Information Systems ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Information Systems

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

Shaping City Neighborhoods Leveraging Crowd Sensors

Giuseppe Rizzo [a,c,b,*], Rosa Meo [b], Ruggero G. Pensa [b], Giacomo Falcone [b], Raphaël Troncy [c]

[a] ISMB, Turin, Italy
[b] Università di Torino, Turin, Italy
[c] EURECOM, Sophia Antipolis, France

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form
19 June 2016
Accepted 28 June 2016

Keywords:
Geographic summarization
Thematic maps
Social media
Clustering
Urban data
Smart city

ABSTRACT

Location-based social networks (LBSN) are capturing large amount of data related to whereabouts of their users. This has become a social phenomenon, that is changing the normal communication means and it opens new research perspectives on how to compute descriptive models out of this collection of geo-spatial data. In this paper, we propose a methodology for clustering location-based information in order to provide first glance summaries of geographic areas. The summaries are a composition of fingerprints, each being a cluster, generated by a new subspace clustering algorithm, named GeoSubClu, that is proposed in this paper. The algorithm is parameter-less: it automatically recognizes areas with homogeneous density of similar points of interest and provides clusters with a rich characterization in terms of the representative categories. We measure the validity of the generated clusters using both a qualitative and a quantitative evaluation. In the former, we benchmark the results of our methodology over an existing gold standard, and we compare the achieved results against two baselines. We then further validate the generated clusters using a quantitative analysis, over the same gold standard and a new geographic extent, using statistical validation measures. Results of the qualitative and quantitative experiments show the robustness of our approach in creating geographic clusters which are significant both for humans (holding a F-measure of 88.98% over the gold standard) and from a statistical point of view.

1. Introduction

When planning a visit to a new city or when exploring a new area, travelers usually look for landmarks, sightseeing places, nightlife districts and pleasant restaurants, while avoiding areas that are known for a high crime rate when searching for an accommodation. Such an understanding of social, cultural, political, and economic aspects of an area goes beyond the physical structure of a city as defined by blocks and districts that are usually represented in thematic maps. City thematic maps are largely used by travelers and widely sponsored by travel agencies so far, but they generally offer static views of city parts delimited by too rigid boundaries. In addition, the accuracy of these thematic maps is proportional to their freshness: the more recently published, the better, but, for the dynamic aspects of a city, this requires regular updates, thus making the thematic map quickly outdated. This results in a mismatch between city thematic maps and the living city topologies.

The human generation of living city topologies follows a workflow in which one or more domain experts are involved. The forces that shape the dynamics of a city are manifold and thus complex to be tracked, making the expert task extremely difficult and error prone [1]. Generally, such a process requires: (i) a comprehensive

[*] Corresponding author at: ISMB, Via P. Carlo Boggio, 61 10138, Turin, Italy.
E-mail address: giuseppe.rizzo@ismb.it (G. Rizzo).

http://dx.doi.org/10.1016/j.is.2016.06.009
0306-4379/© 2016 Elsevier Ltd. All rights reserved.

knowledge of the city life character for shaping the right textures while considering numerous city aspects such as social, cultural and economic; (ii) a significant set of observations of those city aspects.

With the advent of the Open Data movement, many public actors such as municipalities, districts, and governments have started to release datasets that report public information such as employment rate and GDP per capita. This opens new perspectives for generating in an automatic fashion thematic maps: given the distribution of a feature (e.g. GDP) and the shape of a territory (e.g. a district or an entire country), it is possible to automatically aggregate data using intelligent algorithms and to infer the distribution of the feature values in the geographic area in a shape that can be later used by experts and thus travelers. In parallel, the massive involvement of citizens in social media services is constantly generating new sources of location-based data. This data encompasses people's actions, dynamics of cities, so that it instantaneously reports any changes in the city topology [2]. Such amount of data can, therefore, be considered as a crucial source for geo-spatial platforms if taken globally. A disruptive characteristic of this data is given by the fact that it is freely and collaboratively created by users and it often reports fine-grained descriptions of points of interests. Users act as crowd sensors by sharing their whereabouts and further enriching the available information of the territory with annotations such as place categories, tips, and comments.

Leveraging this massive amount of user whereabouts data, the objective of this paper is to define a methodology that automatically adds a layer over the typical cartography geographic maps, creating summaries on what crowd sensors tell about venues and, generally speaking, points of interest. Applying a geographic data-driven approach, our work grounds on using unsupervised descriptive models that take as input crowd sensor annotations and aggregate them to highlight geographic patterns that we refer as summaries. For any map, the mining models are composed of first glance high-level patterns (clusters of geographic annotations) that we name fingerprints. A fingerprint generates a thematic map prototype that summarizes a large amount of spatial annotations. Such a summary is beneficial for the end-user since it allows to better focus the attention on areas in which certain types of annotations are prominent while discarding many details that represent isolated annotations which may distract the user attention.

In extracting these thematic map prototypes, we are able to automatically infer the pattern evaluation parameters that allow the mining algorithms to work effectively on each annotation feature and discard the noise. Finally, we are able to combine the single dimension thematic map prototypes into more complete summaries solving the high-dimensional problem of combining the annotations of different categories in the same spatial area. Our approach works with any location-based data as input. In our experimental settings, we use the Foursquare[1] application since it provides a broad coverage both in terms of users and venues. We focus on the top 10 venue categories[2] (first level of the hierarchy), and we use them to build the feature vector of the proposed descriptive model. The model takes into account both the spatial proximity between venues as given by their geographic coordinates and the semantic feature proximity that is derived from the distribution of venue types created by the crowd sensors. We experiment using the research prototype developed by [3], refining the descriptive model, the logic for the parameter selections as described in this work, and providing a thorough experimental setup.

Fig. 1 shows the output of our approach for a geographic area covering the Milan municipality. The colors indicate different semantic types assigned to the clusters.

The reminder of the paper is organized as follows. In Section 2, we formalize the preprocessing stage meant to sample the input data and to generate the data structure used by the algorithm. Section 3 presents our proposed algorithm, while in Section 4, we report the statistically sound mechanism for the automatic parameter selections. We compare the output of our algorithm with a human manually created gold standard in Section 5, and we further validate the generated clusters over two corpora using two statistical validation tests in Section 6. We then describe prior works (Section 7) and we conclude outlining future research directions in Section 8.

2. Grid sampling and feature set

The input data of the summarization process is a set P of geographic points p, each characterized by a semantic feature that is usually listed in a taxonomy or controlled vocabulary associated with the dataset (e.g. the Foursquare taxonomy).[3] We represent the point p by the tuple $(lat, long, f)$, where the variables respectively represent the latitude, longitude and semantic feature, such as the category label used for classifying the venue according to a taxonomy.

We map P to a square-shaped spatial area named bounding box (*BBox*). Then, we split it into geographic subareas (also called cells) of uniform surface forming a regular grid. The number of cells depends on the dimension of the *BBox*. In order to have a statistical significance of the sampled set, the number of cells is greater than 100. Each cell of the grid is then described by the frequency of the categories that occur in the cell and is geographically represented by its focal point (or centroid). This aggregation of the observations occurring in each cell results in generating a set O of geographic objects o, each composed of:

lat: the latitude of the focal point of o;
long: the longitude of the focal point of o;

[1] http://foursquare.com.

[2] https://developer.foursquare.com/categorytree.

[3] https://developer.foursquare.com/categorytree.