What you are is When you are: The Temporal Dimension of Feature Types in Location-based Social Networks

Mao Ye
Department of Computer Science and Engineering
Pennsylvania State University, USA
mxy177@cse.psu.edu

Krzysztof Janowicz
Department of Geography, University of California, Santa Barbara USA
jano@geog.ucsb.edu

Wang-Chien Lee
Department of Computer Science and Engineering
Pennsylvania State University, USA
wlee@cse.psu.edu

Christoph Mülligann
Institute for Geoinformatics, University of Münster, Germany
cmuelligann@uni-muenster.de

ABSTRACT

Feature types play a crucial role in understanding and analyzing geographic information. Usually, these types are defined, standardized, and controlled by domain experts and cover geographic features on the mesoscale level, e.g., populated places, forests, or lakes. While feature types also underlie most Location-Based Services (LBS), assigning a consistent typing schema for Points Of Interest (POI) across different data sets is challenging. In case of Volunteered Geographic Information (VGI), types are assigned as *tags* by a heterogeneous community with different backgrounds and applications in mind. Consequently, VGI research is shifting away from data completeness and positional accuracy as quality measures towards attribute accuracy. As tags can be assigned by everybody and have no formal or stable definition, we propose to study category tags via indirect observations. We extract user check-ins from massive real-world data crawled from Location-based Social Networks to understand the temporal dimension of Points Of Interest. While users may assign different category tags to places, we argue that their temporal characteristics, e.g., opening times, will show distinguishable patterns.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial databases and GIS; H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.5.3 [**Group and Organization Interfaces**]: Web-based interaction

General Terms

Theory, Measurement, Standardization

[Final Draft]

Keywords

Social Networks, Temporal Data, Feature Types, Semantics

1. INTRODUCTION AND MOTIVATION

As argued by Harnad cognition is categorization [15]. Our behavior towards different types of entities varies based on what they *afford* [12]. Categorization itself does not require language, but assigning terms to categories plays a major role in communication. However, as we cannot compare categories in our minds directly and also assign different labels to them, our understanding of terms differs [19]. While addressed in everyday life by situated simulation [4], defining the meaning of domain vocabulary is among the main challenges for interoperability, information re-usage and retrieval, as well as recommendation systems.

Ontologies specified using formal knowledge representation languages are a promising approach to restrict the interpretation of terms towards their intended meaning and have a long tradition in GIScience [22, 11, 18]. While several methodologies and tools have been introduced over the last years, ontology engineering is a difficult task. It requires a common agreement among domain experts and a deep understanding of the logics-based knowledge representation languages [10]. INSPIRE[1], for instance, provides an interesting example for the difficulties arising when trying to arrive at a common and formal agreement for the definition of mesoscale feature types, e.g., rivers [9]. The need for richer semantic annotations to improve information retrieval has recently been acknowledged by Google, Yahoo, and Microsoft by launching their common schema.org platform.

Feature types also play a crucial role for Volunteered Geographic Information (VGI), e.g., for projects such as OpenStreetMap. These projects often maintain a semi-formal vocabulary with dozens or even hundreds of feature types[2]. The categorization of Points of Interest (POI) is especially

[1]The Infrastructure for Spatial Information in the European Community: http://inspire.jrc.ec.europa.eu/
[2]See OSM wiki at http://wiki.openstreetmap.org/wiki/Map_Features as an example.

difficult for several reasons. For instance, a single building can accommodate several different types of stores, those can change frequently over time, and offer multiple functionality. ATMs are a typical example discussed in the OpenStreetMap community. For instance, many post offices or banks are annotated as ATMs, i.e., taxonomic and partonomic relations are confused. The OSM community proposed to use the combination of *amenity=bank* and *atm=yes* as workaround. Pubs and bars are another example and led to the introduction of the *similar features* list in the OSM wiki. While the need for ontological definitions of such feature types is largely acknowledged and may not only benefit retrieval but also data cleaning and integration, attempts to develop POI ontologies have failed so far.

Approaches which try to make a heterogeneous and global community agree on definitions for types, e.g., *Bar* or *Bank*, by defining them in terms of walls, tables, menus, or guests are not likely to be successful. A narrow definition of *Bar* would exclude many places that locals perceive as bars, while too broad definitions would fail to distinguish *Bar* from other feature types such as *Café*. Based on our previous work [24, 34, 17], we propose to study massive real-world data from Location-based Social Networks (LBSN) to extract ontological primitives out of user behavior. These primitives do not define bars in terms of walls or tables, but their temporal characteristics, e.g., whether they are *weekend* or *weekday* locations, visited during *daytime* or in the *evening*. In this study, and complementary to our previous work on the semantic annotation of POIs [34], we are not interested in finding suitable tags for untagged POIs but in finding unique temporal characteristics for feature types. Together with our work on Spatial-Semantic Interaction [24], we aim at introducing *Semantic Signatures* as analogy to spectral signatures from remote sensing. Similarly to multiple spectral bands, these signatures can combine spatial, temporal, and thematic bands, and thereby identify feature types bottom-up. In other words, we use space and time as fundamental ordering principles for knowledge organization [17].

The remaining paper is structured as follows. First, we introduce Location-based Social Networks and Volunteered Geographic Information in more detail. We point out how we applied them in previous work to mine for spatial and temporal patterns. Next, we introduce our approach to Temporal-Semantic Interaction, present the used data, and discuss new measures. We then outline how our work can be applied for tag recommendation, place selection, and data cleaning – all of them being major challenges for LBSN and VGI. To do so, we point out how algorithms could use our findings and show brief examples for each case. Finally, we summarize our work and highlight directions for future work.

2. BACKGROUND AND RELATED WORK

This section reviews works necessary for the understanding of our research. While we will use category tags from LBSN as feature types, one has to keep in mind that volunteers do not follow rigid typing schemata. Hence, partonomic, taxonomic, and activity-related terms may be mixed. In order to stay as close as possible to the real data (and the labeling used by the community), we will use the terms feature type and category tags interchangeably. However, we will refer to activity related tags, e.g., *cocktail*, as category tags.

2.1 Location-based Social Networks

Geography and especially location play an increasing role in social online networks such as Facebook, Foursquare, or Whrrl, and have been analyzed in several studies [6, 29, 25, 35]. Work on feature types in Location-based Social Networks was presented in [34, 20], where [34] exploited the regularity of user behaviors in LBSN to assign category tags to untagged places, and [20] explored the place naming preferences of users in Location-based Social Networks.

Facebook researchers analyzed the distance between the users' social relations, and utilized locations of *friends* to predict the geographic location of specific users [3]. Cheng et al., modeled the spatial distribution of words in *Tweets* to predict the user's location [5]. Characterizing network properties in relation to local geography is studied in [33]. The behavior of users with respect to the location-field in their Twitter profiles has been studied in [16]. How and why people use location-sharing services and the privacy issues related to those services have been discussed in [21, 32, 31]. Finally, applications such as place recommendations [35, 36, 37, 38], content delivery services [28], and friend recommendations [30, 7] have been proposed.

2.2 Volunteered Geographic Information

Volunteered Geographic Information [13] describes the phenomenon of volunteers contributing geographic data and making them accessible under an open license. Projects such as OpenStreetMap (OSM)[3] or Wikimapia[4] provide platforms to publish and access VGI. Tags, representing feature types in VGI, are much more volatile than their counterparts defined by professional authorities. They are subject to frequent changes that emerge from informal discussions within the VGI community. Their usage is largely based on individual experience, cognition, as well as the used tagging and rendering software. So far, most research on VGI has focused on data completeness and positional accuracy [39, 23]. Assessing data quality with respect to non-spatial attributes and especially feature types is difficult as reference data is missing. The methods to determine that a certain POI tagged as *Café* is semantically less accurate than another POI marked as *Bar* if both are, in fact, identified as *Nightclub* are largely missing [24]. Additionally, this would require a ground truth feature type. Our work aims at setting the ground for such a semantic measurement framework.

Studying feature type definitions in OSM[5], shows that there is no explicit account for time. The key-value pair based formalization of types only distinguishes between, for example, amenities and shops on the higher level, and bars and cafés on the lower level. Temporal references are rare and can only be found on the informal description pages. For instance they are used to highlight regional differences:

> *In Mediterranean countries, the word "bar" has a different meaning [...] You go there in the morning to have breakfast, at lunch they serve simple meals, all day long (if not closed after*

[3] http://www.openstreetmap.org
[4] http://wikimapia.org/
[5] http://wiki.openstreetmap.org/wiki/Map_Features

lunch) people use them to get a quick coffee and in the evening it's a meeting place to get an apéritif before dinner. [6]

Without a more formal approach to temporal aspects, VGI cannot be used for place recommendations. The above example also indicates that regional differences matter and should be accounted for.

2.3 Spatial-Semantic Interaction

The *spatial bands* of semantic signatures, i.e., defining *what* you are by *where* you are, is equally important as the temporal bands introduced in this paper. However, both aspects differ substantially. Of most relevance for our work is the cyclicality of time in contrast to space, i.e., periods play a major role in our perception of time. In our everyday lives, we refer to discrete and reoccurring partitions of time, e.g., *evening*, *weekend*, or *New Year*. In contrast, space cannot be partitioned in such a way. Consequently, the above formulation has to be rephrased as *what* you are is *where* you are *with regard to other geographic features*. Spatial-Semantic Interaction models this relationship [24]. Other features types are not included explicitly as categorical variables but implicitly through a similarity measure. Thereby, it is possible to apply statistical measures such as concept variograms [2] or variations of Diggle's D_0 statistic [8].

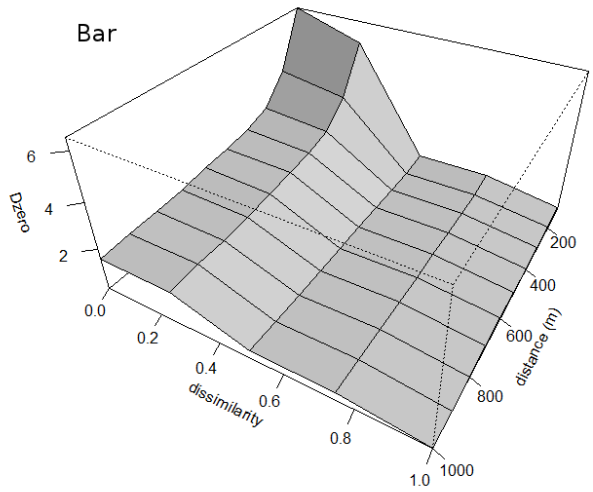


Figure 1: Spatial semantic interaction of bars in London computed from Open Street Map data; based on [24].

Figure 1 depicts an example of a Spatial-Semantic-Interaction D_0 plot. The z-value is interpreted as the likelihood of features of type *Bar* to co-occur with other features (of various types) within a certain semantic and spatial range. Up to a distance of 300 meters and an inter-type similarity value of 0.5, a significant clustering can be observed. In other words, bars tend to co-occur with other bars or features of similar types, e.g., nightclubs, within a close proximity. These two thresholds (space and type similarity) strongly vary among feature types and, therefore, can be used as one band for unique semantic signatures. Together with the temporal bands discussed in this paper, they can be employed to disambiguate feature types. Ontological primitives such as *clumped* or *regularly distributed* on the spatial side, and *evening*, *weekend*, or *weekday* on the temporal side can be computed and used to construct data driven, bottom-up POI ontologies [1]. Application areas will be discussed in section 4 in more detail.

[6] http://wiki.openstreetmap.org/wiki/Tag:amenity=bar

3. STUDYING TEMPORAL-SEMANTIC INTERACTION

In this section, we introduce the crawled data, discuss the weekly and daily temporal bands, as well as a notion of semantic feature type similarity derived from the check-in behavior of users in Location-based Social Networks. Finally, we compare the temporal patterns using a measure inspired by classical point-pattern analysis. Our study aims at demonstrating that a behavioristic approach can be used to discriminate types of places by observing user activities; see also the algebraic approach in [27]. For instance, the crawled check-in patterns to colleges differs significantly from those for cocktail-related places. Hence, our work can serve as basis for various services by predicting the type of an untagged place based on the check-in time of a specific user. With respect to the example above, a untagged POI visited regularly during the evening on weekends is most likely not a college. In this work, we focus solely on temporal aspects – the targeted semantic signatures will combine spatial, temporal, and thematic bands.

We conduct our research based on a dataset crawled from the Whrrl[7] platform in spring 2011. Whrrl was a representative Location-based Social Network and the first to display check-in times, i.e., the timestamps of users entering a certain place. Meanwhile, Whrrl has been acquired by Groupon. Further popular LBSN include Foursquare, Gowalla, or Facebook Place. However, they do not offer access to the temporal data required for this study. We have used two kinds of data, the feature types of places as well as their check-in times. During crawling Whrrl for one month, we extracted 35,745 users and their 440,939 check-in activities to 150,300 different places; see Table 1.

Number of users	35,745
Number of places	150,300
Number of check-ins	440,939
Number of tags	408

Table 1: Data crawled from Whrll in spring 2011.

From those places, we extracted 408 unique category tags, such as *restaurant*, *shop*, or *bar* used for feature typing. It turns out that types are not uniformly represented in Location-based Social Networks. POIs with tags related to dining, food, shopping, and nightlife are the most frequently checked-in places; about 74% of all check-ins are related to them. This observation confirms our common sense about daily life and also the kind of data that users make public. Intuitively, people would like to share their experience about

[7] http://www.whrrl.com

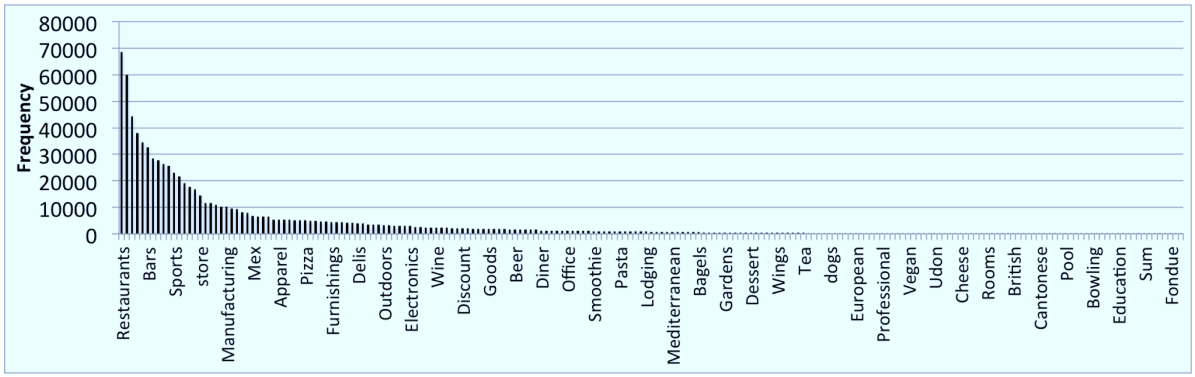


Figure 2: Check-in frequency distribution for selected geographic feature types in Whrrl.

activities such as dining, shopping, or their nightlife. Consequently, they are more likely to expose their footprints when visiting such places. As more users join Location-based Social Networks, the check-in activities to those places as well as their number increases. In other words, places of certain types are over-represented in terms of check-ins and their appearance in cities.

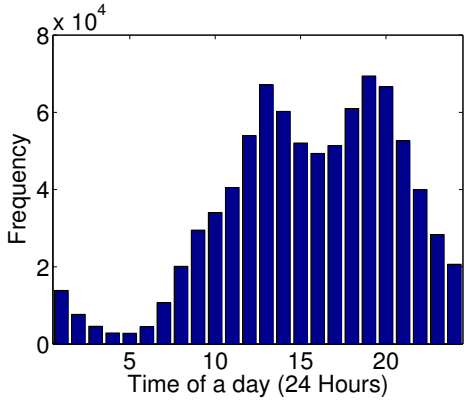


Figure 3: Daily temporal distribution of check-ins.

Figure 2 depicts a histogram of check-in frequencies for selected geographic feature types in Whrrl. Our findings confirm the power law decay shown in other, non-spatial studies on collaborative tagging [14]. Please note that due to the large number of check-ins and unique places most of the 408 types have hundreds or thousands check-ins related to them. As shown in Figure 2, places with tags such as *Restaurants*, *Bars* or *Sports* have the highest amount of check-ins. As Whrrl was based in the US, some types of restaurants are better represented in the data than others, e.g., Mexican restaurants or places offering Pizza. The collected data contains numerous linguistic variations, e.g., we have summarized *Mexican* and the more popular *Mex*. Besides food and drinks, e.g., represented by tags such as *Wine*, *Beer*, or *Cocktail*, shopping is another important daily activity in Whrrl. Examples for such places include those tagged with *Store*, *Electronics*, or *Furnishings*. Finally, travel related terms such as *Outdoors* or *Hotels* form another well represented group in our dataset. This may be due to the habits of Whrrl users. Typically, they are interested in documenting their trip by checking-in at new places while traveling.

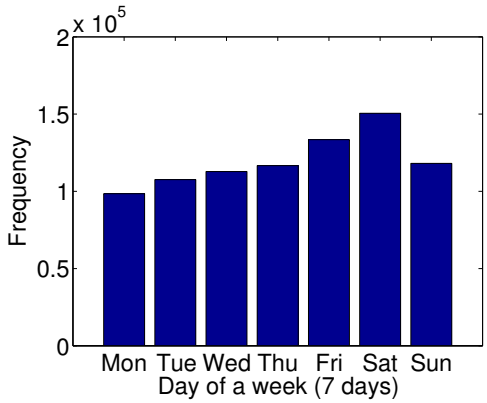
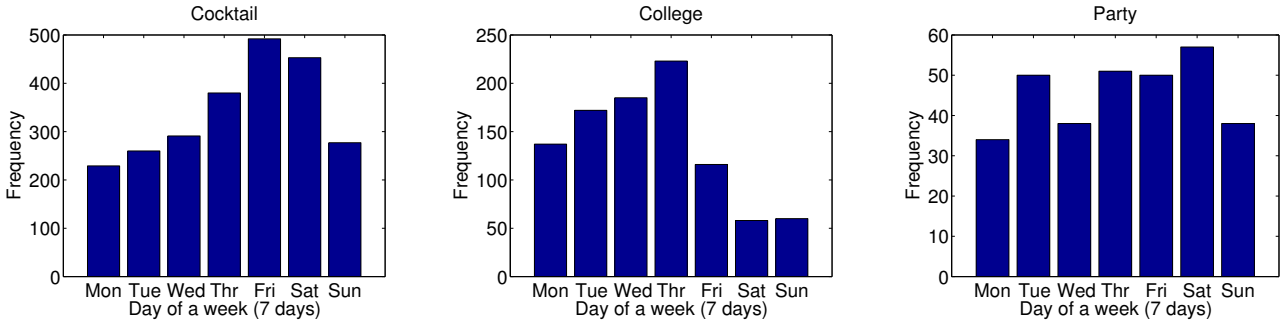


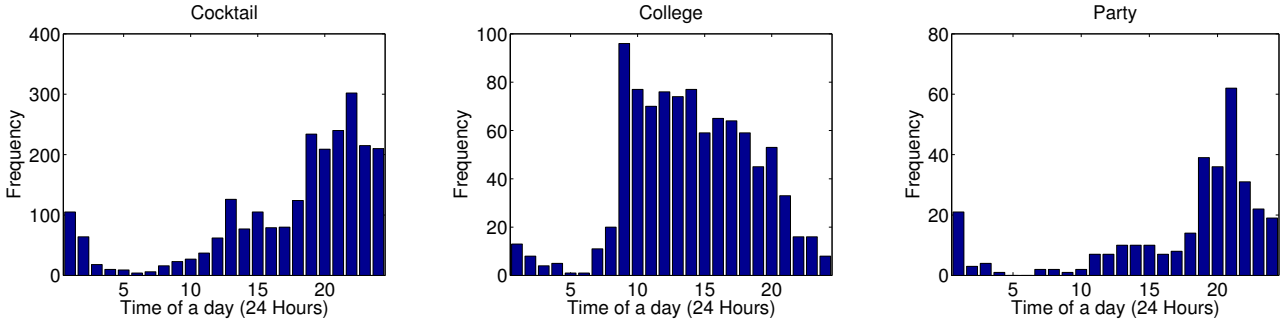
Figure 4: Weekly temporal distribution of check-ins.

Before describing the check-in activities to specific types of places and how they can by used to distinguish place types, we need to discuss the overall distribution of check-ins over days and hours. This is important to show that the patterns found in the behavior of Whrrl users is representative for how people interact with Points of Interest. If the crawled check-in data would not obey common sense, bands and signatures extracted from them would not be meaningful for tag recommendations or place selection; see section 4. Figures 3 and 4 provide a general overview of how users interact with places in Whrrl. As depicted in Figure 3, people interact with places frequently at around noon and in the evening. Most activities happen between 9am and 11pm, with two peaks at around 1pm and 7pm. This is due to the fact that most check-ins are related to restaurants and food. The check-ins mirror the daily lunch and dinner cycles. A related observation can be made for the weekly data. As activities related to dining, shopping, and nightlife are over-represented in the data, we find the highest volume of check-ins on Fridays and Saturdays; see Figure 4. Overall, we could not find evidence for distortions in our data as they would be expected for highly specialized user communities, e.g., a low number on check-ins during weekdays or high activities during late evenings.



(a) Cocktail (b) College (c) Party

Figure 5: Temporal bands of different geographic feature types (weekly band).



(a) Cocktail (b) College (c) Party

Figure 6: Temporal bands of different geographic feature types (daily band).

3.1 Temporal Band

The primary goal of the presented work is to understand geographic feature types based on how, i.e., when, people interact with places of those types. For instance, as discussed above, people usually go to restaurants at around noon and in the evening for lunch and dinner, respectively; while they visit nightclubs during the late evening or night. We argue that by analyzing the check-in pattern for a specific place, we can make predictions about its type. We will call methods to study the temporal dimension of types *Temporal-Semantic Interaction* and name different temporal pattern as *temporal bands*. While we have shown that temporal data can be successfully applied to automatically learn feature types for untagged places [34], we do not argue that one or more temporal bands will always be enough to distinguish feature types. To compute unique semantic signature may require additional bands such as those extracted from analyzing *Spatial-Semantic Interaction*; see [24].

Figure 5 plots the weekly check-in patterns to three different category tags: *cocktail*, *college* and *party*. As can be seen, cocktail related places are visited during the whole week; however, they are most frequently visited during the weekend. For instance, within one month of crawling, we collected more than twice as many check-ins on Fridays compared to Mondays. We have observed similar distributions for other category tags as well, e.g., for the tags *Bars* and *Beer*. It is important to keep in mind that category tags assigned by volunteers differ from feature typing schemata of professional authorities. The tag *cocktail* simply describes those types of Points Of Interest that serve cocktails.

In contrast to the feature types discussed above, places tagged as *college* show a significant check-in decay during the weekend, i.e., they are *weekday* features. Moreover, we can also observe a drop on Fridays which is well known by faculty members teaching on this weekday. Places tagged by *party* do not show such significant patterns. While Saturdays are preferred and the lowest number of check-ins takes place on Sunday and Monday, the distribution rather indicates that users also tagged private apartments with *party*, e.g., for a birthday party (which can take place during every day of the week). This example demonstrates that a single band, in this case the weekly band, may not be sufficient to identify unique patterns for each feature type.

Therefore, Figure 6 adds the daily check-in patterns for the three category tags. Users visit places tagged with *cocktail* in the late evening, typically after 6pm. In contrast, they check-in at colleges during the typical working hours, i.e., from 9am to 6pm. As depicted in Figure 6c, party related check-ins take place during the evening, while almost no check-ins can be observed in the morning. While the number of total check-ins is low for the *party* tag in general, the daily band demonstrates how multiple bands can be used if a single temporal band does not provide enough discriminatory power. Hence, we argue that multiple temporal bands can be combined to provide a robust and meaningful descrip-

tions of different geographic feature types.

Finally, while we only investigate the daily and weekly temporal bands here, other bands can be generated from from seasonal or and even yearly data. However, as our crawling covers only one month, we do not explore these bands in our work. Bands can also be combined, e.g., using kernel density estimation.

3.2 Semantic Similarity

Typically, geographic feature types have been defined *intensionally*, i.e., by necessary and sufficient conditions for membership. While more expressive, this approach faces the problems introduced in section 1. It is unlikely, that a global and highly heterogeneous community will agree on a set of characteristics for types such as *Bar* or *College*. Studying the check-in data from Location-based Social Networks, we have the opportunity to provide an alternative, *extensional* approach. We propose to *learn* the semantics of different feature types, by taking advantages of massive user behavior – in this study, their check-ins. However, the lack of declarative knowledge asks for alternative approaches to reasoning as well. To implement recommendation systems, data cleaning, or to integrate data from heterogeneous sources, and as argued in section 2, we need measures for the proximity of types. In this section, we devise a new semantic dissimilarity function for geographic feature types, measured based on the differences of temporal bands.

Let $G = \{g_1, g_2, ...\}$ denote the domain of geographic feature types, where is $|G| = 408$ corresponds to the number of different category tags in our dataset. The temporal band of geographic feature type $g_i \in G$ is denoted by tb_i. Note that in order to ease the computation of semantic dissimilarity, we propose to unify all the temporal bands of different geographic feature types, i.e., the temporal band is transformed into a probability density function (called normalized temporal band here). For example, given a geographic feature type g_i, the normalized *daily* temporal band is present as $tb_i^d = \langle p_{i,1}, p_{i,2}, \cdots, p_{i,24} \rangle$, where $\sum_{j=1}^{24} p_{i,j} = 1$ and $p_{i,j} (\geq 0)$ is proportional to the frequency of check-ins to the geographic features of type g_i at hour j_{th}. Similarly, we get the normalized weekly temporal band of geographic feature type $g_i \in G$ as $tb_i^w = \langle p_{i,1}, p_{i,2}, \cdots, p_{i,7} \rangle$, where $\sum_{j=1}^{7} p_{i,j} = 1$ and $p_{i,j} (\geq 0)$ is proportional to the frequency of check-ins to the geographic features of type g_i at day j_{th} of a week.

The semantic dissimilarity function is denoted by dis-sim(g_i, g_j) $(g_i, g_j \in G)$ and defined as follows; see 1.

$$\text{dis-sim}(g_i, g_j) = d(tb_i, tb_j) \tag{1}$$

where $d(\cdot, \cdot)$ denote a function to measure the distance between two probability distribution, i.e., tb_i and tb_j.

There are several candidate measures for such distance function, e.g., Kullback-Leibler divergence, Hellinger distance, Total Variation Distance, Energy Distance, or Bhattacharyya distance. Among them, Kullbak-leibler divergence and Hellinger distance are non-symmetric, while the analysis in this paper requires a symmetric semantic dissimilarity function, i.e., dis-sim(g_i, g_j) = dis-sim(g_j, g_i). We have tested Total Variation Distance, Energy Distance, and Bhattacharyya distance, and found that their results are very close with respect to our data. We will use Total Variation Distance in our study; see Equation 2.

$$\text{dis-sim}(g_i, g_j) = \frac{1}{2} \sum_{k=1}^{|tb|} |p_{i,k} - p_{j,k}| \tag{2}$$

For example, if we have two temporal bands $tb_1 = \langle 0.3, 0.4, 0.3 \rangle$ and $tb_2 = \langle 0.2, 0.3, 0.5 \rangle$, then the dissimilarity between g_1 and g_2 is $\frac{1}{2}(|0.3-0.2|+|0.4-0.3|+|0.3-0.5|) = 0.2$.

Whether temporal bands can be used to compute semantic dissimilarity between two feature types depends on the robustness of those bands. More specifically, the fact that Friday 11pm is close to Saturday 1am is not modeled in our approach so far. Consequently, we have to take the circularity of the temporal dimension into account to derive an appropriate dissimilarity measure. Therefore, we propose a smoothing function based on a classical moving-window and kernel approach that operates on the original temporal band and transforms it in to a more robust form.

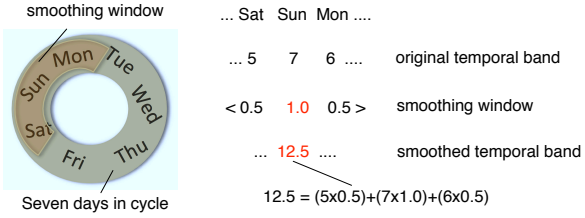


Figure 7: An example for smoothing a temporal band; in this case the values for Sunday.

For instance, Figure 7 depicts a part of a weekly temporal band together with a 3-cell window and the kernel $\langle 0.5, 1.0, 0.5 \rangle$. The central weight of the window is usually defined as 1, and the adjacent weight define the degree of smoothing. Higher values increase the impact of the neighboring days or hours, respectively. Smoothing is achieved by moving the window step-by-step (i.e., day-by-day or hour-by-hour) to determine the new value for the center cell. For example, in Figure 7, the new value for Sunday is the weighted sum of the original temporal band after applying the kernel weights. As depicted, the smoothed value is $(5 \times 0.5) + (7 \times 1.0) + (6 \times 0.5) = 12.5$.

In the following, we use the geographic feature type *stadium* for demonstration purpose as it highlights benefits and shortcomings of our approach at the same time. Table 2 shows the top-10 similar geographic feature types to stadium, where similarity is measured as inverse distance $(sim(g_i, g_j) = 1 - dis - sim(g_i, g_j))$ on the daily band. Compared to the values derived without smoothing, among all category tags *Museums*, *Live Performance*, and *Entertainment* are considered to be highly similar to *Stadium*. These rankings have to be interpreted with care. They do not mean that museums are similar with respect to specific characteristics, such as building structure, but that Whrrl users interact

original		smoothed	
GFT	dis-sim	GFT	dis-sim
Arenas	0	Arenas	0
Apparel	0.019	Apparel	0.019
Outdoors	0.062	French	0.019
Beer	0.064	Beer	0.021
Dogs	0.064	Frozen	0.023
Improvement	0.065	Museums	0.024
Frozen	0.070	Rental	0.027
Adult	0.071	Live Performance	0.027
Indies	0.075	Outdoors	0028
Furnishings	0.076	Entertainment	0.030

Table 2: Sorted top-10 semantic dissimilarity scores between *Stadium* and other geographic feature types (GFT) according to their daily bands. The kernel is set to $\langle 0.5, 1.0, 0.5 \rangle$.

original		smoothed	
GFT	dis-sim	GFT	dis-sim
Arenas	0	Arenas	0
European	0.019	Live Performance	0.134
Live Performance	0.090	Vegetarian	0.153
Fitness	0.101	Theaters	0.155
Theaters	0.109	Drink	0.173
Drink	0.117	Baseball	0.180
Food	0.119	Sports	0.183
Chefs	0.120	Arts	0.185
Caterers	0.120	Recreation	0.187
Vegan	0.121	Entertainment	0.188

Table 3: Sorted semantic dissimilarity scores between *Stadium* and the top-10 geographic feature types (GFT) according to their weekly bands. The kernel is set to $\langle 0.2, 0.5, 1.0, 0.5, 0.2 \rangle$.

with them at similar times. The *Beer* tag shows the benefits as well as drawbacks of such an behavioristic approach. Following an intensional approach to ontology engineering, beer (related places) would share no (or just a few) characteristics with stadiums. However, people often go to sports events to drink beer. Therefore, an approach that is based on user behavior is able to capture such hidden relations. To a certain degree, stadiums are *beer drinking locations*.

As argued before, a single band may not be able to discriminate feature type which makes the combination of different bands necessary. We only consider temporal bands here. Adding a spatial band improves the rankings by removing dining places from the stadium list due to their different distribution in space [24]. Table 3 shows similar geographic feature types, such as *Theaters*, *Baseball*, *Sports*, *Arts*, *Recreation* and *Entertainment* for the weekly band. All these activities/types share a certain weekly pattern with stadiums and have their peaks during weekends.

In both cases, the smoothed results provide the better results with respect to common sense – as argued above, a gold standard is missing. Combining and grouping the results shows similarities between *Stadiums* and sports related places (*Baseball*, *Sports*, *Recreation*, *Outdoor*) as well as those providing entertainment (*Live Performance*, *Theaters*, *Entertainment*, *Museums*).

3.3 Temporal Analysis of Feature Types

So far, we discussed how temporal patterns from massive user check-ins can be explored to understand the semantics of category tags assigned for place typing in Location-based Social Networks. We introduced a moving window based similarity measure using the probabilistic Total Variation Distance to compare individual feature types via their temporal bands. While we will outline how to apply the results to recommendation services and data cleaning in section 4, similarity is restricted to a binary comparison. To understand the relations between different types, we have to analyze their temporal clustering patterns as well.

We abstract each feature type to a point in a one-dimensional space to reveal whether they are clumped, randomly, or regularly distributed with respect to other, similar types; compare to our spatial analysis in [24]. We propose a statistics (called M here) which is inspired by Ripley's spatial point-pattern analysis K [26]. Note that the space is constructed in a relative way, i.e., we can only compute the dissimilarity score between any two geographic feature types. Therefore, given a target geographic feature type $g_i \in G$, the temporal-semantic similarity space can be formed by applying Equation (2) to calculate the dissimilarity scores between g_i and any $g_j \in G$. For example, given the target feature type *cocktail*, we compute the corresponding space as depicted in Figure 8. Based on the user check-in behavior, the *liquor* category tag is more similar to *cocktail* than *pubs*.

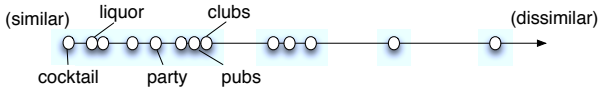


Figure 8: The ray depicts the distribution of feature types from similar to dissimilar with respect to the target type cocktail (dissimilarity = 0).

Next, we devise a $M_{S_i}()$ function to study the resulting clustering patterns in the space S_i for the given target feature type $g_i \in G$; see Equation (3).

$$M_{S_i}(d) = \lambda^{-1} E_{S_i}(d) \tag{3}$$

$E_{S_i}(d)$ is the number of points within distance d to the given original point (i.e., the target type) in S_i, and λ is the intensity, i.e., the expected number of points in a regular distribution. Figure 9 demonstrates how this statistics can be inspected to understand the relation between feature types in Location-based Social Networks. Five geographic feature types are shown: *cocktail*, *shopping*, *stadium*, *food*, and *bars*, as well as a plot for a regular distribution. At a given distance d, the higher $M(d)$, the more geographic feature types are similar to the target type. For instance, there are more similar feature types to *shopping* (40% of all types are within the [0,0.2] similarity interval) than to *cocktail* (10% in the same interval). The reason is that our semantic similarity measure for geographic feature types is defined based on human behavior, i.e., timestamps from check-ins. Shopping is

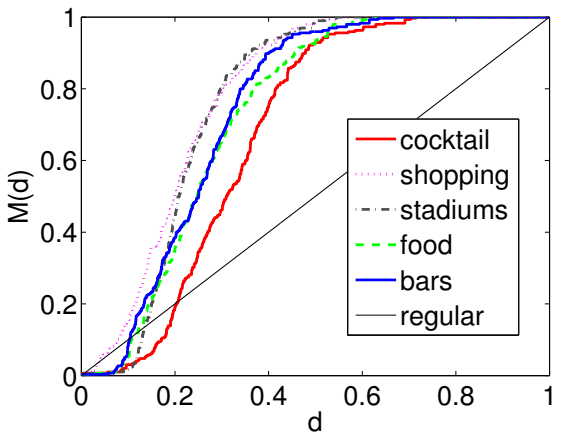


Figure 9: M function calculated based on the smoothed daily band, with a moving window of $\langle 0.5, 1.0, 0.5 \rangle$.

a more generic activity and can be associated with other activities such as renting a movie, dining, buying electronics, and so forth. In contrast, *cocktail* is more specific and restricted to nightlife; thus, there are not as many similar geographic feature types. In other words, the patterns generated by *cocktail* and related category tags are more unique in comparison to the other types; many of them show a distribution in the daily and weekly band that is similar to the one from *shopping*.

In the future, we plan to explore methods to discover subsumption relations between feature types to be used in data-driven ontology engineering.

4. APPLICATIONS

Several applications could benefit from the research on Temporal-Semantic Interaction in Location-based Social Networks. For example, service providers could support users with tag and place recommendations. Moreover, data in Location-based Social Networks is volunteered, therefore the data quality is not guaranteed. A better and data-driven understanding of feature types can foster data cleaning, data integration, and on the long term also assist in ontology engineering. As argued before, categorization is a key to decision support and recommendation. In the following, we provide an overview on how our findings could contribute to such services.

Tag Recommendation. When users check-in at a place, they may want to publish their opinions about this place. Tag recommendations can assist users to assign meaningful tags that are already circulated. Common algorithms for tag recommendation are based on the tag usage frequency, i.e. the most frequently used tag ranks at the top. However, based on the additional temporal bands, we can make context, i.e., time, driven tag recommendations. Two examples for new, prototypical recommendation algorithms based on temporal bands are:

-*Temporal Band Similarity*. In this algorithm, the feature type that has the most similar temporal band to the check-in temporal band of a given place should be ranked top. Let tb^* denote the temporal band of check-ins to a given place, and tb_i be the temporal band of check-ins to some specific geographic feature types $g_i \in G$. Then g_i will be ranked higher if the value of dis-sim(tb^*, tb_i) is smaller.

-*Check-in Probability Maximization* In this case, we exploit the check-in time t of a user to a given place to infer the probability of a category tag to a place, i.e., $Pr(g_i|t)$.

$$Pr(g_i|t) = \frac{Pr(t|g_i)Pr(g_i)}{Pr(t)} \propto Pr(t|g_i)Pr(g_i) \qquad (4)$$

where $Pr(t|g_i)$ can be estimated from the temporal band corresponding to the geographic feature type $g_i \in G$, and the prior $Pr(g_i)$ can be estimated from the raw data. The higher $Pr(g_i|t)$, the higher the corresponding geographic feature types will be ranked.

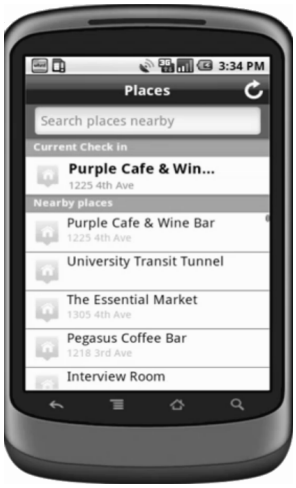


Figure 10: Place Selection in the Whrrl Android application.

Place Selection. To improve the usability of mobile Location-based Social Network applications, our work could assist in choosing the user's current place based on the time and weekday. When a user want to check-in to some place, mobile applications usually provide a list of nearby places. So far, these are ranked according to distances as shown in Figure 10. Though distance is a good criteria to rank places, temporal information is also important and instructive to tell where the user may be. Our study can assist in re-ranking the candidate places by considering the temporal band. For example, if a user checks-in at a given place at 1am, we can to rank places with *pubs* or *nightlife* tags higher than a nearby grocery store. Note that a place may afford multiple activities and have multiple tags. Formally, we calculate the check-in probability $Pr(p_i|t)$ to a place p_i, given the current time t. The higher $Pr(p_i|t)$, the higher the corresponding place p_i will be ranked.

$$Pr(p_i|t) = \max_{g_x \in G_i} Pr(g_x|t) \qquad (5)$$

where G_i is the category tag set of place p_i, and $Pr(g_x|t)$ can be estimated according to Equation (4).

Data Cleaning. A place may be assigned multiple category tags by different people. Since the data is volunteer, noise is very likely, i.e., users may make mistakes when assigning tags to a place. In general, the temporal bands of places of the same (or similar) geographic feature types should be similar as they offer the same activities and share similar time constraints, e.g., opening hours. However, if such places show a very distinctive temporal band, the data may have to be double-checked or cleaned. More formally, given a place p_i and its tags $g_x \in G_i \subset G$, we define the maximum dissimilarity among the tags associated with the place as the clean score (CS_i) of the corresponding place,

$$CS_i = \max_{g_x, g_y \in G_i} \{\text{dis-sim}(g_x, g_y)\} \quad (6)$$

If CS_i exceeds a threshold value θ, we would propose to investigate whether the assigned tags match the place. Larger θ may pass more noisy data; while smaller θ can improve the data quality at the cost of high overhead for data reviewing. In the future, we will use a training dataset to learn a proper θ for specific category tag combinations.

5. SUMMARY AND OUTLOOK

In this paper, we presented a study on the temporal dimension of places and their types in Location-based Social Networks. Starting from the behavioristic assumption that *what* you are can be determined by *when* you are, we crawled the Whrrl platform for one month to extract data about users, places, category tags, and check-in timestamps. For each of 408 category tags, we investigated two kinds of patterns, those from differences in the distribution of daily check-in times and those based on weekly differences. We then applied a moving window based smoothing to account for the circularity in temporal data. Next, we introduced a dissimilarity measure to compare feature types based on the different check-in distributions by computing the probabilistic Total Variation Distance. To extend our study beyond binary comparison, we introduce a statistics to visualize the check-in time-based clustering between different geographic feature types.

The presented work forms one of three pillars, the others being spatial [24] and thematic [1], to introduce *semantic signatures* as a data-driven methodology to uniquely identify and distinguish feature types such as *Bar* or *College*. In analogy to multiple electromagnetic bands used for spectral signatures in remote sensing, we showed how temporal bands can be defined based on the daily and weekly check-in behavior of Whrrl users. We outline how these bands and the introduced measures can be used for application areas such as tag recommendation, place selection, and data cleaning – all of them being major challenges in research on Location-based Social Networks, Volunteered Geographic Information, and Mobile Spatial Decision Support Systems. We also argued that our methodology can be used to derive ontological primitives that are directly extracted from observations, i.e., user behavior. These primitives could be used to define places based on whether they are *weekend*, *weekday*, *evening*, or *daytime* places, instead of defining them in terms of common characteristics such as walls, tables, chairs, or menus. This work directly contributes to our work on Spatial-Semantic-Interaction [24]. Nevertheless, it is important to note that we do not aim at replacing intensional, declarative ontology engineering but propose to combine top-down with bottom-up approaches. Research on geographic ontology design patterns will be required to compose such ontologies out of observation-based primitives.

Temporal-Semantic-Interaction can only set the ground for further research and a lot of work remains to be done. So far, we have not decided on how to combine multiple spatial, temporal, and thematic bands to form the envisioned semantic signatures. Moreover, while we have outlined how our work supports the introduced application areas, these services need to be implemented and tested with human participants. While we envision to use our findings for data cleaning in the future, our own bands require cleaning as well. Volunteered Geographic Information contains noise, is highly heterogeneous, and inconsistent. In this study, we deliberately used unfiltered data[8] to demonstrate the necessity of multiple bands and smoothing. However, often we had to rely on common sense as there are no reference sets or gold standards for non-spatial attributes in VGI. In fact, our work is intended to set the ground and provide measures for further studies. It can assist in answering the question whether a dataset declaring a feature as *Bar* is more or less semantically accurate than another dataset tagging the same location with *Nightclub*, while, according to ground truth, it is of type *Restaurant*. Therefore, the proposed measures are a way to investigate semantic accuracy in addition to feature completeness and positional accuracy studied so far. This will require to further semantify geostatistics.

6. REFERENCES

[1] B. Adams and K. Janowicz. Constructing geo-ontologies by reification of observation data. In *ACM GIS*, 2011.

[2] O. Ahlqvist and A. Shortridge. Characterizing land cover structure with semantic variograms. *Progress in Spatial Data Handling*, pages 401–415, 2006.

[3] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*, pages 61–70, 2010.

[4] L. Barsalou. Situated simulation in the human conceptual system. *Language and Cognitive Processes*, 5(6):513–562, 2003.

[5] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *ACM CIKM*, pages 759–768, 2010.

[6] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. In *AAAI ICWSM*, 2011.

[7] J. Cranshaw, E. Toch, J. I. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *ACM UbiComp*, pages 119–128, 2010.

[8] P. Diggle, A. Chetwynd, R. Häggkvist, and S. Morris. Second-order analysis of space-time clustering. *Statistical methods in medical research*, 4(2):124, 1995.

[9] S. Duce and K. Janowicz. Microtheories for spatial data infrastructures - accounting for diversity of local conceptualizations at a global level. In *International*

[8] However, we summarized lexical variations and removed feature types with very few check-ins.

Conference on Geographic Information Science (GIScience), pages 27–41, 2010.

[10] M. Fernandez-Lopez, A. Gomez-Perez, and N. Juristo. Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI'97 Spring Symposium*, pages 33–40, Stanford, USA, 1997.

[11] A. Frank. Tiers of ontology and consistency constraints in geographical information systems. *International Journal of Geographical Information Science*, 15(7):667–678, 2001.

[12] J. Gibson. The theory of affordances. In R. Shaw and J. Bransford, editors, *Perceiving, Acting, and Knowing - Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum Ass., Hillsdale, New Jersey, 1977.

[13] M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

[14] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *ACM WWW*, pages 211–220. ACM, 2007.

[15] S. Harnad. To cognize is to categorize: Cognition is categorization. In C. Lefebvre and H. Cohen, editors, *Summer Institute in Cognitive Sciences on Categorisation*. Elsevier, 2005.

[16] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *ACM CHI*, pages 237–246, 2011.

[17] K. Janowicz. The role of space and time for knowledge organization on the semantic web. *Semantic Web Journal*, 1(1-2):25–32, 2010.

[18] W. Kuhn. Geospatial Semantics: Why, of What, and How? *Journal on Data Semantics III*, pages 1–24, 2005.

[19] S. Lehar. *The World in Your Head: A Gestalt View of the Mechanism of Conscious Experience*. Lawrence Erlbaum, 2003.

[20] J. Lin, G. Xiang, J. I. Hong, and N. M. Sadeh. Modeling people's place naming preferences in location sharing. In *ACM UbiComp*, pages 75–84, 2010.

[21] J. Lindqvist, J. Cranshaw, J. Wiese, J. I. Hong, and J. Zimmerman. I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application. In *ACM CHI*, pages 2409–2418, 2011.

[22] D. M. Mark. Toward a theoretical framework for geographic entity types. In *International Conference in Spatial Information Theory: A Theoretical Basis for GIS*, pages 270–283, 1993.

[23] P. Mooney, P. Corcoran, and A. C. Winstanley. Towards quality metrics for openstreetmap. In *ACM GIS*, pages 514–517. ACM, 2010.

[24] C. Mülligann, K. Janowicz, M. Ye, and W.-C. Lee. Analyzing spatial-semantic interaction of points of interest in volunteered geographic information. In *International Conference on Spatial Information Theory*, 2011.

[25] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *AAAI ICWSM*, 2011.

[26] B. Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2):255–266, 1976.

[27] R. D. Rugg, M. J. Egenhofer, and W. Kuhn. Formalizing behavior of geographic feature types. *Geographical Systems*, 4:159–179, 1997.

[28] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft. Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades. In *WWW*, pages 457–466, 2011.

[29] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *AAAI ICWSM*, 2011.

[30] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *ACM SIGKDD*, 2011.

[31] E. Toch, J. Cranshaw, P. H. Drielsma, J. Springfield, P. G. Kelley, L. F. Cranor, J. I. Hong, and N. M. Sadeh. Locaccino: a privacy-centric location sharing application. In *ACM UbiComp (Adjunct Papers)*, pages 381–382, 2010.

[32] E. Toch, J. Cranshaw, P. H. Drielsma, J. Y. Tsai, P. G. Kelley, J. Springfield, L. F. Cranor, J. I. Hong, and N. M. Sadeh. Empirical models of privacy in location sharing. In *ACM UbiComp*, pages 129–138, 2010.

[33] S. Yardi and D. Boyd. Tweeting from the town square: Measuring geographic local networks. In *AAAI ICWSM*, 2010.

[34] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In C. Apté, J. Ghosh, and P. Smyth, editors, *ACM SIGKDD*, pages 520–528, 2011.

[35] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *ACM GIS*, pages 458–461, 2010.

[36] M. Ye, P. Yin, W.-C. Lee, and D. L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *ACM SIGIR*, 2011.

[37] Y. Zheng and X. Xie. Learning travel recommendations from user-generated gps traces. *ACM TIST*, 2(1):2, 2011.

[38] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma. Recommending friends and locations based on individual location history. *ACM TWEB*, 5(1):5, 2011.

[39] D. Zielstra and A. Zipf. A comparative study of proprietary geodata and volunteered geographic information for Germany. In *13th AGILE International Conference on Geographic Information Science*, 2010.