Travel Diaries Analysis by Sequential Rule Mining

Submitted to *Journal of Travel Research*

Huy Quan Vu, Ph.D.[1], Gang Li, Ph.D. [2], Rob Law, Ph.D. [3], and Yanchun Zhang, Ph.D. [4]

[1]Center for Applied Informatics
Victoria University, Vic 3011, Australia
Phone: (613) 9919 9318
E-mail: huyquan.vu@vu.edu.au

[2]School of Information Technology
Deakin University, Vic 3125, Australia
Phone: (613) 9251 7434 Fax: (613) 9251 7604
E-mail: gang.li@deakin.edu.au

[3]School of Hotel & Tourism Management,
The Hong Kong Polytechnic University
Phone: (852) 3400 2181 Fax: (852) 2362 9362
E-mail: rob.law@polyu.edu.hk

[4]Center for Applied Informatics
Victoria University, Vic 3011, Australia
Phone :(613) 9919 5060 Fax : (613) 9919 4908
E-mail: yanchun.zhang@vu.edu.au

Travel Diaries Analysis by Sequential Rule Mining

ABSTRACT

Due to the inefficiency in analyzing the comprehensive travel data, tourism managers are facing the challenge of gaining insights into travelers' behavior and preferences. In most cases, existing techniques are incapable of capturing the sequential patterns hidden in travel data. To address these issues, this paper proposes to analyze the travelers' behavior through geotagged photos and *sequential rule mining*. *Travel diaries*, constructed from the photo sequences, can capture comprehensive travel information, and then sequential patterns can be discovered to infer the potential destinations. The effectiveness of the proposed framework is demonstrated in a case study of Australian outbound tourism, using a data set of more than 890,000 photos from 3,623 travelers. The introduced framework has the potential to benefit tourism researchers and practitioners from capturing and understanding the behaviors and preferences of travelers. The findings can support destination-marketing organizations (DMOs) in promoting appropriate destinations to prospective travelers.

Keywords: Data Mining, Geotagged Photo, Sequential Rule Mining, Travel Diary.

1. INTRODUCTION

Tourism plays an important role in the growth of the global economy. In 2013, international tourism generated a total of US$1,075 billion (ATTF 2013). Among this amount, US$102 billion were received from the expenditure of Chinese travelers, making China the number one source market. Germany and the United States both ranked the second with US$84 billion. The Australian tourist expenditure overseas ranked the ninth globally, with US$28 billion (ATTF 2013). An accurate and insightful understanding of travel behavior is thus vital to utilize the great economic benefits of the tourism industry (Edwards et al. 2009). By better understanding travel behavior, tourism practitioners can formulate more appropriate business strategies and travel service/products to meet travelers' needs, which in turn, make a remarkable return on business investment.

Tourism researchers and managers have been pursuing insights into travel behavior to support strategic planning and decision-making in product development and destination management (Li, Meng and Uysal 2008). Knowledge about travelers' location preferences helps tourism managers refine existing attractions, planning new ones, and proposing effective marketing strategies (Lew and McKercher 2006). Understanding of movement patterns of traveler is valuable for tourism organizations in identifying bottlenecks and unnecessary barriers in the flow among tourism destinations (Prideaux 2000), or in segmenting the tourism market to identify suitable travel packages that well align with the characteristic of travelers (Xia et al. 2010).

An analysis of travel patterns is usually performed based on the travel history recorded by travelers during their trips, which are referred to as *travel diary* (Leung et al. 2012; Sheng and Chen, 2013; Vu et al. 2015). Spatial information and temporal information are important components of travel diaries for describing the travel events, so that their behavioral patterns can be inferred. Due to the complex nature of travel behavior, efforts have been made on developing techniques to analyze the travel diaries to extract useful patterns. For instance, a method based on dominant movement patterns was introduced for segmenting tourism market for the case of Phillip Island in Australia (Xia et al. 2010). An anisotropic dynamic spatial lag panel Origin–Destination travel flow model was proposed to analyze Australian domestic and international travel patterns (Deng and Athanasopoulos 2011). Both spatial and temporal dynamics were incorporated for tourism demand modeling from the perspective of origin-destination travel flows to the discovery of useful temporal and spatial patterns. To demonstrate, content and social network analyses were used to examine the travel diaries and map movement patterns of travelers during Beijing Olympics (Leung et al. 2012). Other works adopted Geographic Information System to facilitate the analysis of the movement patterns of travelers (Li et al. 2008; Zakrisson and Zillinger 2012; Orellana et al. 2012). Since traditional data collection methods, such as surveys, opinion polls and questionnaires usually require direct contact with travelers, the collected data are limited in the number of responses and the scale of geographical area included (Zheng, Zha and Chua 2012). Vu et al. (2015) overcome this limitation by utilizing the geotagged photos taken by travelers to capture the spatial and temporal information effectively.

Despite the efforts from researchers, tourism managers are still facing challenges in gaining insights into the complex travel behavior of travelers. Travel diaries usually comprise multiple travel events to different locations/destinations (Leung et al. 2012). The sequential association of the visited locations can reflect travel behaviors and preferences, especially in case of international travel. For instances, some travelers who visited France are also visiting Italy for their trips to Europe; whereas other travelers would visit the United

States after their visit to Canada during their trips to North America. Such sequential associations are useful for agencies in creating more appropriate and promising travel packages. Special offers to visit both the United States and Canada can then be presented to travelers, especially those who want to visit Canada. Such sequential associations are often embedded in the complex sequential travel data, which the existing methods in travel behavior analysis are incapable to account for multiple sequential travel events simultaneously. Traditional approaches using descriptive statistics focus on identifying popular destinations (Furmanov, Cohen and Yan 2012; TRA 2014). A travel sequence has been considered but is limited to a few subsequent travel events (Leung et al. 2012; Barchiesi et al. 2015; Vu et al. 2015). Prior works were unable to discover sequential association in travel diary data for insightful understanding of travelers' behavior.

Recently, a branch of data mining specifically for sequential patterns has emerged due to the increasing availability of sequential databases (Mabroukeh and Ezeife 2010). Sequential patterns and subsequences that appear frequently in sequential data sets can be effectively discovered. For instances, Shie et al. (2012) mined user behavior patterns in mobile environments for planning mobile commerce environments and managing online shopping websites. Aloysius and Binu (2013) mined user-buying patterns to improve shelfing of products based on order of purchasing patterns. Lately, Zheng et al. (2016) attempted to extract sequential behavioral patterns between compliant and non-compliant taxpayers in the financial service industry. Cheng et al. (2016) mined sequential risk patterns from diagnostic clinical records to provide potential clues for physicians for early detection of diseases. Since the travel events in travel diaries can be treated as sequential patterns in a temporal order, it is therefore beneficial to adopt techniques for mining sequential data to analyze the travel diaries.

Aiming to address the limitations in prior works, this paper attempts to incorporate data mining techniques for sequential patterns into travel behavior analysis. A method, named *sequential rules mining* (SRM), is introduced to extract the sequential patterns from travel diaries. SRM is able to reveal the complex travel behavior of travelers and infer the potential associated travel destinations (Cheng et al., 2016). The advantage of the proposed method is demonstrated in a case study of international travel patterns of Australian. We utilize geotagged travel photos available on social media sites as a data source as they are available at a large scale and effective in capturing travel behavior of tourists (Barchiesi et al. 2015; Vu et al. 2015). The geotagged photos are taken by travelers during their trips through digital photo-capturing devices, such as smartphone, smart camera, and tablet. These devices have built-in a global positioning system (GPS) to record geographical information automatically. The travel history of travelers can be extracted from the sequence of posted photos as travel diaries. The study reveals sequential travel patterns of Australian travelers to popular destinations in Asia, Europe and America, to offer insights to tourism managers for destination marketing and travel package development. It is important to mention that the focus of this paper is on the sequential travel patterns of travelers to demonstrate the capability of the travel diary and SRM, other influencing factors of travel behavior are beyond this study's scope of coverage. The introduced framework with the SRM technique has the potential to benefit tourism researchers and practitioners from capturing and understanding the complex travel behaviors and preferences of travelers.

The rest of the paper is organized as follows. Section 2 provides the background on travel diary for travel research and methods for sequential pattern analysis, which is followed by a recap of pattern mining techniques for sequential data. Section 3 presents our framework to process geotagged photos for travel diary construction, which is followed by a description of the SRM technique. Section 4 describes case study and result analysis for Australian travelers, as well as discusses the practical implications of the research outcome. Section 5 concludes the paper and envisages some future research directions.

2. LITERATURE REVIEW

2.1. Travel Diary for Travel Research

Breakwell and Wood (1995, p. 294) defined diary as "*a record of information in relation to the passage of time*". Early attempts in tourism have made used of diaries to record and analyze traveler behavior and expenditure on entertainment, food and shopping (Breen et al. 2001), as well as to explore their experiences, emotions, and satisfaction (Coghlan and Pearce 2010). Travel diaries were used to capture movement of travelers (Ian et al., 2011), or to address transportation problems at tourism destinations (McKercher and Lau 2008).

Travel diaries can be recorded in various forms such as *handwriting on paper* (McKercher and Lau 2008), *video recording* (Pocock and McIntosh, 2013), and *online blog posts* (Leung et al. 2012). Recently, GPS-enabled handheld devices, such as GPS loggers, have been employed by researchers to analyze activities of travelers due to the development and widespread use of GPS technology (Orellana et al., 2012; Birenboim et al., 2013). In these works, direct contact with participants is required to obtain their travel diaries. The collected data are, thus, limited in terms of the number of responses or the scale of the included geographical areas.

Several forms of location data have been utilized to passively capture the travel pattern of travelers. For instance, Sobolevsky et al. (2014, 2015) used bankcard transaction data, which are captured via bankcard terminals, to model the spatial and temporal mobility pattern of travelers. Further, Versichele et al. (2014) adopted Bluetooth tracking of data to determine the visiting patterns of travelers to attractions. Raun, Ahas, and Tiru (2016) measured the visitor flows for destination management using mobile tracking data. Although these data effectively capture the mobility patterns of travelers, they are not freely available for public use. Researchers have resorted to data that are available online, such as the geotagged travel photos (Vu et al. 2015) and geotagged tweets (Chua et al., 2016). The photos were captured by travelers' GPS-enabled photo capturing devices, and then shared publicly on photo-sharing sites, such as Flickr (www.flickr.com) and Panoramio (www.panoramio.com). Geotagged tweets are short messages in a social media platform known as Twitter (https://twitter.com), that are generated by users through their mobile devices with built-in GPS function.

Tourism researchers have used geotagged travel photos to analyze travel behavior at destination. For instances, Kádár (2014) used geotagged photos to study tourist activities in several European cities. Onder et Al. (2014) analyzed Flickr photos in Austria to determine their usefulness in indicating tourism demand. Vu et al. (2015) utilized geotagged photos to discover the travel behavior and preference of inbound tourists to Hong Kong. Recently, the capability of geotagged photos in modeling international travel behavior has received increasing attention. Barchiesi et al. (2015) used large-scale geotagged photos to quantify international travel. Yuan and Medel (2016) focused on the interactions among countries in tourism economics by modeling international travel behavior and inter-country travel flows. Social media strongly influence the tourism industry as people today are becoming heavily dependent on virtual communities in searching for and sharing travel information (Xiang and Gretzel, 2010) given that social media are available at large volumes and have up-to-date content for most locations worldwide. Social media data are significant in studying the movement of tourists as well as in understanding their travel preferences (Chua et al., 2016).

2.2. Travel Pattern Analysis

In the context of tourism, travel patterns are referred to as the movements or travel flows from one tourism attraction to another. A popular approach to study travel patterns is to present the flows in the form of Origin-Destination matrix (Hwang and Fesenmaier 2003).

The values in the matrix can be the actual count of the transitions (Leung et al. 2012), or the proportions of movements from one destination to another, which were computed by Markov Chain technique (Hwang and Fesenmaier 2003). The matrix was also used to represent the changes in the probability of visiting a destination given the changes of attraction at other destinations (Yang et al., 2013). Vu et al. (2015) used Markov chain to examine the flows of tourist in Hong Kong metropolitan area. The Origin-Destination matrix was also visualized using network graph to facilitate the travel analysis for many tourism destinations (Leung et al. 2012; Zach and Gretzel 2012). In these works, the flows are usually represented for two locations at a time, the origin and the destination.

Travel sequences with more than two locations were considered in the work of Xia et al. (2010) for mining dominant movement patterns. The patterns were identified manually from visitor survey for a small-scale tourism attraction with nine destinations within the attraction. The proposed approach is not practical for a large-scale study, especially international travel with many possible destinations. Orellana et al. (2012) used an automatic method, named Generalized Sequential Patterns, to examine visitor movement in natural recreational areas. Their method can extract sequential patterns in a relative order rather than an absolute order. However, Generalized Sequential Patterns focused on identifying popular travel paths from sequential data rather than assessing the sequential association between visited destinations.

A set of techniques for processing sequential pattern discovery has been used in tourism literature, which includes time-series analysis, associate distance measure method, sequence alignment method, and high-frequency pattern methods (Shao and Gretzel, 2010). Among them, sequence alignment is frequently used for determining the movement pattern of tourists in different destinations. They are unable to represent the sequential associations between events in sequential data.

2.3. Mining Pattern from Sequential Data

Discovering a temporal relationship from data is important because it enhances the understanding of the data and provides a basis for making predictions. In data mining, various techniques have been proposed to mine different types of sequential patterns, such as *closed sequential patterns* (Yan, Han and Afshar 2003), *maximal sequential patterns* (Fournier-Viger Wu and Tseng 2013), *compressing sequential patterns* (Chang et al. 2006), and *sequential generator patterns* (Fournier-Viger et al. 2014a). Although these approaches can discover frequent sequences in the travel data set, they are insufficient to make meaningful predictions (Fournier-Viger et al. 2012). For instance, a travel event *c* may appear frequently after travel events *a* and *b*, but there are cases that events *a* and *b* are not followed by event *c*. Predicting that *c* will occur after $\langle a, b\rangle$ according to sequential pattern $\langle a, b, c\rangle$ is not possible. To assess the association between $\langle a, b\rangle$ and $\langle c\rangle$, patterns indicating how many times *c* appears after $\langle a, b\rangle$ and how many times it does not should be available. As such, SRM was proposed as an alternative approach (Fournier-Viger et al., 2012).

A sequential rule is represented in the form $X \Rightarrow Y$, where *X* and *Y* are unordered item sets. The interpretation of the rules is that if some event(s) *X* occur/s in a sequence, item *Y* will occur afterward in the same sequence. Applications of SRM have been found in *stock market analysis* (Yang, Hsieh and Wu 2006), *weather observation* (Hamilton and Karimi 2005), and *e-learning* (Faghih et al. 2010). Sequential rules are related to, but different from, association rules (Tan, Michael and Kumar 2005; Law et al. 2011). The former accounts for the temporal order of items in a sequence, whereas the latter does not. An association rule of the form $X \Rightarrow Y$ does not necessarily mean that event(s) *Y* occurs after other event(s) *X*, which is not suitable for assessing the sequential associations between destinations as the case of SRM.

2.4. Summary

Travel diary is an effective form to capture comprehensive information on the behavior of travelers (Ian et al., 2011). The key component of the travel diary is the spatial-temporal information, which is usually captured using GPS enabling devices. Existing travel diary construction approaches are time consuming and with limited information. Recently, researchers have shifted their attention to user-generated data on social media sites considering its large volume and availability for public use. However, an issue with geotagged photo data is noise. For instance, many geotagged photos were taken in transit rather than at the destinations (Vu et al. 2015). It is also possible that many photos were taken in a tourist destination. Similar issues exist in other types of geotagged social media content available on popular platforms, such as Twitter, Facebook, and Instagram, given that travelers can post content on social media via their mobile devices while traveling. In the application of sequential travel behavior, especially for international travel, the sequences of visited destinations such as cities in the world are of interest, rather than the raw spatial and temporal information embedded in the social media content. Prior works have not presented an effective approach to transform the geotagged social media content into travel diaries for efficient analysis of travel sequential pattern (Kádár abd Gede 2013; Onder et al. 2014; Vu et al. 2015; Garcia-Palomares et al. 2015).

In existing attempts for analyzing travel patterns, the flow are usually limited to two locations at a time (Leung et al. 2012; Zach and Gretzel 2012; Vu et al. 2015), which is inadequate to extract complex sequential patterns from travel. Other works used sequential pattern mining techniques but they focused on identifying popular travel paths (Xia et al. 2010; Orellana et al. 2012), rather than the sequential association between destinations. If DMOs know that travelers are likely to travel to a destination *c* after visiting destinations *a* and *b*, they can design travel packages that promote travelers to visit destinations *a, b,* and special offer for visiting *c*. However, prior approaches in the tourism literature have not been able to identify such sequential associations.

This paper aims to address the aforementioned shortcomings through the following specific objectives:

- present a processing framework for geotagged social media content to construct travel diaries that capture the international travels in the form of sequences;
- introduce SRM into the analysis of travel diaries to identify complex sequential association between destinations; and
- demonstrate the effect of the proposed method for travel behavior analysis by using a case study of Australian outbound travelers.

3. METHODOLOGY

This section presents our method for *sequential travel behavior analysis* that involves geotagged photos that are available from online databases such as Flickr. The main advantage of Flickr over other popular social media platforms is that photo databases are publicly available. The retrieval of all available photos is convenient at any given point in time; whereas it is not the case for Twitter, Facebook, or Instagram because either a quota limit or a fee applies. Moreover, Flickr is known for its reliable data source that can provide useful indicators for tourism demand (Barchiesi et al., 2015). Therefore, we use geotagged photos as a representative geotagged social media content to demonstrate our method. Our framework consists of three stages: 1) *travel data extraction from geotagged photos*, 2) *travel diary construction*, and 3) *sequential rule mining*.

3.1. Travel Data Extraction from Geotagged Photos

The geotagged photos can be retrieved from the Flickr server through its *application programming interface* (API). Full documentation is available at www.flickr.com/services/api. One challenge of data extraction is the identification of users whose photos should be analyzed. For example, we would like to retrieve photos posted by people living in *Melbourne*, but none of the API functions directly supports this operation. We propose to retrieve lists of users and their location of residence initially through specific Flickr groups using the Group Search function. Flickr allows users to create or actively associate with a group. For example, we search for groups whose names contain the keyword *Melbourne*. Some group members are possibly *Melbourne* local residents. This approach allows for a quick access to many users who likely belong to our group of interest.

Let us use $G = \{g_1^w, g_2^w, \ldots\}$ to denote a list of groups returned by the Flickr Group Search function with keyword *w*. For each group $g_1^w \in \mathrm{G}$, we retrieve members' UserID list

$M_{g_i} = \{u_1^{g_i}, u_2^{g_i}, \dots\}$ and their associated location of residence data. One user may have registered into more than one group. Thus, UserID lists $\{M_{g_1}, M_{g_2}, \dots\}$ for the groups are merged to remove any duplicated records. Users belonging to the group of interest, such as *Melbourne* local residents, are grouped together. The entire photo collection for each user is then retrieved using Photo Search with UserID specified as a search parameter. A bounding box covering the entire country of residence for the targeted user group is used to identify photos taken during domestic or international trips. Photos taken outside of the bounding box are assumed to be taken during outbound trips to other countries and kept for further analysis. The photo collection of each user is sorted according to the temporal order of the time taken from the oldest to newest. The travel information can be inferred from the sequence of geographical data associated with the photos.

3.2. Travel Diary Construction

This stage converts the photo collections of users into sequences of visited destinations, which we call *outbound travel diaries*. The issue with the geotagged photo data is that the geographical information is in the form of raw GPS data (*latitude* and *longitude*). The data must be converted into a suitable format, presenting sequences of visited destinations. We propose to process the data by adopting Geocoding API service provided by Google Map, where the GPS data of each photo are mapped to its corresponding location or region. Documentation of Geocoding API is available at http://developers.google.com/maps. The labels of the mapped locations can be at multiple levels such as city or country to represent tourism destinations. It should be noted that the photos collected for each user could be taken during multiple outbound trips. The taken time between the photos is examined to determine their corresponding outbound trips.

The next step is to convert the travel diaries into sequences of destinations. Let $L = \{l_1, l_1, \ldots, l_m\}$ be a set of *m* distinct items, each representing a travel destination. A travel sequence is defined as an ordered list of travel events $\langle(\hat{l})_1 \rightarrow (\hat{l})_2 \rightarrow \ldots \rightarrow (\hat{l})_n\rangle$ where $\hat{l}$ can be any item $l_i \in L$, and $(\hat{l})_t$ represents a travel event at time *t* to location l_i. For example, given the travel sequence $\langle(l_2)_1 \rightarrow (l_4)_2 \rightarrow (l_2)_3\rangle$, the traveler visited destination l_2 first, then destination l_4 second, and revisited l_2 later in the trip. The relative order of the items in the sequence is more important than the positions of the items. For sequence $\langle(l_2)_1 \rightarrow (l_4)_2\rangle$, l_2 does not necessary mean the first destination, and l_4 does not mean the last destination. The following paragraphs demonstrate the construction of travel sequence from travel diary using an example.

Example 1: Table 1 shows a sample travel diaries of a traveler during two outbound trips. One trip is to Europe and the other is to Asia as indicated by the Trip ID. The GPS information was mapped to its corresponding city and country using Geocoding API. The traveler may take many take many photos in each location. We only show the information of several photos here for demonstration purpose, but it still preserves the sequential information of visited destinations. Table 2 shows the travel sequences generated from the travel diary at both country and city levels. It is important to note that the sequence for the trip to Asia shows two destinations at the country level, Thailand and Singapore. The sequence at the city level, containing three items, which provides greater details on the visited cities. As such, detailed insights can be obtained. Travel sequence can also be constructed at a more detailed level such as district or street based on Geocoding API. In this paper, we only consider the city and country levels for analyzing international travel patterns.

***** Please place Table 1 here *****

***** Please place Table 2 here *****

3.3. Sequential Rule Mining

Given a sequential data set $S = \{s_1, s_2, \dots, s_m\}$, where each travel sequence is an ordered list of item sets $s = \langle (\hat{l})_1 \rightarrow (\hat{l})_2 \rightarrow \dots \rightarrow (\hat{l})_n \rangle$. Each item set $(\hat{l})_i$ can comprise one or more items $l_i \in L$ to represent events happening simultaneously. In our application, we assume that item set $(\hat{l})_i$ contains only one item to represent each outbound travel event to a specific destination.

Sequential rule r: $X \Rightarrow Y$ is a relationship between two unordered item sets $X, Y \subseteq L$, such that $X \cap Y \neq \emptyset$ and $X, Y \neq \emptyset$. An item set *X* or *Y* occurs in a sequence *s* if *X* or $Y \subseteq \bigcup_{i=1}^{n} (\hat{l})_i$.

A rule $r : X \Rightarrow Y$ occurs in *s* (written as $r \sqsubseteq s$) if a number k $(1 \leq k \leq n)$ exists, such that $X \subseteq \bigcup_{i=1}^{k} (\hat{l})_i$ and $Y \subseteq \bigcup_{i=k}^{n} (\hat{l})_i$.

Example 2: The rule $\{l_2, l_3, l_5\} \Rightarrow \{l_6, l_7\}$ occurs in sequence $\langle (l_2)_1 \rightarrow (l_3)_2 \rightarrow (l_5)_3 \rightarrow (l_6)_4 \rightarrow (l_7)_5 \rangle$ but not the rule $\{l_2, l_5, l_7\} \Rightarrow \{l_6\}$ because l_6 does not occur after l_7. For simplicity, we omit the index indicating the order of the event without the loss of order meaning.

Sequential rule is defined based on two metrics: *support*, denoted as $supp(r)$, reflects how often the rule *r* appears in the sequential database *S*; and *confidence*, denoted as $conf(r)$, reflects how certain an item set *X* is followed by item set *Y* in the sequential data set.

$$(3.1) \qquad supp(r) = \frac{|\{s | s \in S \wedge r \sqsubseteq s\}|}{|S|}$$

$$(3.2) \qquad conf(r) = \frac{|\{s | s \in S \wedge r \sqsubseteq s\}|}{|\{s | s \in S \wedge X \sqsubseteq s\}|}$$

Traditionally, the process of mining sequential rules starts by finding all frequent sequences in S, whose support is greater than a user defined threshold $min_{(supp)} \in [0,1]$ (Fournier-Viger et al. 2012). Then, the rules that describe the relationships between different sequence items are constructed. The confidences of the generated rules are computed based on Equation 3.2. A rule is considered as representing a strong sequential association, if its confidence is greater than an user defined threshold $min_{(conf)} \in [0,1]$. Such rules are kept for further analysis. Let us demonstrate the concept of *support* and *confidence* using a simple example.

Example 3: Suppose we have a sequential database as shown in Table 3. SRM is applied to this data set; with $min_{(supp)} = 0.5$, and $min_{(conf)} = 0.6$ .Some sequential rules are identified as shown in Table 4. For instance, the rule r_1: $\{l_1, l_2, l_3\} \Rightarrow \{l_5\}$ has a support of $2/4 = 0.5$ because the item set $\{l_1, l_2, l_3\}$, $\{l_5\}$ appears twice out of four sequences in the data set. The confidence of r_1 is $2/2 = 1$ because the antecedent $\{l_1, l_2, l_3\}$ appears twice and is always followed by the consequent $\{l_5\}$. The support and confidence values of other rules are calculated in a similar manner.

***** Please place Table 3 here *****

***** Please place Table 4 here *****

In practice, setting $min_{(conf)}$ is easier than $min_{(supp)}$ (Fournier-Viger and Tseng 2011). The $min_{(conf)}$ can be defined by users based on how confident the user wants the rules to be, whereas the $min_{(supp)}$ should be selected based on the characteristics of the data set. A small value for $min_{(supp)}$ can result in a large amount of rules, while a high value for $min_{(supp)}$ can lead to no or few rules. Therefore, we adopt a recently developed approach, named *top-k*

SRM (Top-K SRM) (Fournier-Viger and Tseng 2011), to discover k rules with the highest support, such that their confidences are higher than a user-specified $min_{(conf)}$. Formally, *Top-K SRM* discovers a set of k rules $R = \{r_1, r_2, \ldots, r_k\}$, such that for each rule $r_m \in R$, $conf(r_m) \geq min_{(conf)}$, and no other rule $r_n \notin R$ with $supp(r_n) > supp(r_m)$ and $conf(r_n) \geq min_{(conf)}$ exists. As such, users only need to provide the $min_{(conf)}$ in the case study and the top rules returned are examined by the algorithm. A practical application of *Top-K SRM* will be demonstrated in Section 4.

4. A CASE STUDY

This section presents a case study of travel diary analysis for *Australian outbound tourism*. The data collection process is initially presented, which is followed by the construction of travel diary. SRM is then applied to identify sequentially associated destinations. The capability of the travel diary in capturing travel behavior is further demonstrated through an analysis of sequential patterns. A discussion of the results is provided with practical implications.

4.1. Data Collection

The data set used in this study was collected from *Flickr* through the method described in Section 3.1. A list of user IDs was first retrieved from *Flickr*'s group together with their locations of residence. Users of interests were identified, and their entire photo collections were retrieved subsequently. Our study focused on users residing in *Sydney*, *Melbourne*, *Brisbane*, *Perth*, and *Adelaide*, the top five most populated cities in Australia (ABS 2015). A bounding box covering the entire geographical area of Australia was specified, with coordinates $min_{latitude} = -45.38122,\ max_{latitude} = -11.044189$ and $min_{longitude} = 110.678793,\ max_{longitude} = 153.845616$. A photo was treated as taken during outbound travel if its location is outside the bounding box; otherwise, it was treated as taken during domestic travel and excluded from further analysis. User accounts with no photo posted were excluded from the data collection. The final data set comprised 809,313 photos taken by 3,623 users during outbound travel. The earliest photos were taken in 2001 and the latest photos were taken in 2015. Table 5 describes our data set with respect to different cities.

***** Please place Table 5 here *****

Sydney has the highest number with 1,435 users. *Melbourne* places the second with more than 1,000 users. *Brisbane* and *Perth* have fewer users. *Adelaide* has the least number with 213 users. This order is similar to the popularity ranking for these cities (ABS 2015); *Sydney* is the most populated city, and *Adelaide* places fifth. The number of photos taken per user is similar across the groups. The average time span for photo collections of Melbourne travelers is the highest with an average of around three years, while the photo collections of Sydney travel group has the least time span. The time span of the photo collection is not a major factor in our study, because our analysis focuses on the sequential travel pattern for each outbound trip rather than the travel history of travelers. In addition, Australia is located far away from countries in other continents; the travel patterns from different Australian cities to other continents would be relatively similar given the limited number of air routes. Therefore, we treat users from different cities as the same group to represent Australian travelers in the subsequent analysis.

We acknowledge the variety of travel styles and preferences among the travelers, such as for businesses, holiday or family visits. This study, however, does not consider such differences due to scope of coverage. Instead, this study presents an approach, which focuses on extracting patterns reflecting sequential associations among visited destinations, embedded in the travel photo sequences.

4.2. Travel Diary Construction

The geographical information of photos was mapped to its corresponding city and country using Geocoding API, as described in Section 3.2. We made an assumption that photos taken more than 30 days apart were likely in different outbound trips. The photo collection of each user was sorted in temporal order, and separated into different trips. Totally, 17,188 travel diaries were constructed from the collected data set. The travel diaries were then converted into sequences of visited destinations. The number of travel diaries in this study is much more than the travel diary data set used in prior studies (Xia et al. 2010; Orellana et al. 2012; Vu et al. 2015).

Among the travel diaries, we noticed that 12,819 travel diaries comprise single country and 4,369 travel diaries involve two or more countries. Table 6 shows the proportions of visited continents, single country trips vs. two countries or more trips. Please be noted that destinations in Oceania refer to countries other than Australia, as the collected data are for the outbound trips of Australian residents. We can see that the majority of trips with a single country was in Asia with 33.72%. Travelers were more likely to travel to Europe in trips spanning two or more countries with 43.16%, significantly higher than trips to a single country. Z-test with $p\ value\ \leq\ 0.05$ verified statistical significance. Little difference was noticed for trips to Africa, America, and Asia. Travelers were less likely to travel to two different countries in Oceania.

***** Please place Table 6 here *****

We further examine the capability of the geotagged photo in capturing travel behavior of Australian travelers via Table 7, which shows the top 20 visited as identified from the collected data set. The top countries in our list are among top 10 destinations according to national outbound survey by Tourism Research Australia (TRA 2014b). The most popular countries in both lists are United States, United Kingdom and New Zealand. In particular, travelers are likely to visit multiple times to the United States and the United Kingdom as shown by high values for average number of trip per travelers. These destinations are in fact the home countries of many Australian residents (TRA 2014b), where they probably visited frequently. We also noticed that our list does not include Fiji, a popular destination of Australian travelers (TRA, 2014b). Fiji ranked 22nd in our data set; as a result, Fiji was not listed in Table 7. Nevertheless, the geotagged photos can still capture the general travel behavior of the travelers. Table 7 presents and examines the popular destinations; we included all of the destinations to construct the travel sequences in the subsequent analysis.

***** Please place Table 7 here *****

4.3. Travel Sequence Analysis

4.3.1. Sequential Rules of Visited Countries

The data sets of the constructed travel sequence at country level were input into the *Top-K SRM* algorithm (Section 3.3), whose implementation is available as an open-source data mining library (Fournier-Viger et al. 2014b). Only those 4,369 travel diaries involving two countries were considered in this analysis, as there is no sequence in the travel diaries to a single country. The minimum confidence was set to $min_{(conf)} = 0.6$, and the k value was set to 50. The *Top-K SRM* scans the data sets for sequential rules greater or equal to 0.6 and returns those rules with top support. Some rules may have similar items in the antecedent and same item(s) in the consequent. Such rules contain redundant information, and thus only those rules with the top supports are reported. Twenty-three sequential rules were selected as shown in Table 8. The countries are denoted by the three-letter country code defined in ISO 3166 published by the International Organization for Standardization (ISO) (www.iso.org/iso/country_codes). The rules are grouped based on the continent of the countries in the consequence part for convenience of interpretation. All rules satisfy the minimum confidence threshold of 0.6. We summarize the findings as follows:

- Australian travelers have a high chance to travel to the *United States* (*USA*) if they plan to visit *Canada* (CDN) or *Mexico* (MEX), as indicated by rules r_1 and r_2 respectively. The confidences of both rules are above 0.74. If they travel to Bolivia (BOL), they are likely to also visit Peru (PER) with a confidence of 0.871 (rule r_3).
- For destination in Asia, a relatively strong sequential association was found between *Lao* (LAO) and *Thailand* (THA) as in rule r_4. If a traveler plans to visit *Lao*, he/she is likely to visit *Thailand* during the trip. An explanation for the low number of rules is that Australian travelers are likely to visit a single country in Asia as discussed in Section 4.3. Thus, a low number of sequential patterns exist in the trips to two or more countries in Asia. Analysis at the city level would provide further insights in the later sections.
- Quite a number of rules were found for destinations in Europe. Namely, if travelers visited Czech (CZE), France (FRA) and/or Austria (AUT), they have a high possibility of visiting Germany (DEU) as well as indicated by rules r_{5-7}. Some travelers are likely to visit Italy (ITA) after visiting Austria, France, or Greece (GRC) (rules r_{8-9}). Bosnia and Herzegovina (BIH) are likely to be visited after Croatia (HRV) as shown in rule r_{10}. Rules r_{11-23} show strong sequential associations between the countries of the United Kingdom (GBR) and other European countries. The

combinations of the visited countries are varied but the United Kingdom is often the last destination. A possible explanation for these patterns is that the United Kingdom is the home country of many Australian residents. Therefore, they are likely taking advantage of their trips back home to visit other European country on the way.

Table 8 shows that most countries in the identified sequential rules are the most visited destinations by travelers considering that Top-K SRM returns sequential rules with high supports also indicate the frequent items. In the next section, we examined the sequential pattern between cities for more insights into the travel patterns.

***** Please place Table 8 here *****

4.3.2. Sequential Rules of Visited Cities

This section focuses on demonstrating the capability of travel diaries in capturing the travel patterns at the micro level between cities. We examine the multi-city trips to destinations in America, Asia and Europe in this analysis. Only those travel sequences with two or more cities in each continent are input into the SRM algorithm. We notice that some rules at the city level are redundant to rules at country level as reported in Table 8. For example, the rule $Dublin \Rightarrow London$ would provide similar patterns as the rule $Ireland \Rightarrow United\ Kingdom$. We report rules that provide new information, as shown in Table 9. The city names are shown together with their corresponding country codes.

Some sequential associations are found between cities in American countries. For instances, travelers are likely to visit *Los Angeles* if they visited *Chicago* and/or *Denver* (rules c_{1-3}). Travelers visited *La Paz* have a high chance to also visit *Lima* next (rule c_5). Rules

c_{6-8} shows some sequential associations between cities in Asia such as *Hebron* and *Jerusalem*, *Kathmandu* and *Kolkata*, *Jakarta – Kuala Lumpur* and *Singapore*.

Although quite a number of rules are found for European cities, many of them provide redundant information. The rules c_{9-11}show some associations not yet discovered in previous analysis at the country level. Namely, travelers who visited *Monaco*, *Amsterdam* or *Madrid* are likely to visit *Paris*. $Zagreb$ is likely to be the next city after *Sarajevo*.

***** Please place Table 9 here *****

Aside from the well-known tourism destinations listed in Table 9, DMOs would be interested in the travel patterns of travelers between second- or third-tier destinations to gain more insight. As a demonstration, we examine the sequential rules between cities in the *United Kingdom*, except for London. Table 10 shows the top 10 sequential rules with 0.6 or more confidence.

Rules e_1 and e_2 show the possibility for Australian travelers to visit *Edinburg* after *Dunfermline*, and *Inverness* after *Elgin*, respectively. Travelers who have visited *Greenock* and *Inverness* are also likely to visit *Isle of Lewis* (Rule e_3). Furthermore, *Kendal* is likely to be visited after *Barrow in Furness* (Rule e_4). Travelers are likely to visit *Oxford* after *Cowley* (Rule e_6), or after *Edinburgh* and *Killington* (Rule e_5). Travelers have a high possibility of visiting Perth if they have visited nearby cities, such as *Dundee*, *Alloa*, *Arbroath*, and *Glentorhes* (Rules e_{7-10}). Most cities in the identified rules are at a close distance, which is convenient for land travel.. Except for Rule e_5, *Edinburg* is far away from *Oxford* but is frequently visited probably because of popularity and convenience of air transportation.

***** Please place Table 10 here *****

The SRM aims to assess how certain it is for a destination to be visited after other destinations based on the confidence. For example, people may travel frequently between *Los Angeles*, *Chicago,* and *Denver*. If DMOs are certain that Los *Angeles* will be the next destination after *Chicago* using SRM, more focused travel packages can be developed to promote those who visit *Chicago* to travel to *Los Angeles*. Nevertheless, travel diaries constructed from the geotagged photos can be used to identify popular sequential patterns to support for the construction of travel itinerary. We demonstrate such capability in the next section.

4.3.3. Travel itinerary analysis

This section demonstrates the capability of travel diaries in capturing popular travel pattern through an analysis of sequential travel pattern among Asian cities. Only travel sequences with two cities or more in Asia are considered. We applied *top-k* sequential pattern mining algorithm to extract the frequent patterns (Fournier-Viger and Tseng 2011). Top 50 patterns with highest support are returned, and patterns with similar items are removed as they provide redundant information. We are left with 24 frequent patterns as shown in Table 11.

***** Please place Table 11 here *****

We can see that the identified patterns contain major tourism cities in Asia such as *Bangkok*, *Ho Chi Minh City*, *Hong Kong, Kuala Lumpur,* and *Singapore*. For instances, sequential patterns s_{1-7} show several patterns starting from *Bangkok* to other nearby cities.

The popular cities to visit after Bangkok are *Kuala Lumpur* and *Singapore* with the supports of around 0.05, respectively. Some travelers visited all three cities as shown in pattern s_3 and s_5. Popular sequential patterns starting from *Ho Chi Minh City* are shown in patterns s_{8-12}. Travel pattern starting from *Hong Kong* tends to go to other destinations in China, especially for *Shanghai* with the support of around 0.1. The sequence s_{18} has the highest support (0.134) among all patterns, which reflects the fact that *Kuala Lumpur* to *Singapore* is a very popular travel path, perhaps due to their close distance. It is interesting to note that although Singapore is a major destination and is close to Australia, travelers are more likely to visit Singapore after other cities in multiple destination trips in Asia. Besides, frequent patterns were found for less popular cities such as *Kathmandu* to *Kolkata* (sequence s_{23}) or *Kolkata* to *Singapore* (sequence s_{24})

Due to long distance between Australia and other continents, Australian travelers often travel to Europe or America via Asian cities due to more options of airlines. It is beneficial for DMOs to identify the hub destinations in Asia for better development of travel itinerary for long haul travel. We examine the travel diaries and identify any transition from a city in Asia to the next city in Europe and America. Transition frequency is visualized using a heat map (Krentzman et al., 2011), as shown in Figure 1. Asian cities are on the vertical axis. European and America cities are listed on the horizontal axis; the prefix of the city names indicates the corresponding continent. Due to the large number of cities, only cities visited by at least 1% of the travelers are included in the figure. A darker cell indicates high frequency, and a lighter cell indicates otherwise.

***** Please place Figure 1 here *****

We can see that *Dubai*, *Hong Kong* and *Singapore* are the most popular destinations for Australian to travel to *London* as shown with dark cell. This is consistent with the fact that those cities are major hub destination with large airports and major airlines. Hong Kong is also a popular destination for traveling to Paris. Travelers are likely to travel to *Paris* via *Hong Kong*. *Shanghai* is a popular hub destination for traveling to *Berlin*. Few direct transitions from Asia to America are shown in Figure 1, which is consistent with the fact that direct routes from Australia to America are more convenient. Tokyo to Los Angeles is a commonly used path from Asia to America by Australian travelers. We further examined the travel diaries and found that around 70% of Australian travelers spent more than one day in Tokyo before traveling to Los Angeles. This result suggests that Tokyo is usually visited for other purposes rather than simply for connecting flights.

4.4. Discussion

The analysis using SRM (Section 4.3.1) has identified some strong sequential associations between visited destinations of Australian outbound travelers. DMOs can advertise specific travel packages that promote travelers to visit multiple destinations in their trips. For instance, special offers to visit the *United Kingdom* can be created if the travelers also visit *Germany, Italy* and *Spain* (rule s_{23} in Table 8), due to the strong sequential associations between them. In this way, DMOs can encourage travelers travel to more destinations and purchase higher value travel packages. The analysis of travel patterns can also be done at city level for detailed information as indicated in Section 4.3.2. Domestic travel packages between *Chicago*, *Denver* and *Los Angeles* can be offered for people who visited the *United States*, as indicated by strong associations in rules s_{1-3} (Table 9). In addition, detailed insights into the travel patterns of travelers between second- and third-tier destinations can be obtained based on our proposed approach. Table 10 shows some strong rules between nearby cities in the United Kingdom. DMOs may offer different means of transportations other than airlines to travelers to promote their travel packages, such as *Elgin* to *Inverness* (Rule e_2), and *Arbroath* to *Perth* (Rule e_9). The analysis of sequential pattern can be done at more fined grained level depending on the practical application through the mapping of GPS data to locations using Geocoding API.

The analysis in Section 4.3.3 shows that the travel diaries constructed from geotagged photos can capture effectively the international travel behavior for the case of Australia. The travel diaries have captured popular travel sequences between Asian cities, as shown Table 11. *Bangkok*, *Ho Chi Minh City* and *Kuala Lumpur* are major based destinations to travel to other places in South East Asia, while travelers usually travel from *Hong Kong* to cities in Northern Asia such as *Shanghai* and *Tokyo*. It is interesting to see that Singapore is frequently visited after other cities despite being a major destination in Asia. DMOs can then advertise suitable travel itineraries for Australian travelers following such frequent patterns. The heat map of the transition in Figure 1 confirmed that the travel diaries could capture the popular travel paths from Asia to Europe via some major hub destinations. Researchers can adopt the proposed travel diary construction approach in further analysis of travel behavior.

It should be noted that the sequential rules are different from traditional approaches of sequential pattern analysis as SRM aims to identify strong sequential association between the visited destinations based on the confidence. The approaches used in prior works (Xia et al. 2010; Orellana et al. 2012) may be able to identify some frequent sequential patterns as show in Section 4.3.3, but they are incapable of extracting the sequential association as in case of SRM shown in Sections 4.3.1 and 4.3.2.

This study is not without limitations. Although some sequential rules have been identified which reflect certain travel patterns of Australian, other factors influencing the travel decision was not considered in this study. The findings should be considered as a demonstration of how sequential patterns can be extracted from travel diaries. The travel pattern should be considered together with other factors, such as user demographic profile or travel motivation, in practical applications. Besides, gaps may exist between the findings and the actual travel behaviors. A combination of multiple data sources of geotagged photos is suggested for a specific practical application. Demographic factors were not considered to explain specific travel patterns. Besides, travel pattern of first time and repeat visitor will likely differ, as examined in prior studies (Hwang, Gretzel, and Fesenmaier, 2002; Kempermann, Joh, and Timmermans, 2004). The travel diaries constructed from the geotagged photos can capture the travel sequences, but it is uncertain if the first trip in the travel diaries is the actual first trip of the travelers, and the first trip might be prior to the data collection period. Therefore, we were unable to investigate the difference between the first and the repeated trips. The analysis of travel patterns was based on the observed behavior of travelers through the geotagged photo data. Given the limited scope of this study, we were unable to investigate the relationship between the observed travel patterns and the availability of airlines, which had shown significant influence on travel connections (Hwang, Gretzel, and Fesenmaier, 2006). Apart from the travel patterns, the actual photos taken can provide comprehensive information about the activities of travelers in a destination, which has not been considered in this study.

5. CONCLUSIONS

Insight into sequential travel patterns of travelers is important to identify preferred destinations and future travel intentions. This understanding is crucial for tourism managers and industry practitioners to design suitable travel packages and make appropriate offers. Unfortunately, such knowledge has not been fully obtained given the difficulty of capturing the complex travel behavior. Travel events usually occur over a long period, especially in the case of international travel, which makes collecting sequential travel information difficult. Traditional approaches to travel pattern analysis were unable to capture sequential association between visited destinations. To address these shortcomings, this paper presented an approach to construct travel diaries from geotagged photos and introduced SRM technique to extract sequential association of destination from travel sequences.

The effectiveness of the proposed approach was demonstrated in a case study of Australian outbound tourism, using a large data set of more than 890,000 photos from 3,623 outbound travelers. Travel diaries are constructed from geotagged photos, which contain comprehensive past travel information of travelers. The case study confirmed that the travel diaries constructed from geotagged photos are effective in capturing travel patterns. The analysis of travel diaries reveals interesting sequential association that can assist DMOs in developing better travel packages. DMOs can promote proper destinations to prospective travelers to achieve a high purchasing rate. The introduced framework with SRM technique has the potential to benefit tourism researchers worldwide from improving their understanding of travel behaviors.

One potential extension of this work is to incorporate other information reflecting the context of the travel into the analysis, in addition to the spatial and temporal information. For example, the textual meta-data and the visual content of the actual photos taken at destinations can be examined for additional insights into the activities of tourists. Other influencing factors, such as travel styles, preferences, and travel purposes, can be incorporated for more detailed insight into travelers' behavior. Photo-taking behavior is important in understanding the geotagged photo data and thus should be the focus of future research. The construction of the travel diary presented can be applied to other geotagged social media content such as those on Twitter, Facebook, and Instagram, which we shall investigate in future studies. SRM is a general-purpose approach for mining sequential associations. Aside from social media, SRM is beneficial to investigating its applicability in analyzing travel diaries constructed from other data sources, such as GPS loggers, bank transactions, and mobile tracking data. Airline availability is one of the influencing factors of travel connections. Indeed, future studies can incorporate airline network data into the analysis of sequential association for more detailed insight.

REFERENCES

Aloysius, G., and D. Binu. 2013. An approach to products placement in supermarkets using PrefixSpan algorithm. *Journal of King Saud University – Computer and Information Sciences*, 25(1): 77-87.

ABS. 2015. *3218.0 - regional population growth, Australia, 2013-14*. Technical report, Australian Bureau of Statistics, Canberra. http://www.abs.gov.au/ausstats/abs@.nsf/mf/3218.0 (accessed May 20, 2016)

ATTF (2013). *Tourism industry trend update*. Technical Report Q1 edition, Australia Tourism and Transport Forum, Sydney, Australia. http://tra.gov.au/documents/State-of-the-industry/Tourism_Update_March_Qtr_2014_FINAL.pdf (accessed May 24, 2016)

Barchiesi, D., H. S. Moat, C. Alis, S. Bishop, and T. Preis. 2015. Quantifying international travel flows using Flickr. *PLOS ONE*, 10(7): e0128470.

Bargeman, B., C.-H.. Joh , and H. Timmermans. 2002. Vacation behavior using a sequence alignment method. *Annals of Tourism Research*, 29(2):320–337.

Birenboim, A., S. Anton-Clave, A. P. Russo, and N. Shoval. 2013. Temporal activity patterns of theme park visitors. *Tourism Geographies*, 15(4):601-619.

Breakwell, G., and P. Wood. 1995. Diary Techniques. In G. M. Breakwell, S. Hammond, and C. Fife-Schaw (Eds.) *Research Methods in Psychology* , (pp. 293-301). Sage, London.

Breen, H., A. Bull, and M. Walo. 2001. A comparison of survey methods to estimate visitor expenditure at a local event. *Tourism Management*, 22(5):473-479.

Chang, L., D. Yang, S. Tang, and T. Wang. 2006. Mining compressed sequential patterns. In *Proceedings of the 2nd International Conference on Advanced Data Mining and Applications*, 761-768, Xian, China.

Cheng, Y.-T., Y.-F. Lin, K.-H. Chiang, and V. S. Tseng. 2016. Mining disease sequential risk patterns from nationwide clinical databases for early assessment of chronic obstructive pulmonary disease. In *Proceedings of the International Conference on Biomedical and Health Informatics (BHI)*, 324-327, Las Vegas, Nevada, USA.

Chua, A., L. Servillo, E. Marcheggiani, and A. V, Moere. 2016. Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management*, 57:295–310.

Coghlan, A., and P. Pearce. 2010. Tracking affective components of satisfaction. *Tourism and Hospitality Research*, 10(1):42-58.

Deng, M., and G. Athanasopoulos. 2011. Modelling Australian domestic and international inbound travel: A spatial-temporal approach. *Tourism Management*, 32(5):1075-1084.

Edwards, D., T. Griffin, B. Hayllar, T. Dickson, and S. Schweinsberg. 2009. *Understanding tourist experience and behavior in cities: An Australian case study. Technical report. Sustainable Tourism.* http://www.sustainabletourismonline.com/31/destination-access/understanding-tourist-experiences-and-behaviour-in-cities-an-australian-case-study (accessed April 10, 2016)

Fournier-Viger, P., U. Faghihi, R. Nkambou, and E. M. Nguifo. 2012. CMrules: Mining sequential rules common to several sequences. *Knowledge-Based Systems*, 25:63-76.

Fournier-Viger, P., A. Gomariz, M. Sebek, and M. Hlosta. 2014a. VGEN: Fast vertical mining of sequential generator patterns. In *Proceedings of the 16th International Conference on Data Warehousing and Knowledge Discovery*, 476-488, Munich, Germany.

Fournier-Viger, P., A. Gomariz, T. Gueniche, A. Soltani, C.-W. Wu, and V. S. Tseng. 2014b. SPMF: a java open-source pattern mining library. *Journal of Machine Learning Research*, 15:3389-3393.

Fournier-Viger, P., and V. S. Tseng. 2011. Mining top-k sequential rules. In *Proceedings of the 7th International Conference Advanced Data Mining and Applications*, 180-194, Beijing, China.

Fournier-Viger, P., C.-W. Wu, and V. S. Tseng. 2013. Mining maximal sequential patterns without candidate maintenance. In *Proceeding of the 9th International Conference on Advanced Data Mining and Applications*, 169-180, Hangzhou, China.

Furmanov, K., O. Balaeva, and M. Predvoditeleva. 2012. Tourism flows from the Russian federation to the European Union. Anatolia - *An International Journal of Tourism and Hospitality Research*, 23(1):17-31.

Garcia-Palomares, J. C., J. Gutierrez, and C. Minguez. 2015. Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography*, 63:408-417.

Hamilton, H. J., and K. Karimi. 2005. The TIMERS II algorithm for the discovery of causality. In *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 744-750, Hanoi, Vietnam.

Hwang, Y.-H., and D. R. Fesenmaier. 2003. Multidestination pleasure travel patterns: Empirical evidence from the American Travel Survey, *Journal of Travel Research*, 42:166–171.

Hwang, Y.-H., U. Gretzel, and D.R. Fesenmaier. 2002. Multi-city pleasure trip patterns: An analysis of international travelers to the U.S. *City Tourism,* edited by Wöber, K., pp. 53-62, Springer Verlag, Vienna, Austria.

Hwang, Y.-H., U. Gretzel, and D.R. Fesenmaier. 2006. Multicity trip patterns: Tourists to the United States. *Annals of Tourism Research,* 33(4):53-62.

Ian, P., P. Shane, and L. Jillian. 2011. Grey nomads on tour: A revolution in travel and tourism for older adults. *Tourism Analysis*, 16(3): 283-294.

Kádár, B., and M. Gede. 2013. Where Do Tourists Go? Visualizing and Analysing the Spatial Distribution of Geotagged Photography. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 48(2):78–88.

Kádár, B. 2014. Measuring tourist activities in cities using geotagged photography. *Tourism Geographies*, 16(1):88–104.

Kempermann, A., C.-H. Joh, and H. Timmermans. 2004. Comparing First-time and Repeat Visitors' Activity Patterns in a Tourism Environment. *Consumer Psychology of Tourism, Hospitality and Leisure*, edited by G. I. Crouch, R. R. Perdue, H. Timmermans, & M. Uysal, pp.103-119, Cambridge, MA: CABI Publishing

Krentzman, A. R., E. A. R., Robinson, J. M. Jester, and B. E. Perron (2011). Heat maps: A technique for classifying and analyzing drinking behavior. *Substance Use & Misuse*, 46(5), 687-695.

Law, G., and B. McKercher. 2006. Understanding tourist movement patterns in a destination: a GIS approach. *Tourism and Hospitality Research,* 7:39-49

Law, R., J. Rong, H. Q. Vu, G. Li, and H. A. Lee. 2011. Identifying changes and trends in Hong Kong outbound tourism. *Tourism Management,* 32:1106-1114

Leung, X. Y., F. Wang, B. Wu, B. Bai, K. A. Stahura, and Z. Xie. 2012. A social network analysis of overseas tourist movement patterns in Beijing: the impact of the Olympic Games. *International Journal of Tourism Research,* 14:469-484.

Lew, A., and B. McKercher. 2006. Modelling tourist movement a local destination analysis. *Annals of Tourism Research,* 33(2):403-423.

Li, X., F. Meng, and M. Uysal. 2008. Spatial pattern of tourist flows among the Asian Pacific countries: An examination over a decade. *Asia Pacific Journal of Tourism Research,* 13(3):229-243.

Mabroukeh, N. R., and C. I. Ezeife. 2010. A taxonomy of sequential pattern. *ACM Computing Surveys,* 43:1-41.

McKercher, B., and G. Lau. 2008. Movement patterns of tourists within a destination. *Tourism Geographies*, 10(3):355-374.

Onder, I., W. Koerbitz, and A. Hubmann-Haidvogel . 2014. Tracing tourists by their digital footprints: The case of Austria. *Journal of Travel Research*, 55:566-573.

Orellana, D., A. K. Bregt, A. Ligtenberg, and M. Wachowicz. 2012. Exploring visitor movement patterns in natural recreational areas. *Tourism Management*, 33:672-682.

Prideaux, B. 2000. The role of the transport system in destination development. *Tourism Management*, 21(1):53-63.

Pocock, N., and A. McIntosh. 2013. Long-term traveler return, 'home'? *Annals of Tourism Research*, 42:402-424.

Raun, J., R. Ahas, and M. Tiru. 2016. Measuring tourism destinations using mobile tracking data. *Tourism Management*, 57: 202–212.

Sheng, C.-W., and M.-C. Chen. 2013. Tourist experience expectations: questionnaire development and text narrative analysis, *International Journal of Culture, Tourism and Hospitality Research*, 7(1):93 – 104.

Shie, B.-E., H.-F. Hsiao, P. S. Yu, V. S. Tseng. 2012. Discovering Valuable User Behavior Patterns in Mobile Commerce Environments. In *New Frontiers in Applied Data Mining*, edited by L. Cao, J. Z. Huang, J. Bailey, Y. S. Koh and J. Luo, 77-88. Springer Berlin Heidelberg.

Shao, J., and U. Gretzel. 2010. Looking Does not Automatically Lead to Booking: Analysis of Clickstreams on a Chinese Travel Agency Website. In Proceedings of *Information and Communication Technologies in Tourism*, pp. 197-208, Lugano, Switzerland.

Shoval, N., B., McKercher, A., Birenboim, and E. Ng. 2015. The Application of a Sequence Alignment Method to the Creation of Typologies of Tourist Activity in Time and Space. *Environment and Planning B: Planning and Design,42*:76–94

Sobolevsky, S., R. Sitko , R. T. D. Combes, B. Hawelka, J. M. Arias, and C. Ratti. 2014. Money on the Move:Big Data of Bank Card Transactions as the New Proxy for Human

Mobility Patterns and Regional Delineation. The Case of Residents and Foreign Visitors in Spain. *IEEE International Congress on Big Data*, pp. 136-143, Anchorage, AK.

Sobolevsky, S., I. Bijic , A. Belyi, I. Sitko, B. Hawelka, J. M. Arias and C. Ratti. 2015. Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity. *IEEE International Congress on Big Data*, pp. 600-607, New York, NY.

Tan, P.-N., S. Michael, and V. Kumar. 2005. Introduction to Data Mining, Chapter 6. *Association Analysis: Basic Concepts and Algorithms*. Addison-Wesley, Boston, USA.

TRA 2014. *Tourism update - updated results to 'state of the industry 2013'*. Technical report, Tourism Research Australia, Canberra. https://www.tra.gov.au/documents/State-of-the-industry/TRA_State_of_the_Industry_2014_FINAL.pdf (accessed May 10, 2016)

TRA 2014b. *Australian Traveling Oversea*, Tourism Research Australia. Viewed on 22 September 2016 < http://www.tra.gov.au/statistics/australians-travelling-overseas.html>

Versichele, M., L. de Groote, M.C. Bouuaer, T. Neutens, I. Moerman, and N. Van de Weghe. 2014. Pattern mining in tourist attraction visits through association rule learning on bluetooth tracking data: A case study of Ghent, Belgium. *Tourism Management,* 44: 67–81.

Vu, H. Q., G. Li, R. Law, and B. H. Ye. 2015. Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tourism Management,* 46:222-232.

Xia, J., F. Evans, K. Spilsbury, V. Ciesielski, C. Arrowsmith, and G. Wright. 2010. Market segments based on dominant movement patterns of tourists. *Tourism Management*, 31(4):464-469.

Xiang, Z, and U. Gretzel. 2010. Role of social media in online travel information search. *Tourism Management*, 31(2): 179–188.

Yan, X., J. Han, and R. Afshar. 2003. CloSpan: Mining closed sequential patterns in large datasets. In *Proceeding of SIAM International Conference on Data Mining*, San Francisco, 166-177, California, USA.

Yang, D.-l., Y. L. Hsieh, and J. Wu. 2006. Using data mining to study upstream and downstream causal relationship in stock market. In *Proceedings of the 9th Joint Conference on Information Sciences,* Kaohsiung, Taiwan.

Yang, Y., T. Fik, and J. Zhang. 2013. Modeling sequential tourist flows: Where is the next destination? *Annals of Tourism Research,* 43:297-320.

Yuan, Y., and M. Medel.2016. Characterizing international travel behavior from geotagged photos: A case study of Flicr. *Plos ONE*, 11(5): e0154885.

Zach, F., and U. Gretzel. 2012. Tourist-activated networks: implications for Dynamic Bundling and EN Route Recommendations. *Information Technology & Tourism*, 13(3):229–238.

Zakrisson, I., and M. Zillinger. 2012. Emotions in motion: tourist experiences in time and space. *Current Issues in Tourism*, 15(6):505–523.

Zheng, Y.-T., Z.-J. Zha, and T.-S. Chua. 2012. Mining travel patterns from geotagged photos. *ACM Transactions on Intelligent Systems and Technology*, 3(3):1–18.

Zheng, Z., W. Wei, C. Liu, W. Cao, L. Cao, and M. Bhatia. 2016. An effective contrast sequential pattern mining approach to taxpayer behavior analysis. *World Wide Web,* 19(4):633-651.

Table 1: Travel Diary Example

ID	Date	Trip ID	Latitude / Longitude	City	Country/Continent
P1	21-Jun-14	1	48.8529 / 2.2992	Paris	France/Europe
P2	21-Jun-14	1	48.8734 / 2.2953	Paris	France/Europe
P3	22-Jun-14	1	48.8410 / 2.3207	Paris	France/Europe
P4	23-Jun-14	1	40.4913 / -3.5920	Madrid	Spain/Europe
P5	24-Jun-14	1	40.4108 / -3.7073	Madrid	Spain/Europe
P6	24-Jun-14	1	40.4334 / -3.7042	Madrid	Spain/Europe
P7	25-Jun-14	1	40.4175 / -3.7143	Madrid	Spain/Europe
P8	26-Jun-14	1	41.7949 / 12.2506	Rome	Italy/Europe
P9	26-Jun-14	1	41.8913 / 12.4918	Rome	Italy/Europe
P 0	28-Jun-14	1	41.9105 / 12.4764	Rome	Italy/Europe
P11	1-Dec-15	2	13.6922 / 100.7512	Bangkok	Thailand/Asia
P12	1-Dec-15	2	13.7403 / 100.5090	Bangkok	Thailand/Asia
P13	2-Dec-15	2	13.7509 / 100.4984	Bangkok	Thailand/Asia
P14	3-Dec-15	2	13.7551 / 100.5115	Bangkok	Thailand/Asia
P15	5-Dec-15	2	12.9292 / 100.8770	Pattaya	Thailand/Asia
P16	5-Dec-15	2	12.9307 / 100.8781	Pattaya	Thailand/Asia
P17	7-Dec-15	2	1.2950 / 103.8583	Singapore	Singapore/Asia
P18	7-Dec-15	2	1.2809 / 103.8638	Singapore	Singapore/Asia
P19	7-Dec-15	2	1.2905 / 103.8455	Singapore	Singapore/Asia
P20	8-Dec-15	2	1.3620 / 103.9906	Singapore	Singapore/Asia

Table 2: Travel Sequences

Trip	Sequences (country level)	Sequences (city level)
1	$\langle France \rightarrow Spain \rightarrow Italy \rangle$	$\langle Paris \rightarrow Madrid \rightarrow Rome \rangle$
2	$\langle Thailand \rightarrow Singapore \rangle$	$\langle Bangkok \rightarrow$ Pattaya $\rightarrow Singapore \rangle$

Table 3. A Sequential Database

ID	Sequences
s_1	$\langle l_1 \to l_2 \to l_3 \to l_6 \to l_7 \to l_5 \rangle$
s_2	$\langle l_1 \to l_4 \to l_3 \to l_2 \to l_1 \to l_2 \to l_5 \to l_6 \rangle$
s_3	$\langle l_1 \to l_2 \to l_6 \to l_5 \rangle$
s_4	$\langle l_2 \to l_6 \to l_7 \to l_8 \rangle$

Note: symbol "→" represents transition

Table 4. Sequential Rules

ID	Rule	Support	Confidence
r_1	$\{l_1, l_2, l_3\} \Rightarrow \{l_5\}$	0.5	1.0
r_2	$\{l_1\} \Rightarrow \{l_3, l_5, l_6\}$	0.5	0.66
r_3	$\{l_1, l_2\} \Rightarrow \{l_5, l_6\}$	0.75	1.0
r_4	$\{l_2\} \Rightarrow \{l_5, l_6\}$	0.75	0.75
r_5	$\{l_1\} \Rightarrow \{l_5, l_6\}$	0.75	1.0
r_6	$\{l_3\} \Rightarrow \{l_6\}$	0.5	1.0
r_7	$\{l_1\} \Rightarrow \{l_2\}$	0.75	1.0

Note: symbol "⇒" represents rule

Table 5. Geotagged Photo Data Collection

Travel Group	# of Users	# of Photos	# of Photos/User	Average Time Span (Year)
Sydney	1,435	367,207	255.89	1.66
Melbourne	1,033	248,056	240.13	3.01
Brisbane	479	101,940	212.82	2.90
Perth	463	125,049	270.08	2.96
Adelaide	213	48,061	225.64	2.44
Total	3,623	890,313	245.74	

Table 6. Destinations of Outbound Trips

Continent	Proportion (%)			*z*-Score	*p*-Value*
	Single Country	**Two Countries or More**	Difference		
Africa	3.44	3.07	-0.37	1.225	0.219
America	18.33	13.71	-4.62	8.5072	*0.000*
Asia	33.72	34.33	0.61	-1.2974	0.194
Europe	29.98	43.16	12.18	-26.715	*0.000*
Oceania	14.54	5.29	-9.25	18.122	*0.000*
Number of Trips	12,819	4,369			

*Significant at $p \leq 0.05$

Table 7. Popularly Visited Countries

Country (Code)	# of Travelers	# of Trips	# Trips / Traveler
United States of America (USA)*	1,119	2,535	2.27
United Kingdom (GBR)*	1,040	2,967	2.85
New Zealand (NZL)*	818	1,544	1.89
France (FRA)	702	1,184	1.69
Italy (ITA)	596	1,012	1.70
Japan (JPN)	503	941	1.87
Singapore (SGP)*	496	762	1.54
Malaysia (MAL)*	464	754	1.63
Thailand (THA)*	452	703	1.56
China (CHN)*	440	834	1.90
Germany (DEU)	417	753	1.81
Hong Kong (HKG)*	412	652	1.58
Indonesia (IDN)*	386	631	1.63
Spain (ESP)	350	514	1.47
Canada (CDN)	304	580	1.91
India (IND)	271	454	1.68
Viet Nam (VNM)	255	363	1.42
Switzerland (CHE)	232	343	1.48
Netherland (NLD)	230	425	1.85
Austria (AUT)	185	246	1.33

*Among Top 10 Destinations according to National Outbound Survey by Tourism Research Australia (TRA 2014b)

Table 8. Sequential Rules by Country

Sequential Rules	**Support**	**Confidence**	***Rule ID***
America			
$CDN \Rightarrow USA$	0.051	0.743	r_1
$MEX \Rightarrow USA$	0.014	0.768	r_2
$BOL \Rightarrow PER$	0.006	0.871	r_3
Asia			
$LAO \Rightarrow THA$	0.009	0.661	r_4
Europe			
$CZE, FRA \Rightarrow DEU$	0.009	0.621	r_5
$AUT, FRA \Rightarrow DEU$	0.013	0.663	r_6
$AUT, CZE \Rightarrow DEU$	0.008	0.642	r_7
$AUT, FRA \Rightarrow ITA$	0.012	0.639	r_8
$FRA, GRC \Rightarrow ITA$	0.008	0.733	r_9
$BIH \Rightarrow HRV$	0.008	0.673	r_{10}
$ITA, ESP \Rightarrow GBR$	0.015	0.638	r_{11}
$AUT, FRA \Rightarrow GBR$	0.012	0.639	r_{12}
$FRA , DEU \Rightarrow GBR$	0. 030	0.638	r_{13}
$FRA , NLD \Rightarrow GBR$	0.015	0.653	r_{14}
$DEU , ESP \Rightarrow GBR$	0.011	0.667	r_{15}
$IRL \Rightarrow GBR$	0.024	0.682	r_{16}
$ISL \Rightarrow GBR$	0.009	0.684	r_{17}
$FRA, ITA, CHE \Rightarrow GBR$	0.011	0.696	r_{18}
$FRA, ITA, ESP \Rightarrow GBR$	0.011	0.716	r_{19}
$BEL, DEU \Rightarrow GBR$	0.009	0.717	r_{20}
$FRA, DEU, CHE \Rightarrow GBR$	0.010	0.729	r_{21}
$FRA, DEU, NLD \Rightarrow GBR$	0.010	0.763	r_{22}
$DEU, ITA, ESP \Rightarrow GBR$	0.009	0.810	r_{23}

Table 9. Sequential Rules by City

Sequential Rules	**Support**	**Confidence**	***Rule ID***
America			
$Chicago\ (USA) \Rightarrow\ LosAngeles\ (USA)$	0.146	0.646	c_1
$Denver\ (USA) \Rightarrow\ LosAngeles\ (USA)$	0.139	0.775	c_2
$Chicago\ (USA), Denver\ (USA) \Rightarrow\ LosAngeles\ (USA)$	0.042	0.754	c_3
$Edmonton\ (CDN) \Rightarrow\ Vancouver\ (CDN)$	0.051	0.722	c_4
$La\ Paz\ (BOL) \Rightarrow\ Lima\ (PER)$	0.027	0.871	c_5
Asia			
$Hebron\ (ISR) \Rightarrow\ Jerusalem(ISR)$	0.009	0.661	c_6
$Kathmandu\ (NPL) \Rightarrow\ Kolkata(IND)$	0.008	0.733	c_7
$Jakarta\ (IDN), Kuala\ Lumpur\ (MYS) \Rightarrow\ Singapore(SGP)$	0.008	0.673	c_8
Europe			
$Monaco\ (MCO) \Rightarrow\ Paris\ (FRA)$	0.029	0.932	c_9
$Amsterdam\ (NLD), Madrid\ (ESP) \Rightarrow\ Paris\ (FRA)$	0.012	0.815	c_{10}
$Sarajevo\ (BIH) \Rightarrow\ Zagreb\ (HRV)$	0.017	0.702	c_{11}

Table 10. Sequential Rules by City in the United Kingdom

Sequential Rules	**Support**	**Confidence**	***Rule ID***
$Dunfermline \Rightarrow Edinburgh$	0.031	0.622	e_1
$Elgin \Rightarrow Inverness$	0.022	0.833	e_2
$Greenock, Inverness \Rightarrow Isle\ of\ Lewis$	0.023	0.600	e_3
$Barrow\ in\ Furness \Rightarrow Kendal$	0.013	0.600	e_4
$Edinburgh, Kidlington \Rightarrow Oxford$	0.013	0.600	e_5
$Cowley \Rightarrow Oxford$	0.012	0.846	e_6
$Dundee \Rightarrow Perth$	0.049	0.677	e_7
$Alloa \Rightarrow Perth$	0.020	0.783	e_8
$Arbroath \Rightarrow Perth$	0.018	0.889	e_9
$Glenrothes \Rightarrow Perth$	0.017	0.833	e_{10}

Table 11. Sequential Patterns for Destinations in Asia

Sequential Pattern	**Support**	***ID***
$Bangkok \rightarrow Kuala\ Lumpur$	0.055	s_1
$Bangkok \rightarrow Singapore$	0.049	s_2
$Bangkok \rightarrow Kuala\ Lumpur \rightarrow Singapore$	0.018	s_3
$Bangkok \rightarrow Ho\ Chi\ Minh$	0.041	s_4
$Bangkok \rightarrow Ho\ Chi\ Minh \rightarrow Phnom\ Penh$	0.020	s_5
$Bangkok \rightarrow Vientiane$	0.031	s_6
$Bangkok \rightarrow Hong\ Kong$	0.024	s_7
$Ho\ Chi\ Minh \rightarrow Phnom\ Penh$	0.053	s_8
$Ho\ Chi\ Minh \rightarrow Singapore$	0.035	s_9
$Ho\ Chi\ Minh \rightarrow Kuala\ Lumpur$	0.027	s_{10}
$Ho\ Chi\ Minh \rightarrow Kuala\ Lumpur \rightarrow Singapore$	0.024	s_{11}
$Ho\ Chi\ Minh \rightarrow Vientiane$	0.024	s_{12}
$Hong\ Kong \rightarrow Shanghai$	0.101	s_{13}
$Hong\ Kong \rightarrow Macau$	0.051	s_{14}
$Hong\ Kong \rightarrow Macau \rightarrow Shanghai$	0.021	s_{15}
$Hong\ Kong \rightarrow Singapore$	0.031	s_{16}
$Hong\ Kong \rightarrow Tokyo$	0.035	s_{17}
$Kuala\ Lumpur \rightarrow Singapore$	0.134	s_{18}
$Kuala\ Lumpur \rightarrow Kuching$	0.026	s_{19}
$Kuala\ Lumpur \rightarrow Makassar$	0.024	s_{20}
$Shanghai \rightarrow Tokyo$	0.034	s_{21}
$Singapore \rightarrow Tokyo$	0.031	s_{22}
$Kathmandu \rightarrow Kolkata$	0.024	s_{23}
$Kolkata \rightarrow Singapore$	0.036	s_{24}

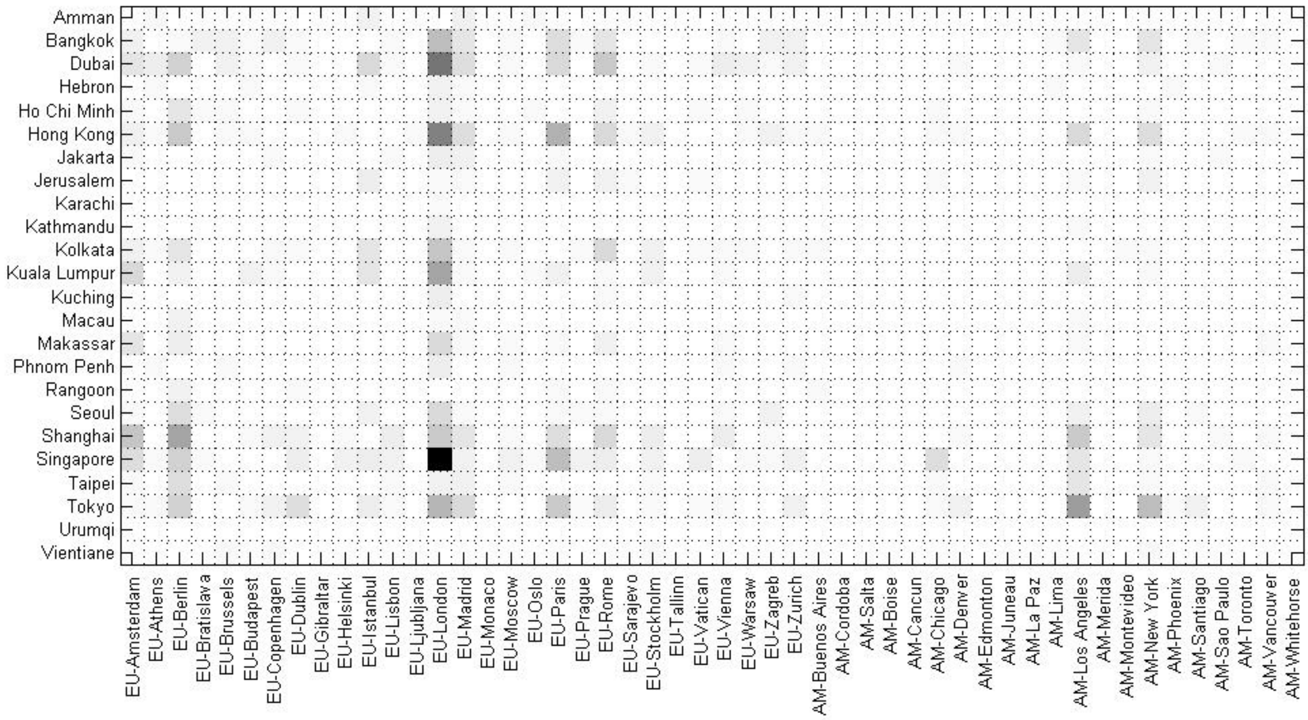

Figure 1. Transition from Asia to Europe and America