Analysis of Instagram Users' Movement Pattern by Cluster Analysis and Association Rule Mining

Zehui Wang[1(✉)], Luca Koroll[1], Wolfram Höpken[1], and Matthias Fuchs[2]

[1] University of Applied Science Ravensburg-Weingarten, Weingarten, Germany
{zehui.wang,luca.koroll,wolfram.hoepken}@rwu.de
[2] Mid-Sweden University, Östersund, Sweden
matthias.fuchs@etour.se

Abstract. Understanding the characteristics of tourists' movements is essential for tourism destination management. With advances in information and communication technology, more and more people are willing to upload photos and videos to various social media platforms while traveling. These openly available media data is gaining increasing attention in the field of movement pattern mining as a new data source. In this study, uploaded images and their geographic information within Lake Constance region, Germany were collected and through clustering analysis, a state-of-the-art k-means with noise removal algorithm was compared with the commonly used DBCSCAN on Instagram dataset. Finally, association rules between popular attractions at region-level and city-level were mined respectively. Results show that social media data like Instagram constitute a valuable input to analyse tourists' movement patterns as input to decision support and destination management.

Keywords: Movement pattern · Big data · Instagram · Crawling · DBSCAN · NK-MEANS · Association rule mining

1 Introduction

Movement pattern analysis as a systematic approach is widely used in decision making processes in many fields, especially in the tourism industry. Understanding the tourists' movements between Points of Interest (POIs) plays a fundamental role for destination management activities and is directly applicable to advance the restaurant and hotel industry in the local area [1].

Accurate and valuable analysis of movement patterns requires a large amount of high-quality data. Most previous research obtained people's travel data by surveying individuals' location history or by using automatic location-sensing devices [2], which were neither scalable nor cost-effective to cover numerous individuals [3]. Due to the development of Internet technologies, social media platforms are becoming increasingly popular, in which an enormous number of photos and videos are voluntarily generated and also contain geographic information.

J. L. Stienmetz et al. (Eds.): ENTER 2022, *Information and Communication Technologies in Tourism 2022*, pp. 97–109, 2022.
https://doi.org/10.1007/978-3-030-94751-4_10

Among social media platforms, Instagram, with over 5 million daily active users, contains a large amount of potentially valuable information, i.e. user-generated content which is geographically tagged, to be mined for the tourism industry. However, sophisticated procedures are necessary to systematically retrieve and store this data, especially since the huge amount of data and access restrictions of the Instagram platform make ordinary data retrieval processes difficult.

Facing massive data volume, new approaches and novel learning techniques are needed to fully make sense of them. Machine learning as a rapidly developing technology over the past decade provides potential solutions to mine information and knowledge hidden in the data [4]. The goal of this paper is twofold. One is to introduce an efficient crawling framework for collecting geo-tagged photos from Instagram. The other is, with the help of two different clustering algorithms to group photo-upload locations into different popular POIs and to explore association rules between clusters from different geographical scales to obtain the movement patterns from Instagram users. Compared to previous research, the main contributions of this work are the following:

1. Develop an efficient high-performance crawling algorithm which extracts geographically tagged social media posts and activities from Instagram.
2. Analyze the movement patterns of users in the Lake Constance area using uploaded geographic information from Instagram photos.
3. Recurrent a state-of-the-art clustering algorithm Noise Removal for k-means (NK-MEANS) and compare with the commonly used Density Based Spatial Clustering of Applications with Noise (DBSCAN).

The rest of this paper is organized as follows: in Sect. 2, the up-to-date development of movement pattern research is reviewed; Sect. 3 briefly introduces the methodology involved in this study; Clustering results and association rules obtained from the dataset are presented in Sect. 4; Sect. 5 draws the conclusions about this study and outlines some possible lines of future work.

2 Literature Review

The research of spatial movement patterns of tourists between destinations has been discussed as early as the 1990s [5]. However, the limitations of data volume, research methods, and computing power at that time led to the understanding of the tourist mobility problem as a black-box problem, which was difficult to explore and express [6]. Since movement within a destination played a fundamental role in understanding tourist behaviour, research on tourist's movement patterns received increasing attention. However, these studies with traditional survey-based approaches were always limited by issues of cost, scalability, data volume, and privacy.

With the development of web technologies, more and more people can upload photos and videos to photo sharing platforms to share their journey with others.

Thanks to various social media platforms, a large amount of openly available data with geographic information proves to be available through crawling and scraping mechanisms. Nevertheless, collecting social media data is challenging because of the mix of structured and semi-structured data. Erlandsson et al. [7] crawled Facebook data by using the API (Application Programming Interface) and defined major requirements regarding the crawling results; Jalal et al. [8] scraped Instagram data by using a keyword and location-based approach, which were both utilized in this paper as well. While many studies, like Chu et al. [9] used the Python scraping framework *Scrapy*, an own framework was created in this study due to lack of functionality of existing ones as later described.

Huge amount of data with geographic information provides alternative data sources for many geospatial and social media applications. However, utilizing these data for analysis poses a new challenge. Facing this problem, Arefieva et al. [10] used images from Instagram to cluster tourist's destination; Mukhina et al. [11] analyzed tourists' attraction points using Instagram profiles. However, Instagram, the largest photo-video sharing platform, has a wealth of information about movement patterns in tourist attraction areas, which have not been deeply explored. To bridge this gap, this paper investigated tourists' movement patterns at and between POIs in the Lake Constance region using data crawled from Instagram, based on the framework from Höpken et al.'s study [12] and compares the performance of NK-MEANS with DBSCAN with regard to the geographic information clustering problem.

3 Methodology

3.1 Data Extraction and Preparation

The ETL (extract, transform, load) process forms the first step of this study. It describes how data is retrieved, transformed and stored in a data warehouse in preparation for social media data of the geographical region of Lake Constance [13].

Extract. Extraction of public social media data can be achieved by using web scrapers [8] or an API (Application Programming Interface) provided by the platform [7]. Both techniques were used in this study to obtain the best coverage and depth of data. Instagram provides an overview page for each city in Germany, which was iterated by an ordinal city identifier, in order to collect locations for the respective region. The crawling procedure first collected all published posts connected to a location by iterating through the paginated Instagram API, and second gathered deeper information such as comments or an accessibility caption by crawling each post one-by-one. Since this study is non-commercial and the crawled user information is non-inclusive of any personal private information, there are no ethical and legal issues involved.

Transform. The transformation step focuses on aggregating semi-structured JSON data which was retrieved by using the Instagram API, complemented by unstructured browser-based scraping data as well as image media.

Load. Several hundred parallel crawlers were used to retrieve the data performantly and write them synchronously to the database by using the object-relational mapping software *SQLAlchemy*. To realize this immense crawling volume, rotating IPs, deployment of crawlers in a container infrastructure and multiple concurrent VPN connections, as well as redundancy on multiple server systems were introduced.

Since the data preprocessing is essential to reduce computation time, a geo-based filtering approach was used to limit the dataset to the relevant POIs and a threshold was set for the minimum number of posts whose location was classified as relevant. To ensure that the crawled information is tourism relevant, the posts from local residents and commercial accounts can be distinguished and discarded by taking advantage of the high frequency of tourists' uploading behavior in a short time period and mostly occurring on weekends and holidays.

3.2 Clustering

Clustering plays an important role in the realm of unsupervised learning [14]. The clustering of uploaded photos from Instagram by geographic information is prominent for further analysis. POI information predefined by the platform, i.e. city and location names, can be directly used as input for the cluster analysis. However, for roughly 30% of the photos the city name of the uploaded locations is missing, and the same POI can have different name variants which also causes difficulties when trying to group photos uploaded from the same location into the same cluster. Therefore, in this study, the precise geographic information from uploaded photos, i.e. latitude and longitude, was used for accurately identifying POIs by a cluster analysis.

For the problem of clustering uploaded photos to corresponding POIs, two clustering algorithms based on different principles, namely DBSCAN and NK-MEANS, were implemented and compared for their suitability in identifying meaningful clusters. A brief description of these two algorithms follows:

DBSCAN. DBSCAN is the first density-based clustering algorithm which was proposed by Ester et al. in 1996. It was designed to cluster data of arbitrary shapes in the presence of noise, both for data in 2D or 3D Euclidean space and for data in some high-dimensional feature space [15]. Since DBCSCAN has the advantage of identifying arbitrary shaped clusters and automatically removing outliers, this property is perfect for grouping uploaded photos into relevant POIs. However, DBSCAN has some drawbacks that cannot be ignored when dealing with data from Instagram. More specifically, the algorithm requires a priori knowledge to obtain satisfying clustering results, i.e. *Eps*, the radius of a neighborhood with respect to a core point and *MinPts*, a minimum number of neighboring points, which a core point within *Eps* has. But for Instagram datasets, this a priori knowledge is usually unknown. Secondly, the distribution of uploaded photos on the map is non-uniform. Dealing with datasets with varying densities, DBSCAN could be prone to dilemma in deciding meaningful clusters [16]. Finally, DBSCAN has a time complexity $O(n \log n)$ [14], which

incurs a relatively higher computational complexity than some other clustering algorithms. Therefore, for comparison, a partition-based algorithm with fast noise removal, namely NK-MEANS, was also employed in this study to group the uploaded photos in this study.

NK-MEANS. NK-MEANS is an improved k-means clustering algorithm with automatic noise removal. K-means is one of the most well-known clustering algorithms, whose core idea is to iteratively build and improve clusters by assigning each data point to its nearest cluster centroid (central point of the cluster) and recompute the cluster centroid, until some criteria for convergence is met. Not all uploaded photos from Instagram are related to nearby POIs. Therefore, they can be defined as noise in the dataset. However, the k-means algorithm is highly sensitive to noise, and if k-means algorithm is directly used for clustering, neither satisfying results nor a precise comparison with the results from DBSCAN can be obtained. Therefore, this study uses a method proposed by Im et al. in 2020 [17], which extends the k-means algorithm by a preprocessing step removing outliers in a way suitable for the k-means algorithm.

To quantify the improvement of clustering results after noise removal, this study employed the Hopkins statistic H to measure the clustering tendency of a dataset. A value close to 1 tends to indicate the data is perfectly clustered [18].

Regarding the parameters of NK-MEANS, the appropriate number of clusters k and the proportion of outliers z found by DBSCAN will be used for NK-MEANS to ensure the comparability of the two approaches.

Although the results obtained after the clustering are already meaningful at the geographical level, they are not sufficient for the following association rule analysis. Some clusters contain so few locations or locations have so few uploaded photos that valuable motion patterns cannot be mined. They seem to appear randomly and, thus, cannot be considered as a POI. Therefore, the popularity of clusters and locations was investigated. When a cluster or a location contains less than a certain percentage of uploads, it was discarded as noise as well.

3.3 Association Rule Mining

After popular POIs are identified by the clustering algorithm, mining the user's movement patterns among POIs is the final task in this study. Association rule mining is one of the most popular pattern discovery methods in Knowledge Discovery in Databases (KDD), since its introduction in 1993 [19]. In this study, all uploads by one user were considered as transaction and the visitation of a popular POI as an item (multiple visits to the same popular POIs were counted as one item).

Based on a transaction matrix, spanned by transactions in the one dimension and items (i.e. popular POIs) in the other dimension, frequent itemsets (i.e. combinations of POIs often visited together by the same user) were generated by the FP-Growth algorithm, and based on these frequent itemsets the association rule mining was implemented. Some criterions involved are briefly explained next:

If the set of all user-based transactions is given as $\mathcal{T} = t_1, \ldots, t_n$, then the frequency of itemset X, which is a combination of popular POIs, is defined as: $\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in \mathcal{T}\}|$. Let$X$, Y be an antecedent and consequent itemset $(X \cap Y = \phi)$, $rule\,(X \rightarrow Y)$ denotes an association rule from X to Y. Accordingly, Support $s(X \rightarrow Y)$, which indicates how frequently the itemsets appears together; Confidence $c\,(X \rightarrow Y)$, which indicates how often the rule has been found to be true; Lift $lift\,(X \rightarrow Y)$, which indicates the ratio of the observed support to that expected if itemsets were independent, can be defined [20]. Notably, the size of the Instagram database in this study is relatively large, even a valuable association rule would not have a significantly large Support. Therefore, Lift is used to filter valuable association rules. The larger the lift of a rule is, the more is the rule potentially useful for predicting the consequent in future data sets.

To ensure that the mined association rules are indeed valuable, the p-value of each rule was calculated in this study as well. Suppose X and Y are two itemsets that are independently and identically distributed. Assuming this null hypothesis is correct, the p-value of $rule\,(X \rightarrow Y)$ indicates the probability of obtaining test results, i.e. frequencies at least as high as the results actually observed. A very small p-value means that the observed frequency of an itemset would be quite unlikely, if there is no association between X and Y.

4 Results and Discussion

4.1 Data Extraction and Visualization

The data extraction process is proved to be able to capture several million posts published in the Lake Constance region. It extracted all Instagram posts tagged by a location within the analyzed region. The database consists of 9.6 GB of textual data and 146.6 GB of media files, which contain 46,658 locations, 1,215,063 users and 2,490,640 posts during the period from May 2013 to September 2020. These objects represent highly interesting information such as geographic coordinates or image classification tags added by Instagram's computer vision algorithm.

When looking in detail at the time distribution of the established Instagram database in Fig. 1, an exponential growth of Instagram usage in the analyzed region can clearly be observed. While less than 500 posts were published each week in early 2014, that number grew exponentially to over 65,000 by the end of 2019. Notably, more posts are published in summer than in winter by a steep slope in the middle of each year. This result also confirms that the Lake Constance region is especially popular as a summer vacation spot.

When plotting all locations in Fig. 2 as a heat map weighted by the number of media uploaded at each location, a distribution with a rough trend of clustering can be observed. The results of the heat map clearly demonstrate that the data as is does not constitute an appropriate input to association rule analysis due to a large amount of noise. As indicated above, this data was utilized for knowledge mining at two different geographical scales.

Fig. 1. Temporal distribution of posts amount in the years 2014 to 2019

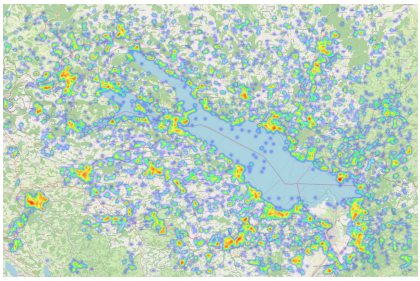


Fig. 2. Heat map of raw data spatial distribution weighted by post amount

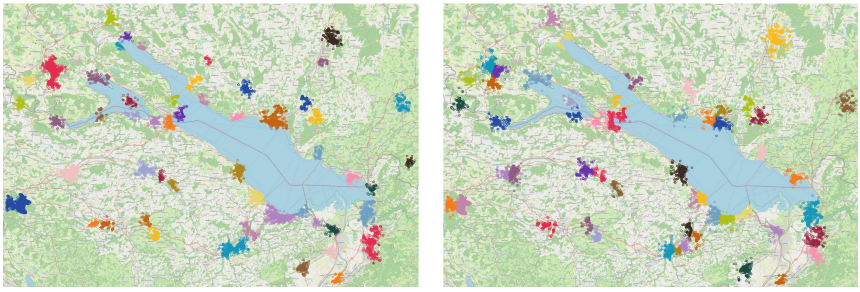
4.2 Region-Level Association Rule Analysis

Based on the data obtained from Instagram, outliers were processed as a first step. Then, the retained data in the whole Lake Constance area were clustered, so that the users' movement patterns can be analyzed on a large geographic scale. Two methods introduced in Subsect. 3.2 were implemented to remove noise from raw data and cluster single uploads to POIs. The comparison before and after noise removal is shown in Table 1. To ensure comparability of clustering results, approximately the same amount of data is removed as noise. Clustering results are shown in Fig. 3, which presents 53 popular POIs (color-coded) from upload locations distributed in the Lake Constance region.

Table 1. Data information comparison before and after noise removal

	H	Amount of locations	Amount of photos	Avg. locations per cluster	Avg. photos per cluster
Raw data	0.936	28,186	4,392,525	–	–
DBSCAN	0.989 (+5.36%)	18,211 (−35.39%)	1,501,478 (−65.82%)	344	28,330
NK-MEANS	0.987 (+5.17%)	18,270 (−35.18%)	1,511,688 (−65.58%)	345	28,522

To quantitatively compare the results from the two clustering algorithms, the Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index were used to evaluate the clustering performance without ground truth. Results in Table 2 demonstrate that, except the Calinski-Harabasz Index, for all criteria DBSCAN performs better than NK-MEANS. This is probably the case because Calinski-Harabasz Index is generally higher for convex clusters like those produced by k-means clustering than other concepts of clusters, such as density-based clusters like those obtained from DBSCAN. By observing the distribution



(a) Clusters by DBSCAN (b) Clusters by NK-MEANS

Fig. 3. Distribution of 53 popular POIs in the Lake Constance region

of clusters in Fig. 3, DBSCAN achieves better clustering results compared to NK-Means without over-clustering (e.g. Friedrichshafen region) or under-clustering (e.g. Ravensburg and Weingarten region) problems. Thus, in terms of the region-level dataset from Instagram, DBSCAN outperforms NK-MEANS for clustering of 2D geographic information, and therefore the clustering results from DBSCAN were used in the consecutive association rule mining.

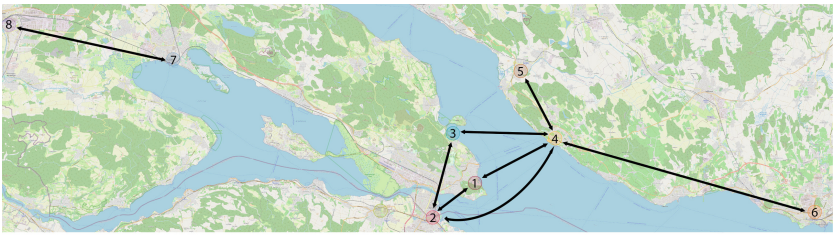
Table 2. Comparison of clustering performance by DBSCAN and NK-MEANS

	Silhouette coefficient	Calinski-Harabasz index	Davies-Bouldin index
Optimum value	1	$+\infty$	0
DBSCAN	0.741	2.250×10^5	0.359
NK-MEANS	0.649	3.213×10^5	0.516

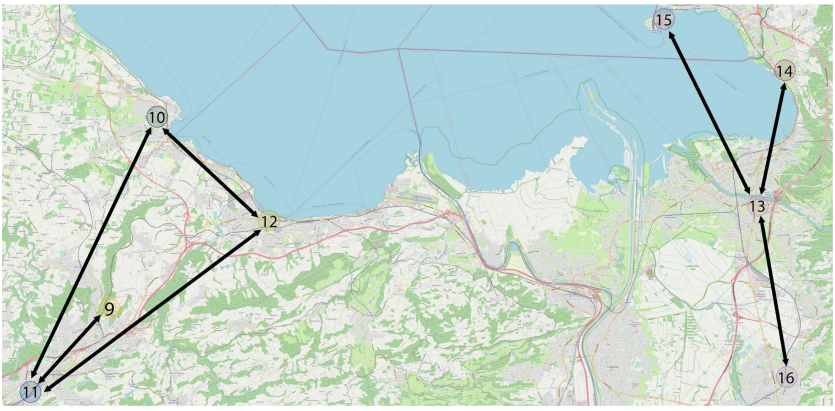
Based on the 53 popular POIs derived from DBSCAN, association rule mining was performed for the frequent items filtered by $s(X \rightarrow Y) = 0.004$ as a threshold. Meaningful association rules ($lift(X \rightarrow Y) > 1$) between 16 region-level POIs are shown in Fig. 4. Since the association rules for the same POIs often appear in both directions, Table 3 only presents the rule that has higher Support, Confidence or Lift between two POIs. Moreover, it is verified that the p-value of each underlying frequent itemset indicates significance (i.e. p-value $< 10^{-5}$). In conclusion, the retained clusters basically match the actual range of popular tourist cities in the Lake Constance region, and the mined association rules reflect the movement patterns of users between cities around Lake Constance.

4.3 City-Level Association Rule Analysis

To explore the movement patterns of users at a smaller geographical scale, this study selected data within the city of Friedrichshafen to mine the association



(a) Association rules centered on Kreuzlingen (2) and Meersburg (4)



(b) Association rules centered on St. Gallen (11) and Bregenz (13)

Fig. 4. Illustration of bi-directional movement patterns within the Lake Constance region

Table 3. Association rules in the Lake Constance region based on Lift descending order

Antecedents	Concequents	Support	Confidence	Lift
Lochau (14)	Bregenz (13)	0.00511	0.35124	3.94735
Radolfzell (7)	Singen (8)	0.00707	0.18854	3.68051
Meersburg (4)	Uhldingen-Mühlhofen (5)	0.00444	0.20264	3.62238
Konstanz (1)	Konstanz-Altstadt (2)	0.00559	0.19807	3.47863
Mörschwil (9)	St. Gallen (11)	0.00454	0.37551	3.21520
Konstanz (1)	Meersburg (4)	0.00431	0.15272	2.72999
Arbon (10)	Rorschach (12)	0.00609	0.15192	2.71288
Dornbirn (16)	Bregenz (13)	0.01364	0.18563	2.08614
Mainau (3)	Meersburg (4)	0.00548	0.11122	1.98812
Rorschach (12)	St. Gallen (11)	0.01247	0.22264	1.90627
Arbon (10)	St. Gallen (11)	0.00879	0.21923	1.87711
Mainau (3)	Konstanz-Altstadt (2)	0.00503	0.10201	1.79164
Meersburg (4)	Konstanz-Altstadt (2)	0.00473	0.08463	1.48641
Meersburg (4)	Friedrichshafen (6)	0.00666	0.11907	1.17481
Bregenz (13)	Lindau (15)	0.01335	0.15000	1.08449

rules between POIs. Due to the relatively small amount of data, there is little difference between different clustering algorithms. DBSCAN, that performs better in Subsect. 4.2, was used for clustering at city-level as well. Eventually, 499 locations were grouped into 28 clusters after noise removal. It contains a total of 21,755 geo-tagged photos and their distributions are shown in Fig. 5.

Fig. 5. Distribution of 28 popular POIs within Friedrichshafen

Fig. 6. Illustration of movement patterns within Friedrichshafen

These 28 popular POIs, which are mainly located near the lakeshore, basically cover the tourist attractions in Friedrichshafen. Based on these clusters, a threshold of $s(X \rightarrow Y) = 0.001$ was set to filter frequent items and then meaningful association rules where identified, which are listed in Table 4.

Table 4. Association rules within Friedrichshafen based on Lift descending order

Antecedents	Concequents	Support	Confidence	Lift	p-value
Mini-golf Course (8) Beach (7)	Promenade (1) Boat Rental (2)	0.00131	0.43750	9.19008	$<10^{-5}$
Promenade (1) Club House (12)	Shopping Center (13)	0.00103	0.18033	5.37066	$<10^{-5}$
Promenade (1) Railway Station (6)	Beach (7)	0.00103	0.23404	3.85575	$<10^{-5}$
Restaurant Area (5)	Shopping Center (13)	0.00131	0.11667	3.47465	$<10^{-5}$
Restaurant Area (5)	Promenade (1) Boat Rental (2)	0.00159	0.14167	2.97583	$<10^{-5}$
Promenade (1) City Garden (3)	Beach (7)	0.00131	0.15730	2.59151	0.00025
Yacht Club (9)	Promenade (1) Boat Rental (2)	0.00103	0.11957	2.51157	0.00172
School Museum (10)	Old Town (4)	0.00122	0.05285	2.26917	0.00233
University Campus (11)	Club House (12)	0.00159	0.08095	2.02230	0.00261

The mined association rules are distributed between 13 popular POIs shown in Fig. 6, which are centered on lakeside promenade (1) and spanning over railway station (6), club house (12) and shopping center (13). This results impressively reflect the user's movement trajectories among tourist attractions within Friedrichshafen.

5 Conclusion and Outlook

Nowadays, with the increasing popularity of social media, this study is the first to collect media containing geographic information from Instagram through an efficient crawler framework and to use this data to mine movement patterns of users within the scenic area of Lake Constance. It can be concluded that big data from social media contains valuable knowledge for the local tourism industry and should, therefore, be given more attention in the future.

When association rules were mined, it has been found that the volume of data plays a significant role in determining the reliability of association rules. The movement patterns at the city-level are less reliable than those at the region-level because of the relatively small amount of data. This means that future research on movement patterns may heavily depend on the availability of big data from various social media platforms.

However, it is worth noting that the data crawled from Instagram also contains plenty of noise, which must be cleaned before analysis, e.g. to remove geographic outliers or to manually label location-names which are useless or incorrect for the study of movement patterns.

Based on the database built from the social media data of Instagram, there is still a great potential for future research. Considering the temporal order of each user's photo upload, the sequential patterns of tourists can be explored in combination with geographic information, which can lead to a more precise recommendation for local tourism. Furthermore, in addition to geographical information, the content of users' uploaded photos, related comments and account profiles can be analysed with Natural Language Processing (NLP) or Computer Vision (CV) techniques to discover more feedback-based knowledge and, thus, to propose highly individualized travel advice.

References

1. Mckercher B, Lau G (2008) Movement patterns of tourists within a destination. Tour Geogr 10(3):355–374
2. Shoval N, Isaacson M (2007) Tracking tourists in the digital age. Ann Tour Res 34(1):141–159
3. Hu F, Li Z, Yang C, Jiang Y (2019) A graph-based approach to detecting tourist movement patterns using social media data. Cartogr Geogr Inf Sci 46(4):368–382
4. Qiu J, Wu Q, Ding G, Xu Y, Feng S (2016) A survey of machine learning for big data processing. EURASIP J Adv Signal Process 2016(1):1–16
5. Mings RC, McHugh KE (1992) The spatial configuration of travel to yellowstone national park. J Travel Res 30(4):38–46

6. Haldrup M (2004) Laid-back mobilities: second-home holidays in time and space. Tour Geogr 6(4):434–454
7. Erlandsson F, Nia R, Boldt M, Johnson H, Wu SF (2015) Crawling online social networks. In: 2015 second European network intelligence conference. IEEE. https://doi.org/10.1109/enic.2015.10
8. Jalal M, Wang K, Jefferson S, Zheng Y, Nsoesie EO, Betke M (2019) Scraping social media photos posted in Kenya and elsewhere to detect and analyze food types. In: Proceedings of the 5th international workshop on multimedia assisted dietary management - MADiMa 2019. ACM Press. https://doi.org/10.1145/3347448.3357170
9. Chu D, Shen Z, Zhang Y, Yang S, Lin X (2017) Real-time popularity prediction on instagram. In: Huang Z, Xiao X, Cao X (eds) Databases Theory and Applications, vol 10538. Lecture Notes in Computer Science. Springer, Cham, pp 275–279. https://doi.org/10.1007/978-3-319-68155-9_21
10. Arefieva V, Egger R, Yu J (2021) A machine learning approach to cluster destination image on instagram. Tour Manage 85:104318
11. Mukhina KD, Rakitin SV, Visheratin AA (2017) Detection of tourists attraction points using instagram profiles. Procedia Comput Sci 108:2378–2382
12. Höpken W, Müller M, Fuchs M, Lexhagen M (2020) Flickr data for analysing tourists' spatial behaviour and movement patterns: a comparison of clustering techniques. J Hosp Tour Technol 11(1):69–82
13. Walha A, Ghozzi F, Gargouri F (2017) ETL4social-data: modeling approach for topic hierarchy. In: Proceedings of the 9th international joint conference on knowledge discovery, knowledge engineering and knowledge management. SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0006588901070118
14. Xu D, Tian Y (2015) A comprehensive survey of clustering algorithms. Ann Data Sci 2(2):165–193
15. Ester M, Kriegel HP, Sander J, Xu X, et al (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol 96, pp 226–231
16. Khan K, Rehman SU, Aziz K, Fong S, Sarasvady S (2014) DBSCAN: past, present and future. In: The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014). IEEE, pp 232–238
17. Im S, Qaem MM, Moseley B, Sun X, Zhou R (2020) Fast noise removal for k-means clustering. In: International conference on artificial intelligence and statistics. PMLR, pp 456–466
18. Cross G, Jain A (1982) Measurement of clustering tendency. In: Theory and application of digital control. Elsevier, pp 315–320
19. Hipp J, Güntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mining-a general survey and comparison. ACM SIGKDD Explor Newsl 2(1):58–64
20. Larose DT, Larose CD (2014) Discovering Knowledge in Data: An Introduction to Data Mining, vol 4. Wiley, Hoboken