Travel History: Reconstructing Travelers Semantic Trajectories based on Heterogeneous Social Footprints

Amon Veiga Santana[1,2], Jorge Campos[1,3]

[1] Salvador University – UNIFACS
[2] Heidelberg University
[3] Bahia State University - UNEB
Salvador, Bahia, Brazil
amoncaldas@yahoo.com.br jorge@unifacs.br,

Abstract. Travel specialized websites have increased their sociability and usage by adopting mechanisms that facilitates content sharing in real time between users. These web applications, however, lack tools that allow travelers to share their experiences, such as places they have visited, itineraries they have performed, and other activities of a typical touristic trip. These kinds of information, when available, are insufficient and incomplete. The process of generating structured and semantic rich datasets based on recommended trips, routes and destinations usually requires high effort to be generated. This task is frequently manual, cumbersome, inaccurate, time-consuming, and depends on user's willingness to cooperate. This work proposes a solution for reconstructing travel histories using heterogeneous social sources, such as posts in social networks, GPS positioning data, location history data generated by cloud services or any digital footprint with an associated geographic position. The solution encompasses a conceptual model; a methodology to reconstruct travel histories based on heterogeneous social tracks sources; and an application to present the reconstructed travel itinerary in a graphical and interactive fashion. An experiment conducted with real travelers showed that the proposed solution is a reasonable way to reconstruct semantic-rich travel histories in an automatic fashion.

Categories and Subject Descriptors: H.2 [**Database Management**]: Database Applications; H.3 [**Information Storage and Retrieval**]: Online Information Services; H.4 [**INFORMATION SYSTEMS APPLICATIONS**]: Communications Applications; H.5 [**INFORMATION INTERFACES AND PRESENTATION**]: Multimedia Information Systems

Keywords: Crowdsourcing, Moving Object, Semantic Enrichment, Social Networks, Trajectories, Travel History

1. INTRODUCTION

The popularization of Online Social Network (OSN) and User Generated Content (UGC) have modified the way people search, find, read, access, and share information on the Internet [Ye et al. 2011]. OSNs have an important role in the production and search for information. OSN users', for instance, are frequently involved in activities to find relevant contents, advices, opinions, or to simply interact with their mates to have fun [Lange-faria and Elliot 2012]. UGCs (e.g., posts in social networks and comments in websites and forums) have become an important and recognized source of information in the tourism domain [Akehurst 2009]. Travel specialized websites, for instance, have increased its sociability and usage by adopting mechanisms that facilitates content sharing in real time between users. A 2011 PhocusWright report[1] shows that nine of ten cyber travelers read and trust online reviews in touristic related sites. Unfortunately, there are far more people willing to consume this

[1] http://www.researchandmarkets.com/reports/1866967/phocuswrights_social_media_in_travel_2011

This work was partially supported by CAPES.

kind of content, than people disposed to generate them [Tobergte and Curtis 2013]. It is because most people see UGC as a time consuming and boring task, but they will not mind to contribute if there exists some kind of application or service that captures their contribution in an automatic fashion.

A special kind of information incorporated by most OSNs that has attracted the attention of the travel and tourism community is the users' position while they are moving. The increasing number of location-enabled devices opens the possibility of making the position of the user a mandatory piece of information to virtually any kind of social interaction or user generated content. Moreover, the capability of keeping track of the position of a user at high detailed levels opens the possibility to combine traveler's trajectory data and georeferenced social interactions to produce, in an automatic fashion, a structured and semantic rich dataset of traveler's preferences and behaviors.

The growing habit of travelers using social networks as a mechanism to publish georeferenced events and information about their travels combined with the large number of devices able to capture the user position at different levels of granularity opens the possibility to rebuild the complete traveler history, including paths performed, places visited, means of transportation used and even personal impressions and opinions regarding the points of interest and the way to go between two places.

This article introduces Travel History, a conceptual model and a methodology to reconstruct the trajectory of travelers based on records of their position and their interactions posted on social networks. Position information may vary from the usual detailed GPS logs to any evidence of places visited by the traveler and recovered from the traveler social network repository. Thus, Travel History model supports the representation of trajectories with different levels of granularities mixed and interleaved with travelers' social interactions. The proposal of a generic conceptual model for describing travels based on heterogeneous sources of information, together with the presentation of a data model, a methodology, and algorithms that use multivariate digital footprints in the reconstruction of semantic-rich traveler's trajectories helps to fill a gap in trajectory analysis of the tourism domain.

The remainder of this article is structured as follows: Section 2 discusses related work. Section 3 presents the Travel History Model. Section 4 presents the details of the heterogeneous source of social footprints. Section 5 discusses the techniques used to instantiate entities of the model. Section 6 introduces a prototype tool and presents some results of an experimental evaluation with real travelers' volunteers. Section 7 presents conclusions and indicates future work.

2. RELATED WORK

The analysis of trajectories of moving objects has been intensively discussed by the GIS community over the last decade. Fed initially by the profusion of data captured from sensors and location devices, studies in this field have evolved from the generation of trajectories using GPS raw trajectory data to the use of novel means to enrich trajectories semantically. One salient new source of information comes from the growing habit among users to interact in social networks, posting, commenting, or sharing contents that contain geographic references. This source of information has proven its value for many different fields and purposes. It is of special interest of this work the combination of trajectory semantic enrichment techniques and georeferenced post in social networks to produce semantic rich set of information about travelers and their visits.

Semantic enrichment and annotation in trajectory data are very active research topics. Spaccapietra et al. [2008] proposed the first model that treats trajectories of moving objects as a spatiotemporal concept. The *Stop*s and Moves model is one of the most accepted model to represent trajectory of moving objects. A *Stop* is part of the trajectory in which the traveling object did not move and a Move corresponds to the dynamic part of the trajectory, i.e., it is a segment of the trajectory in which a spatio-temporal evolution of the traveling object is observed. Spaccapietra's work has inspired different initiatives aiming to understand and represent the semantic of *Stop*s and Moves in many different fields. Alvares et al. [2007] was the first initiative to instantiate the *Stop*s and Moves model. This work presents a semantic model of annotation of trajectories collected through GPS. The model is extensible and is intended to be adaptable to different application domains. The basic functionality of this model is to identify the elementary elements of a trajectory, that is, *Stops* and *Moves*. The strategy is to add a preprocessing phase in which occurs the semantic enrichment of the trajectories parts with geographic information to facilitate the queries, analysis, and data mining of the mobile objects. The preprocessing phase of the work is based on an algorithm called SMoT (*Stop*s and Moves of Trajectories). This algorithm is based on the semantic assignment considering the geographic information along trajectory. In Tietbohl et al. [2008] was presented a model to discover relevant places on trajectories. This work, however, focuses on the stationary part of the trajectory (i.e., *Stop*s) and its major contribution is an algorithm called CB-SMoT (clustering-based stop and moves of trajectory). CB-SMoT is a spatio-temporal clustering method used to identify *Stops* based on the speed of the vehicle. Palma's work shows the efficacy of the CB-SMoT method in the identification of *Stops* in an urban transportation context.

[Andrienko et al. 2007] used the time for adding semantic annotation to the stationary part of a trajectory and argued that the more time is spent in a place more important it is to a person. In the Zheng et al. [2009] work, it was exposed a technique that considered, beyond the spending time, the geographic coincidence with Points of Interest (POI) defined in the application. Zheng et al. [2010] proposed a technique based on speed, acceleration and the orientation of the user to detect the transportation mode used to move from one place to another. Even though there are other approaches in the detection of means of transportation based on GPS records, a deeper discussion about this topic is outside the scope of this article.

Most of the mentioned works are based on the development of algorithms that allow the identification of parts of the trajectories (i.e., *Stop* and Move) and powered with some mechanism to annotate these parts with semantics of the application domain. Other approaches have evolved to the design of generic frameworks capable of performing the semantic enrichment of different types of trajectories. Yan et al. [2013] presented SeMiTri, a framework that deals with semantic annotation of trajectory based on background geographic information. This framework is generic enough to deal with heterogeneous trajectories and cover a wide range of applications. Yan's work specialized the concept of *Stop*s and Moves and introduces the concepts of semantic episodes, trips and regions of interest. [Bogorny et al. 2013] presented a conceptual and data model for the representation of the semantic trajectories of the mobile objects called CONSTAnT (CONceptual model of Semantic TrAjecTories). The CONSTAnT model considers that a trajectory may have different behaviors during its course. The model uses the Behavior entity to identify the behavior for each sub trajectory. In the CONSTAnT model, the behavior of the mobile object can be specialized in simple or collective behaviors. The simple behavior analyzes the displacement of an object without considering the displacement of other objects close to it. Collective behaviors aim to identify patterns of displacement of a group of objects, such as avoidance, chasing, and leadership. A comprehensive set of solutions for semantic trajectories modeling and analysis can be found in [Parent et al. 2013].

Researches in the trajectory domain provide a solid base for the development of effective solutions to extract information from raw trajectory data. In another front, several initiatives focus in pattern and knowledge discovery from User Generated Georeferenced Content (UGGC). In our context, UGGC is defined as a UGC that carries some kind of information that allows the identification of the geographic location of the related content, not necessarily the location of the user. A georeferenced picture of Copacabana beach posted in Instagram and a Web review made by someone in New York about the Copacabana Palace Hotel, for instance, are both examples of UGGC of the same geographic region. UGCC do not have the same spatial granularity of positioning devices, such as GPS, but allow a more refined semantic extraction about the content being described.

Related with initiatives that deal with UGCC for semantic enrichment, Ji et al. [2009; Hao et al. [2010] proposed a solution for mining city attractions from touristic blogs posts, Rattenbury et al.

[2007] proposed an approach to extract semantic from georeferenced picture posted in the Flickr social network, and Gao et al. [2010] proposed a method to identify touristic attractions from Flickr's georeferenced pictures and to enrich the description of such attractions with information extracted from collaborative websites like Yahoo Travel Guide[2] and WikiTravel[3]. Lu et al. [2010] proposed a picture-based customized trip planning. This system allows trip planners to specify personal preferences and generates travel routes from geo-tagged photos. The proposed solution is limited to surrounding attractions in a given city or region and does not support travel plans lasting more than one day and involving multiple destinations. Yoon et al. [2012] proposed a framework for itinerary social recommendation using trajectories generated by local residents and expert travelers.

Despite the enormous potential of aforementioned initiatives, few works have combined the use of trajectory data and UGCC in the process of trajectories reconstruction and semantic enrichment. Fileto et al. [2013] proposed a method for trajectory annotation based on the spatiotemporal compatibility of Twitter posts. Although the spatial component of the post is mentioned in the work of Gil et al. [2014], the methodology introduced in the article takes into account only the temporal compatibility between trajectory data and Twitter posts, that is, they do not use the posts' content or location to enrich the trajectory semantically.

Analyzing early related work, it is noticeable that most solutions used detailed logs of position devices to analyze people's movements. This tendency is switching to incorporate location information embedded in social interactions and stored in the cloud. In a social post, for instance, the position comes with some kind of information or even a personal opinion about the place visited. Thus, trajectory reconstruction using georeferenced social interactions can be the strategy to recover context elements and semantics of trajectory. As the process of trip planning involves information search and retrieval, it is natural that travelers also look for this kind of information among their friends and people from their social circle. At the best of our knowledge, however, there is no service that, considering previous users' experience registered as social tracks, offers efficient means for travelers to access information about structured travel itineraries, including attractions and transportation means. Next section introduces an attempt to treat and represent this kind of information.

3. TRAVEL HISTORY

Travel History conceptual model encompasses all information needed to represent relevant actions and movements of a traveler. The central entity of the model is *Travel History*, which is an entity that aggregates *Stays* and *Trails* traveled by an individual during a given time interval. Figure 1 shows the relationship among models entities using UML notation.

A *Trail* is an entity that captures the traveler movement. Each *Trail* has an associated path and a transportation mode. The path is a collection of geographic points that represents the geometry of the movement. The path may vary from a pair of points indicating only the endpoints of the movement up to a collection of points representing the detailed path fulfilled by the traveler. The transportation mode indicates how the *Traveler* goes from one place to another (walking, by train, etc.).

Stays represent the places where the *Traveler* remained for a while or changed the transportation mode. Each *Stay* occurs at a *Place*. A *Place* is considered as a geographic location together with semantic information (e.g., a description, political categorization, or a combination of them). Consider, for instance, the place where the traveler stops walking and takes a cab. This place is considered as a *Stay*, since a change in the transportation mode was detected. If the traveler remains around a place for a while, this place also becomes a *Stay* in our model.

[2] https://www.yahoo.com/travel/guides

[3] http://wikitravel.org

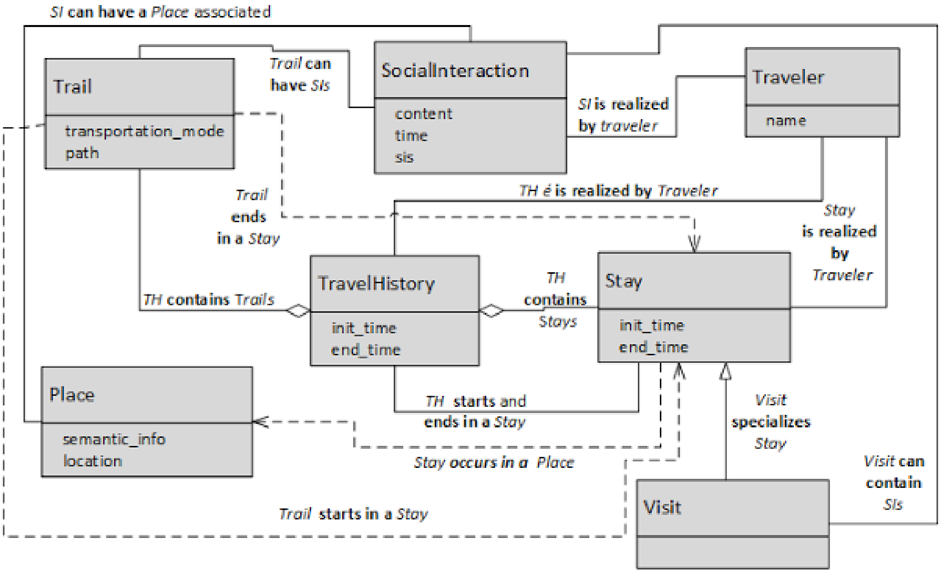


Fig. 1. Travel History Conceptual Model.

The terms *Stops* and *Moves* from the seminal work of Spaccapietra et al. [2008] are generic and aim at representing parts of a trajectory at high level of abstraction. Many work have specialized these concepts creating entities more suitable for the application domain. In this work, a *Trail* is a specialization of a Move and a *Stay* is a specialization of a *Stop*. A *Stay* is further specialized as a Visit in our model. We believe that *Trail*, *Stay* and *Visits* are more appropriate entities to represent the semantic rich parts of a traveler trajectory.

A *Stay* may or may not have a special meaning for the trip. When the Traveler's permanency at some place is relevant for the trip, the *Stay* is specialized and becomes a *Visit*. The relevance of the *Stay* takes into account the amount of time spent at the place or the amount of social interactions related to the place. Thus, a Visit represents a place where the *Traveler* has made and registered some social interaction or stayed for certain amount of time above a given threshold.

Visits and *Trails* may have one or more *Social Interactions*. These interactions are contents that help to understand Traveler's intentions or activities. The association of *Visits* and *Trails* with *Social Interactions* considers both temporal and geographical matches. Temporal matches consider the time of the realization of the *Social Interaction*, (i.e. the interaction is associated with a *Stay* or a *Trail* that is going on at the time of the occurrence of the post). Geographical matches occur when a *Social Interaction* has some geographic information associated to its content (i.e., it is an UGGC). In this case, the location of the interaction is determinant to establish the association between the *Social Interaction* and the respective *Stay* or *Trail*. The amount of Social Interactions related to a *Stay* or *Trail* is also an indicator of the relevance for the trip.

Travel History model was conceived aiming at handling multiple types of UGGCs retrieved from different OSNs and combined with any sort of positioning data about the user movement. Next section discusses the mechanism of converting these kinds of heterogeneous data into entities of the model.

4. HETEROGENEOUS SOCIAL FOOTPRINTS

The Travel History reconstruction process is based on heterogeneous sources of data. Sometimes it is available a very fine set of registers of a traveler's movement captured by some kind of position device. Other times there is a less fine position records, but the data comes with some semantics attached, and

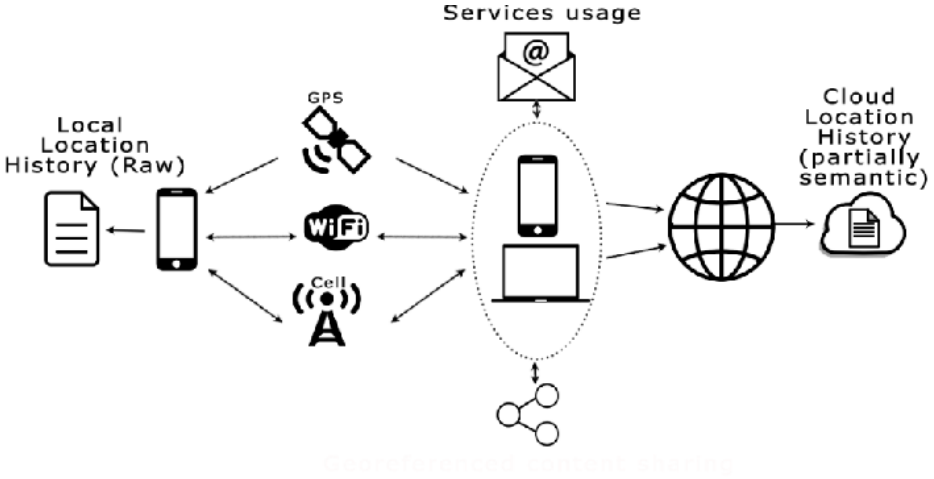

Services usage
Local Location History (Raw)
GPS
WiFi
Cell
Cloud Location History (partially semantic)
Georeferenced content sharing

Fig. 2. How the data of RTD and STD are generated/captured.

not rarely, there is a social interaction that can be used as a source of information about the travel. Even coming from different sources, these datasets share common concepts and structures. Thus, they were grouped in three main categories: 1) Raw Trajectory Data (RTD), 2) Semantic Trajectory Data (STD), and 3) Social Interactions. When a Social Interaction has an associated geographic location, it is called Georeferenced Social Interaction (GSI).

RTD and STD are sequences of spatiotemporal records. Although they share the same basic structure, they differ significantly considering both the spatial-temporal granularity and content. RTD are generated by positioning devices and contains only registers with the position and the timestamp of each reading. RTD is usually obtained from a single device and store the data as they are produced. STD, on the other hand, is a result of preprocessed data acquired from many different sources that goes through a process of semantic enrichment (Figure 2). STD come from some cloud service such as "Google Takeout", which allows users to recover information about their movement. Users have to authorize Google to keep track of their whereabouts and Google uses this information in lots of different services. Although RTD records are denser than STD records, the semantic of the later is much more relevant to the trajectory reconstruction than the former.

The last category of source of information used in the travel reconstruction is GSI. GSI records are even sparser than RTD and STD records. Thus, GSI alone contributes little to reconstruct the detailed trajectory geometry, but they are very important to enrich the trajectory semantically.

Differently from the location history files (RTD and STD), the data coming from social interactions are not spatially or temporally structured; are spread in different social networks; and come in a myriad of formats and data structures. Moreover, to use information from social interactions it is necessary to use a different API for each social network and to deal with different privacy policies regarding the access and usage of this kind of information. Finally, it is necessary to analyze semantic and geographic information social interaction can provide. For these reasons, we have decided to use only the three main social networks used in travel context (i.e., Facebook, Instagram and Twitter). For these social networks, we have mapped all relevant types of social interactions and their associated content. On **Facebook**, for instance, we identified four types of relevant *Social Interaction*:

- *Simple Posts* - Social Interactions with text content that can contain references (tags) to other users and to a geographic place (*check-in*).
- Photos - Specialization of a *Simple Post* that contains one or more pictures associated.
- Videos - similar to the Photos, but containing videos.

—Albums - collection of Photos and Videos (the album can contain a geographic location and the date in which the medias were captured) grouped by any criteria defined by the *Traveler.* In travel context, it often represents the visit to a place (attraction, city, state, country).

There are also variations for each type of *Social Interaction* mentioned above. It is possible to retrieve social interactions that were published by the users themselves and by the users' friends (using the tagging functionality). It is possible also to retrieve social interactions with geographic data attached. For each of these variations, it is necessary to make different requests to the APIs.

On **Instagram**, we identified two types of relevant *Social Interaction*:

—Photos a publication with a photo content and that can also contains a geographic location associated and a description of the place or content.

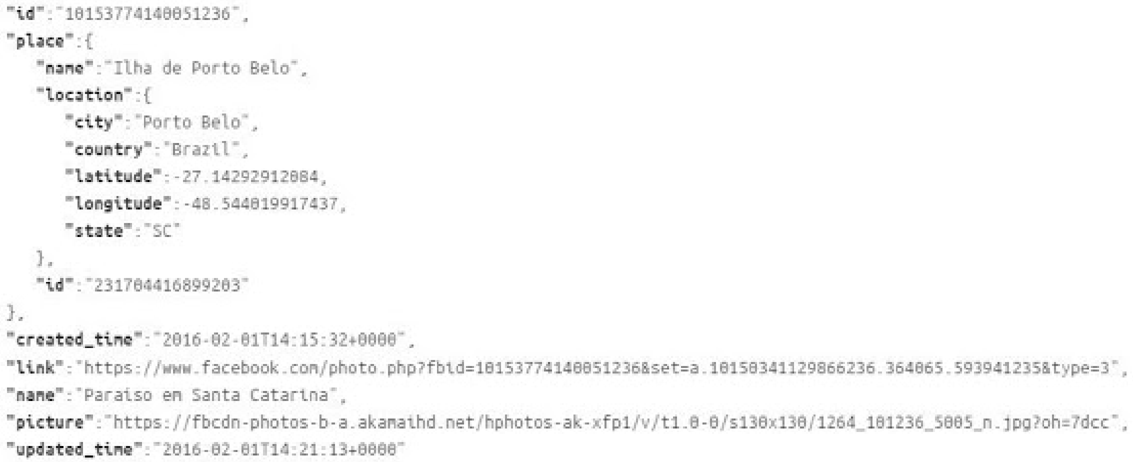
—Videos - similar to the Photos, but containing videos limited up 60 seconds of duration.

In the case of Instagram, even having distinct types of content, a single request can retrieve both types of content.

On **Twitter**, a *microblog* with supports to graphical medias, there is only one type of content, called tweet. Each tweet can contain medias associated (like photos and videos), mention to other users and also geographic tags referring to a place (geotweet).

Facebook, Instagram, and Twitter APIs return the requested data using JavaScript Object Notation (JSON) format. Figure 3 shows an excerpt from a Facebook response. It is an example of a geographic social interaction. Among the attributes, there are the name of the place (Island of Porto Belo), the country and the geographical coordinates.

Even using a small, but significant, number of social network, the process of gathering social interactions is not trivial. There are several optional attributes for each API request, there is legacy code from previous versions of the API, changes in the company data policy, or even adaptation to new legislations. Thus, it is always a challenge to build applications that are both fault tolerant and resilient to changes in this ever-evolving scenario.



```
"id":"10153774140051236",
"place":{
   "name":"Ilha de Porto Belo",
   "location":{
      "city":"Porto Belo",
      "country":"Brazil",
      "latitude":-27.14292912084,
      "longitude":-48.544019917437,
      "state":"SC"
   },
   "id":"231704416899203"
},
"created_time":"2016-02-01T14:15:32+0000",
"link":"https://www.facebook.com/photo.php?fbid=10153774140051236&set=a.10150341129866236.364065.593941235&type=3",
"name":"Paraiso em Santa Catarina",
"picture":"https://fbcdn-photos-b-a.akamaihd.net/hphotos-ak-xfp1/v/t1.0-0/s130x130/1264_101236_5005_n.jpg?oh=7dcc",
"updated_time":"2016-02-01T14:21:13+0000"
```

Fig. 3. Segment of a JSON response using the Facebook API.

5. REBUILDING TRAVEL HISTORIES

The process of rebuilding *Travel Histories* goes from gathering all pieces of information to the instantiation and semantic enrichment of models' entities. This process can be split in three phases: data acquisition; data processing; and entities generation (Figure 4). In the first phase, data are acquired from different sources, like social networks, location history web services, and location's tracks recorded in the user device. In the second phase, the data are processed to identify *Stay* candidates and transportation modes. In the last phase, *Stays*, *Visits* and *Trails* are generated and semantically enriched.

The data acquisition phase starts with the definition of the time window when the travel happened. After defining the temporal window of the trip, the reconstruction process continues by gathering all relevant information about user movement and his/her *Social Interactions*. The data acquisition strategy depends on the category of the sources available. RTD and STD are collected as a single file. RTD records come from mobile applications that continuously record the position of the travelers over time. These records are stored in the device internal memory and can be imported at any time. STD records come from cloud services (like Google Takeout). These records are requested by the owner of the data, the only person able to retrieve them. RTD and STD files are traversed and relevant information is extracted and stored in a local database. The process of gathering social interactions, on the other hand, can be fully automated. OSN's users authorize a computer application to search and retrieve all social interactions of a given period. These data are also stored in a local database.

In order to illustrate the reconstruction process, consider a hypothetical travel with samples of information gathered from the traveler's digital footprints. Figure 5 shows a travel timeline with samples of information where, in the first segment there is only intermittent RTD records. In the second segment, RTD and GSI (content posted in social networks like Facebook, Instagram, and Twitter) are available. In the last segment, a combination of RTD, STD and GSI is available and during some periods they overlap.

With all pieces of information in place, the second phase of the reconstruction process starts by identifying *Stays* candidates and transportation modes between these candidates. *Stays* candidates represent the locations where the traveler hangs around for a while or change the transportation mode.

Stays candidates are generated considering the category of the data source to be processed. To

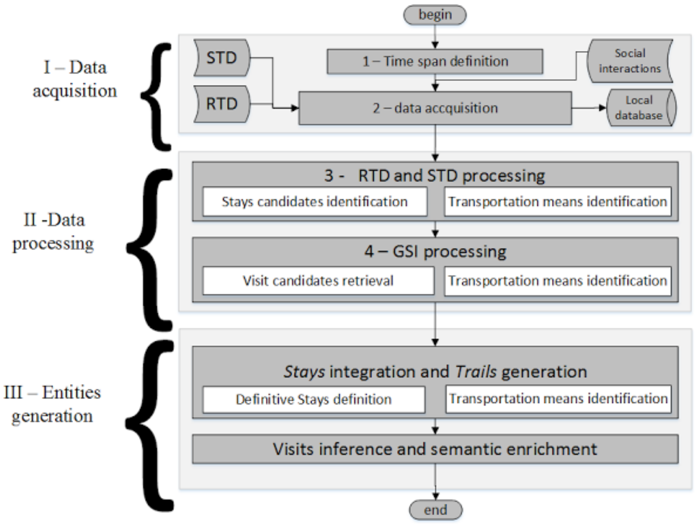

begin
I – Data acquisition
STD
RTD
1 – Time span definition
Social interactions
2 – data accquisition
Local database
II -Data processing
3 - RTD and STD processing
Stays candidates identification
Transportation means identification
4 – GSI processing
Visit candidates retrieval
Transportation means identification
III – Entities generation
Stays integration and Trails generation
Definitive Stays definition
Transportation means identification
Visits inference and semantic enrichment
end

Fig. 4. The three phases of the Travel History generation process.

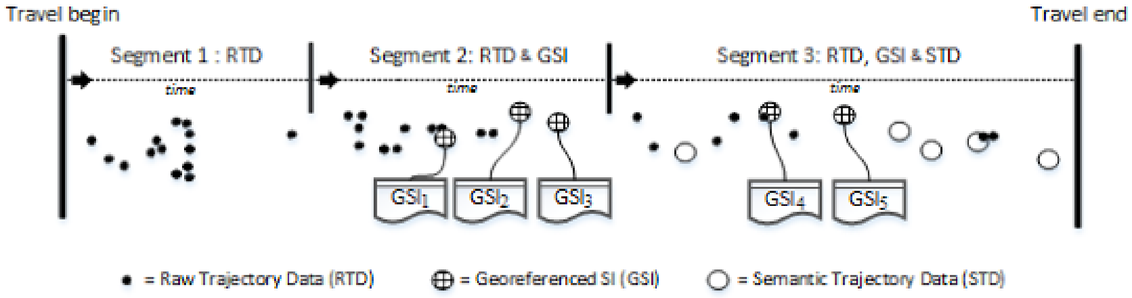

Travel begin
Travel end
Segment 1 : RTD
Segment 2: RTD & GSI
Segment 3: RTD, GSI & STD
time
time
time
GSI1
GSI2
GSI3
GSI4
GSI5
• = Raw Trajectory Data (RTD)
⊕ = Georeferenced SI (GSI)
○ = Semantic Trajectory Data (STD)

Fig. 5. Graphical presentation of RTD, STD and GSI records.

detect *Stays* from RTD and STD sources, it was developed an algorithm capable of recognizing these entities based on the geometric configurations of the track (clusters of points or isolated points) and based on some semantic information already present in the data. This algorithm is a combination and improvement of the techniques defined by Spaccapietra et al. [2008; Zheng et al. [2010]. While processing RTD files, the transportation mode used between two *Stays* candidates is also computed. The definition of the transportation mode takes into account the following aspects: speed, speed variation, acceleration, orientation variation and continuity. Each transportation mode has a single combination of these factors. By taking them together, it is possible to infer how the *Traveler* moved between *Stays*.

Considering RTD and STD sources, there are four methods to identify *Stays* among the data. The first method considers that when a traveler stays in a place, a cluster of points (i.e., a dense formation of points) is formed. The algorithm identify this formation of points and group these points to form a *Stay* candidate (Figure 6 - case A). On the other extreme, isolated points also becomes a *Stay* candidate (Figure 6 - case B). This case occurs when there is a record distant from both the previous and the next point in the sequence and if it is not considered an outlier. *Stays* are also defined at every location where a transportation mode change occurs (Figure 6 - case C). Finally, a *Stay* can be inferred from the semantic already embedded in STD files. These data sometimes have semantic information like "still" or "tilting" associated with a place. (Figure 6 - case D). These places always become a *Stay* candidate.

The four different ways of detecting *Stays* Candidates in the data can detailed as following:

—**By density**: by measuring the distance and the speed of travel between the points of a trajectory, we identify if there is a *Stay*. A representation of this case is illustrated in Figure 6, column A. It is shown in that column that there are segments in which the relative distance between the points is lower than the general average. In such cases, if the speed between these points is below a given value, we consider that these points (two or more) are a *Stay* candidate.

—**By isolation**: when a point has as distance from the nearest points greater than the average of the general distance by a percentage higher than a defined threshold, this point is considered in an isolation situation. This situation may that the positioning devices capability was restricted at that moment (user option, signal failure, low battery, etc.). No matter the reason, this point is considered a *Stay* candidate, as it may be crucial to the trajectory spatial description. This case is illustrated in Figure 6, in column B.

—**By transportation change**: Once the means of transportation is identified, when the *Traveler* changes the transportation mode (for example, after walking and taking a bus), the transition location is identified as a *Stay* candidate. Whether this *Stay* is relevant or not will be decided at a later stage. This type of event is shown in Figure 6 (column C).

—**By semantic information**: Semantic information may be available for Tracks from Semantic Trajectory Data (STD) files and are analyzed to infer if they describe a *Stay*. In the case of files

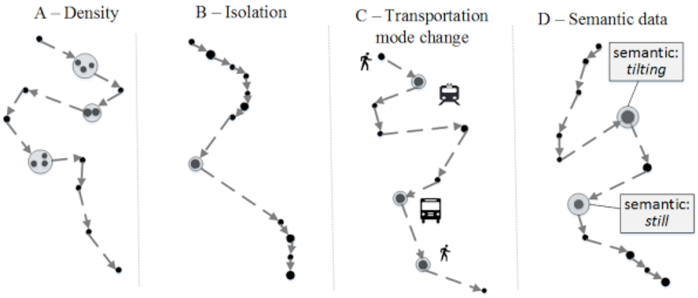


Fig. 6. *Stays* identification techniques based in RTD and STD and the result when it is integrated with Social Interactions.

coming from the Google Takeout service, for instance, there are definitions of activities associated with points that may be *Stay*, such as the terms Still and Tilting attached to a point. This type of event is shown in Figure 6 (column D).

Outliers are treated during the pre-processing phase of the reconstruction process. Most outliers are disregarded based on the physical unviability for a *Traveler* being at a certain place, considering, for example, the maximum speed of known transportation means. Other aspects considered during the outliers detection process is the transportation means continuity. It is not usual that a *Traveler* changes from transportation alternately several times. In these case, the segment that does not fit in in the average pattern is replaced by the most recurrent one. At the end, all the outliers are disregarded from the dataset and are not used in the reconstruction process.

Stay candidates are also generated considering GSI information. In this case, the rule is simple, that is, every GSI generate a *Stay* candidate. Later, some of these *Stays* will be grouped, becoming a single *Visit*, others will not be confirmed as a *Stay* and will become a social interaction of a *Trail*. All *Stays* candidates, no matter the source of information, are stored in a common persistence entity. Figure 7 illustrates the result of the *Stays* candidate generation from the hypothetical data records showed in Figure 7. Each circle in segments 1, 2 and 3 represents a *Stay* candidate.

During the last phase of the reconstruction process, all high-level entities of the model representing parts of the travel history are generated, integrated and semantically enriched. Throughout the integration step, issues related to the duplicity and overlaps are solved. At this point, each *Stay* candidate is processed, confirmed as a definitive *Stay*, promoted to a *Visit*, or merged with others *Stays*. Since *Stays* candidates are generated from different sources separately, it is possible that some *Stays* candidates refer the same event of the trip. The *Stays* merge process occurs when the distance between two *Stays* is less than a given threshold.

The next step in the reconstruction process is the *Trails* generation. *Trails* connect two *Stays* and

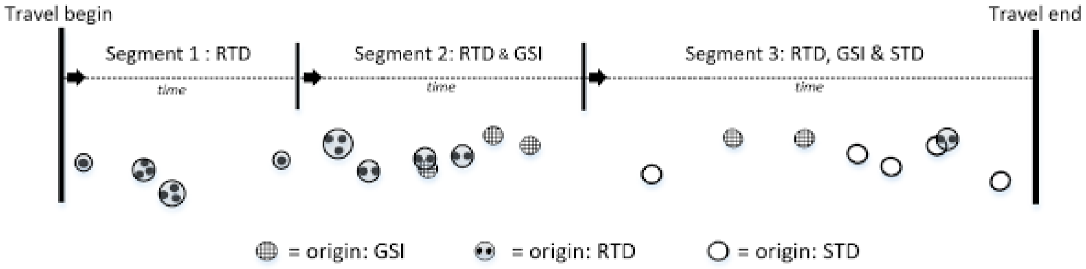


Fig. 7. *Stays* extracted from RTD, STD and GSI records.

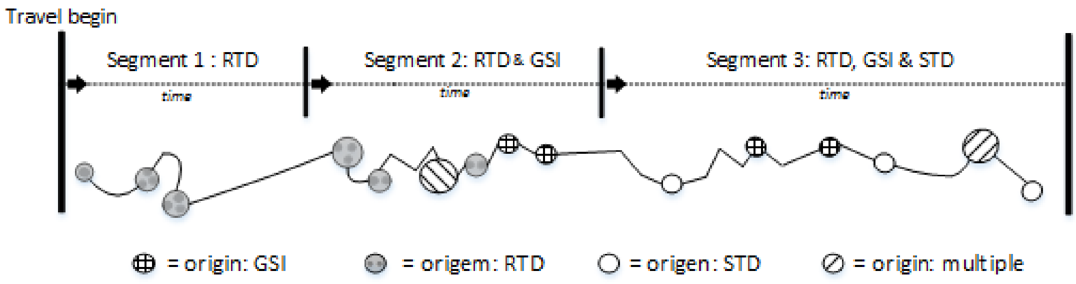


Fig. 8. Result after *Stays* integration and *Trails* generation.

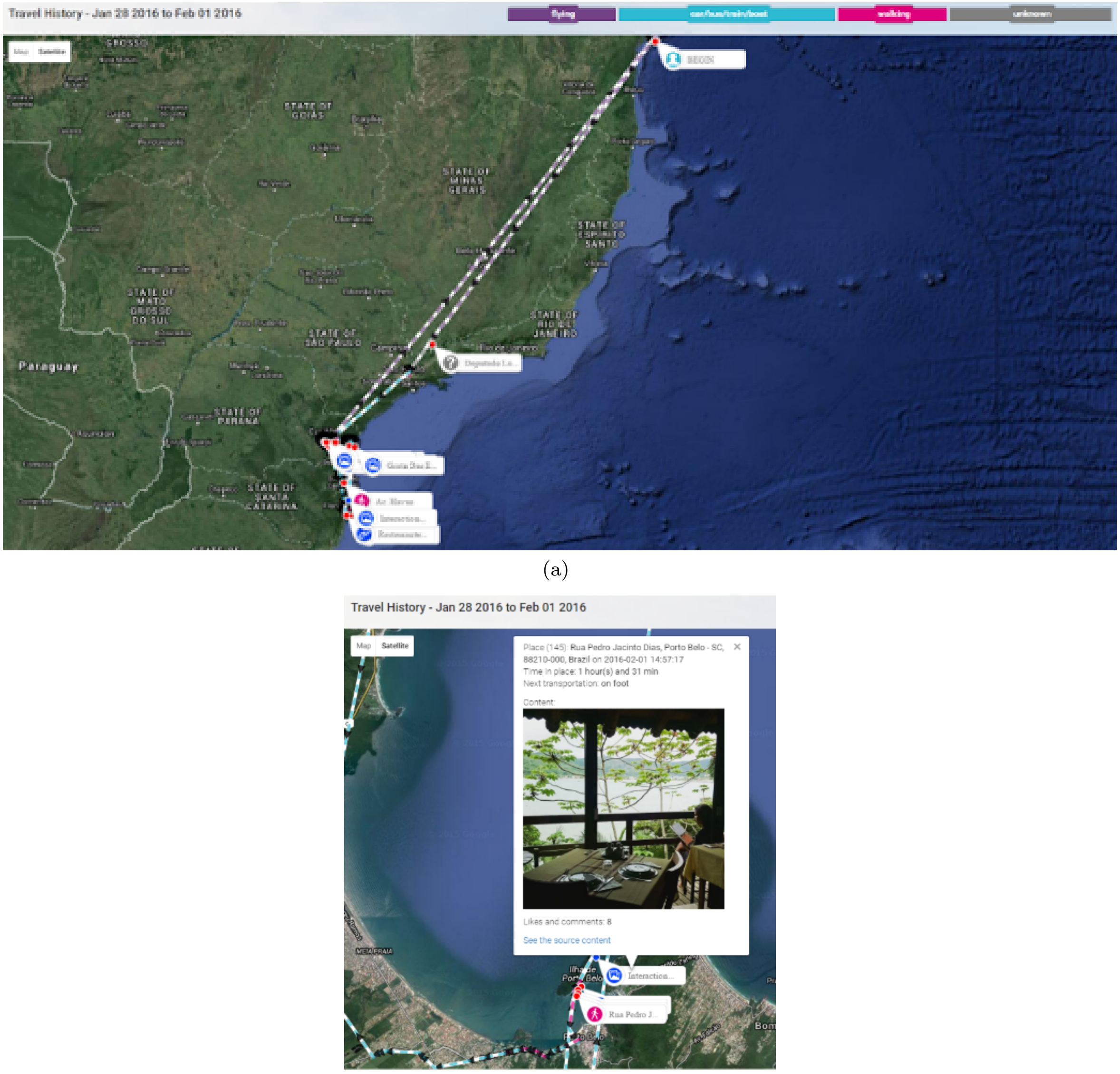
describe the *Traveler* movement between them. During *Trails* generation, it is necessary to identify the transportation mode. If the *Stays* were generated from the same source, the transportation mode between them has been already defined in the second phase, but if the *Stays* were generated based on different sources, the same algorithm to detect transportation mode discussed earlier is used. The last step of the phase of the reconstructing process is the semantic enrichment of *Trails* and *Visits*. *Visits* are specialized versions of *Stays*. For a *Stay* to become a *Visit*, it is considered the amount of time spent on the site and the number of *Social Interactions* carried out by the traveler. Once all model entities are instantiated, integrated and compatibilized (Figure 8), an application using geovisualization techniques can easily depicts the graphical realization of the reconstruct travel history.

To evaluate the Travel History model, it was developed a prototype tool that employs all techniques presented in this article (available at http://th.fazendoasmalas.com). The prototype allows the acquisition, processing and generation of Travel Histories. At the end, the tool shows the user travel history in an interactive map. *Stays*, *Visits*, and *Trails* are presented in a graphical and user-friendly web application. Figure 9(a) shows the reconstruction of a Travel History generated using different and rich sets of information. The travel history depicted occurred between January 28 and February 1, 2016. It was a five days' trip in the south part of Brazil, including a visit to the capital of Paraná, a train trip between the capital and some places along the state coastline, a visit to Mel Island in the Paranaguá Bay, and a visit to the state of Santa Catarina, including the capital Florianópolis and other small towns nearby. To reconstruct this travel history, it was used as source of information GPS log files, location history files generated by Google and the online social networks Facebook, Instagram and Twitter. When RTD, STD, and GSI are all available to reconstruct the travel, it is possible to zoom in the map presentation to analyze the detailed path of the traveler and to visualize comments and social interactions posted along the path. Figure 9(b) shows a detailed view of part of the trip to Paraná and Santa Catarina and a box with the one of the GSI posted about the trip.

We used the application to support the realization of an experimental evaluation of the proposed model and methodology. Next section discusses the result of such an experiment.

6. EXPERIMENT AND RESULTS

Aiming at validating the data model and the proposed methodology to reconstruct travel histories based on social footprints we have developed an experiment with a small but very specialized group of volunteers. Volunteers from RBBV (acronym, in Portuguese, for Brazilian Travel Bloggers Network) have been invited to use the tool, submit their data, reconstruct their travels and evaluate the travel history generated by the application. A total of 58 volunteers started the experiment, but only 23 travelers completed the entire process successfully. Some volunteers did not submitted data, others submitted inconsistent data, and some did not perform the evaluation. The volunteers were oriented to answer a questionnaire after analyzing and exploring their reconstructed travel. The questions



(a)

Travel History - Jan 28 2016 to Feb 01 2016

Place (145): Rua Pedro Jacinto Dias, Porto Belo - SC, 88210-000, Brazil on 2016-02-01 14:57:17
Time in place: 1 hour(s) and 31 min
Next transportation: on foot

Content:

Likes and comments: 8

See the source content

(b)

Fig. 9. Graphical presentation of a Travel History generated using RTD, STD and GSI data sources: a) an overview of the entire trip and b) a detailed view of part of the trip.

presented to the volunteers aimed at verifying the level of satisfaction with the accuracy and similarity of the Travel History created based on their digital footprints when compared with the events and destinations of the real trip they have made. An interactive map allowed volunteers to check visited *Places*, analyze the performed *Trails*, verified transportation means used, and examine associated semantics. The tool used to present the graphical realization of the trip was not evaluated. To the best of our knowledge, however, there is no application or tool that reconstructs the path of travelers based on their social footprints in an automatic fashion.

In the process of evaluating the travel reconstruction process, the volunteers have answered five questions. For each question, volunteers were asked to give a grade ranging from 0 to 10. A compilation of the experiment results are presented in the Figure I, showing, for each aspect analyzed, the average grade given by the volunteers, the standard deviation and the percentage of grades that represents a

Table I. Travel History reconstruction evaluation results

Aspect analyzed	Avg. Grade	Standard Deviation	Totally accurate or in most of the cases
Visits identification accuracy	7.71	2.7	78.25%
Visits and *Trails* order accuracy	7.82	2.03	82.6%
Transportation means identification accuracy	7.39	2.19	69.56%
Activities and semantic identification accuracy	7.06	1.79	74%
Travel History rebuilt represents the real travel made.	8.69	2.14	95.65%
Averages considering all aspects	**7.73**	**2.17**	**80.01%**

positive evaluation (grade more than 7). The last row summarizes the grades considering the grades given for all aspects.

The first question aims to evaluate the accuracy of *Visits* identification. The result for this question indicates that most of the evaluators ($\approx$ 79%) considered that the identification of *Visits* was completely accurate or precise in most cases, while $\approx$ 13% felt it was accurate in some places, and 8% found the process of *Visit* identification was slightly or completely inaccurate.

The second question measures the satisfaction with the temporal order of visits and trails. This question is related to the integration process of *Stays* and *Visits*, which is responsible for identifying overlaps, to perform merges, and to sort these entities. The result indicates that most evaluators (82.6%) considered that the order of visits and trail order has been completely precise or precise in most cases, 13.05% found that the identification of the order was accurate in some places and only one evaluator (4.31%) considered that the order of *Visits* and *Trails* was imprecise.

The third question evaluates the accuracy of identifying the transportation mode used in each *Trail*. Although the level of satisfaction with the identification of the transportation mean is close to 70%, this aspect has the worst evaluation on the survey. The identification of the transportation mode is directly linked to the existence and granularity of RTD and STD sources and the accuracy of the location of GSI.

The identification of transportation means is directly linked to the level of granularity of digital footprints (in the case of GPS records) and to the recovery of *Social Interactions* with reliable geographical information, which can still be a problem with some social networks. In addition, the identification of the transportation means improves when the interactions in OSNs are made in real time during the trip and not published after the end the travel. The popularization of mobile always connected devices is increasing every day, allowing people to interact in real time even when they are traveling, which tends to improve the accuracy of the process of transportation means identification.

The fourth question is more subjective and it is related to the accuracy of the semantic enrichment process. In this regard, 73.9% of the evaluators answered that the identification of activities and interactions was completely accurate or accurate in most cases. The semantic enrichment process can be improved by incorporating the capability of including textual content of Social Interactions and with the ability to access structured information about users' activities. Facebook, for example, has such kind of information, but, at the time of writing, it is not possible to access such kind of information using third-party applications.

The fifth question evaluated the overall perception of the reconstruction process. It is by far the best-rated item on the survey. Almost 96% of the evaluators considered that the Travel History reconstructed represents, totally or in the major part, the travels they have made. Taking all aspects together, about 80% are satisfied with the proposal of reconstructing semantic trajectories based on heterogeneous social tracks sources.

7. CONCLUSION

Considering the digital socialization growth and the search for online social recognition, travelers begin to demand ways to share their travel experiences in a systematic and intuitive way. This article proposes a conceptual and a data models and a methodology to reconstruct semantic-rich traveler's trajectories. The central entity of the model is Travel History, which is an entity that aggregates *Stays*, *Visits* and *Trails* traveled by an individual during a given time interval. A *Trail* is an entity that captures the traveler movement and the transportation mode used. A *Stay* represent places where the *Traveler* remained for a while or changes the transportation mode. A *Stay* becomes a Visit if it is a place of intense online social interaction or if it is a place where the traveler spent a considerable amount of time.

Model's entities are instantiated based on a myriad of sources of information, varying from detailed low-level GPS registries and going up to high-level georeferenced social interactions. Thus, the proposed methodology used to generate models entities and to identify transportation mode is based on techniques to process, analyze, and integrate data with different levels of semantic and spatial-temporal granularity. The ability to reconstruct the trip successfully is directly related to the quality and quantity of the sources of information available. Different, reliable, and abundant sources of information will produce rich and accurate travel histories. On one hand, RTD and STD are good sources of information for detailed analysis of the trip. On the other hand, GSI generates semantic richer entities. As expected, the combination of all sources produces the best result.

In order to evaluate the proposed model and methodology, an experiment with travelers from a social traveler's network was designed and run. The results of the experiment show an overall level of satisfaction of 80%, considering the identification of the model's entities (i.e., *Stays*, *Visits* and *Trails*), the temporal order of these entities, the identifications of the transportation mode used by the traveler, the identification of activities performed by the traveler during the trip, the semantic enrichment of travelers' activities, and the level of adherence of the modeled travel history with the real trip.

As future work, there are several aspects that can be investigated. Semantic enrichment, for instance, can be improved by incorporating text-mining algorithms. Moreover, a travel social media ontology can be developed to improve semantic identification. To improve the data accuracy and granularity, mobile applications can be used to collect other kinds of social interaction, like offline media capture or any other type of interactions on the device. Algorithms for transportation means identification can be improved to become more accurate and to support the identification of other kinds of transportation.

The use of the proposed model and methodology in Web application in the tourism domain will allow the reconstruction of large number of Travel Histories, which in turn, can be a way to generate a knowledge base for travel itineraries, preferences, attractions and other aspects and events inherent to travels. This knowledge can be used as the base of a travel recommendation system or other initiatives such as urban planning, demographic and behavioral studies, intelligent transportation systems, social recognizing research, among others. Despite the fact that the model is generic and that it can be, in principle, used in several domains to describe semantic trajectories, the usage for other domains requires further investigations

REFERENCES

Akehurst, G. User Generated Content: the use of blogs for tourism organisations and tourism consumers. *Service Business* 3 (1): 51–61, 2009.

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., and Vaisman, A. A Model for Enriching Trajectories with Semantic Geographical Information. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems (GIS '07)*. Seattle, USA, pp. 22:1–22:8, 2007.

Andrienko, G., Andrienko, N., and Wrobel, S. Visual Analytics Tools for Analysis of Movement Data. *ACM SIGKDD Explorations Newsletter* 9 (2): 38, 2007.

Bogorny, V., Renso, C., Aquino, A. R., Siqueira, F. L., and Alvares, L. O. CONSTAnT - A Conceptual Data Model for Semantic Trajectories of Moving Objects. *Transactions in GIS* 18 (295179): 66–88, 2013.

Fileto, R., Krüger, M., Pelekis, N., Theodoridis, Y., and Renso, C. Baquara: A Holistic Ontological Framework for Movement Analysis Using Linked Data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8217 LNCS. Hong Kong, China, pp. 342–355, 2013.

Gao, Y., Tang, J., Hong, R., Dai, Q., Chua, T.-S., and Jain, R. W2Go: a travel guidance system by automatic landmark ranking. In *Proceedings of the International Conference on Multimedia - MM '10*. Firenze, Italy, pp. 123, 2010.

Gil, R., Nabo, B., Fileto, R., Nanni, M., and Renso, C. Annotating Trajectories by Fusing them with Social Media Users ' Posts. In *Geoinfo Synposiun 2014*. Campos do Jordão, Brazil, pp. 21–33, 2014.

Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.-M., Pang, Y., and Zhang, L. Equip Tourists With Knowledge Mined from Travelogues. In *Proceedings of the 19th International Conference on World Wide Web - WWW '10*. Raleigh, USA, pp. 401–410, 2010.

Ji, R., Xie, X., Yao, H., and Ma, W.-Y. Mining City Landmarks from Blogs by Graph Modeling. In *ACM International Conference on Multimedia*. Beijing, China, pp. 105–114, 2009.

Lange-faria, W. and Elliot, S. Understanding the Role of Social Media in Destination Marketing. *Tourismos: an International Multidisciplinary Journal of Tourism* 7 (1): 193–211, 2012.

Lu, X., Wang, C., Yang, J.-M., Pang, Y., and Zhang, L. Photo2Trip: generating travel routes from geo-tagged photos for trip planning. In *Proceedings of the International Conference on Multimedia - MM '10*. Firenze, Italy, pp. 143–152, 2010.

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., Theodoridis, Y., and Yan, Z. Semantic Trajectories Modeling and Analysis. *ACM Computing Surveys* 45 (4): 42:1–42:32, 2013.

Rattenbury, T., Good, N., and Naaman, M. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. pp. 103–110, 2007.

Spaccapietra, S., Parent, C., Damiani, M. L., Antonio, J., De Macedo, J. A., Porto, F., and Vangenot, C. A Conceptual View on Trajectories. *Data & Knowledge Engineering* 65.1 (May 2007): 126–146, 2008.

Tietbohl, A., Bogorny, V., Kuijpers, B., and Alvares, L. O. A Clustering-based Approach for Discovering Interesting Places in Trajectories. In *SAC '08 Proceedings of the 2008 ACM Symposium on Applied Computing*. Fortaleza, Brazil, pp. 863–868, 2008.

Tobergte, D. R. and Curtis, S. *Social Media in Travel, Tourism and Hospitality: Theory, Practice and Cases*. Vol. 53. Ashgate Publishing, Ltd., 2013.

Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., and Aberer, K. Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology* 4 (3): 49:1–49:38, 2013.

Ye, Q., Law, R., Gu, B., and Chen, W. The Influence of User-Generated Content on Traveler Behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior* 27 (2): 634–639, 2011.

Yoon, H., Zheng, Y., Xie, X., and Woo, W. Social Itinerary Recommendation from User-Generated Digital Trails. *Personal and Ubiquitous Computing* 16 (5): 469–484, 2012.

Zheng, Y., Chen, Y., Li, Q., Xie, X., and Ma, W.-Y. Understanding Transportation Modes Based on GPS Data for Web Applications. *ACM Transactions on the Web* 4 (1): 1–36, 2010.

Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. Mining Interesting Locations and Travel Sequences From GPS Trajectories. In *Proceedings of the 18th International Conference on World Wide Web - WWW '09*. Madrid, Spain, pp. 791–800, 2009.