A Novel Approach to Mining Travel Sequences Using Collections of Geotagged Photos

Slava Kisilevich[1], Daniel Keim[1], Lior Rokach[2]

[1] University of Konstanz, Konstanz, Germany
{slaks, keim}@dbvis.inf.uni-konstanz.de
[2] Department of Information Systems Engineering and
The Deutsche Telekom Laboratories
Ben-Gurion University of the Negev, Beer-Sheva, Israel
liorrk@bgu.ac.il

Abstract. In this paper we present a novel approach for analyzing the trajectories of moving objects and of people in particular. The mined data from these sequences can provide valuable information for understanding the surrounding locations, discovering attractive place or mining frequent sequences of visited places. Based on geotagged photos, our framework mines semantically annotated sequences. Our framework is capable of mining semantically annotated sequences of any length to discover patterns that are not necessarily immediate antecedents. The approach consists of four main steps. In the first step, every photo location is semantically annotated by assigning it to a known nearby point of interest. In the second step, a density-based clustering algorithm is applied to all unassigned photos, creating regions of unknown points of interest. In the third step, a travel sequence of every individual is built. In the final step, travel sequence patterns are mined using the semantics that were obtained from the first two steps. Case studies of Guimarães, Portugal (where the conference takes place) and Berlin, Germany demonstrate the capabilities of the proposed framework.

Keywords: sequence mining, trajectories, sequence patterns

M. Painho et al., (eds.), *Geospatial Thinking*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-642-12326-9_9,

1 Introduction

Location acquisition technologies and web-centric information sharing are ubiquitous in today's world and have become a focus for research in a variety of fields, data mining in particular, due to the vast quantity of data involved. Existing works on analyzing people's mobility mainly concentrate on the trajectories obtained by GPS-enabled devices. Such trajectories usually consist of many space-and-time referenced points measured at a constant interval where the foremost nontrivial task is to extract (semantically) important parts or stay points.

Several approaches exist to find the important elements: (1) applying density functions to find regions where intersections of trajectories are high or (2) finding parts of a trajectory where the object stayed for a significant period of time. After the stay points are found, data mining algorithms can be applied to mine frequent sequences.

These approaches involve several issues. (1) Important intersection sites for various individuals may seem to be the same but in fact correspond to different sites that were visited. For example, one person visited a bank and another entered a shop. The bank and the shop are situated close to each other and these regions were defined as one stay point in the trajectory of these two persons. (2) Since the stay points are defined mainly using characteristics of the trajectory, without any background knowledge, there is a need to interpret the obtained sequences. The first issue can be tackled by assuming that the regions visited by people are important, making no distinction between the sites visited in these regions. The second issue can be resolved by using external databases of points of interest (POIs) to explain the important places. However, a POI database may be unavailable, inapplicable to the data (shopping, work) or incomplete.

Large-scale, GPS-based datasets of people's trajectories are still unavailable partly because of data acquisition problems. For example, Zhen et al. (2009) reported that a *large* GPS dataset was created from data collected by 107 users carrying GPS-enabled devices with them for one year. The regions that these users covered included 36 cities in China and various areas in the USA, South Korea, and Japan. Without regard to the difficulty of data acquisition, the question of whether 107 users are enough to mine travel sequences in different parts of the world remains open. We will also try to answer this question in this paper.

A recent trend in analyzing people's activity and travel behavior is the use of geotagged photos shared by people and publicly available on such photo-sharing sites like Flickr[1] or Panoramio[2]. The data from these geo-

[1] http://www.flickr.com

tagged photos differ technically and semantically from raw GPS-based type trajectories. Unlike trajectories recorded by GPS devices and measured at a constant time, photo data can be regarded as a private case of raw trajectories in which an individual is capturing an important event. Using time and location of photos taken by a person, it is possible to construct event-based trajectories, which can then be used to analyze travel activity. The act of sharing the photo with others through photo-sharing sites reveals important information, including time, location, title, tags and the photo itself. Therefore, this data can be directly used in retrieving interesting places, providing us with the opportunity to discover travel sequences and understand in what order people visit such places.

In this paper, we address the problem of automatically finding semantically annotated sequences. For instance, consider the following sequences:

1. A → B →C
2. A → * → D

The first sequence can be interpreted as a route followed by people from place A to place B and from place B to place C. It is important to note that those who reached C from B are the same persons as those who reached B from A. In the second sequence, those who started from A and reach D, did not necessarily visit a particular place, rather they may have visited any possible place before visiting D.

Our approach to mining travel sequences consists of four main parts. In the first part, we automatically assign every geotagged photo to a nearby POI using an external POI database. Since we do not perform any image analysis, we cannot really know what was photographed. However, the fact that the photo was taken near some known POI assumes the presence of the photographer in that place. After step one, there are photos that were not assigned to any POI. There are two reasons for this. (1) A photo was taken in an area where there are no POIs (for example a forest or parking lot near someone's house). (2) A photo was taken in an attractive place but a POI is missing in the database. Therefore, there is still a need to analyze these locations and artificially create points of interest using several constraints. For this purpose, we apply a density-based clustering algorithm in order to find dense areas (Rokach and Maimom 2005). This allows us to filter out outliers – sparse areas, where the number of people who took photos is less than a predefined threshold. The dense regions that are obtained are new, unknown points of interest which are added to the areas acquired in the first step. The automated process annotates these areas with symbolic names and stores the boundaries of these regions for future access. In the third step, the travel sequence of each person is con-

[2] http://www.panoramio.com

structed using the notion of a session: a time frame in which a person takes photos in a particular area. In the fourth step, travel sequence patterns are mined using semantics obtained from the first two steps.

The goal of this paper is to suggest an automatic approach for mining semantically annotated travel sequences using geotagged photos by searching for sequence patterns of any length. The sequences obtained may contain patterns that are not necessarily the immediate antecedents. Moreover, the approach that we propose can examine sequences in which the same pattern is repeated more than once in the same sequence.

The main contribution of this paper is the development of a new data mining process that employs concepts that have been developed in various other fields such as bioinformatics and artificial intelligence.

2 Related work

The mining of frequent sequential patterns in databases of customer transactions was first presented by Agrawal and Srikant (1995). The method adopts an a-priori-like approach (Agrawal and Srikant, 1994) where the idea is to find subsets that are common to at least a minimum number of sequences, termed itemsets. The method uses the following observation: if the sequence of length k is not frequent, then neither can the sequence of length $k+1$ ever be frequent. The algorithm can be applied to generic items provided they can be sorted using transaction time. Time, however, is not considered in pattern mining. The limitation of the approach is that it cannot find sequences with repeating patterns and sequences in which patterns are not necessarily immediate antecedents.

There are application domains where time duration between adjacent events is also important. This issue was addressed in MiSTA, a generic algorithm for mining temporally annotated sequences, where frequent patterns are mined using sequence and temporal similarity (Giannotti et al., 2006). As an extension to MiSTA, Giannotti et al. (2007) presented three different approaches to mining trajectory sequences that are reflecting site visits at approximately the same time. These two approaches share the idea that the transformation of a trajectory into a sequence of significant parts and the application of semantic meaning are done as a preprocessing step prior to mining the sequence patterns. Since the trajectory is transformed into a sequence of generic events, the MiSTA algorithm can be directly applied to them.

The MiSTA authors suggested two general methods for performing preprocessing. In the first case, background knowledge should be applied to

trajectories. To perform this task may require an additional database of POIs or a domain expert. In the second case, significant parts are found without using background knowledge, only the properties of the trajectories themselves. Specifically, the authors proposed to divide the area of investigation into grids and to count the density of trajectories in every grid. Thus, the significant places are defined in terms of frequency of visits by different persons. In contrast to temporal annotated sequences, we define sequences as a frequent move from one place to another without regard to time similarity.

Alvares et al. (2007a) proposed a generic model for semantically annotating trajectories and representing a moving pattern in the geographic database. This approach has two main parts. In the first part, the significant places in a trajectory are found by identifying moves and stops (Spaccapietra et al., 2007). Stops are significant places that are also called *stay points,* sites where a person stayed for a certain period of time. The extraction of stops depends on time and distance thresholds. Moves are transitions between consecutive stops. In the second part, stops are integrated into the database along with geographic data like POIs. This makes it possible to perform spatial queries on stop regions by annotating them with semantically meaningful information. They demonstrated this approach for mining frequent trajectory patterns between two stops of conference attendees (Alvares et al., 2007b).

Zheng et al. (2009) mine travel sequences by inferring interesting places from trajectories and the person's experience. The method is based on calculating probabilities that a person will take a specific path using information about how many people move from one place to another. The most interesting sequences of length n can be found by summing the probabilities of every two-length sequence comprising the larger sequence and selecting sequences with high score. However, the notion of such sequences differs from classical sequences based on the frequency of patterns. The authors report that finding sequences of length n is possible but a time consuming process and hypothesize that people would not likely visit many places in a trip. Thus, two-length sequences were only considered in their paper.

The main differences between our work and existing state-of-the-art methods can be summarized as follows:

1. We work with trajectories on the semantic level instead of trajectories as raw points.
2. We introduce a concrete approach for semantically annotating points within a trajectory.

3. Interesting places are found before mining sequence patterns in contrast to existing approaches where interesting places are found using characteristics of trajectories such as density, frequency, stay time, stop points.
4. The sequence patterns can be of any length in which patterns are not necessarily immediate antecedents.
5. We evaluate our algorithm on a real-world database obtained from Flickr.

3 Framework

Fig. 1 presents the proposed framework. First, we try to match photo coordinates with known POIs. Then the remaining unassigned photos are clustered and new POIs are identified. This is followed by converting the individual's trajectory into sequences of POIs. These sequences are analyzed and new sequence patterns are discovered. The following subsections describe each step in more detail.

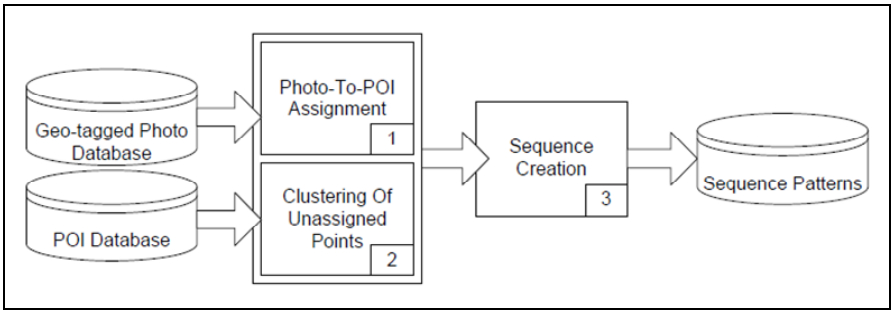


Fig. 1. Steps of the framework

3.1 Datasets

We collected metadata of geotagged photos from the Flickr photo-sharing Web site using its publicly available API. The Flickr API does not allow downloading metadata by providing exact boundaries of the area of interest. Therefore, we used an approach similar to Web crawling. We downloaded all the photo metadata of arbitrarily selected subjects and obtained the list of their contacts as well as the list of groups their photos belonged to. The same procedure was iteratively applied on other retrieved users.

We began collecting the data from the beginning of June. By the end of October 2009, we managed to collect 64,975,609 entries from 2,617,271 users. In the preprocessing step, we converted coordinates expressed in degrees into meters based on Universal Transverse Mercator (UTM) coordinates. This was done in order to enable us to apply distance-related functions. In total, 56,305 entries with wrong or missing dates were removed. These entries included 6,229 with incorrect dates; 50,076 photos were taken after October 1, 2009,

We used the Wikipedia database[3] as a source for POI data. This database is an on-going community project aimed at applying geographic annotation to articles describing interesting sites around the world. The database we obtained contains 450,637 entries of various geotagged sites such as cities, landmarks, monuments, buildings, towers, etc. For our purposes, the most important information that the entries contained were *id, title,* and *coordinates.*

3.2 Photo to POI assignment

In this step, every geotagged photo from the database is matched to a nearby POI using a distance threshold called *photo-to-POI.* If the distance between the photo and a POI is not longer than the *photo-to-POI* distance threshold, the photo is assigned to that POI. If there are several POIs within the distance threshold, the photo is assigned to the closest POI.

3.3 Discovery of new POIs based on unassigned photos

In this step, we use a clustering algorithm to create regions of unknown POIs using photos that were not assigned during the previous step. In our previous work (Kisilevich et al., 2010), we showed that density-based clustering can be used in finding attractive areas. In general, the density based clustering algorithms has several advantages over other types of clustering algorithms: Density based clustering algorithms require minimum domain knowledge to determine the input parameters and can discover clusters with arbitrary shape. In addition, density-based clustering algorithms can filter outliers and work effectively when applied to large databases.

[3] http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Wikipedia-World/en

3.4 Sequence Creation

In this step, we assemble the POIs visited by a person into a sequence of places using the time stamp of the photo. If two consecutive photos are assigned to the same POI, only one photo is taken into consideration. We discard sequences that have only one POI since they do not contribute to discovering new sequence patterns. In general, sequences of any length can be built in this step. However, sequence creation can be constrained using such criteria as a time interval between every two consecutive photos or a total time interval between first and last photo. For example, Girardin et al. (2009) applied a 30-day interval threshold to differentiate between tourists, whose photo sessions lasted less than 30 days and locals whose sessions were longer. This heuristic approach can be used for differentiating between travel patterns of various groups of visitors. In our experiments, we implemented the same idea.

3.5 Sequence Patterns

The term "sequence pattern" usually refers to a set of short sequences that is precisely specified by some formalism. As is the practice in bioinformatics research, we are also adopting a regular expression in order to represent sequence patterns. A pattern is defined as any string consisting of a letter of the alphabet and the wild-card character '*'. The wild-card (also known as the "don't care" character) is used to denote a position that can be occupied by any letter of the alphabet.

In this paper, we consider the Teiresias algorithm (Rigoutsos and Floratos, 1998) which was originally developed as a combinatorial pattern discovery algorithm in bioinformatics for analyzing DNA sequences. The algorithm identifies recurrent maximal patterns within sequences. Although the method is combinatorial in nature and able to produce all patterns that appear in at least a (user-defined) minimum number of sequences, it achieves a high degree of efficiency by avoiding the enumeration of the entire pattern space. The algorithm, which has also been successfully used for information retrieval and intelligent manufacturing (Rokach el al., 2008A and Rokach et al., 2008B), performs a well-organized exhaustive search. In the worst case, the algorithm is exponential, but works very well for usual inputs. Furthermore, the reported patterns are maximal; any reported pattern cannot be made more specific and still keep on appearing at the exact same positions within the input sequences. Teiresias searches for patterns that satisfy certain density constraints, limiting the number of wild-cards occurring in any stretch of pattern. More specifically, Teiresias

looks for maximal <L,W> patterns with support of at least K (i.e. in the corpus there are at least K distinct sequences that match this pattern). A pattern P is called <L,W> pattern if every sub-pattern of P with length of at least W operations (combination of specific operations and "." wild-card operations) contains at least L specific operations. For example, given the following corpus of 6 trajectory sequences:

1. Reichstag → Der Bevölkerung → Brandenburg Gate → Memorial to the Roma and Sinti Holocaust Victims → Pariser Platz

2. Reichstag → Marienviertel → Memorial to the Murdered Jews of Europe → Brandenburg Gate

3. Reichstag → Berliner Dom → Liebknecht Bridge → Checkpoint Charlie → Brandenburg Gate → Treptower Park → Pariser Platz

4. Reichstag → 18th of March Square → Brandenburg Gate

5. Potsdamer Platz → Zoological Garden → Marienviertel → Reichstag → 18th of March Square → Brandenburg Gate

6. Sony Center → Pleasure Garden → Reichstag → Der Bevölkerung → Unter den Linden → Memorial to the Murdered Jews of Europe → Brandenburg Gate

The Teiresias program (L=K=2 and W=3) discovers 5 recurring patterns shown in the following table. The first column represents the support of the pattern.

Table 1. Illustrative results of the Teiresias algorithm

#	Sequence patterns
2	Reichstag → 18th of March Square → Brandenburg Gate
2	Reichstag → Der Bevölkerung
2	Memorial to the Murdered Jews of Europe → Brandenburg Gate
3	Reichstag → * → Brandenburg Gate
2	Brandenburg Gate → * → Pariser Platz

4 Experimental Evaluation

In this section, we present an experimental evaluation using two case studies of areas in Guimaraes, Portugal (where the conference takes place) and Berlin, Germany. In particular, this experimental study has the following goals:

1. To examine whether the proposed method can be applied to regions with different scales, number of persons and their photos, and several points of interest.
2. To examine the effect on travel patterns of such parameters as the photo-to-POI threshold (Sect. 3.2), the distance threshold for density-based clustering and the minimum number of people in a cluster (Sect. 3.3), and session length (Sect. 3.4), .

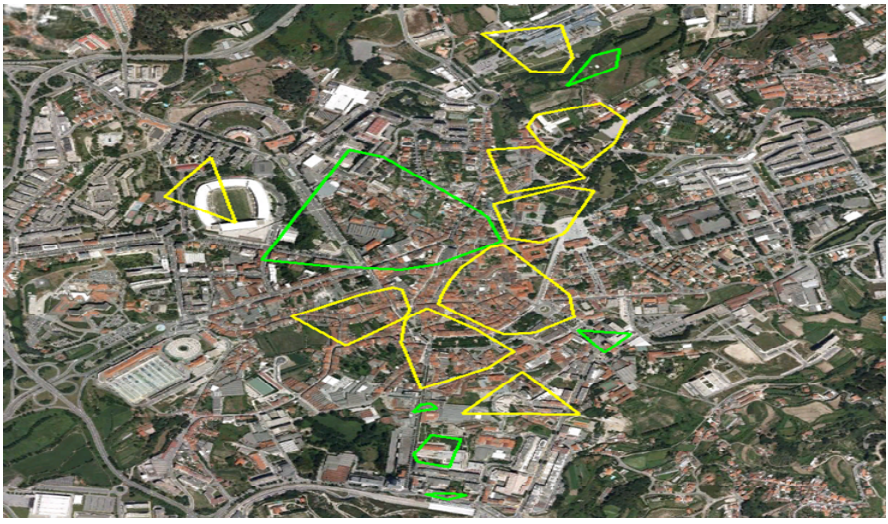
Throughout the entire experimental process, we observed a constant session time of ten days, a cluster threshold of three people and a minimum support K=3 of sequence patterns . We used session time as a heuristic for classifying people into locals and tourists. We classified a person as a tourist if she took photos during a period of no more than ten days. Otherwise, she was considered as a local resident and her sequences were discarded. The following subsections describe the experimental study in detail.

Case 1. Guimarães, Portugal

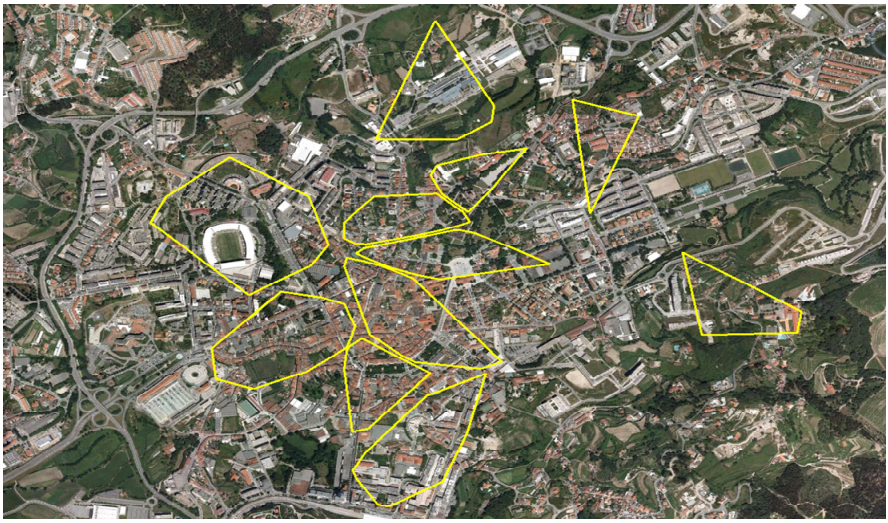
Guimaraes is a relatively small city with historical roots going back to the 9th century. The city was the first capital of Portugal and is often called "the birthplace of the Portuguese nationality". UNESCO declared its historical section as a World Heritage site. In spite of its historical importance, only a very small number of people shared their photos on Flickr compared to the sharing of photos that is generally derived from other cities.

We defined an area of approximately 8.5 square kilometers around the center of Guimaraes with the following boundaries: longitude = 8.318° West and 8.276° East; latitude = 41.435° South and 41.457° North. From 2005 until October 2009, we were able to obtain only 391 photos from 152 people. The Wiki database contains only 11 POIs in the defined area: *Nossa Senhora da Oliveira, Guimaraes Castle, Palace of the Dukes of Braganza, Church of Sao Miguel do Castelo, Guimaraes Historical Center, Sao Paio, Dom-Afonso-Henriques-Stadion, Azurem University, Sao Sebastiao, Pousada de Santa Marinha, Oliveira do Castelo.*

We used 200 and 400 meters as a distance for a photo-to-POI assignment (Sect. 3.2) in order to obtain the sequence patterns. We applied DBSCAN (Ester et al., 1996) on unassigned photos using a distance threshold of 100 meters and identified unknown POIs (Sect. 3.3). These new POIs were added to the existing POIs. A total of 342 photos from 127 individuals were assigned to existing and unknown POIs. Fig. 2 shows regions of existing and unknown POIs using a photo-to-POI threshold of 200 (Fig. 2a) and 400 meters respectively (Fig. 2b).



(a)



(b)

Fig. 2. Guimaraes, Portugal. Cluster boundaries of photos assigned to existing POIs (yellow) using a photo-to-POI distance threshold of (a) 200 meters and (b) 400 meters. Cluster boundaries forming new areas of POIs were obtained using a distance threshold of 100 meters and a density threshold of three people in a cluster (green)

The Teiresias algorithm discovered frequent sequence patterns of length two only. The general statistics pertaining to sequences and patterns are presented in Table 2. It can be seen that only 18 out of 127 sequences for a photo-to-POI threshold of 200 meters and 24 out of 138 sequences for a photo-to-POI threshold of 400 meters were created. There are two reasons for this. Firstly, the majority of people took photos in only one place. Secondly, some of the sequences were discarded because their length exceeded the 10-day threshold. Teiresias discovered 8 patterns using a photo-to-POI threshold of 200 and 7 patterns using 400 meters respectively. Table 3 shows five most frequent sequence patterns for every photo-to-POI threshold, where three generated sequences do not differ in two cases. The sequences that are different for 200 and 400-meter threshold are marked in bold.

Table 2. Guimaraes, Portugal. General statistics

Photo-to-POI threshold	<L,W>	# of people in sequences	# of valid sequences	# of sequence patterns
200	<2,3>	127	18	8
400	<2,3>	138	24	7

Table 3. Guimaraes, Portugal. Sequence patterns using L=2, W=3

Photo-to-POI threshold	# of input sequences	Sequence patterns
200	5	Guimaraes Historical Center → Nossa Senhora da Oliveira
	3	Guimaraes Castle → Church of Sao Miguel do Castelo
	3	**Nossa Senhora da Oliveira → Church of Sao Miguel do Castelo**
	3	Church of Sao Miguel do Castelo → Nossa Senhora da Oliveira
	2	**Church of Sao Miguel do Castelo → Nossa Senhora da Oliveira**
400	4	Guimaraes Historical Center → Nossa Senhora da Oliveira
	4	**Guimaraes Castle → Nossa Senhora da Oliveira**
	3	**Nossa Senhora da Oliveira → * → Nossa Senhora da Oliveira**
	2	Church of Sao Miguel do Castelo → Nossa Senhora da Oliveira
	3	Guimaraes Castle → Church of Sao Miguel do Castelo

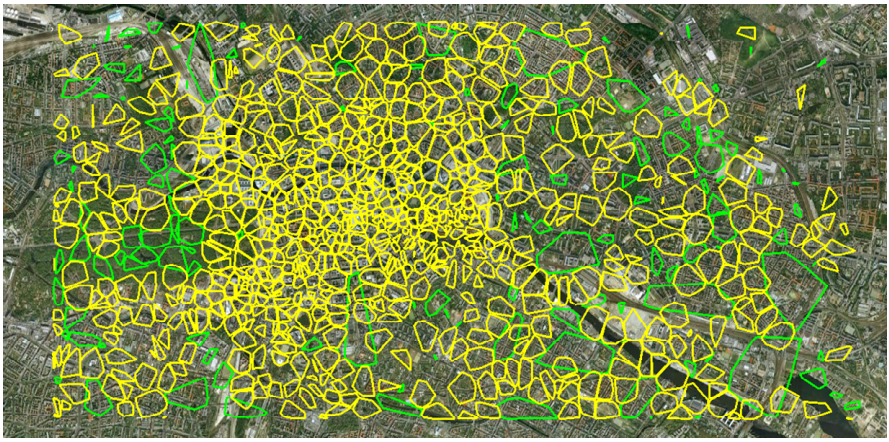
Case 2. Berlin, Germany

Berlin is the capital of Germany and its largest city. It is one of the most popular tourist destinations in the EU. In 2008, a total of 17,758,591 persons visited Berlin according to European Cities Tourism Site[4]. Of this total, 7,033,593 people were classified as foreign visitors.

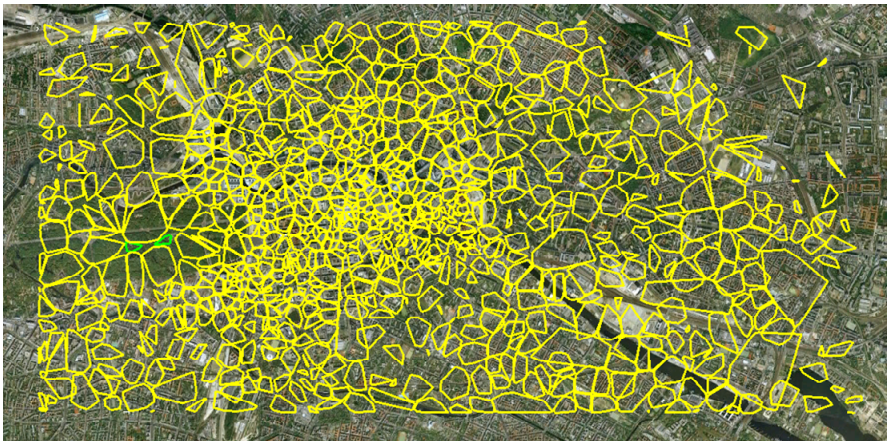
We defined an area of approximately 46.7 square kilometers around the center of Berlin with the following boundaries: longitude = 13.341° West, 13.483° East; latitude = 52.495° South and 52.537° North. We retrieved 71,821 photos from 9,505 people between 2005 and October 2009. The Wiki database contains 857 POIs in the defined area.

We used 200 and 400 meters as a threshold for a photo-to-POI assignment. These new POIs were added to the existing POIs. A total of 68,624 photos from 8,952 users were assigned to existing and unknown POIs. Fig. 3 shows regions of existing and unknown POIs using a photo-to-POI distance threshold of 200 (Fig. 3a) and 400 meters (Fig. 3b) respectively. When the photo-to-POI threshold was 400 meters, almost all the photos were assigned to existing POIs and only two clusters of unknown POIs were created (see Fig. 3b).

[4] http://www.europeancitiestourism.com/



(a)



(b)

Fig. 3. Berlin, Germany. Cluster boundaries of photos assigned to existing POIs (yellow) using a photo-to-POI distance threshold of (a) 200 meters and (b) 400 meters. Cluster boundaries forming new areas of POIs were obtained using a distance threshold of 100 meters and a density threshold of three people in a cluster (green)

While the Teiresias algorithm has the potential for discovering patterns of up to length four if applied on Berlin data, we only present patterns of length 2 and 3 in keeping with the editorial limitations of this paper. The

general statistics pertaining to sequences and patterns are presented in Table 4. Tables 5 and 6 present the five most frequent patterns discovered by the Teiresias algorithm.

Table 4. Berlin, Germany. General statistics

Photo-to-POI threshold	<L,W>	# of people in sequences	# of valid sequences	# of sequence patterns
200	<2,3>	8952	2844	2047
	<3,4>			186
	<4,5>			9
400	<2,3>	8968	2845	2086
	<3,4>			195
	<4,5>			11

From Table 3 we can see that using 2,844 sequences from a total of 8,952 sequences and a photo-to-POI distance threshold of 200 meters, the algorithm discovered 2,047 patterns of length 2; 186 patterns of length 3; and 9 patterns of length 4. Using 2,845 sequences from a total of 8,968 sequences with a photo-to-POI distance threshold of 400 meters, the algorithm discovered 2,086 patterns of length 2; 195 patterns of length 3; and 11 patterns of length 4. The first four sequence patterns of length 2 and 3 are identical for two photo-to-POI distance thresholds (Tables 4-5). The first three sequence patterns of length 2 (Table 4) suggest that people began photographing at Brandenburg Gate and then continued to other places. The third sequence pattern in Table 4 contains a wild character indicating that that people started from Brandenburg Gate, then visited any POI and finished at the Reichstag.

We should also note that unknown POIs created by applying density-based clustering (Sect. 3.3) are not part of the most frequent sequence patterns.

Table 5. Berlin, Germany. Sequence patterns using L=2, W=3

Photo-to-POI threshold	# of input sequences	Sequence patterns
200	74	Brandenburg Gate → Reichstag
	53	Brandenburg Gate → Memorial to the Murdered Jews of Europe
	46	Brandenburg Gate → * → Reichstag
	41	Reichstag → Brandenburg Gate
	36	**Pariser Platz → Brandenburg Gate**
400	71	Brandenburg Gate → Reichstag
	51	Brandenburg Gate → Memorial to the Murdered Jews of Europe
	47	Brandenburg Gate → * → Reichstag
	43	Reichstag → Brandenburg Gate
	34	**Reichstag → * → Reichstag**

Table 6. Berlin, Germany. Sequence patterns using L=3, W=4

Photo-to-POI threshold	# of input sequences	Sequence patterns
200	13	Reichstag → Der Bevölkerung → Reichstag
	10	Brandenburg Gate → Memorial to the Roma and Sinti Holocaust Victims → Reichstag
	8	Pariser Platz → Brandenburg Gate → 18th of March Square
	8	Reichstag → Brandenburg Gate → Memorial to the Murdered Jews of Europe
	7	**Der Bevölkerung → Reichstag → Der Bevölkerung**
400	14	Reichstag → Der Bevölkerung → Reichstag
	10	Brandenburg Gate → Memorial to the Roma and Sinti Holocaust Victims → Reichstag
	9	Pariser Platz → Brandenburg Gate → 18th of March Square
	8	Reichstag → Brandenburg Gate → Memorial to the Murdered Jews of Europe
	7	**Zeughaus →Alte Kommandantur → Lustgarten**

5 Discussion

We demonstrated how an automatic data mining process could be used in finding travel patterns from a collection of geotagged photos. However, geographical data mining is far more complex process than its "classical" counterpart. There are several reasons for this:
(1) Data quality, spatial precision and uncertainty play a crucial role in a spatio-temporal analysis.
(2) Many spatial problems are ill-defined. This makes it impossible to apply fully automatic data-mining process to solving particular problems (Andrienko et al., 2007).
(3) The geographical analysis is very sensitive to the length or area over which an attribute is distributed (Miller and Hand, 2009).

Data quality (spatial and temporal) and precision depends on the way the data is generated and should be taken into consideration during analysis and validation of results. Movement data is usually collected using GPS-enabled devices attached to an object or by geotagging images shared on the Web. For example, when a person enters a building a GPS signal can be lost or the positioning may be inaccurate due to a weak connection to satellites. These concerns are valid for geotagged photo data as well. Specifically, there are two ways to geotag a photo and upload it on the Web. One way involves attaching a GPS to a camera. In this case, the geotagging is performed automatically and the person can face the same problems as with conventional GPS devices described above. Alternative solution would be to manually annotate a photo during upload. In this case, several possibilities exist: the individual photographer may geo-annotate the object being photographed instead of the exact place where it was taken or the exact place could be geotagged with a different level of precision. In addition, the timestamp of a taken photo may not correspond to the correct time at which the photo was taken because of: (1) time zones differences between the user's country of origin and the visiting country (2) careless setting of the camera's clock to some unrealistic time or (3) a software failure reading the timestamp of a photo.

In regard to the second issue raised in this section, there are two basic approaches for discovery of interesting sequence patterns: user-driven and data-driven. The user-driven approach is based on an expert's knowledge. However, it is not always efficient when an expert is required to find interesting sequences from thousands of sequence patterns such as was the case of Berlin. In our examples, we used frequency of patterns as a selection

measure. However, frequent sequences do not necessarily constitute the most interesting patterns. In fact, frequent sequences usually represent the obvious patterns. Therefore, different interestingness measures for ranking patterns (Piatetsky-Shapiro, 1991) can be combined with expert's knowledge to find some new unexpected patterns.

The difficulties associated with spatio-temporal data mining indicate that an analyst should select the parameter values very carefully. Unfortunately, we could not cover all the possible combinations of parameter values in our experiments. However, we demonstrated that changing only the distance threshold of the *photo-to-POI* while keeping all other parameters constant, may produce slightly different pattern sequences. Changing parameters at every step of our approach could lead to completely new sequence patterns. While background knowledge of an analyst or domain expert could help overcome the weakness of the automatic process, some degree of human involvement is necessary for inspecting the data, tuning the parameters, controlling the analysis process and revising the obtained results. For example, an unknown POI can be discovered using the procedure presented in Sect. 3.3. The newly discovered POI may be adjacent to the region of an existing POI. An automatic process treats these two regions as distinct. However, visual inspection might reveal that the unknown POI belongs to the existing POI and that the two regions should be merged into one. Therefore, the solution to this issue would be incorporation of data mining techniques into geovisual analytics systems.

6 Conclusion

In this paper, we presented a novel approach for mining travel sequences using geotagged photo data. We showed that the method is capable of mining semantically annotated sequences of any length with patterns that are not necessarily immediate antecedents. We demonstrated the feasibility of our approach on two different cities using real data. We showed that the approach could be applied to different spatial scales -- to places that have a great number of visitors (Berlin) and POIs, and to sites that have relatively few visitors (Guimaraes) and POIs.

In our future work, we intend to integrate our approach within a visual analytics framework. We shall investigate in detail sequence patterns based on: user profiles (locals/tourists); activity (night/day); and seasonal changes. We shall also concentrate on analyzing sequences with specific parameter settings, and then validating and comparing the resulting patterns to existing solutions. In addition, we shall apply different interesting-

ness measures to help the analyst in discovering interesting sequence patterns.

References

Agrawal, R. and Srikant, R. (1995) Mining sequential patterns. Proceedings of International Conference on Data Engineering (ICDE'95), pp. 3-14, 1995.

Agrawal, R. and Srikant, R. (1994) Fast algorithms for mining association rules. Proceedings of International Conference Very Large Data Bases, pp. 487-499, 1994.

Alvares, L.O., Bogorny, V., Kuijpers, B., de Macedo, J.A.F., Moelans, B. and Vaisman, A. (2007a) A model for enriching trajectories with semantic geographical information. Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems, pp. 22, 2007.

Alvares, L.O., Bogorny, V., de Macedo, J.A.F., Moelans, B. and Spaccapietra, S. (2007b) Dynamic modeling of trajectory patterns using data mining and reverse engineering. Tutorials, posters, panels and industrial contributions at the 26th international conference on Conceptual modeling, volume 83, pp. 149-154, 2007.

Andrienko, G., Andrienko, N., Jankowski, P., Keim D., Kraak M.-J., MacEahren A., Wrobel S. (2007) Geovisual analytics for spatial decision support: Setting the research agenda. International Journal of Geographical Information Science, 21/8, pp. 839–857.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings ACM KDD, pp. 226-231, 1996.

Giannotti, F., Nanni, M. and Pedreschi, D. (2006) Efficient mining of temporally annotated sequences. Proceedings of the 6th SIAM International Conference on Data Mining (SDM'06), pp. 346-357, 2006.

Giannotti, F., Nanni, M., Pinelli, F. and Pedreschi, D. (2007) Trajectory pattern mining. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 339, 2007.

Girardin, F., Vaccari, A., Gerber, A., Biderman A., and Carlo R. (2009) Quantifying urban attractiveness from the distribution and density of digital footprints. International Journal of Spatial Data Infrastructures Research, 4, pp. 175-200, 2009.

Kisilevich, S., Florian, M., Peter, Bak., Tchaikin, A., Keim, D. (2010) Where Would You Go on Your Next Vacation? A Framework for Visual Exploration of Attractive Places, accepted in Geoprocessing 2010.

Miller H. and Han J. (2009) Geographic data mining and knowledge discovery. Chapman & Hall/CRC.

Piatetsky-Shapiro G., (1991) Discovery, analysis and presentation of strong rules. In: G. Piatetsky-Shapiro and W.J. Frawley Editors, Knowledge Discovery in Databases AAAI (1991), p. 229

Rigoutsos, I. and Floratos, A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics, 14/1, pp. 55-67.

Rokach, L. and Maimon, O. (2005), Clustering methods, In O. Maimon & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer, pp. 321--352.

Rokach, L., Romano R. and Maimon O. (2008A), Negation recognition in medical narrative reports, Information Retrieval, 11(6):499-538.

Rokach, L., Romano R. and Maimon O. (2008B), Mining manufacturing databases to discover the effect of operation sequence on the product quality, Journal of Intelligent Manufacturing, 19(3):313-325.

Spaccapietra, S., Parent, C., Damiani, M.L., de Macedo, J.A., Porto, F. and Vangenot, C.(2007) A conceptual view on trajectories. Technical report, Ecole Polytechnique Federal de Lausanne, 2007.

Zheng, Y., Zhang, L., Xie, X. and Ma, W.Y. (2009) Mining interesting locations and travel sequences from GPS trajectories. Proceedings of the 18th international conference on World Wide Web, pp. 791-800, 2009