Workshop on Visualisation in Environmental Sciences (EnvirVis) (2015)
A. Middel and K. Rink and G. H. Weber (Editors)

Mining Social Images to Analyze Routing Preferences in Tourist Areas

A. Torrisi [1,2], G. Signorello[2], G. Gallo[1], M. De Salvo[3], G. M. Farinella[1]

[1]Department of Mathematics and Computer Science, University of Catania
[2]Centre for the Conservation and Management of Nature and Agroecosystems (Cutgana), University of Catania
[3]Department of Business Administration, University of Verona

Abstract

Social media platforms provide a useful source of data for environmental planning. In the last years these data have been exploited to perform social behaviour analysis. This work uses the huge amount of georeferenced images publicly available on social media as a source of information to infer the behaviour of tourists. Visual analytic mapping tools combined with the Parzen-Rosenblatt non-parametric kernel density estimation give us visual clues to assess the attractiveness of tourist geographical areas. To investigate the preferred combinations of locations visited by the tourists within a time window of few days we propose to mine association rules using the Apriori algorithm. A prototype of an integrated system to visually perform the suggested analysis has been realized and the paper reports about some of case studies performed with it.

Categories and Subject Descriptors (according to ACM CCS):
H.4.2 [INFORMATION SYSTEMS APPLICATIONS]: Types of Systems—Decision support
H.2.8 [DATABASE MANAGEMENT]: Database Applications—Image databases

1. Introduction and Motivation

Tourism behavior analysis is important to the economy of many geographical areas. Researchers have traditionally studied tourism flows by conducting interviews and surveys at entrances of sites of interest, such as museums and national parks. This kind of data collection is expensive and is limited in terms of spatial and temporal coverage. Tickets at the entrance of the attractions are usually used to count the number of visitors. However a huge number of natural and monumental areas are public spaces and there is no possibility to count people through ticketing. For instance, we know that highly recognized public tourist areas inscribed on the World Heritage List [Une] are visited by many people every year, but there is no possibility to know the number of people which have visited such places that do not require any access ticket or permit. The other source of data, the direct interview of tourists, is time consuming and it can be usually performed only for a limited number of days to limit costs. Since stratified information about gender, nationality, ages, etc. of the visiting persons is of special relevance, the size of the significative samples for a direct interview would be prohibitively high. To avoid direct interviews we propose to exploit the huge amount of information spontaneously posted by tourists on social media. Specifically, we consider the wisdom offered by the geolocalized images shared by tourists on social platforms, such as Flickr [Yah], to understand preferences of the tourists and relations among different tourist sites. Mining social data brings to the analyst an extra bonus: it is very useful for planners and policy makers to learn the statistical association rules that model the tourist traffic among neighboring sites, e.g., "is the site A frequently visited by the tourists which usually visit the site B?" We show that standard techniques to mine rules from relational data can be easily applied in this case, and if coupled with some simple graph drawing techniques may grant interesting and non trivial insight.

2. Related Work

Recent studies consider the possibility to exploit the data present in social media to analyse tourist behaviour and preferences [ScA14, WGSL13, CD11, WKC13, DSG*12]. Specifically, considering the photos acquired and shared on social media platforms by tourists through their smartphones, it is possible to have geolocalised data to be ana-

DOI: 10.2312/envirvis.20151093

lyzed for tourism behaviour analysis purpose. Among the photo sharing platforms, Flickr [Yah] is one of the more common and for this reason has been selected for the present research. This social media platform provides an Application Programming Interface (APIs) used by developers to get and analyze public data shared by the users. It should be noted that the proposed technique may be extended to any social media platform that makes geotagged data easily accessible. Since most of the images are georeferenced when acquired by a smartphone, the photo shared in a social media can give an important information about the presence of a person in a site.

Chareyron et al. [CD11] presented a framework to detect tourist areas. In order to localize the most popular areas of Paris, the distribution of photos in social media with respect to geoinformation is considered and a peak-finding algorithm is applied. They propose a method to find a path related the best sites to be visited by a tourist. Wood et al. [WGSL13] proposed to exploit social media images to quantify nature-based tourism and recreation. Specifically, they considered 836 recreational sites around the world and used the geolocalized images gathered from the profiles of Flickr photographers to derive travelers' origins. Their analysis pointed out that the crowd-sourced information allow a reliable estimation of visitation rates. Straumann et al. [ScA14] analyzed spatial and temporal patterns of georeferenced photographs of Zurich to understand user behaviour of foreign versus domestic visitors. The extraction of behaviour patterns has been performed using qualitative and quantitative visual analytical methods. Wang et al. [WKC13] studied the feasibility of observing the state of the natural world by recognizing specific types of scenes (i.e., snowy scenes) and flowers (i.e., California Poppy) in large-scale social image collections. Doersch et al. [DSG*12] used a large repository of georeferenced images and Computer Vision techniques to automatically discover which visual elements (e.g., street signs, balconies, etc) are most discriminative for a certain geo-spatial area (e.g., Paris).

Taking into account the aforementioned works, in this paper we exploit the available data on the social media together with web-based visualization tools [Hig] to analyze the behaviors of tourist in different sites of interest in two paradigmatic situations: a broad region (i.e. Sicily island) and a monumental city (i.e. Rome). We propose to employ the Parzen-Rosenblatt kernel density estimation method [Bis06] to infer the georeferenced probability distribution related to the visitors of the monitored areas. In addition, the Apriori data mining algorithm [AS94] is employed to discover association rules among sites of interest and to highlight general trends of tourists preferences.

3. Proposed Analytical Tool

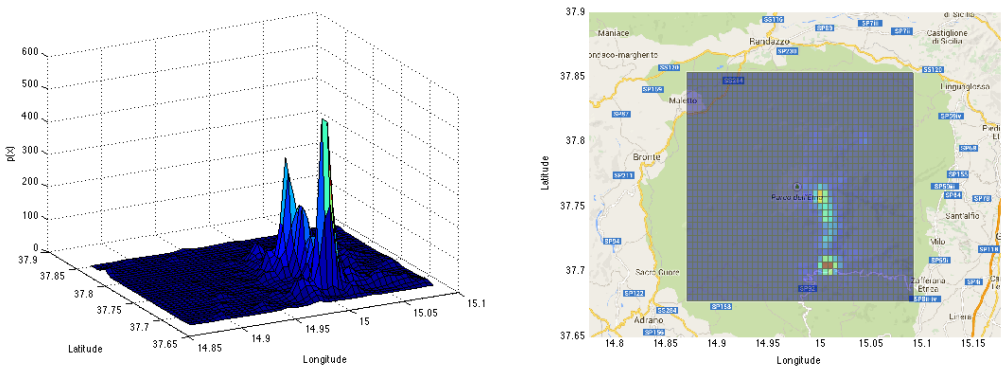
In this work we take into account the geolocalized data that can be retrieved from Flickr [Yah]. Given a set $S = \{s^1_{(lon,lat)}, s^2_{(lon,lat)} \dots, s^m_{(lon,lat)}\}$ of tourist sites identified by their GPS positions, we use the Flickr APIs to download the images and the related EXIF and Flickr metadata for each site. By taking into account the GPS coordinates $s^i_{(lon,lat)}$ of the i^{th} tourist site we consider all the images having GPS coordinates belonging to a square region centered in the tourist site. The size of the square region depends of the site and is set such that the area of interest around the tourist site is covered. Notice that a squared shape has been adopted in this prototypal study for sake of simplicity in implementation. This process produces a set $C = \{I^1_{(lon,lat)}, I^2_{(lon,lat)} \dots, I^n_{(lon,lat)}\}$ of longitude (lon) and latitude (lat) coordinates related to the social georeferenced images which users have uploaded for all the considered tourist sites in S. Given the set C we use the Parzen-Rosenblatt method [Bis06] to perform the estimation of the probability density function $p(lon, lat)$ in the minimal rectangular geographical area that includes all the sites. The Parzen-Rosenblatt method doesn't require any knowledge or assumption about the underlying distribution. Intuitively, this approach counts how many samples fall within a specified square region $R_{(lon,lat)}$ with size $h \times h$ surrounding a geolocalised point of interest with GPS coordinates (lon, lat). In our study, we consider a Gaussian kernel centered on geolocalised point of interest to compute the Parzen-Rosenblatt probability density function. Hence, given the set C of the images' coordinates, the probability density function in a location $x = (lon, lat)$ is estimated as:

$$p(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2\pi h^2} \exp\left(-\frac{||x - x_i||^2}{2h^2} \right) \quad (1)$$

To turn this averaging method into a visual analytic tool the resulting probability density function is overlayed using standard mapping and image processing tools on the monitored tourist geographic area to better understand the most popular tourist sites. An example of such a distribution applied to the data related to the Mount Etna is shown in Figure 1.

Since each image has EXIF and Flickr metadata associated, it is simple to obtain information about the number of photographers which have visited the considered tourist sites, with their gender and nationality, as well as the date in which the image has been shot. These information are used together with Highcharts API [Hig] to visually analyze the distribution of tourists with respect to gender, nationality (e.g., domestic vs foreign) and considering the date of the tourist visit. Two examples of visual charts included in our visual analysis tool are shown in Figure 2.

To learn association rules among tourist sites the Apriori algorithm is employed [AS94]. To do that we refer to an ideal matrix in which the rows correspond to tourists (i.e., the different photographers) and the columns correspond to the different tourist sites. Please observe that in this way we associate to an individual photographer all the sites that he



(a) (b)

Figure 1: Parzen-Rosenblatt kernel density estimation. (a) Density estimated on data related to the Mount Etna. (b) 2D projection of the density superimposed over geographic map. The highest peak occurs at the point of coordinates $(lat = 37.7, lon = 14.99)$ that corresponds to the site "Rifugio Sapienza".

has posted on Flickr within a temporal window of few days. No ordered route information is considered in this project. Every cell in the matrix is a binary variable whose value is one if a tourist has visited a specific site (equivalently the tourist has shot a photo of that site) and zero otherwise. Given this matrix, the algorithm attempts to infer frequent subsets of sites which are common to the different tourists with a minimum support. To this aim Apriori uses a bottom-up approach, where frequent subsets of tourist sites are extended one site at a time (i.e., the candidate), and groups of candidates are tested against the data. This iterative process continues until no further successful extensions of the sites' subsets are found. The frequent tourist sites item sets computed by the Apriori algorithm are then exploited to determine association rules that highlight general trends related to which sites are visited jointly with a specific confidence score. In our settings, for a given rule $s^i_{(lon,lat)} \rightarrow s^j_{(lon,lat)}$, its confidence is proportional to the likelihood that the site $s^j_{(lon,lat)}$ is visited during the same trip of a photographer who has visited the site $s^i_{(lon,lat)}$.

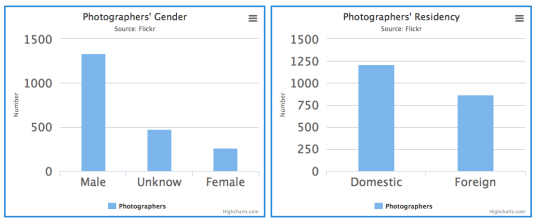


Figure 2: Examples of charts generated with Highcharts API. (a) Gender and (b) residency of photographers corresponding to social images of Mount Etna

4. Case Studies

In this section we report the experiments performed on two tourist areas: Rome and Sicily. The reason why we choose these areas come from the need to test the proposed framework on two kinds of sites: one within a city boundaries and one in a much broader area like a large island.

4.1. Rome

In this case study we consider the geolocalized Flickr images related to 10 tourist sites in Rome, for a total amount of 27963 photographers. To generate association rules the Apriori algorithm requires from the analyst the choice of a *minsupport*, the minimum amount of evidence (the number of photographers) required to consider an association rule as valid. High values of *minsupport* reduce the chance to find association rules observed over a too small set of visitors. One way to choose appropriate *minsupport* is to start with a high value and then gradually decrease it until enough association rules are generated. For the choice of *minsupport* the analyst has to find a trade off between two conflicting issues: include a sufficiently large set of sites in the rule mining and at the same time to skip rarely photographed locations. We choose a *minsupport* of 14.9% to balance between these issues. This choice left us only 8 sites from the original 10 to consider for rule mining.

The Apriori algorithm asks for another parameter: the $minconfidence$. This parameter weights how strong is the association between sites discounting the relative incidence of each single site. Setting $minconfidence$ at 30% lead us to generate the rules in Table 1. Some of these rules are quite intuitive, like the one $Colosseo \rightarrow PiazzaVenezia$ because the two places are really close and it is very likely that tourists visit them both. As shown in Figure 4 a graph representation of the rules found with the Apriori algorithm

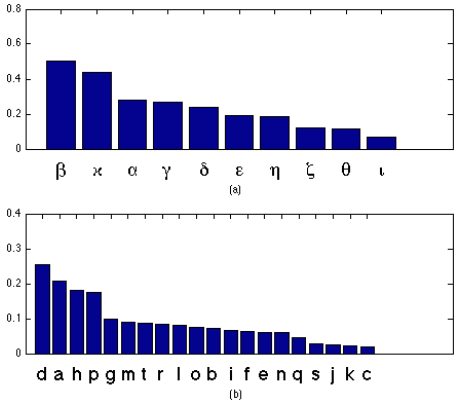


Figure 3: Photographers' distribution for tourist sites in Rome (a) and Sicily (b). Sites name refers to labels on Table 1

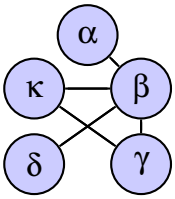


Figure 4: Graph representation of the association rules in Table 1 related to the sites in *Rome*

reveals the centrality of the site β (Colosseo) which is the highest degree node for the graph. Moreover it is interesting to observe in the graph that the largest clique has size 3 and includes *Colosseo*, *Vaticano* and *Pantheon*.

4.2. Sicily Island

A second experiment has been conducted on Sicily island by considering 20 main tourist sites. We consider both naturalistic areas, like "Mount Etna" and "Riserva dello Zingaro", and other sites relevant from cultural and historical motivations, like "Ortigia" and "Palermo". In order to apply the Apriori algorithm on these data, the same procedure described above for the "Rome" case study has been employed. First, we consider a $minsupport = 6.5\%$, so that 12 sites with greater supports are considered for further analysis (Figure 3(b)). We calculate the support for each couple of these candidate sites and the final *minsupport* is such that 10 couples have a support above it. Association rules found with $minsupport = 1.18\%$ are shown in Table 1. The *minsupport* value used here is lower than that used for the experiment related to Rome. This is due mainly to two reasons: the Sicilian sites are arranged on a much larger area compared to the context of a single city, like Rome. For this reason, the possibility that a photographer has visited two faraway locations of Sicily, taking photos and later uploading them to

Rome (Photographers = 27963 - *minsupport* = 14.89%)		
Site	**Label**	**Association rules**
Piazza Venezia	α	$\beta \rightarrow \kappa$ $(sup = 23.22\%, conf = 46.06\%)$
Colosseo	β	$\kappa \rightarrow \beta$ $(sup = 23.22\%, conf = 53.00\%)$
Pantheon	γ	$\alpha \rightarrow \beta$ $(sup = 17.69\%, conf = 63.40\%)$
Fontana di Trevi	δ	$\beta \rightarrow \alpha$ $(sup = 17.69\%, conf = 35.08\%)$
Piazza di Spagna	ε	$\beta \rightarrow \gamma$ $(sup = 17.12\%, conf = 33.95\%)$
Villa Borghese	ζ	$\gamma \rightarrow \beta$ $(sup = 17.12\%, conf = 64.65\%)$
Piazza Navona	η	$\beta \rightarrow \delta$ $(sup = 16.63\%, conf = 32.98\%)$
Piazza del popolo	θ	$\delta \rightarrow \beta$ $(sup = 16.63\%, conf = 69.29\%)$
Campo dei fiori	ι	$\gamma \rightarrow \kappa$ $(sup = 14.89\%, conf = 56.23\%)$
Vaticano	κ	$\kappa \rightarrow \gamma$ $(sup = 14.89\%, conf = 33.99\%)$
Sicily (Photographers = 9945 - *minsupport* = 1.8%)		
Site	**Label**	**Association rules**
Etna	*a*	$a \rightarrow d$ $(sup = 6.82\%, conf = 32.75\%)$
Catania - Cathedral	*b*	$h \rightarrow d$ $(sup = 5.93\%, conf = 32.66\%)$
Catania - Teatro Max	*c*	$e \rightarrow d$ $(sup = 3.73\%, conf = 62.14\%)$
Taormina	*d*	$o \rightarrow p$ $(sup = 3.29\%, conf = 43.61\%)$
Taormina - Isola Bella	*e*	$m \rightarrow p$ $(sup = 2.93\%, conf = 32.55\%)$
Vulcano	*f*	$m \rightarrow d$ $(sup = 2.75\%, conf = 30.54\%)$
Stromboli	*g*	$l \rightarrow d$ $(sup = 2.74\%, conf = 34.25\%)$
Ortigia	*h*	$b \rightarrow d$ $(sup = 2.63\%, conf = 36.08\%)$
Ragusa Ibla	*i*	$o \rightarrow m$ $(sup = 2.60\%, conf = 34.44\%)$
Piazza Armerina	*j*	$b \rightarrow h$ $(sup = 2.55\%, conf = 34.98\%)$
Scala dei turchi	*k*	$b \rightarrow a$ $(sup = 2.54\%, conf = 34.84\%)$
Valle dei templi	*l*	$l \rightarrow h$ $(sup = 2.49\%, conf = 31.11\%)$
Monreale	*m*	$a, d \rightarrow h$ $(sup = 2.45\%, conf = 35.93\%)$
Riserva dello zingaro	*n*	$a, h \rightarrow d$ $(sup = 2.45\%, conf = 52.47\%)$
Palermo - Cathedral	*o*	$d, h \rightarrow a$ $(sup = 2.45\%, conf = 31.35\%)$
Palermo - Teatro Max	*p*	$i \rightarrow h$ $(sup = 2.35\%, conf = 36.00\%)$
Marsala	*q*	$f \rightarrow g$ $(sup = 2.22\%, conf = 34.42\%)$
Erice	*r*	$e \rightarrow a$ $(sup = 1.95\%, conf = 32.49\%)$
Segesta	*s*	
Favignana	*t*	

Table 1: Association rules generated by considering Flickr images of Rome and Sicily

Flickr, is reduced. Moreover, the amount of georeferenced data in Sicily is smaller than that found in Rome. Despite this, the resulting association rules contain interesting elements. Such information can provide useful knowledge to the stakeholders to optimize tourism managament. For example, the rules show an isolated component in the relation graph: *Vulcano* $\rightarrow$ *Stromboli* that testifies that many people visit only the Eolian Island. The rules also confirm the well-known notion that *Taormina* is the focal site for the visitors: indeed many choose to stay there and to take day trips to the other places. Other rules simply reflect the proximity of interest points.

5. Conclusions and Future Work

In this article we demonstrate the soundness of an analysis of the behaviour of tourists based on their georeferenced images uploaded on social media. Through visual analysis of these data it is easy to understand where the concentration of tourists happens also with relation to gender and nationality. We proposed an application of the Apriori data mining algorithm to extract association rules that help to discover relationships between touristic places. Future research will provide insights to the results presented here through further examples and integrating other qualitative features to the Apriori analysis, like gender and nationality of photographers. We also are actively working to enrich the basic information coming from the metadata with Computer Vision techniques to extract relevant visual information from the pictures.

References

[AS94] AGRAWAL R., SRIKANT R.: Fast algorithms for mining association rules in large databases. In *International Conference on Very Large Data Bases* (1994), pp. 487–499. 2

[Bis06] BISHOP C. M.: *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006. 2

[CD11] CHAREYRON G., DA-RUGNA J.: A robust detection of tourism area from geolocated image databases. In *World Congress on Engineering and Computer Science* (2011), pp. 118–122. 1, 2

[DSG*12] DOERSCH C., SINGH S., GUPTA A., SIVIC J., EFROS A. A.: What makes paris look like paris? *ACM Transactions on Graphics 31*, 4 (2012), 101:1–101:9. 1, 2

[Hig] HIGHSOFT AS: Highcharts. http://www.highcharts.com/. [Online; accessed 25-January-2015]. 2

[ScA14] STRAUMANN R. K., ÇÖLTEKIN A., ANDRIENKO G.: Towards (re)constructing narratives from georeferenced photographs through visual analytics. *The Cartographic Journal 51*, 2 (2014), 152–165. 1, 2

[Une] UNESCO: World heritage list. http://whc.unesco.org/en/list/. [Online; accessed 25-January-2015]. 1

[WGSL13] WOOD S. A., GUERRY A. D., SILVER J. M., LACAYO M.: Using social media to quantify nature-based tourism and recreation. *Scientific Reports 3*, 2976 (2013), 1–7. 1, 2

[WKC13] WANG J., KORAYEM M., CRANDALL D.: Observing the natural world with flickr. In *ICCV Workshops on Computer Vision for Converging Perspectives, 2013* (2013), pp. 452–459. 1, 2

[Yah] YAHOO!: Flickr. http://www.flickr.com/. [Online; accessed 25-January-2015]. 1, 2