Supporting Geo-Ontology Engineering Through Spatial Data Analytics

Gloria Re Calegari(✉), Emanuela Carlino, Irene Celino, and Diego Peroni

CEFRIEL – Politecnico of Milano, Via Fucini 2, 20133 Milano, Italy
{gloria.re,emanuela.carlino,irene.celino,diego.peroni}@cefriel.it

Abstract. Geo-ontologies are becoming first-class artifacts in spatial data management because of their ability to represent places and points of interest. Several general-purpose geo-ontologies are available and widely employed to describe spatial entities across the world. The cultural, contextual and geographic differences between locations, however, call for more specialized and spatially-customized geo-ontologies. In order to help ontology engineers in (re)engineering geo-ontologies, spatial data analytics can provide interesting insights on territorial characteristics, thus revealing peculiarities and diversities between places.

In this paper we propose a set of spatial analytics methods and tools to evaluate existing instances of a general-purpose geo-ontology within two distinct urban environments, in order to support ontology engineers in two tasks: (1) the identification of possible location-specific ontology restructuring activities, like specializations or extensions, and (2) the specification of new potential concepts to formalize neighborhood semantic models. We apply the proposed approach to datasets related to the cities of Milano and London extracted from LinkedGeoData, we present the experimental results and we discuss their value to assist geo-ontology engineering.

1 Introduction and Motivation

Imagine you live in the UK but you are currently on vacation in Italy for the first time. In the morning, when you are at home, you are used to buy newspapers and cigarettes in the small store around the corner; where can you get the same items now, since none of the shops in the vicinity of your B&B looks like a convenience store? Then, after having walked all day long throughout the old town, you finally reach the recommended restaurant in your tourist guide to taste Italian cuisine; unluckily it is closed and you seem to be in a residential area. You'd better find a district with a wide choice of eating places and – why not? – also at walking distance to nightlife venues; is there in town such a neighborhood, which is both rich in restaurants and clubs?

When we experience the environment surrounding us, we tend to elaborate what we see through the spatial categories we are used to; however, when we are in a less familiar place, we recognize that our conceptualizations do not perfectly apply and, even when they do, to better get oriented we make use of guides classifying for us the territory.

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 556–571, 2016.
DOI: 10.1007/978-3-319-34129-3_34

This semantic diversity has consequences on the conceptual models that knowledge engineers can build to describe spatial objects. When modeling a geo-ontology of urban points of interest, creating a general-purpose yet correct conceptualization is a challenge, because the ontology engineer has to decide the most suitable level of abstraction [1]. If the geo-ontology is generic, the risk is that it does not provide enough details to describe and distinguish between spatial objects with similar characteristics or functions in different places. On the other hand, if the engineering process produces a very rich and exhaustive conceptualization, some spatial objects' types could be too peculiar because of cultural diversity and apply only to specific geographic areas.

Many popular geospatial ontologies, adopted throughout different geographic regions, prefer to remain at a quite high level of abstraction, focusing on the basic concepts, the primitive notions [2]. A successful example is the LinkedGeoData ontology [3], which results from the ontologization of OpenStreetMap descriptive tags. How would it be possible to easily add location-specific extensions to a general-purpose geospatial ontology? Is there any analytics technique that could help in finding spatial concepts with a similar or diverging meaning at different locations? Would it be feasible to also characterize a territory by identifying its emerging semantic neighborhoods from the proximity analysis of different spatial object types?

Those are the questions that we address in our work, by proposing a set of spatial analytics techniques that, by processing the instances of a general-purpose geo-ontology, can provide hints to ontology engineers to support their modeling activities to create location-specific ontology extensions. The remainder of the paper is organized as follows: Sect. 2 describes the background of our approach in the context of related work; the main objectives and the experimental data set-up is described in Sect. 3; we detail the proposed approach, its empirical application and the obtained results in Sect. 4 at spatial feature level and in Sect. 5 at neighborhood level; finally, in Sect. 6 we offer our concluding remarks.

2 Background and Related Work

Geographic information usually describes locations as simple coordinates which are point-like, ubiquitous and precise. Beyond this straightforward interpretation of geo-information as geographic coordinates, there is the concept of place, which is the human way to understand and refer to space. Places are not point-like and have fuzzy boundaries determined by physical, cultural, and cognitive processes, such as the concept of "downtown" [4]. The identity and meaning of places cannot be captured considering only the spatial component of data, for which the semantic aspect is essential.

In this sense the analysis of geographic information in terms of its semantics is crucial and, nowadays, the modelling of geo-ontologies is becoming a great challenge [2]. Since geographic information spans across different domains, geo-ontology engineering needs to move from the top-down development of a small number of global ontologies, to the creation of a higher number of local ontologies that reflect location-specific perspectives and are developed in a bottom-up fashion [5].

An important aspects in ontology building is the choice of the level of abstraction: ontologies can contain both top-level concepts that apply across many or all domains and bottom-level concepts that apply only within a specific domain. The same holds for geospatial datasets, which can be semantically described at various levels to convey their meaning [1]. Since geo-ontologies are strictly related to the context, domain-specific concepts can mean situated concepts, i.e. conceptualizations dependent on specific processes (natural, social, scientific, or possibly machine) and whose instances are entities within a specific spatio-temporal context [6].

In our experiments we extract data from LinkedGeoData [3], which is described with a general-purpose ontology, and we aim at finding possible location-specific specializations or extensions to its ontology. LinkedGeoData uses the information collected by the OpenStreetMap project with the aim of providing a rich integrated and interlinked geographic dataset for the Semantic Web.

Since OpenStreetMap is a prominent example of volunteered geographic information (VGI) [7], LinkedGeoData knowledge reflects the way in which the environment is experienced [8]. This peculiarity implies that this type of information can be effectively employed to characterize a specific area not only in terms of geolocation [9]. LinkedGeoData information can highlight whether differences between regions exist [10] and can be useful to discover semantic dissimilarities of point features [11].

In this paper, starting from the LinkedGeoData ontology, we use a set of spatial analytics methods on OpenStreetMap data to check how to assist some steps of geo-ontology engineering [12], like evolution, repair and specialization.

The combination of the ontology, which represents the semantics of the data, and the sheer spatial analysis, which considers only the geographic coordinates, has been used for various purposes: to guide the choice of suitable data and clustering method for the task of locating shopping malls [13]; to characterize citizens behavior through location-based social network [14]; to create geographic summaries using social media [15]; to extract urban land use and support smart cities planning activities [16].

In the following, we use this combination of geo-ontologies and spatial analytics to provide a semantic characterization of the territory, in order to give citizens and visitors a thorough knowledge of a region in terms of different semantics areas.

3 Objectives, Data Preparation and Assumptions

In this paper, we present our approach to analyze the instances of a general-purpose geospatial ontology in two different urban environments, Milano and London. We apply a set of spatial analytics methods to derive helpful insights on their “urban semantics” to support the ontology engineers’ re-engineering efforts on that ontology, in the specialization and extension phases (as defined in [12]). The two main objectives of our work are:

O1. Identifying concepts that play a different role in the two cities, thus possibly indicating different cultural or pragmatic meanings; we address this goal, by analyzing the pattern distribution of each spatial feature.

O2. Highlighting new potential concepts to characterize the urban neighborhoods of the two cities; we address this goal by analyzing the co-occurring aggregations of different spatial features.

Our approach is inspired by the observation-driven framework proposed in [5], with the following differences: we do not aim to build a new ontology but to identify improvement points in a pre-existing conceptualization; we base our analysis on volunteered geographic information instead of sensor data; we focus on spatial primitives, touching geo-ontology design patterns only to a limited extent with the neighborhoods characterization.

We apply the approach described in the following Sects. 4 and 5 to a dataset derived from LinkedGeoData [3]. More specifically, our spatial objects are individuals in LinkedGeoData; we focus our investigation on the instances of the `lgdo:Amenity` concept and its sub-classes (shops, restaurants, hotels, offices, etc.). Each spatial object is described in terms of a semantic feature – its LinkedGeoData ontology class, which represent the place's category – and a spatial characterization; regarding the latter, in case of a point (named "node" in OpenStreetMap terminology) we take its latitude-longitude pair, in case of a polygon ("way") we compute its centroid and consider the coordinate pair of that centroid.

The experimental dataset consists of the spatial objects included in two reference zones: the Milano municipality border (around 200 km^2) and the Central London sub-region, the innermost boroughs of the UK capital city (around 130 km^2). The experiments illustrated in this paper are based on a dataset extracted at the end of October 2015; this means that the Milano objects include also the pavilions of the World Exposition in the EXPO 2015 area. In total, we collected around 13,000 spatial objects in Milano and 30,000 in London; those objects are instances of around 180 LinkedGeoData ontology classes (our spatial features).

The assumptions we make on the considered dataset are as follows. LinkedGeoData is derived from OpenStreetMap and OpenStreetMap is an open, collaborative bottom-up effort for collecting this large-scale spatial knowledge base. Therefore, the data cannot be expected to be complete and it is hard to give an estimate of its actual coverage [17]; nonetheless, since the data is the result of manual annotation, we can consider that OpenStreetMap volunteer editors add the most relevant and characterizing spatial objects, which are indeed the features that we wish to analyze to derive some insights on the urban space. Still, the mapping can be inhomogeneous (some zones can be more detailed annotated than others). Since we decided to focus on Milano and London, however, we can discard this potential issue: our direct knowledge of the city of Milano let us affirm that the spatial objects mapping is quite good and homogeneous throughout the city; OpenStreetMap coverage in the London area was evaluated in [18] and shown to be quite accurate in comparison to official sources. Furthermore, according to global OpenStreetMap statistics[1], Italy and UK are ranked 7th and 10th for number of created spatial objects, and 4th and 5th for density of created spatial objects per square kilometer. More details and further experimental results are available at http://swa.cefriel.it/geo/eswc2016.html.

4 Objective 1: Re-Engineering Spatial Features

If a particular spatial feature is condensed in a specific region of a larger area, that feature can be considered as a relevant element to characterize that region. Starting from this point of view, the first step of our experiments consists in conducting a spatial analytics exploration to find which spatial features can be considered prominent to tipify our reference areas and to verify if the same concept plays a different role in the different places. This is the first kind of support we can provide to geo-ontology engineers.

To reach this goal, we resort to two different statistical analysis: the analysis of the spatial patterns and aggregations and the points density analysis of each spatial feature. Naturally, the two analyses move into the same direction and are helpful to discover similarities and dissimilarities between different areas.

4.1 Analyzing Spatial Objects' Distribution

Regarding the first statistical analysis, we adopt some spatial analytics metrics to evaluate the tendency of each spatial feature to show spatial patterns and to create aggregations. The indicators we choose are the Morishita index and the Moran index.

The Morishita index [19] is a statistical measure of dispersion; we use it to detect spatial point patterns based on quadrat counts: the geographic area is divided into a regular grid with cells of equal size and shape and the numbers of points falling in each cell are counted and compared. If the point pattern is completely random, the index should be approximately equal to 1; values greater than 1 suggest that a spatial aggregation exists. The trend of the Morishita diagram and the value of the index indicate if a spatial aggregation can occur. Some examples of Morishita diagrams are reported on the companion website.

The other indicator is the Moran index [20], which is a measure of spatial autocorrelation: this index values range from –1 (perfect dispersion) to +1 (perfect correlation); a zero value indicates a random spatial pattern. The Moran and the Morishita indexes can be jointly used as indicators of spatial autocorrelation and aggregation.

We compute the Morishita and the Moran indexes for all spatial features, i.e. for all selected LinkedGeoData classes. The instances of some spatial features, like `lgdo:FuelStation`, `lgdo:Supermarket`, `lgdo:Police` – which are usually

[1] Cf. http://osmstats.neis-one.org/?item=countries.

homogeneously spread throughout a city, without specific aggregations – present low values for both the Morishita and the Moran Index, while those features that are normally grouped in specific areas, like `lgdo:Clothes` or `lgdo:Cafe` in a big touristic city, show higher values for both indicators. These considerations, although simple and immediate, confirm the reliability of the two adopted indexes, that we use to select the features on which we apply density-based clustering, as explained in the following section.

4.2 Clustering Spatial Objects

In the second step, we start from the consideration that through clustering we can highlight how spatial objects, described only by their latitude-longitude coordinates, form agglomerations in a specific region. Density-based clustering methods [21] are used to this end: they find clusters based on the density of points in regions and they are able to identify clusters of arbitrary shapes. The key idea of density-based clustering is that, for each object of a cluster, its neighborhood within a given radius ϵ has to contain at least a minimum number *minpts* of other objects, i.e. the cardinality of the object's neighborhood has to exceed a threshold.

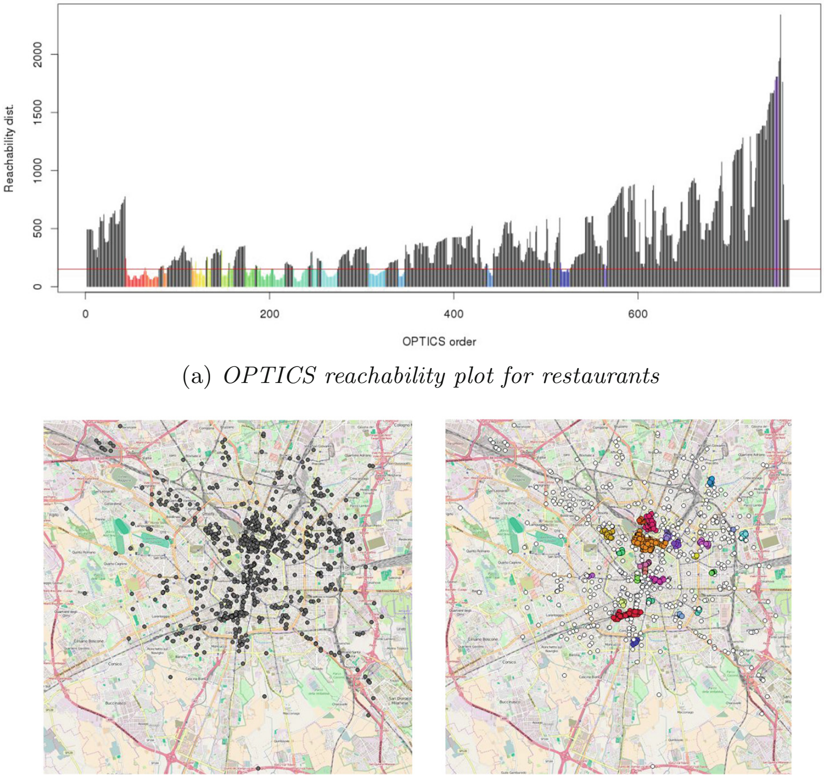
We decided to adopt the OPTICS algorithm [22], which is the extended hierarchical version of the more famous DBSCAN method [21] to detect clusters with different densities. Both methods group points that are closely packed together and mark as outliers those points that lay in low-density regions. We prefer to adopt OPTICS because, since it is a hierarchical clustering, it is possible to "cut" its reachability plot at a given threshold ϵ_{cut}, thus revealing clusters with the same density.

Figure 1a shows the reachability plot of the instances of `lgdo:Restaurant` in Milano (shown in Fig. 1b); cutting the plot at ϵ_{cut} (represented by the red horizontal line) discriminates between the restaurant clusters and the more isolated points. Figure 1c shows the restaurant clusters as colored points on the map, while the white spots represents the outliers that are not clustered (i.e. less than *minpts* restaurants in a radius of ϵ_{cut}).

To apply the OPTICS algorithm, it is necessary to set the three parameters, ϵ, *minpts* and ϵ_{cut}. The first two parameters are used to compute the distances between each point and its nearest one, thus building the reachability plot; than the ϵ_{cut} parameter defines the desired level of clusters density.

We decide to perform OPTICS clustering on each spatial feature separately (i.e., `lgdo:Restaurant`, `lgdo:Pub`, `lgdo:Clothes`, etc.), in order to identify areas characterized by a high density of a specific class. To ensure effective results comparability, we decide to adopt the same parameters' values for all spatial features and for all locations.

Since we are working with points characterized only by their spatial component (the latitude-longitude pair), we choose parameters with a precise physical meaning: we consider that points instances of a specific spatial feature constitute a cluster if there are at least 5 points of the same class in a radius of 150 m (i.e. $minpts = 5$ and $\epsilon_{cut} = 150$). Imposing this constraint implies that not all the



(a) *OPTICS reachability plot for restaurants*

(b) *All restaurants*

(c) *Clustered restaurants*

Fig. 1. Results of OPTICS clustering applied on instances of `lgdo:Restaurant` in Milano, with $minpts = 5$ and $\epsilon_{cut} = 150\,\text{m}$. (Color figure online)

considered spatial features result in clusters. This circumstance is useful to better understand the semantic characterization of each single place and to compare the two environments finding any dissimilarities.

Analyzing the results deriving from this clustering step, London and Milano outcomes are similar for some spatial features, but appear considerably different for other ones. In terms of spatial features which produce at least one cluster, the difference between the two city is remarkable: in London 56 different amenity types are sufficiently dense in at least one region, while in Milano only 25. Results on `lgdo:Hotel` amenities clustering highlight the difference of the considered areas in the two cities: in London we take only the innermost boroughs, characterized by a strongly touristic component, and OPTICS clustering reflects this consideration: 40 % of points in this category were clustered; conversely, in Milano we consider both central and peripheral regions, so the share of clustered hotels is much lower (14 %).

To highlight the possible existence of different semantic interpretations of the same concept in the two cities, let's consider two classes: `lgdo:Telephone` and `lgdo:Pub`. The red telephone box in London is a hallmark, while in Milano it is simply an installation useful for citizens located in the main places of displacement. This distinction is remarked by our clustering results: in London telephone clusters are discovered and mostly in the touristic areas; in Milano the OPTICS algorithm does not identify any cluster, because points are spread all over the city.

Pub is another concept with a different meaning for London and Milano communities: in UK pubs are widespread around a city, because people go there to eat and drink both at lunch- and dinner-time; in Italy people usually go to pubs after dinner and only to have a drink, so they are concentrated in popular nightlife areas. In London only 7 % of points are clustered, while in Milano this percentage reaches 20 %.

4.3 Support to Ontology Engineers

The instance analysis of an existing conceptualization can provide useful suggestions to the ontology engineer, who would like to evaluate the validity and applicability of a general geo-ontology in different locations. The approach outlined above with spatial distribution and spatial density analysis already ends up with some hints.

If a spatial feature corresponding to an ontological concept displays the same behavior in different places, this means that the concept is valid throughout the different territories and does not require any ontology re-engineering intervention (e.g. `lgdo:Restaurant` or `lgdo:Clothes` in our analysis of Milano and London).

On the other hand, whenever a spatial feature shows different values of the Morishita and Moran indicators and/or different tendency to form clusters, that sign can indicate the need to intervene: maybe location-specific extensions to the geo-ontology are required to provide a more precise conceptualization. For example, the `lgdo:Pub` concept could be split to take into account the two different meanings that spatial feature has in UK and Italy; similarly, the `lgdo:Telephone` concept could become a sub-class of `lgdo:TourismThing` in a London-specific ontology extension.

It is worth noting that also a pure numerical analysis of the spatial objects can provide interesting insights: it could reveal concepts whose instances in a place are "outliers" with regards to the instances of the same concept in other areas. This is for example the case of `lgdo:Convenience` in Milano: only a few instances of this class are present (instead of the 750 in central London); indeed, in Italy this type of shop is usually simply considered a supermarket, while cigarettes are only sold by tobacconists. A pure numerical outlier feature could further be investigated through the above spatial analysis: indeed `lgdo:Convenience` instances have a zero Moran index (which means random spatial pattern) and do not cluster in Milano, while they have a Moran index of 0.39 and form 18 clusters in London.

5 Objective 2: Specifying Spatial Neighborhoods

After identifying the features that best characterize the two cities, and those that exhibit different meanings in different places, our second goal is to answer the following questions: which are the spatial features that occur together? Is it possible to semantically characterize a region according to its semantic features co-occurence? The second step of our experiments is organized into three substeps, characterized by different, but strongly interrelated, statistical analysis.

Starting with the spatial feature clusters computed for each category (resulting from the approach illustrated in Sect. 4), first we analyze their spatial co-occurence using again clustering techniques to discover emerging neighborhoods. Then, we further investigate the feature composition of those neighborhoods in terms of a new spatial indicator, inspired by the popular *tf-idf* score, that we introduce to identify the amenities that better define the urban space. Lastly, we characterize neighborhoods with new district-specific concepts, representing shopping, cultural or residential areas, by building spatio-semantic "queries" that incorporate spatial features' co-occurrence.

5.1 Identifying Neighborhoods

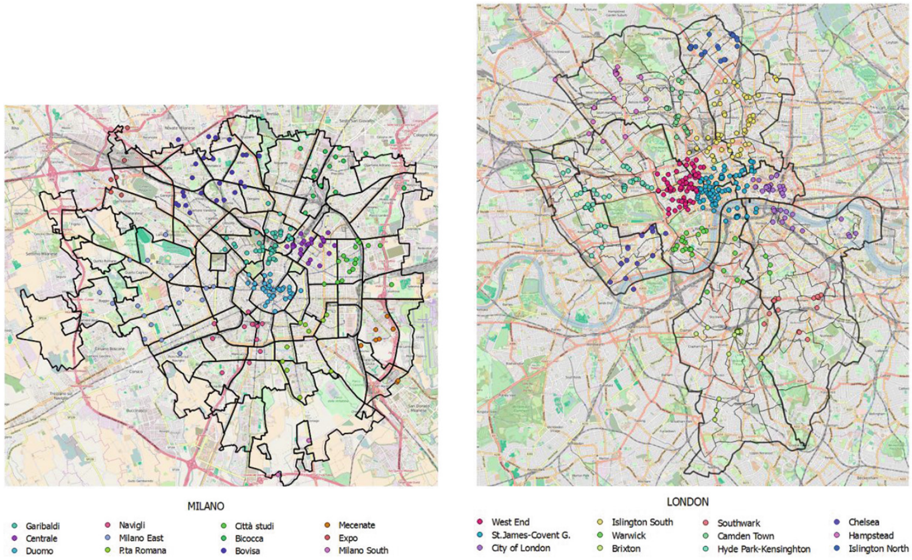
For each cluster of spatial objects obtained with the OPTICS algorithm, we define the convex hull polygon containing all the points belonging to that cluster and we compute its centroid. In this way we obtain a set of points representing all the categories' hotspot locations, which are urban areas with a high number of amenity of a given type.

Then we apply a hierarchical agglomerative clustering [23] technique (with Ward minimum variance method) on these centroids with the aim of dividing the urban space in neighborhoods, according to the simultaneous presence of spatial features. In this case, we do not use a density-based clustering algorithm because the centroids are artificially-created points that actually represent sets of points, hence the centroid density is not a meaningful indicator; consequently, a pure hierarchical clustering is a better fit for our goal. Using this technique, we obtain 12 clusters for both Milano and London, as shown in Fig. 2.

By comparing the obtained clusters with the official boundaries of the two cities' boroughs/districts, we discovered some interesting matches. In Milano (cf. Fig. 2a) we can identify, in the city centre, Duomo area, Centrale and Garibaldi railway stations and the area around Navigli and, in the outer parts, the university area – Città studi –, the most important industrial sites – Bicocca and Bovisa – and the area that hosted EXPO 2015. In London (cf. Fig. 2b) some famous boroughs are highlighted too: the town centre consisting in the City of London, West End, St. James and Covent Garden; the characteristic districts of Camden Town, Hyde Park, Kensington and Chelsea.

5.2 Characterizing Neighborhoods

Starting from the neighborhoods resulting from the latter clustering on both cities, we further investigate our clustering results to discover which spatial features best



(a) *Milano* (b) *London*

Fig. 2. Centroid clustering to identify emerging neighborhoods. (Color figure online)

characterize each city region. To reach this goal, we define a spatial-specific version of the popular *tf-idf* score widely used in information retrieval and text mining. The classical version of this indicator is a numerical statistics used to highlight how important a word is to a document in a collection or corpus, taking into account the number of times a word appears in the document, weighted by its frequency in the corpus (*tf* part) and penalized if it appears very frequently in all documents (*idf* part).

Referring to our experiments, we want to highlight how important a spatial feature is to a specific neighborhood. Therefore, similarly to [24], for each neighborhood and for each feature we define the two components of our index as follows. The spatial object frequency *sof* of a spatial feature in a neighborhood is defined as:

$$sof = \frac{|n \cap f|}{|n|}$$

where n is the set of clustered points in the neighborhood and f is the set of all spatial points with that feature. Similarly, we define the inverse neighborhood frequency *inf* of a spatial feature as:

$$inf = 1 + log\frac{|N|}{|\{n : n \cap f \neq \emptyset\}|}$$

where $|N|$ is the total number of neighborhoods (24 in our experiments, because we have 12 districts in each city) and the denominator is the number of neighborhoods in which the spatial feature is represented by at least a point.

Table 1. Top three *sof-inf* scores in each neighborhood

Garibaldi	Centrale	Duomo	Navigli	MiEast	Pt.Romana	CittàStudi	Bicocca	Bovisa	Mecenate	Expo	MiSouth
office	hotel	shoes	pub	school	bank	university	indust.	indust.	indust.	fast_food	indust.
(1.28)	(3.30)	(4.18)	(4.18)	(4.18)	(0.37)	(2.90)	(0.63)	(1.48)	(0.34)	(1.39)	(0.05)
bar	bar	clothes	bar	parking	parking	bicycle_p	office	office	office	indust.	parking
(1.16)	(1.04)	(2.78)	(0.49)	(0.46)	(0.15)	(0.60)	(0.50)	(1.06)	(0.20)	(0.25)	(0.03)
restaur.	bank	bank	bank	office	office	bar	univ.	bicycle_p	bicycle_p	office	/
(0.71)	(0.70)	(1.90)	(0.41)	(0.25)	(0.15)	(0.49)	(0.24)	(0.31)	(0.06)	(0.08)	

(a) *Milano*

WestEnd	St.James	City	Isl.South	Warwick	Brixton	Southwark	Camden	HydePark	Chelsea	Hampstead	Isl.North
shoes	theatre	pub	school	hotel	greengr.	atm	charity	antiques	embassy	school	comm_c
(2.18)	(4.18)	(1.31)	(2.32)	(0.59)	(3.49)	(0.96)	(4.18)	(2.80)	(2.63)	(1.85)	(3.08)
art	musical	office	indust.	indust.	housew.	university	tattoo	hotel	shoes	parking	conven.
(2.15)	(2.39)	(1.01)	(2.26)	(0.58)	(2.57)	(0.70)	(2.24)	(2.40)	(0.86)	(0.25)	(0.65)
clothes	pub	bank	conven.	atm	butcher	bicycle_p	bar	conven.	clothes	conven.	/
(2.00)	(2.40)	(0.99)	(1.12)	(0.44)	(1.13)	(0.28)	(0.57)	(0.90)	(0.53)	(0.14)	

(b) *London*

Finally, we obtain our spatial object frequency–inverse neighborhood frequency index *sof-inf* by multiplying the two components.

To analyze neighborhoods composition in terms of their most relevant spatial features, we sort the *sof-inf* scores of all features in a neighborhood, thus ranking the distinctive amenity categories. Table 1a and b show Milano and London results respectively.

Considering Milano, we can offer our considerations based also on our direct knowledge of the environment. As highlighted previously, Bicocca and Bovisa are the most important industrial sites and this is confirmed by the highest *sof-inf* scores of the `lgdo:Industrial` spatial feature. The Garibaldi, Centrale and Duomo districts present higher values for those features strongly related to tourists: `lgdo:Bar`, `lgdo:Restaurant`, `lgdo:Clothes`, `lgdo:Hotel`. Navigli, the area known for its dynamic nightlife, is the only neighborhood in which the `lgdo:Pub` spatial feature achieves a considerable *sof-inf* score. The last consideration applies to the university area Città-Studi: the *sof-inf* analysis confirms the youthful look of the area, which is dense of typical structures related to student lifestyle (`lgdo:University`, `lgdo:BicycleParking` and `lgdo:Bar`). It is evident that there is a clear difference between the central areas, characterized by features related to shopping and to the daily/night life, and peripheral region where features as industrial and office are the most frequent ones.

Similar considerations can be made on London, even if only the very centre of the city is explored. West End and St. James are the most touristic areas (higher values for `lgdo:Shoes`, `lgdo:Clothes`, `lgdo:Art` and `lgdo:Theatre`). The City of London business area is marked by clusters of `lgdo:Office` and `lgdo:Bank` instances. The *sof-inf* analysis also highlights two peculiar areas, that we cannot find in Milano: Camden Town and Chelsea. The former area is the only one characterized by an "alternative" category like `lgdo:Tattoo`: it actually is the best known area in London for its "quirks". The latter area is one of the most prestigious, which is distinguished by presence of `lgdo:Embassy` clusters, i.e. structures normally situated in the most rich regions of a city.

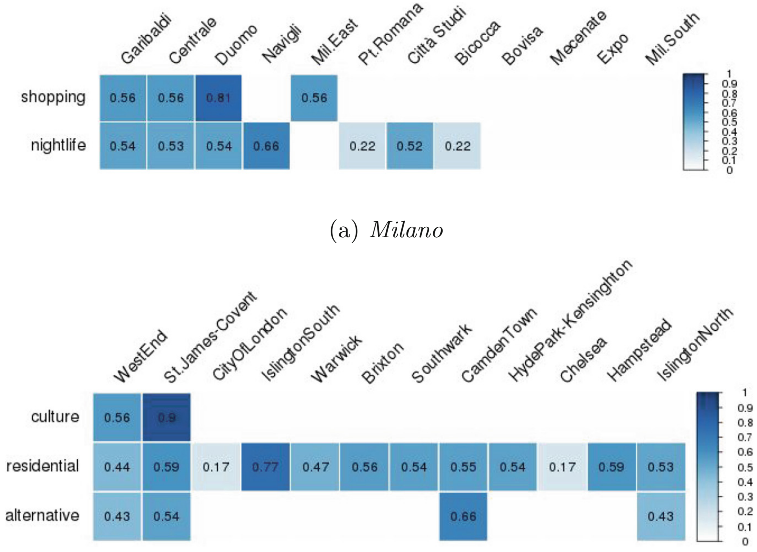
5.3 Semantically Querying Neighborhoods

The third analysis is aimed to create neighborhood concepts that express the co-occurrence of different spatial features in the same area. To this end, we define combinations of LinkedGeoData amenities that we expect to characterize a specific class of districts, as follows (we omit the `lgdo:` prefix for simplicity):

Shopping: `Clothes`, `Shoes`, `Chemist` and `BeautyShop`
Nightlife: `Restaurant`, `Pub`, `Bar` and `Theatre`
Residential: `School`, `Atm`, `Butcher`, `Greengrocer`, `Convenience` and `Parking`
Culture: `Theatre`, `ArtShop`, `BookShop`, `MusicalInstruments` and `College`
Alternative: `Erotic`, `Tattoo`, `Charity`, `CommunityCentre` and `ArtShop`

In a sense, we aim to define neighborhood types with an ontology design pattern [25] that implies the simultaneous presence of a set of spatial features.

Those combinations of spatial features can be considered our *spatio-semantic queries*, similarly to a list of terms to be retrieved in a document corpus. We compute the *sof-inf* score defined in the previous section also for each query; then we compute the cosine similarity between the query *sof-inf* vector and each neighborhood *sof-inf* vector, thus evaluating the match between the classes of districts and the urban neighborhoods. The intrinsic difference of the two cities is reflected in our semantic queries' results, as shown in Fig. 3.



(a) *Milano*

(b) *London*

Fig. 3. Similarity scores between *sof-inf* query and neighborhood vectors.

In Milano it is reasonable that the *shopping* query presents the highest match (0.81) within the Duomo neighborhood, the principal touristic area. Milano *nightlife* is mainly concentrated in the area around Navigli (0.66), which is known as a lively area full of restaurants and pubs. The similarity scores of the same queries get uniform values in London across all districts, indicating that shopping and nightlife are more equally spread; it is worth reminding that in our experiment we analyze only the central boroughs of London, while in Milano area the suburbs are also included.

The cosmopolitan character of London is reflected by the existence of two neighborhoods strongly connoted as *cultural* and *alternative*: respectively, St. James–Covent Garden area (0.90), full of theatres, museums and art centres, and Camden Town (0.66), famous for its peculiarity and eccentric culture. As regards *residential* areas, it is interesting to focus on the districts with the lowest similarity scores: the City of London, the business core, and Chelsea, the most exclusive and prestigious area.

Our method would work on any other city and can prove effective also to compare similar cities; an example to compare Milano and London is reported at http://swa.cefriel.it/geo/eswc2016.html.

5.4 Support to Ontology Engineers

This second set of spatial analytics techniques is oriented to provide ontology engineers with some evidence of emerging neighborhood conceptualizations; a possible result then could be the specification of a new set of concepts to synthesize the "semantics" of urban districts.

While we use the clustering of spatial feature clusters only as a means to split the area of interest in regions with a specific characterization (as opposed to official and administrative boundaries), the definition of the *sof-inf* score, inspired by information retrieval practice, is an important step to allow a knowledge engineer to select the most prominent spatial features that characterize a neighborhood, similarly to [9]. This technique can help in effectively describing spatial regions by summarizing their distinctive categories.

Finally, the semantic query approach presented in Sect. 5.3 is the method we propose to support the ontological specification of neighborhood concepts as a design pattern combining spatial features: adopting this technique, not only it is possible to verify the actual "instantiation" of such concepts in different geographic areas, but the ontology engineer can also test different hypotheses, computing the similarity scores for different feature compositions. A possible extension to the proposed procedure can introduce spatial features' weights in the *sof-inf* computation of the spatio-semantic query.

This approach can prove useful also when the target neighborhood ontology must fit a specific level of abstraction: if the conceptualization is expected to be general-purpose, the semantic query should get homogeneous similarity scores in different geographical areas; conversely, if the new concepts are location-specific, the combination of spatial features can be selected based on the maximization of its similarity score with the desired regions.

6 Conclusions

Even when applying proper methodologies, ontology engineering largely remains an art that requires a deep domain knowledge. In the case of geo-ontologies, the understanding of location-specific knowledge is key; whatever the kind of spatial objects to be described, their geographic distribution can be investigated through spatial analytics to gain hints and suggestions to support the ontology engineering of their thematic characterization.

In this paper we employed a set of state-of-the-art data analytics techniques to study and compare geospatial objects from LinkedGeoData and provide additional insights to guide and support the (re-)engineering of the respective ontology. We showed how the plain analysis of objects coordinates can reveal cultural and location-specific differences between different cities. Our contribution also included the definition of a spatial variant of the tf-idf index, to illustrate how the combined analysis of semantic and spatial information can support the specification of neighborhood characterization.

The adopted methods are generic enough to be applied to different spatial datasets, since our experiments demonstrated that the diverse number of spatial objects and the possible dataset incompleteness do not negatively impact their analysis. Nonetheless, we will apply the proposed approach to a larger set of heterogeneous cities to further test, refine and select the best spatial metrics and indicators to reveal location-specific semantics. We would like to further investigate the generality and applicability of our techniques, by analyzing different types of spatial entities, possibly also within non-urban contexts.

The main limitation of our approach is that we do not provide any automated means to geo-ontology re-engineering, because our approach's findings bring only supporting insights. In addition, our analyses can confirm already-known characteristics, or they can discover unknown specificities that are hard to interpret and that require further investigation by the engineer. The natural next step is therefore to more tightly integrate spatial analytics within ontology engineering processes and tools.

Acknowledgments. This work was supported by the PROACTIVE project (id 40723101) co-funded by Regione Lombardia (POR-FESR 2007–2013).

References

1. Frank, A.U.: Chapter 2: ontology for spatio-temporal databases. In: Sellis, T.K., et al. (eds.) Spatio-Temporal Databases. LNCS, vol. 2520, pp. 9–77. Springer, Heidelberg (2003)
2. Janowicz, K., Scheider, S., Pehle, T., Hart, G.: Geospatial semantics and linked spatiotemporal data-past, present, and future. Semant. Web J. **3**(4), 321–332 (2012)

3. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: Linkedgeodata: a core for a web of spatial open data. Semant. Web **3**(4), 333–354 (2012)
4. Montello, D.R., Goodchild, M.F., Gottsegen, J., Fohl, P.: Where's downtown?: behavioral methods for determining referents of vague spatial queries. Spat. Cogn. Comput. **3**(2–3), 185–204 (2003)
5. Janowicz, K.: Observation-driven geo-ontology engineering. Trans. GIS **16**(3), 351–374 (2012)
6. Brodaric, B., Gahegan, M.: Experiments to examine the situated nature of geoscientific concepts. Spat. Cogn. Comput. **7**(1), 61–95 (2007)
7. Goodchild, M.: Citizens as sensors: the world of volunteered geography. GeoJournal **69**, 211–221 (2007)
8. Brodaric, B.: Geo-pragmatics for the geospatial semantic web. Trans. GIS **11**(3), 453–477 (2007)
9. Tomko, M., Purves, R.S.: Venice, city of canals: characterizing regions through content classification. Trans. GIS **13**(3), 295–314 (2009)
10. Mooney, P., Corcoran, P.: The annotation process in OpenStreetMap. Trans. GIS **16**(4), 561–579 (2012)
11. Mülligann, C., Janowicz, K., Ye, M., Lee, W.-C.: Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In: Egenhofer, M., Giudice, N., Moratz, R., Worboys, M. (eds.) COSIT 2011. LNCS, vol. 6899, pp. 350–370. Springer, Heidelberg (2011)
12. Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A.: Ontology Engineering in a Networked World. Springer, Heidelberg (2012)
13. Wang, X., Hamilton, H.J.: Towards an ontology-based spatial clustering framework. In: Kégl, B., Lee, H.-H. (eds.) Canadian AI 2005. LNCS (LNAI), vol. 3501, pp. 205–216. Springer, Heidelberg (2005)
14. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In: The Social Mobile Web (2011)
15. Rizzo, G., Falcone, G., Meo, R., Pensa, R.G., Troncy, R., Milicic, V.: Geographic summaries from crowdsourced data. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC Satellite Events 2014. LNCS, vol. 8798, pp. 477–482. Springer, Heidelberg (2014)
16. Calegari, R.G., Carlino, E., Peroni, D., Celino, I.: Extracting urban land use from linked open geospatial data. ISPRS Int. J. Geo-Inf. **4**(4), 2109–2130 (2015)
17. Mooney, P., Corcoran, P., Winstanley, A.C.: Towards quality metrics for OpenStreetMap. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 514–517. ACM (2010)
18. Haklay, M., et al.: How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. Environ. Plann. B Plan. Des. **37**(4), 682 (2010)
19. Morisita, M.: Measuring of the dispersion of individuals and analysis of the distributional patterns. Mem. Fac. Sci. Kyushu Univ. Ser. E **2**(21), 5–235 (1959)
20. Gittleman, J.L., Kot, M.: Adaptation: statistics and a null model for estimating phylogenetic effects. Syst. Biol. **39**(3), 227–241 (1990)
21. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, pp. 226–231 (1996)
22. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: Optics: ordering points to identify the clustering structure. In: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM (1999)

23. Rokach, L., Maimon, O.: Clustering methods. In: Maimon, L., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 321–352. Springer, Heidelberg (2005)
24. Walker, A.R., Moody, M.P., Pham, B.L.: A spatial similarity ranking framework for spatial metadata retrieval (2006)
25. Gangemi, A., Presutti, V.: Ontology design patterns. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 221–243. Springer, Heidelberg (2009)