Monitoring social network user density variations in areas of interest

João Rosa André Sabino Armanda Rodrigues
NOVA LINCS, Departamento de Informática
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
Quinta da Torre, 2829–516
Caparica, Portugal
{jp.rosa@campus., amgs@campus., armanda.rodrigues@}fct.unl.pt

Abstract

Mapping the population's density is an important challenge for crisis management, specifically, early warning systems. Its assessment is key to determine the risk of an event and plan the emergency response in crisis situations. One shortcoming of current systems is that the population density is determined statically, mostly based on census data. We propose to dynamically assess the population density using data extracted from social networks with geo-referenced information, such as Twitter and Instagram. Not only do users constantly update their status on social networks, but a growing percentage of the information that is shared is geo-referenced, or can be analysed and related with a geographic location. The methodology we present in this paper will enable a dynamic map of social network user density variation in a given geographical area. We present the methodology, and two preliminary studies that assess the quality of the information on Instagram, and conclude that it enables the detection of a variation in the user density on two different scenarios.

Keywords: Social networks, User density estimation, Emergency management

1 Introduction

Assessing the risk of potential emergency situations caused by natural phenomena requires the study of the consequences of several scenarios on the affected areas. This risk assessment mainly considers the population density of the area, its land use, and natural environment.

Census data is the main source of information to estimate the population density of the area. However, although infrastructures and the local natural environment do not change in a short time interval, the human population density of the affected area may significantly vary, not only throughout the day, but weekly and seasonally.

We propose to develop a methodology to build dynamic social network user density maps for specific areas, using data from social networks like Twitter[1] and Instagram[2], leveraging its popularity and the increasing availability of high precision GSP sensor. This is a step towards the dynamic mapping of population density, which would increase risk assessment accuracy by minimizing the difference between population estimation and the actual density. This is particularly useful for early warning systems, where alerts are issued between a few days and a few hours before the event [1].

Our methodology is focused on the analysis of data for short time intervals, and relies on machine learning methods to relate users with potential future locations [2].

In this paper we present the outline of our methodology, and two preliminary studies that evaluate the quality of data from Instagram, concluding that it is suitable for the application of the methodology.

1 http://www.twitter.com

2 http://www.instagram.com

2 Methodology

After collecting data from the network, we perform a frequency analysis on the posts' keywords to determine which best describe the area. These are suitable candidates to train the classifier. We also select a random set of keywords that have no relationship to the area as a negative set. These two keyword sets are required to train the classifier.

The final step is the result analysis. Our goal is to make sure we detect most crowds in the area, i.e., true positives results, while tolerating some false positives. For early warning systems, crowd detection requires high recall values and acceptable precision values. The exact values of these metrics need to be defined in our future work, and evaluated by emergency management experts.

3 Preliminary Studies

We present two preliminary studies over the Instagram network to evaluate the precision of its geolocated content.

The two events were the Reef Hawaiian Pro Final in Hawaii, which occurred between 12 and 23 of November 2014, and one stage of World Rally Championship in Wales, which occurred between 13 and 16 of November 2014. These specific events were chosen because the areas where they occur are less populated, which is ideal to compare the number of people detected during the event with a regular day.

Each study is initiated by defining the request string to submit to the Instagram API. These are a geographic coordinate point and a radius. Requests are executed during and after the event, storing results in XML and JSON files for analysis, and to generate visual layouts with the results (with

Google Maps overlays), which enable response validation, and the comparison of the two time intervals.

The two studies differ on the type of data collected from the network. In the first study we requested users that have posted content about the event. Each data point, i.e., a user location, represents the location associated with the first post of that user.

In the second study we collected all posts. We expect to have more results with this approach, but it will represent less people. While the goal is to count people, the second study will enable an empirical evaluation of the post content, and verify that it is in fact related with the event.

3.1 Reef Hawaiian Pro Surf Final

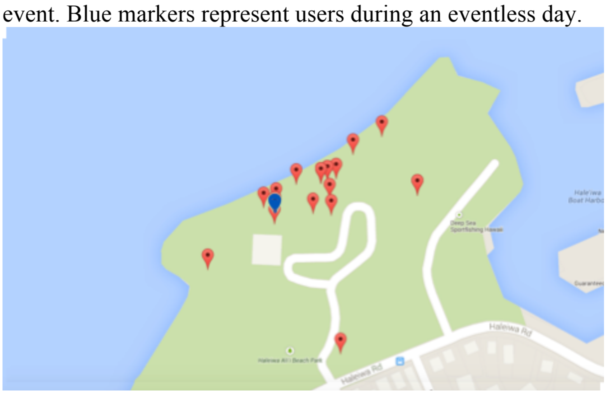
The study area is not populated most of the year. However, while it never gets actually crowded, it attracts people interested in surf and beach activities during the summer. We focus our study in an event that occurs during the time of the year when it is most populated, and try to detect the impact on population density during the surf championship event. Table 1 presents the parameters used for data capture.

Table 1. Instagram API query parameters and results for Reef Hawaiian Pro Surf.

Collection Time	Coordinate (Lat, Lng)	Radius	#Users
15/11/2014 (7h) to 16/11/2014 (23h)	21.592899º, -158.108767º	1 km	25
18/11/2014 (7h) to 19/11/2014 (23h)	21.592899º, -158.108767º	1 km	3

There were more users during the event (25), when compared with a regular day (3). Figure 1 shows the users that were near the location where the event occurred. Results hint that on event day there was a peak of population at the geographical area.

Figure 1. Users' location during the main event of Reef Hawaiian Pro Surf. Red markers represent users during an event. Blue markers represent users during an eventless day.



3.2 WRC Stage

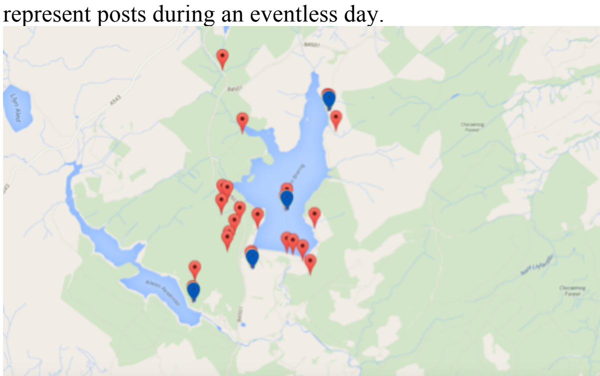
The study area is mostly unpopulated, and only attracts a few people interested in the landscape or water sports during the year. We focus on this area to evaluate the actual posts, and verify the content to determine if they are actually related with the event. Table 2 presents the parameters used in the queries.

Table 2. Instagram API query parameters and results for WRC Stage.

Collection Time	Coordinate (Lat, Lng)	Radius	#Posts
15/11/2014 (8h) to 16/11/2014 (23h)	53.081218º, -3.559491º	5 km	98
20/11/2014 (8h) to 21/11/2014 (23h)	53.081218º, -3.550491º	5 km	20

There are more posts during the event, when compared with the following day. Figure 2 shows posts during and after.

Figure 2. Posts during the event day of the WRC stage. Red markers represent posts during an event. Blue markers represent posts during an eventless day.



We manually classified the posts in order to determine that they are actually related with the event. Results show that only one post had unrelated content. We conclude that the variation in the number of posts is caused by the event.

4 Conclusions

This paper proposes a methodology for social network user density estimation, tailored to enhance crowd estimation for early warning systems. We present two studies to assess the quality of data from social networks, focusing on Instagram. They show that there are enough geo-referenced data on Instagram to map and detect changes in user density at a place, when capturing posts related with an event.

5 References

[1] C. J. E. M. Fortes, M. T. Reis, P. Poseiro, R. Capitão, J. A. Santos, L. V. Pinheiro, J. Craveiro, A. Rodrigues, A. Sabino, S. F. Silva, J. C. Ferreira, P. D. Raposeiro, C. Silva, M. C. Rodrigues, A. Simões, E. B. Azevedo, and F. Reis, "HIDRALERTA Project – A Flood Forecast and Alert System in Coastal and Port Areas," in Proceddings of the IWA World Water Congress and Exhibition, 2014.

[2] A. Sabino and A. Rodrigues, "Indirect Location Recommendation," in Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2014.