ARTICLE IN PRESS

Pervasive and Mobile Computing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Using geosocial search for urban air pollution monitoring

Matteo Sammarco [a], Rita Tse [b], Giovanni Pau [a,*], Gustavo Marfia [c,*]

[a] LIP6, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France

[b] Macao Polytechnic Institute, Rua de Luís Gonzaga Gomes, Macao

[c] Dipartimento di Scienze per la Qualità della Vita, Università di Bologna, 237 Corso D'Augusto, Rimini, Italy

ARTICLE INFO

Article history:
Received 30 September 2015
Received in revised form 5 June 2016
Accepted 7 July 2016
Available online xxxx

Keywords:
Air pollution
Online Social Networks
Text analysis
Pollution forecast

ABSTRACT

While Twitter and other Online Social Networks (OSNs) or microblogs are considered as a source of information for breaking news or uproarious and unexpected events, they could also be exploited as a dense worldwide sensors network for physical measurements. The *corpus* of geotagged posts from OSNs includes people's feedbacks about a wide range of topics, with precise temporal and geographical metadata, that can be used as a support or an improvement to hardware sensors. For instance, if collocated people, independently and at the same time, write posts complaining about high temperatures, it could effectively denote a raise of heat in that place. In this paper, we explore the feasibility to use a geographical search on social networks, that is, a geosocial search, about air pollution related posts, as effective air impureness measurements. We evaluate our assumption in large cities over three continents of the planet, where a minimum increment about the number of air pollution related posts in an area, indeed corresponds to a raise of minimum pollution values in such area. Such a correlation can be exploited to integrate and extend existing air pollution monitoring networks. At the end of the manuscript we propose to further employ the time series of posts returned by the geosocial search to predict next pollution values.

1. Introduction

Photography in Interstellar, Christopher Nolan's last movie, well depicts the current landscape of some of the most populated and industrialized places in the world. During air pollution peaks, the usual blue sky becomes covered by a thick reddish coat due to the massive presence of pollutants, especially sulfur dioxide, particulate matter, and nitrogen oxides (Fig. 1).

By now, this environmental issue is so diffused and evident that countries from all around the world are active to propose and ratify treaties such as the Gothenburg and Kyoto protocols. Such concern dates back in time, as over 40 years ago, in 1974, the United Nation Environment Programme (UNEP) and the World Health Organization (WHO) started the Global Environment Monitoring System for air pollution (GEMS/Air), which deployed air monitoring equipment in over 50 large cities throughout the world [1]. Mega-cities are indeed one of the most affected places for the air pollution plague [2], mainly due to prolonged exposures to high road traffic levels and factory emissions.

Nowadays, although institutional agencies and independent entities have improved the monitoring networks in large cities with the deployment of more sensing stations, they are usually not sufficient. For instance, MonitorAr-Rio, the agency in charge of monitoring air quality in Rio de Janeiro, comprises only nine fixed stations covering a territory of 1255 squared kilometers [3].

* Corresponding authors.
E-mail addresses: giovanni.pau@lip6.fr (G. Pau), gustavo.marfia@unibo.it (G. Marfia).

http://dx.doi.org/10.1016/j.pmcj.2016.07.001
1574-1192/

Please cite this article in press as: M. Sammarco, et al., Using geosocial search for urban air pollution monitoring, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.07.001

Fig. 1. Paris landscape during an air pollution peak (left side) and during a day with under-threshold values (right side). The massive presence of pollutants in the air heavily affects the view. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Although the cost of pollution sensing platforms is progressively decreasing, a pervasive and widespread deployment of such technologies is still far away in time for its costs. Both satellite and terrestrial solutions are being studied for a fine grained pollution estimation, with different results in terms of sensitivity, resolution and accuracy [4–6]. However, the common denominator of all these technologies is the use of specialized hardware devices capable, by different means, of estimating the presence and concentration of specific pollutants. No automatic system has, instead, so far, employed feedbacks received from the public.

In this work we explore the possibility of integrating air pollution sensing networks with measurements based on air pollution related posts, spontaneously generated by users on Online Social Networks (OSNs), through geosocial search. A geosocial search is the operation of finding events and user activities advertised on OSNs, in a specific geographical area [7,8]. OSNs include a huge *corpus* of geotagged posts related to various topics, which may be exploited to discover some knowledge pertaining a specific area. For example, if many co-located posts, written by different users, complain about car noise coming from the street, it is likely that a traffic jam is currently happening in that specific area. Similarly, if a geosocial search concerning high temperatures reveals many protests in a defined area, that place may be far from green spaces, thus more people suffer from heat. Also, geosocial searching for air allergies we may infer which are the most exposed places to air pollution or to spring pollen.

Resorting to geosocial search for urban air pollution monitoring brings several advantages. Mainly, in high density urban territories, post generation and upload is dense. Integrating information gathered from OSNs within traditional monitoring systems, may provide spatially more extensive and finer grained information. Such information could also be integrated in intelligent transport systems (ITS) or in in-car connected systems, as high pollution levels could indicate traffic congestion or particular weather conditions. Thus, traffic, for example, could be rerouted accordingly. In addition, people could decide on the sport where it is best to jog or to have a bike ride. In essence, city administrations and citizens at large could take advantage from the geosocial search for many different urban applications.

Apart from geographic coordinates, posts include sharp temporal information too, while many traditional monitoring networks, like MonitorAr-Rio, for example, only provide pollutant level values on a day-by-day granularity. Finally, gathering public posts from OSNs does not require particular costs as it totally avoids the installation, deployment and maintenance costs that are typical of traditional stations.

Nevertheless, using geosocial search for air pollution monitoring is challenging. Socio-cultural differences, which are substantial when changing language and location, must be accounted for in order to implement an effective search application [9]. People, in addition, very often adopt local expressions or slang to express their feelings in short messages. Thus, selecting a relevant set, from the massive stream of all posts, is, indeed, a very hard task. However, even when a relevant set were found, it may not be sufficient to detect pollution, as it may be possible that different people react in different ways, as they exhibit different tolerance thresholds to air pollution. Moreover, information on OSNs could be biased or untrue [10,11].

The contribution of this work is the design and implementation of a geosocial keyword-based search system to find where people complain most about bad air quality. The proposed system takes for geocultural differences with specialized dictionaries of keywords. When the correlation between the time series of pollutant measurements and related posts is proved, we also provide an algorithm which exploits OSN posts to forecast the most likely pollution level in a given area.

The system that is here proposed is depicted in Fig. 2. As people, independently from each other and from pollution sensors data, express their protests concerning bad air quality on OSNs, their air pollution related posts are selected from the unstructured mass of posts. Filtering is done either *a priori* resorting to a keyword-based searching scheme, or *a posteriori* through a post classification mechanism based on machine learning [12] and natural language processing (NLP) [13]. When many air pollution related posts are found in a given area, this likely indicates an actual raise of pollution in that area. If such information comes from an area where an existing correlation between pollutant levels and OSN posts has been previously verified, it can then be integrated with other pollution sensing network data, whenever available.

In a nutshell, we verify the feasibility of the proposed system as follows. We first show that air pollution related posts, with or without a per-user credibility assessment, can be put to good use as real air quality measurements. Such fact is corroborated by the large scale measurements that have been conducted in mega-cities over three continents. Secondly, we

Please cite this article in press as: M. Sammarco, et al., Using geosocial search for urban air pollution monitoring, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.07.001