Assessing pervasive user-generated content to describe tourist dynamics

Fabien Girardin, Josep Blat

Universitat Pompeu Fabra, Barcelona, Spain
{Fabien.Girardin, Josep.Blat}@upf.edu

Abstract. In recent years, the large deployment of mobile devices has led to a massive increase in the volume of records of where people have been and when they were there. The analysis of these spatio-temporal data can supply high-level human behavior information valuable to urban planners, local authorities, and designers of location-based services. This paper proposes the study of publicly available people generated geo-referenced content to provide novel perspectives on tourist dynamics. Our initial works analyzed these digital footprints people leave behind them as a historic of physical presence when they visit a city. The results provided insights on the density of tourists, the points of interests they visit as well as the most common trajectories they follow. Yet to be able to fully analyze these newly available data, there is a need to understand the diverse circumstances they were generated from. We believe that the understanding of the human practice behind these data and their relation to the urban space could open a new perspective in analyzing tourism.

Introduction

The recent emergence of location technologies and techniques favoured the development of new approaches to capture and analyze people's mobility (see [1], 2004 for a survey). One opportunity has been to replace traditional travel diaries, paper-and-pencil interview, computer-assisted telephone interviews, and computer-assisted-self-interview, by automatically collecting mobility data. Yet these automatic location sensing techniques open new privacy, scalability and longevity issues that limit their deployment. In this paper, we show that volunteered geographic information [2] can be at the source of the solution to these issues to capture tourists' mobility and behaviours during their visits. Recent research showed the potential of the geographically annotated material available on the Web. For instance, some scholars showed that the location and time metadata associated with photos and their tags enable the extraction of "place" and "event" semantics [3]. In this work we exploit a similar dataset in the aim of validating a new approach to describe tourist dynamics.

A first part of this paper report on our preliminary work on the problematic of revealing the presence of tourists during their visits of a city. In the second part of this paper, we suggest that these results must be first assessed with the analysis of how

"volunteers" handle and annotate their measurements and observations (e.g semantic description in geotagging, granularity in georeferencing).

Understanding tourist dynamics

Our approach takes advantage of the unprecedented amount of digital data linked to the physical world generated by the recent explosion in the use of capture devices (e.g. digital cameras) and collaborative web platforms to share their content (see [4] for a review). We produced statistics, geovisualizations and animations to reveal the promises of exploiting user-generated content in urban tourism [5]. Our early case studies took place in the Province of Florence (Figure 1) as well as in the cities of Rome, Italy and Barcelona, Spain. The local authorities of these regions aim at better understanding the tourist flows traveling across their boundaries. So far, they have been using classical survey-based hotel and museums frequentation data to know where tourists of different nationalities prefer to spend their time, hence money. However, they lack observations of the mobility, nationality and quantity of the "day trippers", that is the tourists who visit but are "invisible" in the data, as they do not sleep in town.

Fig. 1. Heatmaps revealing the presence of photographers from their accumulated georeferenced photos, in the Northern part of central Italy, in downtown Florence and around the Basilica di Santa Maria dal Fiore. Photographic imagery copyright Telespazio.

We retrieved large amounts of georeferenced photos taken by thousands photographers and shared via the popular photo-sharing web platform Flickr[1] (see Table 1 for the figures). Our approach took advantage of open and freely-available resources and combines them using de facto standards often based on the extensible Markup Language (XML). We used Google Earth for interactive visual synthesis of encodings generated using a combination of MySQL for data storage and querying to select and aggregate. We developed a software named "Urban Dynamics" to access, process, transform, cluster, sample, filter the raw data stored in the database and to generate geovisualizations.

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

Region	Number of photos	Number of photographers
Barcelona	154,106	5818
Province of Florence	81,017	4280
Rome	144,501	6018

Table 1. Accumulated number of georeferenced photos and number of photographers over a 2 years period (2005-2007) in the three regions of the case study.

Based on the time and the disclosed location of the photos, we extracted records of the people presence and movements such as density of tourist, inbound and outbound trajectories, patterns of flow between points of interest; performed statistical analysis and designed geovisualizations. This exploratory visual analysis was used as a mean of preliminary investigation. In fact, the results go far beyond the initial expectations of collecting clues on "day trippers" activities. The study disclosed new insights for tourism officials, such as the density of the flows of different types of visitors among the main attractions of the city (Figure 2).

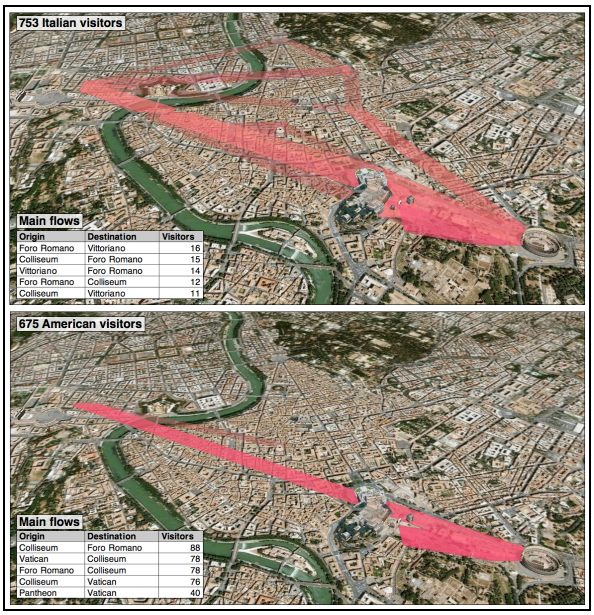

Fig. 2. Geovisualiation of the main paths tourists took to visit the points of interests of the city. 753 Italian visitors (top) are activity off the beaten tracks with a large amount of points of interets visited but little flow. 675 American visitors (bottom) stay on a narrow desire line. Photographic imagery copyright Telespazio.

A user-centered approach to assess the data

At this stage, it can be hypothesized that uploading, tagging and making public the location of a photo can be interpreted as a report of a physical presence in space and

time. However, the data collected suffer similar problems as automatic sensors when it comes to the fluctuating quality in the data and trust. For instance, the city and type of urban landscape impact the use of coarse and fine-grained location information [6]. In addition, at a very general and caricatural level, our early results suggest that German users tend to provide more accurate location information to their photos than Spaniards or that Europeans use more tags than North Americans. These observations highlight the need to understand how people employ the accuracy of location information to link digital information with the physical world.

Understanding the practice

In consequence, prior to assess the significance of user-generated content for tourism, one must first understand people's practices of annotating content with geographical information. We are currently working on providing answers in that direction to the question: Which are the people different georeferencing and geotagging behaviours? We investigate this domain with the following sub-questions and motivations.

Domain: Sub-question	**Motivation and approach**
Semantic: Which ways do people use to label (i.e. geotag and describe) information?	Identify and describe the different strategies of geotagging, by correlating the use of main tags categories to profile the users and the space
Granularity: How do people distinguish and define the levels of granularity to georeference the information?	Identify and describe the different strategies of georeferencing by comparing the categories of tags used and the location accuracy. Profile the types of traces and their accuracy.
Disclosure: How does automatic positioning influence location disclosure?	Describe the impact of automatic georeferencing of information on the geotagging behaviour. For instance, the automatic georeference through GPS could generate a poorer labelling of the information.
Co-evolution: How do the practices of geotagging and georeferencing evolve over time?	Practices might not be static. Describe how they change over time. Users might move from one type of behaviour to another. In that case we would need to describe the circumstances of these changes.

This analysis of the user practice will assess a theoretical framework that describes the human agency in the production of geo-annotated photos. It will be used to study the significance of this user-generated content in urban, tourism, mobility and travel research.

Method

The georeferenced photos stored in Flickr come with an extensive amount of empirical data about themselves (e.g. nationality, type of camera) the language and words the users employ to make public the information location as well as their strategies (e.g. photos uploaded in one batch, accuracy) to do so. The analysis builds on these qualitative data with the generation of categories (e.g. users' perspectives, process, activity, strategy) and positioning it within a theoretical model to describe practices to georeference information. Statistical, data mining and information visualization techniques are used to extract a deeper meaning from individual and collective perspectives in terms of semantic, granularity, disclosure and co-evolution. The validation of the observations of categories and patterns takes place with the comparison within multiple cities and cultures.

Calibration of user-generated content

The accumulated and aggregated records of where and when people were seem to lead to an improved understanding of different aspects of mobility and travel. However these new insights only mirror a specific community of Flickr users. Therefore, there is a strong need to examine their significance in comparison with existing mobility and activity data available in cities. We are currently developing an extension to our current Urban Dynamics application to explore multiple datasets to calibrate and validate the results with secondary order presence, survey and mobility data (e.g. cellphone network usage, tickets sold, manual counting, and Bluetooth scanning). Due to the large difference in the nature of the activity producing the data we compare, it might be that correlating different datasets does not only reinforce observations, but also reveals additional dimensions of user behavior that we might not yet have accounted for. For instance, the presence of photos in one area suggest a space for sightseeing. On the other and a strong cellphone network activity of roamers indicate another type of tourist activity such as relaxing. A current challenge is to understand more precisely the user-activity that is reflected in each of these types of datasets.

Conclusion

The explosion in the use of mobile and captures devices (e.g. mobile phone, digital cameras) and the emergence of content sharing platforms is leading to the emergence of a wealth of publicly available user-generated geospatial data. Our first case study specifically featured the value of Flickr and its geographically reference photos with the goal of performing urban and mobility analysis of visitors. This exploratory analysis enabled to quantify by the amount of photos taken and the presence of individuals the attractiveness over time of the major points of interests of an urban space. This work shows that there might be potential in taking advantage of the digital traces people constantly leave behind them. It could, for instance, reveal the temporal

character of a space, its attractiveness among a certain group of people, and its level of connectivity with other spaces. This type of insight is limited in most travel survey and sensing infrastructure by privacy-sensitive or aggregated information.

Yet these insights have a fluctuating quality as they rely on data generated and disclosed by people. We have observed differences in the accuracy of the location information depending on the region they describe or the cultural background of the volunteers who generated the content. In this paper we have suggested that a detailed study of the individual and social practices behind these data is necessary in order to assess the relevance of the user-generated content to inform on tourist dynamics.

References

1. Wolf, J. Applications of new technologies in travel surveys. In *7th International Conference on Travel Survey Methods, Costa Rica*. (2004).
2. Goodchild, M. F. Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0. *International Journal of Spatial Data Infrastructures Research 2* (2007), 24–32.
3. 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* (New York, NY, USA, 2007), ACM Press, pp. 103–110.
4. Torniai, C., Battle, S., and Cayzer, S. Sharing, discovering and browsing geotagged pictures on the web. Tech. rep., HP Labs, 2007.
5. Girardin, F., Fiore, F. D., Ratti, C., and Blat, J. Leveraging explicitly disclosed location information to understand tourist dynamics: A case study. *Journal of Location-Based Services*. In print (2008).
6. Girardin, F., and Blat, J. Place this photo on a map: A study of explicit disclosure of location information. Late Breaking Result at Ubicomp 2007, September 2007.