level) as well as global phenomena (i.e. graph level). One of an interesting way mostly used to analyze movement networks is to evaluate and detect community structure [20]. Detecting communities in networks provides a means of coarse-graining interactions between entities and offers a more interpretable summary of a complex network.

In literature, studies carried out on social networks use a quantitative dimension of digital traces as a research hypothesis [14]. Models developed focus on

---

[3] International tourism Growth continues outpacing the global economy: edition 2020.

the question of the number of tourists' without considering a qualitative analysis and the interest of users to make a given action. Interest analysis allows us to identify trends that are not perceptible with quantitative analysis. Being able to define another kind of relationship between two places visited in a city will not give the same result as observing flows of tourists between these two places.

To better understand tourists movements and their interest to consider this movements, we propose in this paper to detect tourist communities to perform tourist profiling based not on the quantitative aspect but also on the qualitative aspect i.e. the interest of the tourists to make movements or visits. Our study takes advantage of existing quantitative methods to analyze how of tourists perform an action. Then, we enhance them with qualitative methods to analyze why tourists perform this action. Thanks to a community detection and profiling, stakeholders will be able to better identify behaviors and thus target communities by proposing adapted excursions at the best places.

Our key contributions can be summarized as : **(1)** an automatic `Graph Movement` extraction methodology dedicated to graph of interest, **(2)** the `Measure of Interest` using an interest measure as weighted in a graph. This measure is not dependent on frequency or probability based on number of tourists, **(3)** `Spheres of influence` to compare a neighborhood of each visited place to determine a similarity matrix, **(4)** `Community detection` by a clustering on the similarity matrix provide profiles. A thorough study of tourists' profiles is compared to tourism management works.

In this paper, we will first relate in Section 2 comparable works on community detection network analysis. In Section 3 we formalize our graph data model. In Section 4 we present our approach to perform a community tourist detection and tourist profiling. Our model is implemented and is subject of a case study on a TripAdvisor dataset in Section 5. Finally, the Section 6 concludes this paper.

2   State of the Art

Social networks are a social structure where a set of social actors are connected by relationships. Due to the size of social networks with numerous and heterogeneous relationships, it has been becoming difficult to analyze them. Therefore, several algorithms and graph theory concepts are used to study a structure of social networks [16]. Most researches aiming at analyzing social networks enable to understand different social phenomena including social structures evolution, to measure an importance of nodes, to detect communities of nodes [1] that share some characteristics by looking metrics at the whole network cohesion [17], etc.

Social networks are paradigmatic examples of graphs with communities. The word community itself refers to a social context. The concept of community has no unified definition [18], most of the researchers have reached a consensus that communities in a network indicate groups of nodes, such that nodes within a group are connected more often than those across different groups [30]. Detecting community in networks is very hard and not yet satisfyingly solved, despite an

huge effort of a large interdisciplinary community of scientists working on it over the past few years.

In the literature, several researches have been interested to detect of community in social networks to analyze and understand tourist behaviors. To detect tourists community, social networks are modeled as a directed graph with a weight attached to each edge. These values of weights can be representing, for example, frequencies of movement of tourists, probabilities value, rules metrics, etc. Based on this graph model, a detection community method and approach exposed below have been studied by many research fields.

Graph Partitioning: This method consists to divide a graph into groups of predefined size, such that the number of links in a cluster is denser than the number of edges between the clusters [10]. The disadvantage of this method is to define the size of clusters in a relevant way.

Hierarchical Clustering. Hierarchical clustering techniques are based on a nodes similarity measure [12]. Theses techniques don't need a predefined size and number of communities. Hierarchical clustering techniques can be categorized into two classes. *Agglomerative algorithms* starts by considering each node of a graph as a separate cluster and iteratively merge them based on high similarity and ends up with an unique community. And *Divisive algorithms* start by the entire network as a distinct cluster and iteratively splits it by eliminating links joining nodes with low similarity and ends up with unique communities.

Modularity Optimisation Based. Modality, is a quality function introduced by Newman and Girvan [19] to measure the quality of a partition of graph nodes. The larger the modularity value the better is the partition. The best-known methods within are: *Greedy method of Newman* is an agglomerative method. Initially, each node belongs to a distinct module, then they are merged iteratively based on the modularity gain. And *Blondal's Louvain algorithm* is an heuristic greedy algorithm for uncovering communities in complex weighted graphs [8]. It is based on the modularity optimization.

Statistical inference-based methods. These methods deduct properties of datasets, starting from a set of observations and model hypotheses. If the dataset is a graph, based on hypotheses on how nodes are connected to each other, the model has to fit the actual graph topology [10]. These methods attempt to find the best fit in a model to the graph, where the model assumes that nodes have some sort of classification, based on their connectivity patterns.

All the methods cited above are used to detect communities on network models as a directed weighted graph. Where weights values represent frequencies, probabilities, labels, vectors, etc. The probability weights is deduced from frequencies or is computed, thanks to data mining, by considering relation between nodes as rules. Those methods implement a measure like support, confidence or lift and return rules which satisfy some constraints. To determine communities, one can analyze a graph as a Markov Chain [2] Most advances methods use adapted clustering or partitioning such as spectral clustering [29], label propagation [23].

In the tourism research field, methods used for community detection are focused on detecting community based of the quantitative dimension i.e. frequencies movement of individuals or flows in a graph. These methods form clusters based on clustering metrics (Silhouette, Dunn, etc.). However, the interpretation of clustering contents is unrepresentative of a community sharing a same interest or a profile but rather a mass of data. In this study, we will focus on detection community to extract tourist profiling base on the interest of tourists to make a movement or visit by integrating the notion of overtourism [5].

3   Tourism Movement's Data Model

To model tourism movement, we must consider visited locations, users information and their interactions. To study interactions on locations, we propose to model tourist data by a graph of movement. Graphs rely on links between users and locations through their reviews.

Data Types Our database is composed of users, reviews, and geo-located locations. A location is composed of longitude and latitude coordinates, type of location (hotel, restaurant, attraction) and a rating. Each location has been aligned with administrative areas (GADM)[4]. A user is identified by nationality, age and is described by a timeline. A user timeline represents a chronological set of reviews from its first reviews to its last reviews. This timeline allows computing intermediate properties like time between two consecutive reviews i.e. consecutive visited places. A review represents a note given by a user on a location at a given time.

3.1   Sequences Dataset

To study tourists' movement of a given destination, we need to target tourists. For this, we focus only on users who visit at least one destination. Then we extract all their reviews to gather their circulation all over the world. Tourists review several locations during their trip. A trip is a succession of days when a tourist writ at least one review per day, i.e. as soon as there is a day when a tourist does not write a review, its trip is considered broken. However, a tourist may not review during a limited period during the same trip. The break can be canceled between two trips that took place in the similar country by a user. We consider a sequence is composed of reviews written at most at 7 days apart [13]. The method consists of merging two trips if they satisfy the following conditions: $\Delta B \leqslant \Delta T_i$ and $\Delta B \leqslant \Delta T_j$ and $L_{T_i} = F_{T_j}$. Where $\Delta B \leq 7$ represents break's duration, ΔT_i and ΔT_j present i^{th} and j^{th} trip's duration, F_{T_j} represents the first country visited during j^{th} trip and L_{T_i} represents the last country visited during i^{th} trip. For each tourist, we build a set of trips based on its reviews. Each trip corresponds to sequence and presents a succession of locations in a temporal order of visit during a trip.

---

[4] GADM:https://gadm.org/index.html. $386,735$ administrative areas (country, region, department, district, city, and town).

3.2   Sequential Rule Mining

Based on sequences dataset, we use a sequential rules mining algorithm to discover and to extract all existing rules in the dataset. The input structure is a set of sequences, and the output structure is the rules. A sequential rule is a rule of the form $X \longrightarrow Y$ where X and Y are items. In our case a rule $X \longrightarrow Y$ is interpreted as, if location X occurs then it will be *directly followed* by the location Y. To exploit the sequential rules, two measures are predominantly employed, *support* and *confidence*. The support of a rule $X \longrightarrow Y$ is how many sequences contains the location from X followed directly by the location Y. The confidence of a rule $X \longrightarrow Y$ is the support of the rule divided by the number of sequences containing the location X.

   To discover sequential rules appearing in sequence dataset, we use a *TRule-Growth* algorithm [11] for its performance. The TRuleGrowth algorithm uses a pattern-growth approach for discovering sequential rules such that it can be much more efficient and scalable. After, extracting all the rules from the sequences dataset, we use an *Interest measure* to value interesting and useful rules for an effective decision making.

3.3   Measure of interest

To value interesting rules, there exist several measures developed in divers fields such as machine learning, social science, statistics, data mining, etc. Selecting a right measure for a given application is difficult because many measures may disagree with each other. Each measure has several properties that make them unsuitable for other applications. In our case, the key properties one should consider selecting an accurate measure are:

1. The measure must distinguish the movement between location X and Y with that between location Y and X. The measure must be **asymmetric**.
2. To understand the interest of tourists to perform a given movement, the measure must be **antisymmetry** to express the positive and the negative correlation resulting of the movement between two locations.
3. The measure must have a positive value if there is a positive correlation and equals zero if items are **statistically independent**.
4. Considering size of the dataset, the measure must be easy to implement, readable, editable and context-free with a **limit scale**. The measure must have an upper bound and a lower bound.
5. The quantity dimension mustn't be ignored. The measure will scale in a limited way with it.

   The measure respects and corresponds to all these properties is *Klosgen* [15]. The Klosgen measure is a normalization of the *Added Value measure*, also called *Pavillon Index* or *Confidence*, [25]. For a given rule $X \longrightarrow Y$ Added Value $AV(X \Rightarrow Y)$ measure quantifies how much the probability of Y increases when conditioning on the transactions that contain $X : AV(X \Rightarrow Y) = confidence(X \Rightarrow Y) - support(Y))$. Where, $confidence(X \Rightarrow Y) = \frac{|X \Rightarrow Y|}{|X|}$ and
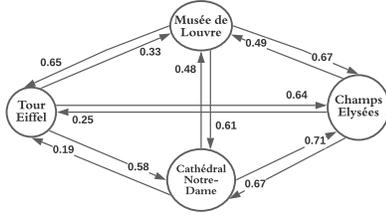
Fig. 1. Directed weighted graph G(V,E,W) with Klosgen values of the four most visited places of Paris.
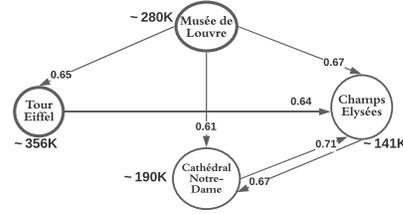
Fig. 2. Weighted graph of the top 2 mainstream nodes with a Klosgen threshold at 0.6.

$support(X) = |X|$. $|X|$ denotes the frequency of X in the database. The Added Value of a rule can be positive, negative or zero. The rules with a high Added Value are considering as interesting rules. Since the Added Value vary greatly depending of the support of the antecedent X and the consequent Y, we normalize the Added Value measure by multiplying it by its squared root which gives Klosgen measure. Then, we compute the Klosgen measure for each rule as follows: $Klosgen(X \Rightarrow Y) = \sqrt{supp(X \Rightarrow Y)} * AV(X \Rightarrow y)$

Using the Klosgen measure has two main advantages. The first is that, the influence of support or confidence is lowered, which means that very low or very high support of the antecedent or the consequent have a low influence on the value of the measure (unlike lift or confidence) [27]. The second is that the Klosgen measure is not a complex measure, which makes its meaning relatively clear, since it is presented as the normalization of the Added Value. The Added Value being itself part of the basic objective interest measures.

3.4   Graph movement model

Based on the set of rules and their interest value calculating by Klosgen, we build a movement graph of tourists. The movement graph $G(V, E, W)$ is a weighted directed graph. The nodes V are a set of the antecedents and the consequent of the rules. These nodes present the locations visited by the tourists. The arcs E are the association of the rules and present two direct consecutive locations. The weighted W of the arcs is the values of the *interest* measure of rules that represent interest of the tourists to move between two places i.e. two nodes. The figure 1 shows an example of graph of movements of tourist fours most visited place in $Paris$.

4   Community Detection

Based of our graph model, to detect community, we first identify a **mainstream nodes**, the nodes having a most influential and strong attractiveness. Secondly, we define a **neighborhood** of mainstream nodes. The neighbors of a node consist

in a set of nodes that are connected to this node up to a *certain distance*, i.e., the number of steps between a source node and its neighbors. Then, for each mainstream node we build a **Sphere of Influence**. Finally, we define a **similarity measure** between spheres of influence, this measure is used to create **clusters** and then detect communities.

4.1   Mainstream nodes

We will define two kinds of nodes, *mainstream nodes* having the most influential on the graph and *secondary nodes* the other remaining nodes. Based on graph of tourist movements $G(V, E, W)$, for each node of V we calculate its support (i.e. number of occurrences) on the sequences dataset, described in section 3.1. We select a subset of node V' having a K maximum value of support. The value of k is fixed according to the number of nodes in the graph and the distribution of the support amount all the nodes. We note $V'' = V \setminus V'$ the secondary nodes as the rest of the nodes which are not mainstream.

Once mainstream nodes and secondary nodes are defined, to perform community detection, we extract a subgraph $G'(V, E', W)$ formed with $V = V' \cup V''$, and arcs $E' \subseteq E \wedge ((v1, v2) \in E \rightarrow W_{v1,v2} > threshold)$. A low weight present a low interest of movement while the most significant weights present the interest of moving from one node to another.

The figure 2 presents subgraph of the graph in figure 1. Where the bold nodes *Musée de Louvre* and *Tour Eiffel* are the mainstream nodes with their respective support value of $280k$ and $356K$. The Klosgen value threshold is taken at 0.6.

4.2   Spheres of influence

In graph theory a sphere of influence of the graph assigns, for each node, a ball centered at that node of neighbors distance D equal to a definite value. In this study, we create a sphere of influence to all mainstream nodes. A sphere of influence of a mainstream node is an aggregation of its neighboring nodes (mainstream and secondary).

The neighbor distance D of an influence sphere equals to the average length of sequences in the sequences dataset minus one. To process a sphere of influence for a mainstream node $S_m i$, we create an array A_i for each mainstream node m_i. The array contains all neighbors of distance 1 to D from m_i. The table 1 shows the result of creating the influence spheres of two mainstream nodes $m_{Tour\ Eiffel}$ and $m_{Musee\ du\ Louvre}$ the graph figure 2.

4.3   Similarity measure

Once all influence spheres are created, we will compare them in pair. Comparing two influence spheres consist to analyze all the nodes composing them and report the percent of similitude into a matrix.

Let M be the matrix of similarity of size $|V'|$, referring to the number of mainstream nodes. The value $M(i, j)$ equal to the percent of nodes in $S_m i$ (i.e.

Table 1. Sphere of Influence of the two mainstream nodes of Figure 2.

Mainstream node	$m_{Musee\ du\ Louvre}$	$m_{Tour\ Eiffel}$
Neighbors at distance = 1	$A_{Musee\ du\ Louvre} =$ {Tour Eiffel, Cathédral Notre-Dame, Champs Elysées }	$A_{Tour\ Eiffel} =$ {Champs Elysées}
Neighbors at distance = 2	none	$A_{Tour\ Eiffel} =$ {Champs Elysées, Cathédral Notre-Dame}

the influence sphere of a mainstream node m_i) that are also present in nodes $S_m j$ (i.e. the influence sphere of a mainstream node m_j). Thus, the matrix is non-symmetric. $M(i,j) = 1$ means that all the nodes in $S_m i \subseteq S_m j$ and $M(i,j) = 0$ means that $S_m i \cap S_m j = 0$. Let us take for example the results in table 1. The matrix of similarity M is as follows:

$$M = \begin{matrix} m_{Musee\ du\ Louvre} \\ m_{Tour\ Eiffel} \end{matrix} \begin{pmatrix} \overset{m_{Musee\ du\ Louvre}}{-} & \overset{m_{Tour\ Eiffel}}{0.66} \\ 1 & - \end{pmatrix}$$

4.4 Profiling

Based on the similarity matrix, we detect community for purpose to create a tourism profile. The matrix is seen as a graph for the profiling process. In this study, we are looking for a community detection based on high similarity values and high density of arcs inside communities compared to that of arcs connecting to outside communities. This distinction between arcs inside the communities and outside is called *modularity*. Then we must optimize the modality in order to have the best possible grouping into communities. To achieve this, we use the *Louvain clustering* [4] which maximizes modularity for each community.

5   Experiments

We conducted experiments on reviews posted by tourists between $1st$ January 2013 to $31th$ December 2016 in `Tripadvisor` focused on Paris, French regions. After cleaning and pre-processing the data, we obtain $1'666'584$ trips of more than 3 reviews, with an average length of stay of 2.5 days and an average number of reviews per stays at 4.14 with close to $40'000$ places of Paris.

5.1   Measure of interest

In our study, values of weights of the graph are computed using *Klosgen* (*Kl*) measure. To compare our results with others methods, we also provide the following measures: *Support* (*Supp*): depends on the number of occurrences of the

rule $X \longrightarrow Y$, $Confidence$ $(Conf)$: is a ratio between the support of the rule on the support of X, as a probability of a Markov chain, $Lift$: provides a value showing how different of a random process the rule is, $Certainty$ $Factor$ (CF): a high Certainty Factor means Y is dependent of X and not another node, $J - Measure$ $(cross - entropy)$ (J): computes the cross-entropy, i.e. the information contain in the rule over all the rules and $Conditional$ $Entropy$ (CE): compute the correlation between X and Y based on the entropy value.

The table 2 presents the top 3 rules based on Support and top 3 rules based on Klosgen. Nodes representing a monuments *Tour Eiffel*, *Musée du Louvre*, *Cathédrale Notre-Dame* and other well-known monuments will be overrepresented with Support. Confidence and Lift highest values are a combination of a high support node with one with a very few support. Same for Certainty Factor, the values between the highest support nodes or node presents in few rules are very low. J-Measure are closed to zero for all rules since rules are numerous for a fixed X or a fixed Y. Moreover, the large majority of Conditional Entropy are closed to zero.

As shown, Klosgen measure is not influence by the support of nodes, nor by the number of nodes (entropy tends to zero). Since the value is between 0 to 1, Klosgen measure is easily readable and understable. This measure fits perfectly with the notion of *interest*.

5.2   Mainstream monuments

To select K mainstream monuments of Paris, we initially order monuments by decreasing order of support and then we compute the cumulative number of reviews per monument of Paris.The curve displays an elbow which distinguishes on the left side mainstream monuments to secondary monuments. We set the number of mainstream monuments when the elbow start at $K = 111$ which represents approximately 73% of all reviews.

5.3   Sphere of influence

To create influence spheres, we first start in mainstream monuments then aggregate the neighbors nodes with distance of $D = 3$, average length of sequences

Table 2. Measures' results on a sample from the database.

Rules	Kl	$Supp$	$Conf$	$Lift$	CF	J	CE
Tour Eiffel $\longrightarrow$ Musée du Louvre	0.33	**49565**	**0.18**	0.58	-0.18	**0.03**	**0.68**
Tour Eiffel $\longrightarrow$ Cathédrale Notre-Dame	0.57	**26129**	0.09	0.51	-0.11	**0.02**	**0.45**
Cathédrale Notre-Dame $\longrightarrow$ Tour Eiffel	0.18	**23092**	**0.20**	0.45	-0.43	**0.04**	**0.71**
Le Bataclan $\longrightarrow$ Place de la République	**0.86**	300	0.09	**2.53**	**0.06**	0	0.44
Hôtel Plaza Athénée $\longrightarrow$ Hotel George V	**0.85**	297	0.05	**4.40**	**0.04**	0	0.29
Arc de Triomphe $\longrightarrow$ Sacré-Coeur	**0.85**	1285	0.05	**1.47**	**0.02**	0	0.27

in the dataset, in our case 4, minus one. The similarity matrix and the Louvain clustering are compute on the 111 corresponding nodes with a Klosgen's threshold = 0.1. Since the graph have about $40'000$ nodes, with 111 mainstream nodes, we do not provide a figure presenting the result. Concerning the similarity matrix, its size is $111 * 111$, thus we can't show it in this paper.

5.4   Clustering Analysis

After applying the Louvain clustering algorithm, we obtain 4 clusters of similar size, around twenty nodes. We do not regard the 5 clusters of size less than 3 and clusters with singleton. We note that tourist's profiles are bounded to an unique cluster and they do not overlap.

   Cluster[1] contains mainly architectural monuments, some can be visited but they are principally known for their aspect like the *Pyramide du Louvre*, *Grand Palais* or *Opéra Garnier*. This cluster contains the main bridges of Paris and typical district like *Montmartre* or *Le Marais*. One monument radically differs is the famous tearoom *Ladurée*. But this restaurant is very close to the monuments of this cluster, which can explain its presence. The profile of this cluster is *Architectural tourism* [26] and *Photography tourism* [22].

   Cluster[2] contains architectural monuments and cultural monuments like museums (mostly about history of the country). Some of those monuments also have a religious context like the *Cathédrale Notre-Dame* or *Basilique Sacré-Coeur*. The outsiders are *Tour Eiffel* and *Jardin du Luxembourg*, those monuments are most recent than the others but have a strong cultural aspect. The profile of this cluster is *Heritage tourism* [21] and *Religious tourism* [24].

   Cluster[3] contains mostly sport complexes, stadium and cabarets. The outsiders are the restaurant *Le Perchoir*, the train stations *Gare de Lyon* and *Gare de l'Est*. The profile of this cluster is *Sport and Event tourism* [6] and *Recreation tourism, Urban tourism* [28].

   Cluster[4] contains the *Grands Magasins de Paris* (Paris department stores), luxury stores and luxury hotels. This cluster clearly refers to shopping and the luxury of Paris, which is a significant attractive aspect of the city. The outsiders are restaurants, but are referred as gastronomic restaurants which refer to luxury. Obviously, this cluster refers to the profiles of *Luxury tourism* and *Culinary tourism* [3].

5.5   Discussions

Impact of Klosgen's threshold. Since the Klosgen's threshold determines if an arc is kept in the graph, an upper value increase the number of clusters and the number of singleton. When increasing the threshold value, the nodes in a cluster remains together or are split into two or more clusters. The threshold does not modify the profiles but only the number of profiles found.

   Clustering without computing spheres of influence. When we compute the Louvain algorithm to the mainstream monuments with Klosgen values

without doing the spheres of influence, we obtain 4 clusters (and some singleton). Some trends can be extract to each cluster, with several outsiders. The trends are in order: *Cultural tourism (Heritage)*; *Cultural tourism (Modern Art)* and *Luxury tourism*; *Sport and Event tourism*; *Luxury tourism*. The profiles are less distinct without computing the spheres of influence and some kind of tourist's profiles are present in many clusters.

Clustering with other measures. We also implement the method with support, confidence and lift measures instead of Klosgen measure. The clusters do not have any main trend but are mostly influenced by the metro lines and other public transports of Paris. Clusters reflect also the overtourism in Paris with a cluster with the top 23 most visited monuments of Paris.

6   Conclusion

When scientists study tourism thanks to social networks data, they used measure based on the frequency to understand the tourists behaviors. We provide a method to extract tourism behaviors using an interest Klosgen measure and by adapting the notion of neighborhood used in social graph to the tourism industry. Our method returns significant results, already observed by tourism management studies in a more limited scale. In future work, we have to enhance our method to refine the profiling, by using a fuzzing clustering to improve the clusters' knowledge discovery. We will additionally use the method to other cities and a vaster area of France to confirm the validity and efficiency of the proposed method.

References

1. Arenas, A., Danon, L., Diaz-Guilera, A., Gleiser, P.M., Guimera, R.: Community analysis in social networks. The European Physical Journal B **38**(2), 373–380 (2004)
2. Baccar, L.B., 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)
3. Batat, W.: The role of luxury gastronomy in culinary tourism: An ethnographic study of michelin-starred restaurants in france. International Journal of Tourism Research **23**(2), 150–163 (2021)
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment (10), P10008 (2008)
5. Capocchi, A., Vallone, C., Pierotti, M., Amaduzzi, A.: Overtourism: A literature review to assess implications and future perspectives. Sustainability **11**(12), 3303 (2019)
6. Chalip, L., Costa, C.A.: Sport event tourism and the destination brand: Towards a general theory. Sport in society **8**(2), 218–237 (2005)
7. Chareyron, G., Cousin, S., Da-Rugna, J., Gabay, D.: Touriscope: map the world using geolocated photographies. In: IGU meeting, Geography of Tourism, Leisure and Global Change (2009)

8. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical review E **70**(6), 066111 (2004)
9. Cooper, C., Hall, C.M.: Contemporary tourism. Routledge (2007)
10. Fortunato, S.: Community detection in graphs. Physics Reports **486**(3-5), 75–174 (Feb 2010)
11. Fournier-Viger, P., Wu, C.W., Tseng, V.S., Nkambou, R.: Mining sequential rules common to several sequences with the window size constraint. In: Canadian Conference on Artificial Intelligence. pp. 299–304. Springer (2012)
12. Friedman, J., Hastie, T., Tibshirani, R., et al.: The elements of statistical learning, vol. 1. Springer series in statistics New York (2001)
13. Gössling, S.: Tourism, tourist learning and sustainability: An exploratory discussion of complexities, problems and opportunities. Journal of Sustainable Tourism **26**(2), 292–306 (2018)
14. Kaufmann, M., Siegfried, P., Huck, L., Stettler, J.: Analysis of tourism hotspot behaviour based on geolocated travel blog data: The case of qyer. ISPRS International Journal of Geo-Information **8**(11), 493 (2019)
15. Klösgen, W.: Problems for knowledge discovery in databases and their treatment in the statistics interpreter explora. International Journal of Intelligent Systems **7**(7), 649–673 (1992)
16. Knoke, D., Yang, S.: Social network analysis, vol. 154. Sage Publications (2019)
17. Kolaczyk, E.D., Csárdi, G.: Statistical analysis of network data with R, vol. 65. Springer (2014)
18. Lancichinetti, A., Fortunato, S.: Community detection algorithms: a comparative analysis. Physical review E **80**(5), 056117 (2009)
19. Nicosia, V., Mangioni, G., Carchiolo, V., Malgeri, M.: Extending the definition of modularity to directed graphs with overlapping communities. Journal of Statistical Mechanics: Theory and Experiment (2009)
20. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. nature **435**(7043), 814–818 (2005)
21. Park, H.Y.: Heritage tourism. Routledge (2013)
22. Picard, D., Robinson, M.: The framed world: Tourism, tourists and photography. Routledge (2016)
23. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical review E **76**(3), 036106 (2007)
24. Rashid, A.G.: Religious tourism–a review of the literature. Journal of Hospitality and Tourism Insights (2018)
25. Sahar, S., Mansour, Y.: Empirical evaluation of interest-level criteria. In: Data Mining and Knowledge Discovery: Theory, Tools, and Technology. vol. 3695, pp. 63–74. International Society for Optics and Photonics (1999)
26. Specht, J.: Architectural tourism: building for urban travel destinations. Springer (2014)
27. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems **29**(4), 293–313 (2004)
28. Tribe, J.: The economics of recreation, leisure and tourism. Routledge (2020)
29. Von Luxburg, U.: A tutorial on spectral clustering. max planck institute for biological cybernetics. Tech Rep (2006)
30. Xia, X., Zhu, S.x.: A survey on weighted network measurement and modeling. TELKOMNIKA Indonesian Journal of Electrical Engineering **11**(1), 181–186 (2013)