ACCEPTED MANUSCRIPT

Quality & Quantity
DOI: 10.1007/•••••-•••-•••-••••-•

SeqPatTour: Sequential Patterns-based Metrics for Better Understanding the Tourism Behaviour

Hugo Alatrista-Salas[1] · **Gaël Chareyron**[1,2] · **Sonia Djebali**[1] · **Imen Ouled Dlala**[1] · **Nicolas Travers**[1]

© **The author(s) ••••**

Abstract Analysing and understanding tourism mobility constitutes a tremendous challenge for states, cities and companies involved in cultural heritage. Traditional methods of tourist flow analysis often rely on survey data and observational studies to map out general trends and movements, focusing on aggregated visitor numbers and basic demographic information without delving deeply into the complex interconnections and sequential patterns of tourist behaviours. This study introduces *SeqPatTour*, a comprehensive methodology that innovatively mines and analyses tourist behaviour through sequential pattern analysis with data from `Tripadvisor`. Distinguished by its ability to perform both quantitative and qualitative analysis, *SeqPatTour* navigates the complexities inherent in analysing sequential patterns on graphs, facilitating a deeper understanding of tourist movements. It critically assesses path extraction and sequential pattern mining techniques to tailor the analysis to specific temporal constraints, thereby introducing specialised metrics for the tourism domain. Our methodology, applied on the city of Lille with 287,429 reviews, leading to a graph of 3,317 nodes and 94,829 relationships, has the potential to enhance analytical capabilities in tourism studies, providing actionable insights for urban planning and policy formulation. Its approach to integrating path extraction with

Accepted manuscript

✉ H. Alatrista-Salas
hugo.alatrista_salas@devinci.fr

G. Chareyron
gael.chareyron@devinci.fr

S. Djebali
sonia.djebali@devinci.fr

I. Ouled Dlala
imen.ouled_dlala@devinci.fr

N. Travers
nicolas.travers@devinci.fr

[1] De Vinci Higher Education, De Vinci Research Center, Paris, France

[2] EA EIREST, Université Paris 1 Panthéon Sorbonne, Paris, France

ACCEPTED MANUSCRIPT

temporal-gap constrained sequential pattern mining effectively uncovers latent behavioural patterns among tourists. This not only underscores the method's value in improving tourist experiences but also its contribution to economic development in tourism-focused areas, navigating through the intricacies of sequential pattern analysis on graph structures to offer profound insights into tourist dynamics.

Keywords: Tourism behaviour; Sequential pattern mining; Directed multigraph; Metrics for tourism analysis

1. Introduction

Tourism is a rapidly expanding sector of the global economy, contributing significantly to GDP growth, job creation, and infrastructure development in developed and developing countries. According to the UNWTO[1] World Tourism Barometer, international tourism generated an estimated USD of 1.7 trillion in 2024. Additionally, total export revenues from tourism, including passenger transport, were estimated at USD 2 trillion in 2024.

France is the world's most popular tourist destination. It offers famous attractions like the *Eiffel Tower*, the *Musée du Louvre*, the Palace of *Versailles*, and the Notre-Dame Cathedral. Beyond Paris, visitors can explore charming villages, mountains, vineyards, and castles across cities such as Bordeaux, Lyon, Marseille, Lille, and Nantes. Tourism in France stimulates economic growth by spending on accommodations, dining, entertainment, and transportation[2]. It can also promote cultural diversity and mutual understanding among countries.

Therefore, policymakers, urban planners, and stakeholders in the tourism industry must understand tourist mobility. Furthermore, the dynamics of tourist movements within metropolitan surroundings constitutes a key idea to tackle this issue. With this in mind, graph pattern mining techniques offer a powerful framework for analysing complex networks of tourist circulation. Tourist movement can be represented as a graph (Chareyron, Quelhas, and Travers, 2020), where vertices represent locations and edges represent a connection between them (a trip, comments on social networks, etc.). We can then "mine" this graph to find hidden patterns and trends that aren't obvious with traditional methods.

By identifying these spatial-temporal patterns, we can see how people interact with a destination. These patterns are shaped by various factors, including the distribution of attractions, availability of amenities, accessibility via transportation networks, and geographic distances. Understanding tourist behavior in detail is essential. It helps identify which sites visitors go to and analyze their movement patterns. This includes how long they stay at each location and how different demographic or national groups interact with the layout of destinations. These insights are key ideas in supporting effective destination management, informing infrastructure development, and enabling personalized

[1] World Tourism Barometer - Volume 23 - Issue 2 - May 2025 (Excerpt)

[2] Tourism represents 7.5% of the French GDP [CampusFrance2023]

Accepted manuscript

ACCEPTED MANUSCRIPT

recommendation systems. These assumptions raises the following question: How can sequential pattern mining on graph-based multidigraphs reveal temporal and behavioral regularities that traditional trajectory analysis fails to detect?

This paper proposes *SeqPatTour*, a comprehensive, data-driven methodology that goes beyond basic trajectory analysis. SeqPatTour extracts temporally aware and behaviorally relevant sequential patterns from tourist activity traces. The goal is to identify how tourists sequentially explore Points of Interest (POI) while considering temporal dynamics, the interconnectedness of places, and tourist profiles. Our approach uses a multigraph model, where nodes represent places (e.g., attractions and landmarks) and edges capture tourist transitions. These edges are annotated with metadata such as timestamps, nationality, and visit frequency. This representation allows for a more expressive, multi-perspective analysis of mobility that captures not only physical movement, but also group-specific behaviors. This paper presents the following main contributions:

- We propose a five-step *SeqPatTour* methodology, that combines trajectory segmentation, path extraction, sequential pattern mining with temporal constraints, and pattern filtering based on tourism-specific relevance criteria, from the `Tripadvisor` dataset,
- We discuss and compare various path extraction and sequential pattern mining techniques to select the best ones for our method
- We developed seven new metrics to characterize the extracted patterns based on their temporal coherence, recurrence, diversity, and spatial dispersion
  These metrics help destination managers answer questions such as: *Which places are visited in consistent sequences? Do certain nationalities prefer specific exploration paths? How long do tourists spend between stops?*
- To showcase our approach, we conducted experiments in Lille, France, to quantify and qualify the proposed methodology and its measures.

The rest of this work is organised as follows. Section 2 presents the state-of-the-art contributions related to data mining in tourism behaviour. Section 3 details the main concepts related to graph theory and pattern mining. Section 4 describes the methodology proposed in this paper, while Section 5 shows experimentation, results and a discussion on the methodology. Section 6 presents the conclusions and outlines directions for future work.

2. Literature Review

Today, tourist data are of great importance. Several studies have been conducted to explore this data by analyzing and extracting interesting insights to help tourist stakeholders to understand tourism behaviors. In sequel, we present several effective studies in pattern mining within social networks and for geolocalized tourism data. We also present some methodologies for tourism data analysis.

Accepted manuscript

ACCEPTED MANUSCRIPT

2.1. Pattern Mining on Social Networks for Tourism

To analyze tourists' behavior, many studies utilize Location-Based Social Networks (LBSN) data. In their work, Vu, Li, and Law (2020) introduced the concept of utility in the analysis of tourist travel diaries. This concept helps to identify highly profitable travel patterns, as evidenced by travel diaries. They created the dataset by combining `Foursquare` check-ins with data from the social network 𝕏 (`Twitter`). Since high-priced venues are likely to lead to higher travel expenditure, the authors measured utility by using the price of dining activities as a proxy for utility. High-utility pattern mining was later used to extract highly profitable travel patterns.

In the same spirit, Talpur and Zhang (2018) used data from `Foursquare` check-ins via the 𝕏 API streaming to study the activity behaviour of tourists visiting Singapore. The experimental dataset includes keys attributes such as check-in ID, user ID, timestamp, Lambert coordinates, category and subcategory of the check-in location. This information was then aggregated by user ID to form sequences of locations types visited by each user. In this study two well-known methods, SPAM (Ayres et al., 2002) and Prefixspan (Pei et al., 2001) are used to analyse the sequential pattern mining. Patterns like *Dining* $\rightarrow$ *Walking* $\rightarrow$ *Shopping* are obtained, representing common behaviour followed by tourists in the afternoon.

Some data from geotagged photos are also used. For example, Lee, Cai, and Lee (2014) combines two techniques - clustering and association rules - to extract Points of Interest (PoIs) associated with tourist destinations in Queensland, Australia, from the `Flickr` platform. To achieve this, the authors propose a complete framework for extracting associative PoIs composed of two main steps: first, he authors identified PoIs using a spatial clustering technique; later, the authors apply the *Apriori* (Agrawal and Srikant, 1994) association rules mining algorithm to detect associative PoI patterns. Cai et al. (2014) also used a spatial clustering and pattern mining technique to extract PoIs from the Queensland `Flickr`. T Unlike the previous study, they used the Prefixspan algorithm is used to extract frequent sequences instead of association rules.

Similarly, Yang et al. (2017) utilizes geo-tagged photos from `Flickr` to mine tourist mobility patterns. To this aim, the authors propose a multi-step process that starts with constructing trajectories, which are sequences of geo-tagged photos. Later, they differentiate native trajectories from touristic ones. Indeed, after crafting individual travel journeys, the authors proceed to identify travel motifs from three points of view. **Topological motifs** which are the subgraphs in mobility networks that describe trip patterns without considering specific place semantics. **Temporal motifs** in which topological motifs are transformed in k-node temporal sequences; and **Semantic motifs**, which include the analysis of preferences of tourists. Finally, motifs are grouped in clusters.

Besides, Farzanyar and Cercone (2015) used a combination of spatial clustering and the Apriori algorithm to extract frequent PoIs from geotagged photos on `Flickr`. The main difference between this paper and the previous ones is the use of Spark to process large amounts of data. Similarly, Höpken et al. (2020) combined pattern mining with clustering techniques to extract Points of

Accepted manuscript

ACCEPTED MANUSCRIPT

Interest (PoIs) from `Flickr` data. However, this study compares two clustering techniques and uses both association rules and sequential pattern mining techniques to identify tourist behaviour patterns.

Despite the variety of existing approaches, current studies often fail to consider key factors like temporal dynamics, tourist segmentation, and behavioural interpretation. Most patterns are too generic to provide practical value. Few studies offer a unified framework or propose metrics that help with decision-making.. These discrepancies underscore the necessity for more systematic and interpretable methodologies to enhance the capture and support of tourist mobility analysis.

2.2. Pattern Mining for Geolocalised Tourism Data

While some papers analyze tourist behavior using GPS data (Orellana et al., 2012; Liu et al., 2022; Choe et al., 2023)they do not use pattern mining to find the most frequent trajectories. In contrast, Park et al. (2020) proposes a four-step methodology for analysing tourist destinations. First, the authors used mobile roaming data from international travelers in three of South Korea's most visited cities. Then, they extracted location points and applied a density-based clustering technique to group nearby points. The researchers then merged individual mobile phone traces into sequences of data points within the same cluster. They used the SPADE algorithm (Zaki, 2001) to extract sequential patterns from these sequences. Semantically, the frequent sequences represent the sequence of clusters most visited by tourists.

In parallel, Liu et al. (2013) studied the tourists' behaviour by incorporating contextual information. Contextual information was defined as features that could influence the tourist's behaviour. They were grouped into social, computational, user, physical, and temporal contexts. The authors generated tourist trajectories using GPS and RFID data, and they collected additional contextual information through surveys. Both data sources were then used to construct a graph. Finally, this study compared three algorithms for extracting association rules to predict interesting places to visit.

Peng and Yamamoto (2020) extracted disjoint sequential patterns. A disjoint pattern is a trajectory that begins and ends at the same location, but takes multiple different paths to get there. To do so, the authors propose a two-step methodology. First, they used a trajectory data mining algorithm to extract sequential patterns from a dataset of street blocks visited by tourists in Japan. Later, in a post-processing step, the authors implement a pruning-based algorithm to extract those patterns that are disjoint.

Although some studies apply sequential pattern mining to GPS data, they often lack a meaningful interpretation of the patterns and fail to provide clear, actionable insights. Also, sometimes they add contextual information, but not done in a consistent way or in a way that can be easily scaled up. Studies rarely discuss the complexity and realism of the extracted patterns, which raises questions about their practical relevance for tourism.

Accepted manuscript

ACCEPTED MANUSCRIPT

2.3. Methodologies for Tourism Data Analysis

Finally, other papers concentrate on exploring and proposing new methodologies for mining data related to tourism. For example, Iorio et al. (2020) introduces a conceptual model called a Digital Tourism System (DTS).

The authors present the main concepts behind this system and propose two models based on the Knowledge Discovery in Databases (KDD) process for extracting knowledge from text and image datasets. This process, widely used in data mining, involves several structured steps such as data selection, preprocessing, transformation, pattern extraction, and interpretation which makes it particularly suitable for handling complex tourism data.

Similarly, Renso et al. (2013) introduced a semantic-enriched KDD process to analyse semantic trajectories, representing tourist movement in a specific context. The authors build a mobility behaviour ontology. Due to the KDD process's popularity in the data mining tasks, several papers also extend the KDD methodology (Fayyad, Piatetsky-Shapiro, and Smyth, 1996) for tourism analysis (Sharma, Meena, and Sharma, 2021; Pyo, Uysal, and Chang, 2002).

Other papers focus on methodologies for analysing specific data. For example, Kim et al. (2017) propose a three-step methodology for extracting knowledge from textual data related to tourism. Also, Li, Deng, and Cai (2020) present a complete framework for analysing factors affecting tourist flow in scenic spots. To this aim, the authors built a multi-source big data platform for collecting tourist flows. They then used this data to analyze the spatiotemporal forecasting of tourist flow using the DA-HKRVM algorithm.

The reviewed methodologies demonstrate the growing sophistication of data analysis in the tourism domain – particularly through the adoption and adaptation of the KDD process – several limitations persist. A significant proportion of the cited works remain focused on framework proposals or isolated case studies, with limited discussion on their real-world scalability, generalisability, or performance under diverse data conditions. Despite its structured pipeline, the KDD process often fails to handle specific challenges in tourism, such as data sparsity, semantic ambiguity in text, and the complexity of integrating multimodal data (*e.g.*, images, GPS traces, and reviews). Additionally, new concepts like semantic trajectories and algorithms like DA-HKRVM occurs without the presence of standardized evaluation benchmarks, thereby complicating the execution of cross-study comparisons.

Accepted manuscript

2.4. Pattern Mining Metrics for Tourist Behaviour

Pattern mining in multigraphs has gained significant attention, particularly in fields such as tourism. Graph structures can model the interactions that take place between tourists, locations, and activities in these fields. Multigraphs can represent multiple types of relationships between the same pairs of nodes. The goal of pattern mining in these graphs is to extract frequent or meaningful substructures that reflect behavioural patterns. The quality and relevance of the extracted patterns largely depend on the metrics used to evaluate them. Traditional pattern mining approaches rely on well-known metrics, such as support,

ACCEPTED MANUSCRIPT

confidence, and lift (Agrawal and Srikant, 1994). These metrics quantify the frequency and strength of the association of the discovered patterns. In the tourism domain, for example, they can identify locations that are frequently visited together. However, these metrics are often inadequate for complex multigraphs. They don't capture the different types of relationships and ignore important context like seasonality, travel motivations, and cultural factors. They mainly focus on the structure, not the meaning behind the connections. Other researches in graph mining has introduced structural analysis techniques, such as motif detection and graphlet decomposition, to help characterise the local topology of networks (Milo et al., 2002). For example, graphlet-based methods (Ahmed et al., 2015) decompose networks into small subgraphs to understand their structure, which can potentially highlight central or popular nodes in tourist itineraries. However, these structural approaches often lack the capacity to incorporate user preferences or semantic information, which limits their applicability to personalised analysis of tourist behaviour.

A critical advancement in pattern mining is the incorporation of temporal and spatial dimensions. Study by Li et al (Li et al., 2008), for example, apply trajectory-based, time-aware metrics to identify sequential travel patterns, which are valuable for analysing movement behaviours. However, these methods rely on the quality of GPS or mobility data and rarely consider contextual factors such as events or weather, which influence tourism behaviour. In tourism specifically, metrics have been adapted to capture complex behaviours. Survey tourism recommendation systems, highlighting metrics that integrate user preferences and context to enable personalised data mining (Gavalas et al., 2014). On the other hand, use geotagged photos to analyse tourist behaviour, introducing measures of place popularity and movement intensity that are similar to the edges of a weighted multigraph (Crandall et al., 2009).

To address the limitations of structural or statistical metrics, knowledge-driven metrics that incorporate ontologies or expert knowledge have emerged. For example, (Chen, Cheng, and Hsu, 2013) used metadata from community photos to infer preferences and recommend destinations, enabling semantic grouping (e.g. museums). However, these approaches face challenges regarding data availability and updating, and require expert input, which limits their adaptability. A classical metrics provide a foundation but are inadequate for complex multigraphs in tourism. The challenge lies in designing context-aware, knowledge-driven, interpretable metrics that provide deeper insights into tourist behaviour and support sustainable tourism decision-making.

Accepted manuscript

2.5. Discussion

The reviewed studies have demonstrated various approaches to extracting patterns from tourist behaviour data using sources such as social media, GPS traces, and geo-tagged photos. However, these approaches have several limitations. They often lack meaningful and practical measures, have limited evaluation standards, and struggle with complex and varied data. They also fail to fully capture the true behavior and context behind the patterns they find.

ACCEPTED MANUSCRIPT

Beyond pattern mining-based solutions, new studies include integrating deep learning models for predictive analysis of tourist behavior (e.g., forecasting using spatial-temporal transformer networks as in Chen et al. (2023), enabling the anticipation of future flows and preferences. Similarly, dynamic recommendation systems based on live tourist flows (such as *R2Tour* that adapts to changing conditions like weather or distance) Yoon and Choi (2023), could be designed to provide adaptive guidance for crowd management and visitor experience optimization. Furthermore, integration with IoT devices and real-time data sources (e.g., the smart-city approach combining IoT with DL Cepeda-Pacheco and Domingo (2022)) would enhance the responsiveness and applicability of the methodology in real-world smart tourism contexts. The present study is not interested in predictions or recommendations, but in understanding tourism behaviour.

Our approach aims to address these gaps by introducing a multigraph model with multidimensional edges, constructed from `Tripadvisor` data. This structure enables a comprehensive representation of tourist behaviour by encoding both the characteristics of trips (edges) and the visited places (nodes), including attributes such as location coordinates, attraction type, user ratings, nationality, and visit dates. To manage the complexity of this graph, we apply a pre-processing step that compresses multiple edges between two nodes, while preserving the underlying information for later detailed analysis. We then apply sequential pattern mining with time gap constraints to extract meaningful and temporally coherent routes taken by tourists. Finally, we incorporate behavioural measures to assess the significance and interpretability of the discovered patterns, ensuring they provide real insights into tourist movement and preferences.

Accepted manuscript

3. General Background

This section presents the formal definitions and core concepts that form the foundation of the proposed approach. These definitions are organized into four primary categories: weighted directed multigraphs, sequential paths, sequential pattern mining, and the time gap-constrained sequential pattern mining problem. By clarifying these elements, the objective is to establish a precise and consistent terminology for the remainder of the paper.

3.1. Directed Labeled Multigraph

A directed multigraph or multidigraph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, consists of a non-empty finite set $\mathcal{V}$ of vertices and a set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ of directed edges. Each directed edge $(u, v) \in \mathcal{E}$ has a start (tail) vertex u and an end (head) vertex v. In a multidigraph, multiple edges with the same endpoints can be represented by multiple directed connections between the same pair of vertices. The out-degree of a vertex $v_i \in \mathcal{V}$ is denoted as $outdegree(v_i)$, the number of edges leaving v_i.

The set $L_{\mathcal{E}}$ represents the labels or attributes associated with the edges in the multidigraph $\mathcal{G}$. Each edge $e_i \in \mathcal{E}$ may have one or more labels from $L_{\mathcal{E}}$,

ACCEPTED MANUSCRIPT

representing properties linked to tourists' information such as weights, types, nationalities, or any other information.

Similarly, the set $L_{\mathcal{V}}$ represents the labels or attributes associated with the vertices in the multidigraph $\mathcal{G}$. Each vertex $v_i \in \mathcal{V}$ may have one or more labels from $L_{\mathcal{V}}$, representing properties linked to locations information such as names, identifiers, etc.

A multigraph can be transformed into a simple directed graph $\mathcal{G}^* \in \mathcal{G}$ by removing duplicate edges and self-loops. This results in a graph where each pair of vertices is connected by, at most, one directed edge.

$$\mathcal{G}^* \subset \mathcal{G} | \mathcal{V}^* = \mathcal{V} \wedge \mathcal{E}^* \subset \mathcal{E} \wedge (\forall e = (u, v) \in \mathcal{E}^* | u \neq v \wedge \nexists e' \neq e \in \mathcal{E}^* \wedge e' = (u, v))$$

3.2. Sequential Paths

A sequential path in a directed graph is defined as a sequence of vertices in which a directed edge connects each consecutive pair of vertices. Formally, let a directed graph $\mathcal{G}^*$, a `sequential path` p in $\mathcal{G}^*$ of length k from vertex v_i to vertex v_k is defined as $p = (v_i, e_1, v_{i+1}, e_2, v_{i+2}, ..., e_{k-1}, v_{i+k})$, where $v_i, v_{i+1}, ..., v_{i+k}$ in $\mathcal{V}$ such that v_i is the starting vertex and v_{i+k} is the ending vertex of the path, $e_1, e_2, ..., e_{k-1}$ are directed edges in $\mathcal{E}^*$, and for each $e_j = (v_{i+j}, v_{i+j+1}) \in \mathcal{E}^*$ $\wedge j \in [1, k[$, the tail of e_j is v_{i+j} and the head of e_j is v_{i+j+1}. In simpler terms, a sequential path in a directed graph is a sequence of vertices connected by directed edges, *i.e.*, $p = (v_i, v_{i+1}, ..., v_{i+k})$, where the direction of the path follows the direction of the edges.

It is important to note that sequential paths follow a partial ordering of the elements based on their positions. Let $\preceq$ denote the partial ordering relation on the sequential path p. For any two elements v_i and v_j in the sequential path, $v_i \preceq v_j$ if and only if $i \preceq j$. In other words, $v_i \preceq v_j$ if v_i occurs before or at the same position as v_j in the sequence.

Also, we define $\mathcal{P}$ as all the paths extracted from $\mathcal{G}^*$, then $|\mathcal{P}|$ represents the number of paths $p \in \mathcal{P}$.

3.3. Sequential Pattern Mining

Before defining the discovering sequential patterns problem, we define the inclusion in the sequential paths notion. Suppose that there are two paths $p = (a_1, a_2, ..., a_n)$ and $p' = (b_1, b_2, ..., b_m)$, where p is a subsequence of p', denoted by $p \sqsubseteq p'$, if it exists $1 \leq j_1 < j_2 < ... < j_n \leq m$ such that $a_1 \subseteq b_{j_1}, a_1 \subseteq b_{j_2}, ..., a_n \subseteq b_{j_n}$ (Pei et al., 2001). Also, p is a super-sequence of p'. In the next paragraph, the support of a sequential path is defined.

A `path database` $\mathcal{D}$ is a set of sequential paths. Each record in $\mathcal{D}$ includes a sequential path identifier (TID) and a path p defined above. The absolute *support* of a sequential path p, denoted by $supp_{abs}(p)$ or simply $supp(p)$, is the number of sequences p' in $\mathcal{D}$ containing p, *i.e.*, $p \sqsubseteq p'$. The relative support of p denoted by $supp_{rel}(p)$, is the ratio of absolute support $supp_{abs}(p)$ to the total number of sequences in $\mathcal{D}$.

A path $s = (v_i, v_{i+1}, ..., v_{i+k})$ is a `sequential pattern` of length k if $supp(s) \geq \sigma$, in which σ is a user-specified minimum support threshold. In this

Accepted manuscript

ACCEPTED MANUSCRIPT

manner, we define $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ the list of sequential patterns extracted from the paths database $\mathcal{D}$. $|\mathcal{S}|$ is the number of sequential patterns and usually they are ordered by related support.

3.4. Time Gap-Constrained Sequential Pattern Mining

The partial order in a sequential path obeys a notion of the time at which the events in the sequence occur. Thus, *time-gap-constrained* sequential pattern mining aims to extract sequential patterns from temporal data, considering temporal constraints between events or elements in the sequences. Formally, a time-gap-constrained sequential pattern s of length k is defined as a sequence of vertices $v_i \in \mathcal{V}$ with $0 \preceq i \preceq k$. For each vertex v_j in the sequence, there exists a vertex v_{j-1} such that v_j follows v_{j-1} in the same sequence s and the time difference between two consecutive vertices v_{j-1} and v_j is less than or equal to the time gap constraint τ. If $\tau = 0$, there can be no element between two elements $v_{j-1}, v_j \in s$ when the pattern s is built.

4. Methodology

This paper aims to extract frequent routes followed by tourist using sequential pattern mining. In order to attempt this goal, a 5-step methodology has been proposed in this paper (see Figure 1). Then, we need to quantify and qualify mined patterns by linking frequent patterns to the underlying multigraph. We need to extract properties from the nodes and relationships in the graph.

Accepted manuscript

4.1. Frequent Pattern Mining in a Multigraph

Step 1 – Multigraph Construction. First, a multidigraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ relies on the fact that tourists can review several locations during their trip. Consequently, the sequence of reviews from user can generate new edges between locations.

In this graph, each vertex $v_i \in \mathcal{V}$ represents a location (a hotel, a restaurant, a tourist spot, etc.). Each edge $e_{ij} \in \mathcal{E}$ represents the interaction between two vertices $v_i, v_j \in \mathcal{V}$. It is important to note that the edges do not necessarily represent the movement of tourists from two vertices, but the comments they post about two places v_i and v_j at two different times. As step 1 of Figure 1 depicts, various edges can be built between two vertices v_i and v_j.

The directed graph is also labeled, *i.e.*, sets $L_{\mathcal{E}}$ and $L_{\mathcal{V}}$ hold information about the edges and vertices in a graph. The characteristics of the edge are, for example, the elapsed time (in days) or the nationality of the tourists for a given trip. As for the vertices, their characteristics can be the name, the *Lambert* coordinates or the rating of the location.

Step 2 – Multigraph Reduction. The second step aims at reducing the complexity of the graph. In fact, the multigraph contains too much various paths that impact directly the computation time of the mining process. Indeed, our goal is to extract frequent patterns and multiple edges between two vertices are not interesting at this stage. Thus, we reduce the multigraph at our disposal into

ACCEPTED MANUSCRIPT

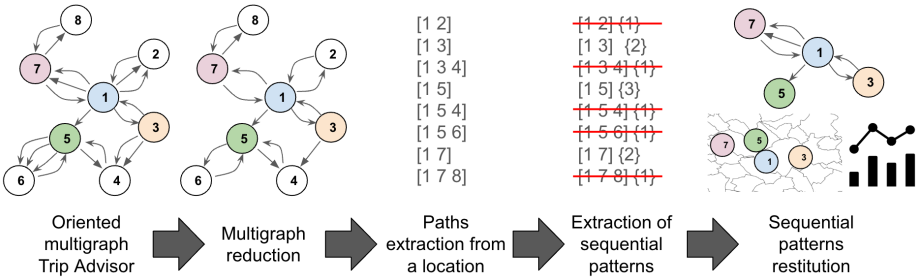


Figure 1. Methodology for extracting sequential patterns from a directed graph of tourist interactions in `Tripadvisor`.

a simple directed graph $\mathcal{G}^*$, *i.e.*, a directed graph without self-loops and multiple edges (step 2 of Figure 1).

Step 3 – Paths extraction. Then, we extract all paths from one vertex to other vertices belonging to the graph. The main idea behind this step is to study how tourists move from a particular place to other places. Step 3 of Figure 1 shows this process starting from vertex #1. It is worth noting that the shortest paths were built from #1 to others by following the edges direction (*e.g.,* $(v_1, v_2, v_3) = (\#1, \#5, \#4)$).

Several approaches were proposed in the literature to extract paths from a directed graph and applied in several applications. The most used are *Breath-Firsts-Search* BFS, which processes the vertices in the graph in the order of their shortest distance from the starting vertex (Levinas, Scherz, and Louzoun, 2022; Saha and Ye, 2024). Another well-known algorithm is *Dijkstra*, which generates a shortest path tree with a given starting vertex (root) (Candra, Budiman, and Hartanto, 2020; Schuberth, Henseler, and Dijkstra, 2018). Finally, the *A-star* algorithm uses heuristics – like edges weight – to find the shortest path between two vertices (Chen, Wu, and Zhou, 2021; Wang et al., 2023). Besides, other path extraction methods were compared in the literature, *e.g.*, Kim et al. (2002). In this work, three techniques were compared, namely *BFS*, *Dijkstra* and *A-star*.

Step 4 – Sequential Pattern Mining. For the fourth step in Figure 1, a sequential pattern mining SPM algorithm was used. SPM involves discovering frequent sub-sequences in a set of sequences and having a temporal order within a sequence (Fournier-Viger et al., 2017). A minimum support value of σ is used to extract sequential patterns. Sub-sequences that occur more than σ times ($supp_{abs}(p) \geq \sigma$) are considered patterns and are shown in the results. In step 4 of Figure 1, σ is set to 2, then all sub-sequences with support less than 2 are not shown in the results (sequences stroked out with a red line).

Several techniques have been proposed to extract sequential patterns, which can be grouped using various strategies. For instance, some algorithms use the *Apriori*-like approach, such as *SPAM* (Ayres et al., 2002) or algorithms using the pattern-growth strategy, such as *Prefixspan* (Pei et al., 2001). On the other hand, algorithms exist to extract sequential patterns under constraints (Alatrista-Salas et al., 2020). In this paper, a sequential pattern mining algorithm with gap

Accepted manuscript

ACCEPTED MANUSCRIPT

temporal constraints has been used (Kadioglu et al., 2023). These sequences should represent consecutive places visited by tourists, *i.e.*, there are no vertices between two consecutive vertices $a_i \in \mathcal{S}$, *s.t.*, $a_j \in \mathcal{S}$. Semantically, a sequential pattern represents the frequent temporal correlation of places visited by tourists. It is important to emphasize that pattern extraction is carried out on the simplified graph, as the sequential pattern extraction algorithm does not account for the detailed structures present in the original graph. Once the patterns have been identified, the analysis returns to the original graph in order to construct tourism-related metrics, thereby leveraging the richer information contained in the multigraph.

Step 5 – Patterns restitution. Finally, step 5 of Figure 1 depicts patterns restitution. This step aims to understand what patterns represent. Patterns were restored from the original graph and the features of the vertices and edges in the graph were used to construct metrics to enrich the patterns obtained. This step is vital in the pattern mining technique and needs domain knowledge (Frawley, Piatetsky-Shapiro, and Matheus, 1992). Knowledge-based metrics are discussed in the next subsection.

4.2. Metrics for Tourists Behaviour Analysis From Sequential Patterns

Once the frequent patterns are mined in the multigraph $\mathcal{G}^*$, we can link node properties to sequences from $\mathcal{G}$. To achieve this we propose the knowledge-based metrics to better understand tourism behavior. Each metric aims at enhancing a distinct feature from tourist digital footprints in order to help to characterize tendencies.

4.2.1. Path Frequency – PF

The `Path Frequency` (PF) represents the absolute support of a sequential pattern s. In other words, PF is the number of sequences from a graph that contain the path sequential pattern (see Equation 1). This measure can take integer values from 1 to the $|\mathcal{D}|$. The relative frequency may be computed by dividing the PF of a traversal pattern by the total number of patterns ($|\mathcal{D}|$).

$$PF(s) = supp_{abs}(s) \tag{1}$$

Patterns with higher PF values indicate more significant tourist interaction. Note that $PF \geq 1$ since a sequential path must exist in sequence database $\mathcal{D}$ when becoming a pattern. In simple terms, PF tells us how often a particular route (a sequence of visits) appears in the dataset.

4.2.2. Path Preference Index – PPI

We are interested in quantifying the decision of tourists to move between vertices belonging to a path sequential pattern while multiple choices remain. Let $s = (a_1, a_2, ..., a_p)$ a sequential pattern $s \in S$, with $a_i \in \mathcal{V}$. The PPI for each frequent

Accepted manuscript

ACCEPTED MANUSCRIPT

sequence in $\mathcal{S}$ is the ratio of the weight between two vertices $\omega(a_i, a_{i+1})$ (number of edges in the multidigraph) divided by a preferred weight w from $L_{\mathcal{E}}$ of all edges starting on a_i, represented by $\omega(a_i)$. The Equation 2 formalises this measure.

$$PPI(s) = \prod_{i=0}^{|s|} \frac{\omega(a_i, a_{i+1})}{\omega(a_i)} \tag{2}$$

High PPI values suggest that tourists prefer more popular destinations over less desirable ones. Conversely, low metric values indicate that only a few tourists have travelled between the vertices of the pattern, even though they could have taken other popular paths. In essence, PPI quantifies the degree to which tourists prefer to traverse from one location to the next along a designated route, in comparison to all potential alternatives.

4.2.3. Path Satisfaction Index – PSI

The PSI measures how much tourist liked the places they visited. To do this, we use the ratings that `Tripadvisor` users added to their comments after visiting a location. We also include the number of comments associated to each place in the calculation. The PSI provides a ratio of reviews by the number of tourists, thus accurately representing how well-received a place is among visitors. Thus, let $s = (a_1, a_2, ..., a_p)$ a sequential pattern $s \in \mathcal{P}$, with $a_i \in \mathcal{V}$, and two weights of vertex a_i, namely $rating(a_i)$ as the rating of vertex a_i, and $nbPosts(a_i)$ the number of posts given a comment about a_i. The PSI of a path sequential pattern s, denoted by $PSI(s)$, is the sum of $rating(a_i) \times nbPosts(a_i)$ for each $a_i \in s$ divided by the sum of the $nbPosts(a_i)$ (see Equation 3).

$$PSI(s) = \frac{\sum_{i=0}^{|s|} rating(a_i) \times nbPosts(a_i)}{\sum_{i=0}^{|s|} nbPosts(a_i)} \tag{3}$$

Therefore, PSI captures how satisfied tourists are with the locations belonging a frequent path. If the value of PSI is high, the tourists liked all places they visited represented by pattern s. Conversely, if the value of PSI is low, this sequence was not satisfactory.

4.2.4. Location-Type Diversity Index – LDI

In this metric, we want to know how different the places visited by tourists are during a tour. To obtain this metric, we compute the *Shannon entropy* of the list of categories – restaurant, hotel, tourist attractions – for each vertex belonging to the sequential pattern. Formally, let $cat(a_i)$ be the category of vertex a_i, $cat(s)$ be a vector mapping all categories $cat(a_i), \forall a_i \in s$. Then, the LDI computes the Shannon entropy $H(s)$ on the sequential pattern s. Equation 4 formally describes this metric.

$$LDI(s) = H(cat(s)) \tag{4}$$

Accepted manuscript

ACCEPTED MANUSCRIPT

where:

$$H(cat(s)) = -\sum_{i=1}^{|cat(s)|} P_i * \log_2(P_i) \quad (5)$$

To summarise, LDI provides insight into the heterogeneity of the categories of visited locations. If LDI is equal to zero, tourists visiting places in the sequential pattern have only one category (*e.g.*, restaurants). More significant values (up to 1) of entropy indicate that the traversal pattern is diverse in terms of the categories of places that belong to it.

4.2.5. Covered Distance by Path – CDP

The CDP calculates the distance (in km) that has been covered by tourists on a tour represented by a sequential pattern. Let s be a sequential pattern, we sum the distance in kilometers from a vertex a_i to a_{i+1}, represented by $dist(a_i, a_{i+1})$, where $a_i, a_{i+1} \in s$. Equation 6 represents this measure.

$$CDP(s) = \sum_{i=0}^{|s|-1} dist(a_i, a_{i+1}) \quad (6)$$

The CDP value in *km* has been calculated using the *Open Street Maps* API (Boeing, 2017) and does not represent a bird's flight distance. Since the CDP quantifies the total distance of a frequent trip, it can be used to assess how far tourists travel to visit points of interest.

4.2.6. Diversity by Distance Covered Index – DDI

The DDI balances the Location-Type Diversity Index (LDI) with the Covered Distance by Path (CDP) metrics. This metric aims to measure the diversity of places visited by tourists, taking into account the distance (in km) covered by a sequential pattern. Since the Shannon entropy can be zero (no variety of places visited), we use the exponential function e for the calculation. In addition, we use a parameter δ which allows us to adjust the weight given to the Shannon entropy in the overall metric (set to 1 by default). Equation 7 shows the calculation of the DDI.

$$DDI(s) = \frac{e^{\delta \times LDI(s)}}{CDP(s)} \quad (7)$$

Small values of DDI represent that tourists belonging a path sequential pattern journeyed long distances and visited places with the same category. Conversely, large values of the DDI show that tourists travelled a few kilometres and visited several categories of tourist attractions. This metric allows us to determine the type of tourists, *i.e.*, those moving locally to visit many diverse attractions (*e.g.*, a city tour) or those displacing considerable distances to visit a particular type of place (*e.g.*, food fairs).

Accepted manuscript

ACCEPTED MANUSCRIPT

4.2.7. Nationality Diversity Index – NDI

We are now interested in measuring the equality/disparity between the values of the frequency distribution of the nationalities of tourists visiting a place along a sequential pattern. To do this, we calculate the *Gini* coefficient for each vertex belonging to the pattern divided by the number of different nationalities presents in the database $|\mathcal{D}|$, as shown in Equation 8.

$$NDI(s) = \frac{\sum_{i=0}^{|s|} gini(i)}{|Countries|} \tag{8}$$

In essence, NDI is a metric that quantifies the heterogeneity of the nationalities of tourists who visit places belonging to a frequent trip. If the NDI coefficient value is close to zero, it means that the frequency distribution of the nationalities is equal. Conversely, if the NDI value is close to 1, the frequency is highly unequal, *i.e.*, one nationality dominates over the others.

5. Experimentation and Results

This section outlines the key findings after implementing the process depicted in Figure 1 on our entire dataset. Three main questions remain to be addressed:

1. Which sequential path construction technique will be used in this study?
2. Which sequential pattern extraction technique best aligns with our goals?
3. What is the impact of the minimum support in extracting patterns using the sequential path construction technique?

The remainder of this section details how we apply the methodology to our dataset and address the aforementioned questions.

Accepted manuscript

5.1. Analysis of Methodology Parameters

First, a labeled directed multigraph $\mathcal{G}$ was built from `Tripadvisor` comments. It corresponds to 287,429 reviews posted from 2013 to 2021 focused on the city of Lille in France. From this two consecutive comments from a given tourist (reducing $287,429$ reviews to $94,829$ trips) provides an edge in the graph associated with the tourist's nationality, rating, date of visit, duration, etc. The corresponding graph contains $3,317$ vertices (locations) and $94,829$ edges (trips). It is worth noting that multiple edges can join two vertices v_i and v_j, as depicted in step 1 of Figure 1. Vertices of the graph represent the places visited by tourists and store features describing a location (*e.g.*, the *Lambert* coordinates or the place's name). Similarly, edges represent a trip between two locations and store information about the trip (*e.g.*, the tourist's nationality or the trip duration). The features used in this study are listed in Table 1.

The comparison between the multigraph and the simplified graph reveals notable differences in edge structure while preserving the same number of vertices (3,317). The number of edges decreases from $94,829$ to $62,164$, a reduction of

ACCEPTED MANUSCRIPT

Table 1. Features used in this study by vertices and edges

Feature name	Description	Attribute of
ID	ID of vertex (place)	vertex
Nom	Name of place visited	vertex
TypeR	Type of place (hotel, restaurant, tourist spot)	vertex
Latitude	Latitude of Lambert coordinates	vertex
Longitude	Longitude of Lambert coordinates	vertex
NbAvis	Total number of reviews by place	vertex
Rating	Evaluation of the place (from 1 to 5)	vertex
MaxDurationDays	Number of days during the trip	edge
Country	Nationality of the tourists who made the trip	edge
Source	ID of the starting vertex of the trip	edge
Target	ID of the ending vertex of the trip	edge
Month	Month of the trip date	edge
Year	Year of the trip date	edge

approximately 34.5% due to the removal of multiple edges. Additionally, the multigraph contains $3,660$ self-loops ($\sim 3.9\%$ of total edges), which are eliminated in the simplified graph. Despite these reductions, all vertices and essential connectivity are preserved, ensuring that paths from any given node remain constructible. To validate connectivity preservation, we computed the connectivity preservation ratio using breadth-first search (BFS), comparing the nodes reachable from each vertex in both graphs. The analysis yielded a ratio of 100%, indicating that the simplified graph provides a more compact representation while maintaining the structural properties relevant for path-based analyses.

Among all edges, 86% are French trips, 5% Belgium, 5% from UK, 1.6% from Netherland, etc. For diversity, we can notice 58% of trips links two restaurants. For time periods, we can say that 60% of trips are made during spring, 30% during summer and 10% during christmas periods. It is worth noting that the characteristics of the dataset, such as the nationality of places visited, do not impact the methodology itself, but rather the metrics calculated for analysing tourist behaviour.

Later, to reduce the computational time of the mining process, the multidigraph $\mathcal{G}$ was reduced to a simple graph $\mathcal{G}^*$ (see Section 4). Weights were also excluded in step 2. It is important to note that this information - multi-edges and weights - is recovered in the pattern restitution step.

Step 3 of Figure 1 has been performed on the formed graph $\mathcal{G}$. This step aims to build paths or trips from the city centre to places visited by tourists (hotels, restaurants or tourist attractions). We have chosen to start from the city centre because the Lille commune has a large number of hotels. This point is confirmed by the high value of the betweenness centrality of the vertex representing the city centre compared to other vertices. Conversely, as discussed in Section 3, several approaches have been used to extract paths from the city centre to other places. We have compared three of them in this paper: A-start, Breath-First-Search (BFS) and Dijkstra. Figure 2 shows a) the execution time in seconds and

Accepted manuscript

ACCEPTED MANUSCRIPT

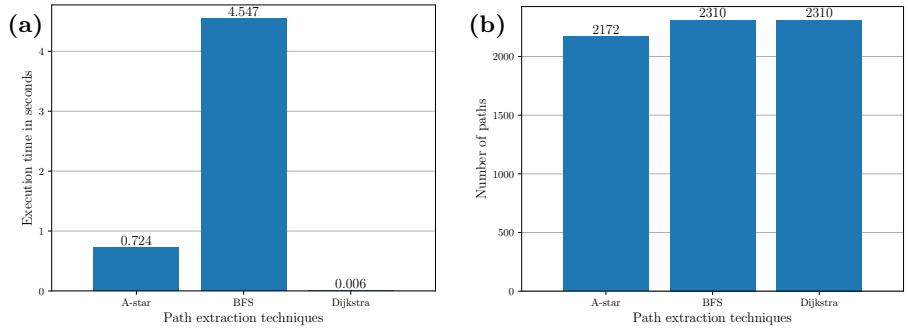


Figure 2. Performances of path extraction techniques: a) in terms of execution time; and, b) in terms of number of paths extracted.

b) the number of paths extracted by each technique. Concerning the execution time, Figure 2(a) depicts that Dijkstra and A-star extract paths faster than the BFS technique. In terms of the number of paths (Figure 2(b)), BFS and Dijkstra extract $2,310$ paths. Conversely, A-star extracts fewer paths due to the heuristic manner in which the technique traverses the graph.

The fourth step of our methodology is to mine the paths to discover the routes frequently used by tourists when visiting the city of Lille, France. Several techniques have been proposed for this purpose. In particular, we are interested in sequential pattern mining techniques, which aim to extract temporal correlations between visited places (vertices of a path). We compared three of the most widely used techniques in the literature, namely `SPAM` (Yang and Kitsuregawa, 2005), `Prefixspan` (Pei et al., 2001) and `Seq2Pat` (Kadioglu et al., 2023). The first two are exhaustive sequential pattern extraction techniques, *i.e.*, given a minimum support σ, the algorithm extracts all possible subsequences that occur in the sequence database $\mathcal{D}$. In contrast, `Seq2Pat` is an algorithm for sequential pattern mining under constraints. This algorithm mines sequences of vertices where each vertex follows the previous one within the same sequence, and the time difference between consecutive events does not exceed the specified time gap constraint. If this constraint is set to 0, the algorithm extracts frequent paths with consecutive vertices.

Figure 3 shows the effect of the time gap constraint. As we can see, `Prefixspan` and `SPAM` extract the same number of patterns 137, but `Seq2Pat` only extracts 78 patterns for a minimal support $\sigma = 10$ (see Figure 3(b)). Thus, patterns of size 1 are omitted by `Seq2Pat`, since we are interested in discovering frequent paths and not Points of Interest (POI – frequent individual vertices), such as Lee, Cai, and Lee (2014); Farzanyar and Cercone (2015); Höpken et al. (2020). Moreover, in the pattern restitution step, as depicted in Figure 1, having fewer patterns saves time analysing patterns that are not interesting for this study. In terms of execution time (see Figure 3(a)), `Seq2Pat` consumes much more time than its two competitors due to the inclusion of constraints in the mining task. It should be noted that `Seq2Pat` is anti-monotonic, a property that is also included in the integration of constraints due to the data structure used by the authors (Multi-valued Decision Diagrams – MDD). In conclusion,

Accepted manuscript

ACCEPTED MANUSCRIPT

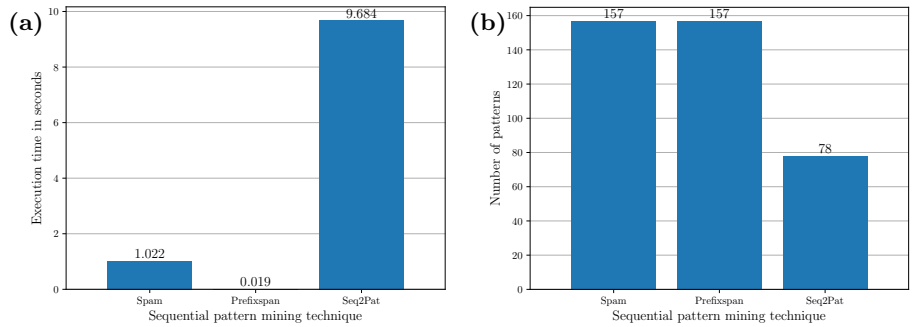


Figure 3. Elapsed time in seconds (a) and the number of patterns extracted (b) by sequential pattern mining technique.

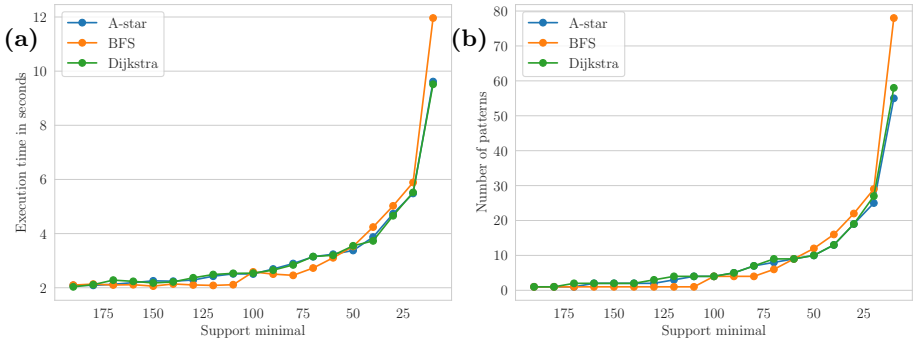


Figure 4. Elapsed time in seconds (a) and the number of patterns extracted (b) by using the `Seq2Pat` algorithm by path extraction techniques

despite the longer execution time, the `Seq2Pat` algorithm produces results that align with the present study's objective, which is to discover and subsequently analyse paths frequently used by tourists in Lille.

On the other hand, pattern extraction algorithms use a threshold called minimum support. All sequences whose number of occurrences in the path database $|\mathcal{D}|$ exceeds this threshold are considered frequent sequences or patterns. Conversely, those whose support is less than the threshold are infrequent and not shown in the results. The threshold can have as its maximum value the number of transactions in the paths database $|\mathcal{D}|$ $(2,310$ in our study). The minimum value that the minimum support can take is 1. If the threshold is close to $|\mathcal{D}|$, the algorithm will extract specific patterns, *i.e.*, most used paths by tourists. On the contrary, if the threshold is low, the algorithm will extract general patterns, *i.e.*, those that few tourists follow.

Figure 4 shows the impact of different thresholds on the mining process regarding the number of patterns extracted and the execution time by different supports by sequential paths databases built from different techniques (A-star, BFS or Dijkstra). The first conclusion we can draw from Figure 4(b) is that patterns start to emerge from a threshold below 200. As the support decreases, the number of patterns grows exponentially. As we mentioned before, the minimal support was fixed to 10 which corresponds to the minimum support value where

ACCEPTED MANUSCRIPT

the highest number of patterns is obtained for any sequential paths database. We can also see that the highest number of patterns was obtained for paths built using the BFS technique. Indeed, the BFS technique starts at the graph's root and visits all the nodes at the current depth level before moving on to the nodes at the next depth level. This fact allows the BFS algorithm to find recurring paths when traversing the tree, resulting in a greater number of patterns *w.r.t.*, Dijkstra and A-star. Finally, exponential growth, both in time and in the number of patterns as the minimum support decreases, is a normal behaviour in pattern mining tasks (Li et al., 2022).

The path extraction technique we will use is BFS. Although BFS is slow, it gives us the most significant number of patterns for a minimum support of 10. In addition, the `Seq2Pat` algorithm extracts patterns with at least two vertices and avoids patterns with a single vertex. After deciding on the path extraction and frequent pattern techniques, we ran `Seq2Pat` on the $2,310$ paths extracted from the city centre to the other locations using the BFS strategy. For a minimum support of 10 we extracted 78 patterns. These patterns will be used in the following for the pattern restitution stage (see step 5 of Figure 1).

It is worth noting that the two main steps—constructing a list of paths or trips and extracting sequential patterns from this list—required relatively little computational time, namely 4.574 and 9.684 seconds, respectively (see Figures 2(a) and 3(a)). These results demonstrate the scalability of the proposed methodology and its applicability to larger graphs without difficulty.

5.2. Applying Metrics for Analysing Tourists' Behaviour

Patterns extracted are then reconstructed using the original graph, *i.e.,* for each pair of vertices that are part of a pattern, all edges found in the original graph between these vertices are recovered, as shown in step 5 of Figure 1. Later, the metrics described in Section 4.2 were applied to the 78 patterns extracted by `Seq2Pat`. To understand how the metrics for analysing tourist behaviour work, only 10 patterns were selected, as shown in Table 2 (location IDs are clickable). Notice that the restaurant `7161082` was closed during the COVID-19 pandemic (it was a good restaurant called "*L'Assiette des Anges*"). Please refer to Table 3 for the IDs' names of the places.

As mentioned above, patterns appear from a minimum support of 200, which can be validated with the first pattern [196911, 196912], whose support is 194. This value represents that around 8.40% of tourists who visit different places leaving the city centre reach through this pattern compulsorily. Furthermore, setting a low support for extracting patterns can be interpreted as the tourists generally taking different paths to discover the city. On the other hand, the CDP value shows that tourists travel $1,354$ *km* to reach the hotel *Mercure Lille Center Grand Place*, represented by the vertex #196912. This value in *km* has been calculated using the Open Street Maps API (Boeing, 2017) and does not represent a bird's flight distance, as depicted in Figure 6. Furthermore, compared to the second pattern [196911, 793722], tourists travel more distances than to visit the restaurant *La Chicoree*, identified by the vertex #793722.

Accepted manuscript

ACCEPTED MANUSCRIPT

Table 2. Examples of sequential patterns obtained for a minimal support of 10

Patterns	Supp	CDP	PSI	LDI	DDI	PPI	NDI
[196911, 196912]	194	1.354	3.834	0.000	0.738	0.075	0.707
[196911, 793722]	109	0.696	3.494	1.000	3.904	0.053	0.746
[196911, 196916]	107	1.542	3.794	0.000	0.648	0.017	0.722
[196911, 6847194]	21	1.690	4.165	1.000	1.608	0.002	0.888
[196911, 196912, 784053]	17	1.279	3.880	0.918	1.957	0.009	0.702
[196911, 196912, 7161082]	12	11.918	3.902	0.918	0.210	0.002	0.797
[196911, 196931, 7218515]	12	12.173	3.659	0.918	0.205	0.023	0.888
[196911, 209260, 8817383]	12	0.566	4.226	0.918	4.425	0.004	0.888
[196911, 196912, 12208448]	11	1.188	3.876	0.918	2.108	0.004	0.742
[2530949, 17689264]	10	1.011	4.164	0.000	0.989	0.002	0.889

Table 3. Correspondence between IDs and place names

ID_place	Place name
196911	Vieux-Lille
196912	Mercure Lille Centre Grand Place
793722	La Chicoree
196916	Ibis Lille Centre Gares
6847194	Rouge Barre
784053	Les 3 Brasseurs
209260	Westfield Euralille
8817383	Team Break Lille
12208448	Papa Raffaele
2530949	Le Barbier qui fume Vieux-Lille
17689264	We Are Ara
7161082	L'assiette des Anges
7218515	McArthurGlen Roubaix
196931	B&B Hotel Lille Roubaix Centre Grand-Place

Accepted manuscript

Regarding the PSI value of the first pattern, the tourists liked the place they visited, but much less than the pattern [196911, 6847194] representing the trip from the city centre to the restaurant *Rouge Barre* (vertex ID #6847194). This value was calculated using the average score given by tourists on Tripadvisor to the places visited during a trip.

The LDI value represents the diversity of places visited, taking into account three possible values: restaurant, hotel and tourist attraction. The value of 0.000 in the first pattern means that the visited nodes have the same category [*hotel, hotel*]. On the other hand, the LDI of 1,000 in the second pattern means that the two visited places are of different categories [*hotel, restaurant*].

From the same perspective, the DDI measure represents the LDI but takes into account the distance travelled by the tourists. For example, the DDI of the first pattern (0.738) is much higher than the DDI of the second pattern (3.904).

ACCEPTED MANUSCRIPT

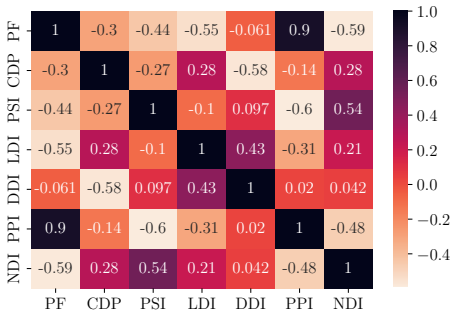


Figure 5. The heatmap of metric correlation.

This can be interpreted as tourists following the first pattern travelled long distances and visited the same type of places, whereas tourists following the second pattern travelled short distances and visited a different type of places.

The PPI represents a group of tourists' choice when visiting one place, having other places to visit. The value is the highest when a tourist chooses the route with a high number of tourist flows. On the contrary, the PPI value is low when the tourist follows one path among several places to visit. It is important to note that this value is calculated taking into account the number of times the places are visited (support), which is why the PPI values change slightly.

Finally, the NDI shows the diversity of the nationalities of the tourists who followed a pattern, taking into account all the nationalities present in the sequence database. For example, the pattern [196911, 196931, 7218515] has a higher value (0.888) compared to the first pattern (0.707) because it has a lower diversity of nationalities [*France*] compared to the first pattern that has a more significant variability of nationalities [*United Kingdom, France, Germany, Belgium*].

Figure 5 shows a heatmap of metric correlation. This matrix can help us better understand the relationship between metrics. For example, the PF or support and the Path Preference Index (PPI) have a highly positive correlation as expected, which means that if the pattern's support is high, the PPI index is also high. Moreover, we can observe a positive correlation between the Nationality Diversity Index (NDI) and the Path Satisfaction Index (PSI), which can be explained by the fact that tourists of the same nationality tend to follow similar paths. On the contrary, the Path Satisfaction Index PSI and the PPI are negatively correlated, *i.e.*, when the (PSI) grows, the PPI decreases; this is due to the variability of weights ω in the PPI which is not the case for the PSI. The same intuition can be used for: the Covered Distance by Path (CDP) and Diversity by Distance covered Index (DDI); the PF and the LDI. Other metrics are less correlated.

From a practical point of view, Figure 5 shows that PF and PPI are strongly and positively correlated ($r = 0.90$). PF reflects how frequently a path occurs in the data, while PPI reflects tourists' preference for that path over an alternative route. This suggests that popular paths are also preferred in practice. PF is also negatively correlated with NDI and LDI. This implies that the most frequently

Accepted manuscript

ACCEPTED MANUSCRIPT

Table 4. Summary statistics for the proposed metrics.

Metric	Min	Max	Mean	Coef. Var.	Skewness	Kurtosis	95% CI
PF	10.000	194.000	27.436	0.0059	3.112	12.173	[21.525, 34.577]
LDI	0.000	1.000	0.658	0.2473	-0.699	-1.501	[0.554, 0.760]
PPI	0.0016	0.1129	0.0108	15.121	3.903	18.239	[0.007, 0.015]
NDI	0.444	0.889	0.807	0.2018	-1.426	1.775	[0.783, 0.829]
DDI	0.0613	8.413	1.996	0.0815	1.403	2.084	[1.616, 2.400]
CDP	0.109	16.777	2.171	0.0750	2.727	6.652	[1.354, 3.166]
PSI	3.310	4.358	3.855	0.0422	-0.049	1.433	[3.819, 3.892]

used and preferred paths are not necessarily the most diverse in terms of location types and/or tourist nationalities. These outcomes highlight the important trade-offs between the popularity and diversity of paths. Moreover, PSI, which measures tourist satisfaction with locations on frequently used paths, has a moderate positive correlation with NDI ($r = 0.54$). This suggests that greater satisfaction with certain routes is associated with higher levels of tourist origin heterogeneity. Figure 5 also shows that CDP, a measure of frequent trips to a location, and DDI which balances distance and location-type diversity, display weaker correlations with the other factors. This suggests that travel distance or balanced diversity is a less salient form of route popularity or preference. These findings imply that PF and PPI are the most informative for applications focused on identifying and promoting preferred tourist routes, while NDI and LDI are particularly useful when the diversity of travellers or destinations is the primary concern. PSI offers an intermediate perspective by connecting satisfaction with diversity, whereas CDP and DDI provide additional, more context-specific insights related to travel distance and balanced diversity.

It is worth noting that the correlation matrix is presented as an exploratory tool to highlight relationships among the proposed metrics. While some measures, particularly distance-based ones, are strongly correlated, others capture distinct aspects of tourist behaviour. This demonstrates that the metrics provide complementary information, supporting their use in broader analytical frameworks such as machine learning.

Table 4 summarises the descriptive statistics for the proposed tourism metrics, offering insight into various aspects of tourist behaviour. For each metric, the minimum and maximum values indicate the observed range of variation, while the mean captures central tendency. The coefficient of variation quantifies relative variability, enabling comparison across metrics with different scales. Skewness and kurtosis describe the shape of each metric's distribution, highlighting potential asymmetries or the presence of extreme values. Finally, the bootstrap-derived 95% confidence intervals provide a measure of the estimated mean's statistical reliability. Together, these statistics provide a comprehensive understanding of how tourists move and interact with points of interest and support the potential use of these metrics in further analyses, such as pattern detection or predictive modelling.

Accepted manuscript

ACCEPTED MANUSCRIPT

Finally, the metrics proposed in the present study were plotted using the `OSMnx` library. Figure 6, for example, shows four metrics for the eight sample patterns presented in Table 2. To improve the interpretability of the plots, the names of the places are listed in Table 3.

Figure 6 (a) shows that most of the sequential patterns start in the city centre, identified by the vertex ID #196911. Patterns [196911, 196912], [196911, 793722], [196911, 196916] of length two are the most frequent (in purple colour), with supports of 194, 109 and 107, respectively. The most frequent pattern [196911, 196912] is used to build other less frequent patterns (in olive colour), such as [196911, 196912, 784053] and [196911, 196912, 12208448], which end with by vertices #784053 and #12208448 respectively. The pattern in light green colour is the largest pattern, which appears 21 times on the sequences dataset ([196911, 6847194]). Finally, the pattern [2530949, 17689264] witnesses a new set of starting points which path traverse the other paths, this less frequent trip connects *Le Barbier qui fume, Vieux-Lille* to *We Are Ara*, shown in yellow.

Similarly, Figure 6 (b) shows the paths coloured by the Path Preference Index (PPI) measure. They vary from 0.0016 to 0.1129 and represent the decision of tourists to move between vertices belonging to a path, given that there are other places to try. The high-valued path [196911, 793722] represents a path chosen by many tourists, knowing that other tourists have chosen different paths (*i.e.,* the tourists' preferred path). On the contrary, the low-valued path [2530949, 17689264] witnesses the fact that many tourists prefer other alternative paths.

Figure 6 (c) shows the PSI metric where frequent trips have high values. Conversely, path [196911, 793722], which represents the trip between the city centre and "*La Chicoree*", has a low value of this metric, indicating that tourists have not appreciated this trip.

Lastly, the Location-type Diversity Index (LDI) values are 1.0, 0.918 and 0.0 (see Figure 6 (d)). This metric represents how different the places are that tourists visit during a trip. A value of zero means that the path has only one type of place (*e.g.*, restaurant → restaurant [196911, 196912]). On the other hand, a value of 1 means that the places are more varied, *e.g.,* hotel → restaurant ([196911, 6847194]).

From a qualitative perspective, it is worth noting that each pattern can be distinctly identified based on the unique combination of measures provided. This multidimensional approach enables domain experts to more precisely characterize each pattern by comparing key factors such as user preferences, pattern length (distance), diversity of nationalities, satisfaction levels, and other relevant indicators. As a result, the analysis supports a more nuanced understanding and differentiation of the observed patterns.

5.3. Discussion

Regarding the state of the art, our work differs from (Jiang, 2021; Jamshidi Nejad, Ahmadi-Abkenari, and Bayat, 2020; Bin et al., 2019), which focuses on extracting traversal patterns. These patterns, generally expressed as frequent itemsets, association rules or sequential itemsets, are extracted by a unique traversal, *i.e.*, a list of edges touching all vertices in a graph. Our work is also not

Accepted manuscript

ACCEPTED MANUSCRIPT

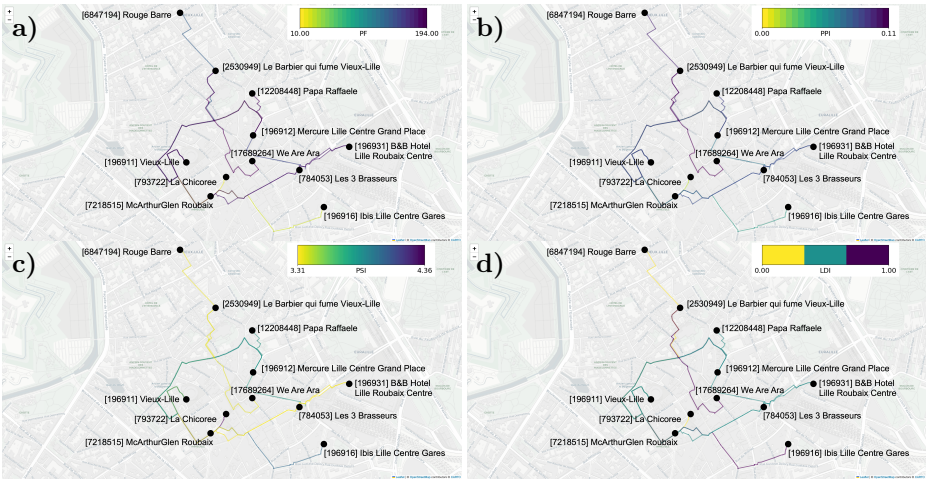


Figure 6. Paths plotted by OSM for PF (a), PPI (b), PSI (c) and LDI (d) metrics for patterns [196911, 196912], [196911, 793722], [196911, 196916], [196911, 6847194], [196911, 196912, 784053], [196911, 209260, 8817383], [196911, 196912, 12208448] and [2530949, 17689264].

focused on the study of PoI (Lee, Cai, and Lee, 2014; Farzanyar and Cercone, 2015; Höpken et al., 2020), which groups POIs into clusters and then construct sequences through which tourists move. Besides, we use a dataset recovered from Tripadvisor, which allows us to deal with multidimensional data, unlike (Orellana et al., 2012; Liu et al., 2022; Choe et al., 2023), which uses GPS datasets. This mixture of features describing tourists and their trips allows us to propose tailored measures for tourism decision makers, different from the work presented in (Prevoteau et al., 2022; Hu et al., 2019; Djebali, Loas, and Travers, 2020), which proposed metrics based on graph theory, such as centrality, betweenness, centrality, degree, among others.

The main contribution of this paper is the proposal and application of a complete methodology for building measures summarising various common behaviours of tourists moving into a city. Based on the extraction and restitution of sequential patterns representing tourists' behaviour, this proposal starts with constructing a multidigraph from comments on the Tripadvisor platform. It ends by extracting tailored measures from extracted patterns.

Another significant contribution of this work is how we have reduced the complexity of a multidigraph to a simple graph whose vertices represent the routes that tourists take from the city centre to other tourist destinations without losing information. In fact, going from a multigraph to a simple graph reduces the complexity of the analysis by avoiding the ping-pong effect (Zaiane, El-Hajj, and Lu, 2001). This strategy is particularly interesting for this project because we are interested in extracting trajectories that tourists frequently used to move from one place to another in a city. We then used an algorithm that allows the extraction of patterns under temporal constraints. That is, we only extract patterns that include places that have been visited consecutively. It is worth noting that the patterns we extracted represent the temporal correlation of places frequently visited by tourists travelling within the city of Lille. The

Accepted manuscript

ACCEPTED MANUSCRIPT

support or the number of times a pattern appears in the sequence database gives us the notion of frequent behaviour.

The metrics have been defined thanks to the collaboration of experts in the field of tourism. These metrics represent aspects related to how tourists move frequently when visiting certain tourist sites. These metrics capture, for example, the distance travelled by tourists, the diversity of place types visited by tourists during a trip, the diversity of the tourist nationalities during a path, the preference a tourist has when choosing a place with respect to other places to visit and the diversity of the places visited per travelled distance.

Our findings can inform a wide range of policy and planning decisions. For urban planners, it is crucial to understand how tourists move within a city. This understanding can help to optimise transportation infrastructure, enhance accessibility to key attractions, alleviate congestion, and improve the overall visitor experience. Similarly, policymakers can use insights from tourist movement analysis to design targeted marketing campaigns, develop tourism policies, and allocate resources more effectively. For example, identifying recurrent movement patterns and combining them with tailored metrics enables personalised tourist itineraries to be designed, underutilised areas to be recommended to balance visitor distribution, and alternative routes to be suggested to mitigate congestion. An example of sequential pattern mining being used for trip recommendations can be found in Zhang et al. (2023). Furthermore, other studies have demonstrated the effectiveness of such analyses in detecting phenomena associated with overtourism, as discussed recently in Alatrista-Salas et al. (2026).

Accepted manuscript

6. Conclusion and future directions

Tourism represents one of the most significant global economic sectors, playing a vital role in the GDP growth of many nations. Understanding how tourists navigate and interact with a destination is essential for policymakers, urban planners, and stakeholders in the tourism industry. This paper introduces a complete methodology for extracting actionable insights from Location-Based Social Network (LBSN) data. The process begins with data acquisition and preprocessing, transforming raw data into a graph structure. The stages involve the extraction of paths and sequential patterns from this graph. Finally, meaningful information is derived through a set of customized metrics that capture tourist movements and preferences, with a focus on the city of Lille, France.

In relation to existing literature, our research stands out as a unique and valuable contribution to the field of tourism mobility analysis. While many studies emphasize the extraction of patterns from graph-based data, our approach goes further by offering a holistic and integrated framework. It combines pattern extraction, reconstruction, and interpretation through domain-specific metrics, providing a richer and more nuanced understanding of tourist behaviour. Notably, the proposed methodology is generalizable and can be applied to any graph that represents movement between tourist locations.

In terms of future research directions, it might be interesting to compare other pattern mining algorithms such as SPADE or CloSpan. The SPADE algorithm

ACCEPTED MANUSCRIPT

adopts a distinct strategy for pattern mining compared to that employed in the present study, while CloSpan focuses on the extraction of closed sequential patterns, i.e., patterns that are not strictly contained within any other sequential pattern with the same support. Similarly, using and comparing techniques for extracting sequences with contiguous locations such as CCSpan or BP-CCSM could be interesting. These techniques could replace the sequential pattern mining with time gap constraints used in the present research. Additionally, we would like to measure the quality of the extracted sequential patterns using metrics such as the coverage or the less temporal contraction.

Another key challenge involves the construction of meaningful paths from a graph structure. In this study, we addressed this by systematically building paths originating from the main square and building paths to all other locations within the graph. Other possibilities could use random walks or subgraphs extraction directly from the initial graph (avoiding the path construction step). In this sense, algorithms such as MuGram or SocMi/ASocMi should be used. Regarding the methodology, it might be interesting to apply it and compare results in different contexts. For example, the metrics obtained from a graph representing tourists travelling in summer might be different from those representing tourists travelling during winter. Furthermore, a null model comparison could assess whether the observed metrics deviate significantly from randomized patterns, thereby providing a statistical baseline for interpretation. Although such an analysis lies beyond the scope of the present study, it constitutes a valuable direction for future research. Finally, we intend to apply the proposed methodology to larger cities with higher tourist numbers to validate its scalability and evaluate its effectiveness in comparing tourist behaviour in cities of different sizes. Similarly, it would be interesting to aggregate patterns at a coarser spatial level, such as districts or communes, enabling comparison with official tourism statistics while complementing the detailed analysis of the proposed metrics.

Funding This work has been funded by the Research & Development Grant "Big Data & Tourism" from the *Metropole Européenne de Lille*

Data Availability The data will be made publicly available upon acceptance of the paper.

Code Availability To support reproducibility, the code will also be released following the paper's acceptance.

Accepted manuscript

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Agrawal, R., Srikant, R.: 1994, Fast Algorithms for Mining Association Rules in Large Databases. In: *Proceedings of the 20th International Conference on Very Large Data Bases*, *VLDB '94*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499. ISBN 1558601538. DOI.

Ahmed, N.K., Neville, J., Rossi, R.A., Duffield, N.: 2015, Efficient Graphlet Counting for Large Networks. In: *2015 IEEE International Conference on Data Mining*, 1. DOI.

ACCEPTED MANUSCRIPT

Alatrista-Salas, H., Guevara-Cogorno, A., Maehara, Y., Nunez-del-Prado, M.: 2020, Efficiently mining gapped and window constraint frequent sequential patterns. In: *Modeling Decisions for Artificial Intelligence: 17th International Conference, MDAI 2020, Sant Cugat, Spain, September 2–4, 2020, Proceedings 17*, 240. Springer. DOI.

Alatrista-Salas, H., Chareyron, G., Djebali, S., Ouled-Dlala, I., Travers, N.: 2026, From Flows to Graphs: Data-Driven Insights on Latent Overtourism with Frequent Pattern Mining. In: Chrysanthis, P.K., Nørvåg, K., Stefanidis, K., Zhang, Z., Quintarelli, E., Zumpano, E. (eds.) *New Trends in Database and Information Systems*, Springer Nature Switzerland, Cham, 327. ISBN 978-3-032-05727-3.

Ayres, J., Flannick, J., Gehrke, J., Yiu, T.: 2002, Sequential pattern mining using a bitmap representation. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD '02*, Association for Computing Machinery, New York, NY, USA, 429–435. ISBN 158113567X. DOI.

Bin, C., Gu, T., Sun, Y., Chang, L.: 2019, A personalized POI route recommendation system based on heterogeneous tourism data and sequential pattern mining. *Multimedia Tools and Applications* **78**, 35135. DOI.

Boeing, G.: 2017, OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* **65**, 126. DOI.

Cai, G., Hio, C., Bermingham, L., Lee, K., Lee, I.: 2014, Mining Frequent Trajectory Patterns and Regions-of-Interest from Flickr Photos. In: *2014 47th Hawaii International Conference on System Sciences*, 1454. DOI.

Candra, A., Budiman, M.A., Hartanto, K.: 2020, Dijkstra's and a-star in finding the shortest path: A tutorial. In: *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, 28. IEEE. DOI.

Cepeda-Pacheco, J.C., Domingo, M.C.: 2022, Deep learning and Internet of Things for tourist attraction recommendations in smart cities. *Neural Computing and Applications* **34**, 7691.

Chareyron, G., Quelhas, U., Travers, N.: 2020, Tourism Analysis on Graphs with Neo4Tourism. In: U, L.H., Yang, J., Cai, Y., Karlapalem, K., Liu, A., Huang, X. (eds.) *Web Information Systems Engineering*, Springer Singapore, Singapore, 37. ISBN 978-981-15-3281-8.

Chen, G., Wu, T., Zhou, Z.: 2021, Research on ship meteorological route based on A-star algorithm. *Mathematical Problems in Engineering* **2021**, 1. DOI.

Chen, J., Li, C., Huang, L., Zheng, W.: 2023, Tourism demand forecasting: a deep learning model based on spatial-temporal transformer. *Tourism Review* **80**, 648. DOI.

Chen, Y.-Y., Cheng, A.-J., Hsu, W.H.: 2013, Travel Recommendation by Mining People Attributes and Travel Group Types rom Community-Contributed Photos. *IEEE Transactions on Multimedia* **15**, 1283. DOI.

Choe, Y., Lee, C.-K., Choi, J., Kim, M., Sim, K.-w.: 2023, Identifying Tourist Spatial and Temporal Patterns Using GPS and Sequence Alignment Method. *Journal of Travel Research* **62**, 1181. DOI.

Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: 2009, Mapping the world's photos. In: *Proceedings of the 18th International Conference on World Wide Web*, *WWW '09*, Association for Computing Machinery, New York, NY, USA, 761–770. ISBN 9781605584874. DOI.

Djebali, S., Loas, N., Travers, N.: 2020, Indicators for Measuring Tourist Mobility. In: Huang, Z., Beek, W., Wang, H., Zhou, R., Zhang, Y. (eds.) *Web Information Systems Engineering – WISE 2020*, Springer International Publishing, Cham, 398.

Farzanyar, X., Cercone, N.: 2015, Trip pattern mining using large scale geo-tagged photos. In: *Proceedings of the International Conference on Computer and Information Science and Technology*, 113.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: 1996, The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* **39**, 27–34. DOI.

Fournier-Viger, P., Lin, J.C.-W., Kiran, R.U., Koh, Y.S., Thomas, R.: 2017, A survey of sequential pattern mining. *Data Science and Pattern Recognition* **1**, 54. DOI.

Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J.: 1992, Knowledge Discovery in Databases: An Overview. *AI Magazine* **13**, 57. DOI.

Gavalas, D., Konstantopoulos, C., Mastakas, K., Pantziou, G.: 2014, Mobile recommender systems in tourism. *Journal of Network and Computer Applications* **39**, 319. DOI.

Höpken, W., Müller, M., Fuchs, M., Lexhagen, M.: 2020, Flickr data for analysing tourists' spatial behaviour and movement patterns: A comparison of clustering techniques. *Journal of Hospitality and Tourism Technology* **11**, 69. DOI.

Accepted manuscript

ACCEPTED MANUSCRIPT

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

Iorio, C., Pandolfo, G., D'Ambrosio, A., Siciliano, R.: 2020, Mining big data in tourism. *Quality & Quantity* **54**, 1655. DOI.

Jamshidi Nejad, S., Ahmadi-Abkenari, F., Bayat, P.: 2020, A Combination of Frequent Pattern Mining and Graph Traversal Approaches for Aspect Elicitation in Customer Reviews. *IEEE Access* **8**, 151908. DOI.

Jiang, Q.: 2021, Analysis of Rural Tourism Demand Characteristics and Experience Differences Based on Association Rule Mining. *Wireless Communications and Mobile Computing* **2021**, 1. DOI.

Kadioglu, S., Wang, X., Hosseininasab, A., van Hoeve, W.-J.: 2023, Seq2Pat: Sequence-to-pattern generation to bridge pattern mining with machine learning. *AI Magazine* **44**, 54. DOI.

Kim, B.J., Yoon, C.N., Han, S.K., Jeong, H.: 2002, Path finding strategies in scale-free networks. *Physical Review E* **65**, 027103. DOI.

Kim, K., Park, O.-j., Yun, S., Yun, H.: 2017, What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. *Technological Forecasting and Social Change* **123**, 362. DOI.

Lee, I., Cai, G., Lee, K.: 2014, Exploration of geo-tagged photos through data mining approaches. *Expert Systems with Applications* **41**, 397. DOI.

Levinas, I., Scherz, R., Louzoun, Y.: 2022, BFS-based distributed algorithm for parallel local-directed subgraph enumeration. *Journal of Complex Networks* **10**, cnac051. DOI.

Li, D., Deng, L., Cai, Z.: 2020, Statistical analysis of tourist flow in tourist spots based on big data platform and DA-HKRVM algorithms. *Personal and Ubiquitous Computing* **24**, 87. DOI.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.-Y.: 2008, Mining user similarity based on location history. In: *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, *GIS '08*, Association for Computing Machinery, New York, NY, USA. ISBN 9781605583235. DOI.

Li, Y., Zhang, S., Guo, L., Liu, J., Wu, Y., Wu, X.: 2022, NetNMSP: Nonoverlapping maximal sequential pattern mining. *Applied Intelligence*, 1. DOI.

Liu, D.-s., Fan, S.-j., et al.: 2013, Tourist behavior pattern mining model based on context. *Discrete Dynamics in Nature and Society* **2013**. DOI.

Liu, W., Wang, B., Yang, Y., Mou, N., Zheng, Y., Zhang, L., Yang, T.: 2022, Cluster analysis of microscopic spatio-temporal patterns of tourists' movement behaviors in mountainous scenic areas using open GPS-trajectory data. *Tourism Management* **93**, 104614. DOI.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: 2002, Network motifs: simple building blocks of complex networks. *Science* **298**, 824. DOI.

Orellana, D., Bregt, A.K., Ligtenberg, A., Wachowicz, M.: 2012, Exploring visitor movement patterns in natural recreational areas. *Tourism Management* **33**, 672. DOI.

Park, S., Xu, Y., Jiang, L., Chen, Z., Huang, S.: 2020, Spatial structures of tourism destinations: A trajectory data mining approach leveraging mobile big data. *Annals of Tourism Research* **84**, 102973. DOI.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.-C.: 2001, PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth. In: *Proceedings 17th International Conference on Data Engineering*, 215. DOI.

Peng, S., Yamamoto, A.: 2020, Mining Disjoint Sequential Pattern Pairs from Tourist Trajectory Data. In: Appice, A., Tsoumakas, G., Manolopoulos, Y., Matwin, S. (eds.) *Discovery Science*, Springer International Publishing, Cham, 645. ISBN 978-3-030-61527-7. DOI.

Prevoteau, H., Djebali, S., Laiping, Z., Travers, N.: 2022, Propagation measure on circulation graphs for tourism behavior analysis. In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, *SAC '22*, Association for Computing Machinery, New York, NY, USA, 556–563. ISBN 9781450387132. DOI.

Pyo, S., Uysal, M., Chang, H.: 2002, Knowledge Discovery in Database for Tourist Destinations. *Journal of Travel Research* **40**, 374. DOI.

Renso, C., Baglioni, M., de Macedo, J.A.F., Trasarti, R., Wachowicz, M.: 2013, How you move reveals who you are: understanding human behavior by analyzing trajectory data. *Knowledge and information systems* **37**, 331. DOI.

Saha, B., Ye, C.: 2024, Faster Approximate All Pairs Shortest Paths. In: *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 4758. SIAM. DOI.

Accepted manuscript

ACCEPTED MANUSCRIPT

Schuberth, F., Henseler, J., Dijkstra, T.K.: 2018, Partial least squares path modeling using ordinal categorical indicators. *Quality & Quantity* **52**, 9. DOI.

Sharma, P., Meena, U., Sharma, G.K.: 2021, Application of Data Mining Algorithms for Tourism Industry. In: Dash, S.S., Das, S., Panigrahi, B.K. (eds.) *Intelligent Computing and Applications*, Springer Singapore, Singapore, 481. DOI.

Talpur, A., Zhang, Y.: 2018, A Study of Tourist Sequential Activity Pattern through Location Based Social Network (LBSN). In: *2018 International Conference on Orange Technologies (ICOT)*, 1. DOI.

Vu, H.Q., Li, G., Law, R.: 2020, Discovering highly profitable travel patterns by high-utility pattern mining. *Tourism Management* **77**, 104008. DOI.

Wang, P., Liu, Y., Yao, W., Yu, Y.: 2023, Improved A-star algorithm based on multivariate fusion heuristic function for autonomous driving path planning. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* **237**, 1527. DOI.

Yang, L., Wu, L., Liu, Y., Kang, C.: 2017, Quantifying Tourist Behavior Patterns by Travel Motifs and Geo-Tagged Photos from Flickr. *ISPRS International Journal of Geo-Information* **6**. DOI.

Yang, Z., Kitsuregawa, M.: 2005, LAPIN-SPAM: An Improved Algorithm for Mining Sequential Pattern. In: *21st International Conference on Data Engineering Workshops (ICDEW'05)*, 1222. DOI.

Yoon, J., Choi, C.: 2023, Real-Time Context-Aware Recommendation System for Tourism. *Sensors* **23**. DOI. https://www.mdpi.com/1424-8220/23/7/3679.

Zaiane, O.R., El-Hajj, M., Lu, P.: 2001, Fast parallel association rule mining without candidacy generation. In: *Proceedings 2001 IEEE International Conference on Data Mining*, 665. DOI.

Zaki, M.J.: 2001, SPADE: An efficient algorithm for mining frequent sequences. *Machine learning* **42**, 31. DOI.

Zhang, X., Pang, X., Wen, X., Wang, F., Li, C., Zhu, M.: 2023, TriPlan: an interactive visual analytics approach for better tourism route planning. *Journal of visualization* **26**, 231.

Accepted manuscript