# Geographic Summaries from Crowdsourced Data

Giuseppe Rizzo[1,2], Giacomo Falcone[1], Rosa Meo[1],
Ruggero G. Pensa[1], Raphaël Troncy[2], Vuk Milicic[2]

[1] Università di Torino, Turin, Italy
{giuseppe.rizzo,rosa.meo,ruggero.pensa}@di.unito.it
[2] EURECOM, Sophia Antipolis, France,
{raphael.troncy,vuk.milicic}@eurecom.fr

**Abstract.** In this paper, we present a research prototype for creating geographic summaries using the whereabouts of Foursquare users. Exploiting the density of the venue types in a particular region, the system adds a layer over any typical cartography geographic maps service, creating a first glance summary over the venues sampled from the Foursquare knowledge base. Each summary is represented by a convex hull. The shape is automatically computed according to the venue densities enclosed in the area. The summary is then labeled with the most prominent category or categories. The prominence is given by the observed venue category density. The prototype provides two outputs: a light-weight representation structured in GeoJSON, and a semantic description using the Open Annotation Ontology. We evaluate the quality of the summaries using the Sum of Squared Errors (SSE) and the Jaccard distance. The system is available at http://geosummly.eurecom.fr.

## 1 Introduction

Social media services 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. This data encompasses people' actions, dynamics of cities, so that it instantaneously reports any changes in the city topologies [6]. Such amount of data can therefore be considered as the new oil for geo-spatial platforms if globally taken. Leveraging on this massive amount of user whereabouts data coming from social media services, we present an approach that automatically adds a layer over the typical cartography geographic maps, creating summaries on what crowd sensors tell about venues and points of interest. Our approach grounds on using unsupervised descriptive models and exposing the results using geospatial data interchange formats that enable reuse on the Web. The prototype makes use of Foursquare, but any location service that exposes venues together with their categories can fit the model.

A few research attempts have been carried out to extract spatial and non-spatial properties that are typical for venues from social platforms. Among them, Tomko *et al.* [7] propose a method to calculate the descriptive prominence of venue categories that are sampled from OpenStreetMap[3] for a particular re-

[3] http://www.openstreetmap.org

gion. They select the most prominent categories for the inclusion in the region characteristic description. The descriptive prominence of a venue is computed using the concept of contrast from background. Meo *et al.* [4] propose a statistical approach to estimate the spatial characterization of an area considering the surroundings without imposing a priori knowledge on the geographic area characterization. An area is then marked depending on the statistical distribution of the observed features gathered from OpenStreetMap. Other research attempts, leveraging on social platforms such as Foursquare and Twitter, focused on spotting and labeling geographic regions according to the user activities ([2], [5]).

The remainder of this paper is organized as follows. The architecture overview of our approach is presented in Section 2. In Section 3, we detail our proposed demonstration, and in Section 4, we conclude and outline future work.

## 2 Architecture Overview

The prototype is composed of four main components, that we further explain in the following subsections. The source code is available at `https://github.com/giusepperizzo/geosummly` together with the API description.

### 2.1 Foursquare Sampling

The first stage consists in collecting the venues metadata from Foursquare. To perform such an operation, we receive as input either the bounding box ($BBox$) coordinates or a GeoJSON structure. A grid division is then applied. To have a statistical significance of the sampled set [8], we make sure to have a number of cells in the grid greater than 100. We also ensure to comply with the limited authorized rate access Foursquare has set in terms of the number of venues that can be retrieved for a given area[4]. For each cell, we collect the surrounded venues and the related metadata (such as the venue category or the number of check-ins). We then represent each cell as a vector, where the feature values ($f_i$) are the category occurrences. We end up having a matrix $NxM$ where $N$ corresponds to the number of cells and $M$ to the number of the categories used [5]). The matrix is then labeled with a timestamp. Hence, depending on the timestamp, we have different slices of the matrix.

### 2.2 Descriptive Models

The input of this stage is the matrix provided by the sampling component. We then consider the problem of computing geographic summaries as clustering geo-referenced objects in different 3-dimensional spaces: $latitude$, $longitude$, and $f_i$. We basically exploit the intrinsic spatial correlation of contiguous cells. For each of the obtained subspaces, we run DBSCAN [1]. As measure of distance among points, we consider the Euclidean distance, while $eps$ and $minPts$ are

[4] `https://developer.foursquare.com/overview/ratelimits`

[5] Depending on the setting, the prototype can make use of the first or the second level of the Foursquare taxonomy`https://developer.foursquare.com/categorytree`

statistically computed using the sampled observations in a particular area. This process brings a set of clusters for each subspace that are then merged according to the objective function. Each cluster is a set of contiguous cells (a region of the space) characterized by having similar distribution in a subset of the venue categories. As an additional side-effect, clusters are potentially overlapping. This approach is a follow-up of the SUBCLU [3] algorithm[6].

### 2.3 Publishing Geographic Summaries

A two-step strategy is proposed for publishing the results of the descriptive models component: *Open Annotation Ontology*[7] and *GeoJSON*[8]. Both strategies are equivalent in terms of the output entropy but they target different audiences, depending on the how the description is re-used. Let's define fingerprint as a cluster and geometry as the shape of the cluster.

*Open Annotation Ontology*: the fingerprint is described with various properties including a name (the dominant category or set of categories for this geometry), a dimension, the absolute number of venues, and the popularity (number of check-ins). The geometry is a polygon described using the GeoSPARQL vocabulary[9]. The annotation is itself identified in order to attach additional provenance information, such as the date when the geographic summary has been computed, described with the PROV[10] vocabulary. The data is available in a SPARQL endpoint at `http://geosummly.eurecom.fr/sparql`. A simple URI design policy has been devised with the three top level objects resulting in the following RDF graph:

```
<http://data.geosummly.eurecom.fr/annotation/UUID>
      a  oa:Annotation ;
      oa:hasTarget <http://data.geosummly.eurecom.fr/geometry/UUID> ;
      oa:hasBody <http://data.geosummly.eurecom.fr/fingerprint/UUID> ;
      prov:startedAtTime "2014-03-19T11:54:13.567Z"^^xsd:dateTime ;
      prov:wasAttributedTo <http://geosummly.eurecom.fr/> .
```

*GeoJSON*: the fingerprint is enclosed in a feature object where the geometry is represented using the MultiPoint class and the metadata is serialized as properties of the object together with the arrays of the enclosed venues.

### 2.4 Visualization

The visualization allows to browse the summaries generated for a spatial area, adding a layer over the typical cartography geographic maps. A zoom interaction enables to explore the venues enclosed in any cluster. This component can use either the GeoJSON or the RDF representations as described above. In addition, the states of different views are persistent through URLs that can be easily shared.

[6] The algorithm technical details are omitted, the focus of this paper being a demonstration.

[7] `http://www.w3.org/ns/oa#`

[8] `http://geojson.org/geojson-spec.html`

[9] `http://schemas.opengis.net/geosparql`

[10] `http://www.w3.org/ns/prov#`

## 3 Demonstration

This section illustrates the proposed framework in action with the geographic data sets released for the 2014 BigData Challenge[11]. Two data sets are used to demonstrate our prototype: *i)* Milan Grid [12] and *ii)* Trentino Grid [13]. Both areas are divided in cells of $d = 200m$, where $d$ is the edge of a squared cell, resulting in having $10K$ cells for Milan, and $33K$ cells for Trentino. The Foursquare sampling stage produced respectively $57,136$ and $21,796$ distinct venues. The probability distribution functions of the categories along the cells show major differences in the two data sets: we mainly observed a major drop in the venue distribution of the Trentino area according to the surface size, that has challenged the performance of our descriptive models algorithm. Figure 1 reports the geographic summary of the Milan extent.

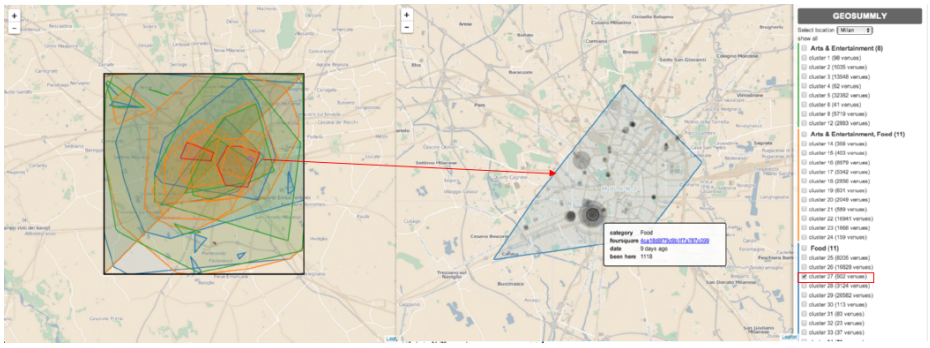


Fig. 1: First glance summary of the Milan extent at left, a zoom in a cluster on the right.

We perform a two-fold evaluation strategy and we report the results for the Milan extent[14]: *i)* a statistical validation where we measured the total Sum of Squared Errors ($SSE_{total}$) of both areas using the original data sets. We randomize them 500 times each, ensuring the same category density distributions. We measure the distance from the $SSE_{total}$ computed from the two grids and the $SSE_{total}$ of the randomly created data sets. The $SSE_{total}$ on the randomized data sets is 68.6637, while the $SSE_{total}$ obtained from the two grids is 2.2175. Hence, we can conservatively claim that there is less than 3% chance that the clusters occur by chance in the real data. *ii)* an output-based evaluation where we perform a 10-fold cross-validation on both data sets and for each fold, we randomly pick up half of the objects (hold-out). We end up with two sets for each fold that constitute two different views of the entire data set. We then compute the clusters from the respective views and we measure the overlap using

[11] `http://www.telecomitalia.com/tit/en/bigdatachallenge.html`
[12] coordinates: $(45.5677, 9.0114, 45.3566, 9.3126)$.
[13] coordinates: $(46.5363, 10.9143, 45.6727, 11.8312)$.
[14] The figures observed for Trentino are in the same order of magnitude (not included for space reason).

the Jaccard distance. We observe an average overlap of 81.30% that satisfies the 70% acceptance threshold.

## 4 Conclusion

This approach provides a first glance summary of a spatial area, exploiting user endeavors collected from Foursquare. To ease the reuse of the summaries on the Web, the prototype generates both a developer friendly (GeoJSON) output and a machine readable one using the Open Annotation ontology. The proposed prototype works on any geographic area from which Foursquare venues are available. As future work, we plan to integrate the categories from OpenStreetMap. We also plan to collect users' feedback for better tuning the descriptive models component. We finally plan to investigate more about the inclusion rate of two or more overlapping fingerprints, and the user zooming level that triggers different visualization scenarios.

## 5 Acknowledgments

This work was supported by the SMAT-F2 project funded by Regione Piemonte, the European Fund for the Regional Development (F.E.S.R.), and the EIT ICT-Labs 3cixty project.

## References

1. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: $2^{nd}$ International Conference on Knowledge Discovery and Data Mining (KDD'96) (1996)
2. Ferrari, L., Rosi, A., Mamei, M., Zambonelli, F.: Extracting Urban Patterns from Location-based Social Networks. In: $3^{rd}$ ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN '11) (2011)
3. Kailing, K., Kriegel, H.P., Kröger, P.: Density-connected subspace clustering for high-dimensional data. In: $4^{th}$ SIAM International Conference on Data Mining (SIAM'04) (2004)
4. Meo, R., Roglia, E., Bottino, A.: The Exploitation of Data from Remote and Human Sensors for Environment Monitoring in the SMAT Project. Sensors 12(12) (2012)
5. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In: ICWSM International Workshop on Social Mobile Web (SMW'11) (2011)
6. Phithakkitnukoon, S., Olivier, P.: Sensing Urban Social Geography Using Online Social Networking Data. In: $5^{th}$ International AAAI Conference on Weblogs and Social Media (ICWSM'11) (2011)
7. Tomko, M., Purves, R.S.: Venice, City of Canals: Characterizing Regions through Content Classification. Transactions in GIS 13(3) (2009)
8. Walpole, R., Myers, R.: Probability and statistics for engineers & scientists (Eighth Edition). Pearson Education International (2007)