Methods for Extracting Place Semantics from Flickr Tags

TYE RATTENBURY
Intel Corporation
and
MOR NAAMAN
Rutgers University

We describe an approach for extracting semantics for tags, unstructured text-labels assigned to resources on the Web, based on each tag's usage patterns. In particular, we focus on the problem of extracting place semantics for tags that are assigned to photos on Flickr, a popular-photo sharing Web site that supports location (latitude/longitude) metadata for photos. We propose the adaptation of two baseline methods, inspired by well-known burst-analysis techniques, for the task; we also describe two novel methods, TagMaps and scale-structure identification. We evaluate the methods on a subset of Flickr data. We show that our scale-structure identification method outperforms existing techniques and that a hybrid approach generates further improvements (achieving 85% precision at 81% recall). The approach and methods described in this work can be used in other domains such as geo-annotated Web pages, where text terms can be extracted and associated with usage patterns.

Categories and Subject Descriptors: H.1.m [**Models and Principles**]: Miscellaneous

General Terms: Algorithms, Measurement

Additional Key Words and Phrases: Tagging systems, tags, places, semantics

ACM Reference Format:

Rattenbury, T., and Naaman, M. 2009. Methods for extracting place semantics from Flickr tags. ACM Trans. Web 3, 1, Article 1 (January 2009), 30 pages. DOI = 10.1145/1462148.1462149 http://doi.acm.org/10.1145/1462148.1462149

1. INTRODUCTION

User-supplied "tags," textual labels assigned to content, are a powerful and useful feature in many social media and Web applications (prominent examples

This article is an extended version of SIGIR 2007 conference paper Rattenbury et al. [2007].

Authors' addresses: T. Rattenbury, People and Practices Research Group, Intel Corporation, 2200 Mission College Blvd., Santa Clara, CA 95054-1549; email: tye.l.rattenbury@intel.com; M. Naaman, SCILS - Department of Library and Information Science, Rutgers University, 4 Huntington Street, New Brunswick, NJ 08901; email: mor@scils.rutgers.edu.

include Flickr, del.icio.us, and YouTube). Tags usually manifest in the form of a freely chosen, short list of keywords associated by a user with a resource such as a photo, Web page, or blog entry. Unlike category- or ontology-based systems, tags have no a priori semantics, and therefore result in unstructured knowledge. The unstructured nature of tags is the basis of their utility. For example, tags are often easier to enter than picking categories from an ontology, allow for greater flexibility and variation, and naturally evolve to reflect emergent properties of their referents [Golder and Huberman 2006].

The information challenge facing tagging systems is to extract structured knowledge from the unstructured set of tags. Despite the lack of ontology and a priori defined semantics, patterns and trends emerge that enable the extraction of structured information from tag-based systems [Golder and Huberman 2006; Marlow et al. 2006; Schmitz 2006]. While complete semantic understanding of tags associated with individual resources is unlikely, the ability to assign *some* structure to tags and tag-based data will make tagging systems more useful.

Broadly, we are interested in the problem of identifying patterns in the distribution of tags over some domain. In this work, we focus on spatial patterns in tags from Flickr. Specifically, we concentrate on Flickr,[1] a popular photo-sharing Web site that supports user-contributed tags and geo-referenced (or, *geotagged*) photos. Tag usage distributions are derived from the metadata of photos associated with each tag. While the correctness of the location metadata for each individual photo is suspect [Bulterman 2004], in large numbers, trends and patterns can be reliably extracted and used [Dubinko et al. 2006; Jaffe et al. 2006], as we show in this work.

Based on the spatial distributions of each tag's usage, we attempt to automatically determine whether each tag has a coherent place semantic. To clarify, by "place semantic" we mean that the tag has a strong associative mapping (i.e., signifier-signified relationship), as demonstrated by its usage, to a specific place—socio-culturally defined location, or set of locations. For example, the tag `Bay Bridge`[2] has a coherent place semantic in the San Francisco Bay area. Our place semantic definition is similar to the "query's dominant location" definition [Wang et al. 2005].

Extraction of place semantics can assist many different applications in the photo retrieval domain and beyond, including:

—improved image search through inferred query semantics;
—automated creation of place gazetteer data that can be used, for example, to improve Web search by identifying relevant spatial regions for particular keywords;
—generation of photo collection visualizations by location;
—support for tag suggestions for photos (or other resources) based on location; and
—automated association of missing location metadata to photos or other resources, based on tags or caption text.

[1] `http://www.flickr.com`.
[2] We use this font to represent tags in the text.

In this work we do not apply our analysis to a specific application, but rather investigate the feasibility of automatically determining place semantics for Flickr tags.

This article explores a number of possible methods for automatically detecting place semantics. In particular, we extend and expand on our work in Rattenbury et al. [2007] by proposing and evaluating new methods to perform this task. Ultimately, we demonstrate the advantage of hybrid methods that mix the output from a number of existing methods.

We note that our general approach to semantics extraction, and the methods we present as instantiations of this approach, can be applied to any information sources with spatial encodings from which we can extract textual terms, like GeoRSS blog data and geo-annotated Web pages. Additionally, the general approach of analyzing a distribution of occurrences over a domain (in our case, geographic coordinates) to infer semantics could be extended to other metadata domains like time, color (hue/saturation), visual features, audio features, and text/semantic features. Indeed, in Rattenbury et al. [2007] we apply the same methods to extract event semantics (using the temporal usage distribution of the photos associated with each tag). The new methods proposed in this work would apply to the time domain as well; however, for simplicity, in this article we concentrate on the spatial domain and detecting place semantics.

To summarize, the contributions of this work are:

—a generalizeable approach for extracting tag semantics based on the distribution of individual tags;

—the modification, application, and analysis of existing methods to the problem of place semantic identification for tag data;

—the demonstration of the superior performance of hybrid (i.e., mixed) methods; and

—a practical application of these methods, evaluating extraction of place semantics from tags associated with geotagged images on Flickr.

We formally define our problem in Section 2. Then we describe the methods (in Section 3) and report on our evaluation (Section 4). Before concluding, we review related work (in Section 5).

2. PROBLEM DEFINITION

In this section, we provide a formal definition of our data and research problem. Our dataset includes two basic elements: photos and tags. We define the set of photos as $\mathbb{P} \triangleq \{p\}$, where p is a tuple (θ_p, ℓ_p, u_p) containing a unique photo ID, θ_p; the photo's capture location, represented by latitude and longitude, ℓ_p; and the ID of the user that contributed the photo, u_p. The location ℓ_p generally refers to the location where the photo p was taken, but sometimes marks the location of the photographed object.

Photos with location metadata can be generated in at least two ways. First, the photographer could use a GPS-enabled camera or carry an external GPS device that automatically associates latitude and longitude coordinates with

photos (in the case of the external device, the association is performed by post-processing software). Location metadata generated automatically is both precise and accurate at the micro-second of degrees resolution, and specifies the location where the photo was taken.

A second method for generating location metadata is to use mapping software to locate photos on a map (e.g., Flickr). This method often allows people to associate location metadata at a range of resolutions, from coarse to precise. Interestingly, people can locate the photo either at the location where the photo was taken or at the location of the primary object(s) in the photo.

In this work, we restrict our dataset to those photos with high-resolution location metadata (i.e., either those photos with automatically associated latitude and longitude coordinates or those photos which people have located with high precision on a map). The distinction on whether the photo location metadata corresponds to the location where the photo was taken or to the location of the photographed object(s) is irrelevant in this work. In terms of place semantics, the geographic location of photographed objects and the geographic location of good vantage points of those objects both carry place semantics (albeit with slightly different meaning) that can be referenced by a single tag. Hence, in terms of the place semantics of tags, both types of geographic location are relevant.

The second element in our dataset is the set of tags associated with each photo. We use the variable x to denote a tag and $\mathbb{X}$ to denote the set of all tags. Note that each photo can have multiple tags associated with it, and each tag is often associated with many photos. We use the notation $\mathbb{X}_S$ to denote the set of tags that appear in any subset $\mathbb{P}_S \subseteq \mathbb{P}$ of the photo set. We also define the subset of photos associated with a specific tag as $\mathbb{P}_x \subseteq \mathbb{P}$. Accordingly, photos with the tag x in a subset $\mathbb{P}_S$ of $\mathbb{P}$ are denoted $\mathbb{P}_{S,x}$. We also define $\mathbb{U}_{S,x}$ as the set of users associated with photos in $\mathbb{P}_{S,x}$; and $\mathbb{U}_S$ as the set of all users associated with photos in $\mathbb{P}_S$. Finally, based on the location metadata associated with photos, we define the location usage distributions for each tag x as $\mathcal{L}_x \stackrel{\triangle}{=} \{\ell_p | p \in \mathbb{P}_x\}$.

Using this data we address the problem of identifying the place semantics of tags.

> *Can place semantics for a tag x be identified from the tag's location usage distribution, $\mathcal{L}_x$?*

In the rest of this work, for simplicity, we refer to those tags that have coherent place semantics as "place tags." Example place tags are `Delhi`, `Logan Airport`, `Notre Dame`, `London`, `Bath`, and `New York Marathon` (interestingly, `New York Marathon` represents both a place and an event). Examples of tags not expected to represent places are `dog`, `party`, `food`, and `blue`.

Some tags have relatively simple, unambiguous place semantics, for example, `Logan Airport` and `New York Marathon`. Other tags might refer to multiple geographic locations, with different semantics depending on the location, for example, `Palace` and `museum`. Still other tags might have place semantics as well as nonplace semantics, for example, `Bath` and `Savannah`. What place semantics we are able to find from a tag's occurrences will depend on the available data.

What semantics did the people who contributed the photos and tags intend? `Bath` outside of the U.K. should not be a place tag, while in the U.K. it will likely occur more often in the city of Bath than in random houses.

The first step in determining whether a tag has a coherent place semantic is to define these terms. We aimed for a definition that addresses both general human perception and the generic (i.e., socially common) notion of place [Aipperspach et al. 2006; Jones et al. 2001]. We propose the following.

> *Place tags* exhibit spatial usage patterns that are significantly geographically localized.

The term "significantly" in this definition is intentionally vague, designed to capture the idea that a tag's place semantic is socially defined and hence ultimately in flux, as are all signifier-signified relationships. More concretely, the definition refers to the fact that a person living in 2008 can expect `New York Marathon` to appear significantly more often in New York City than elsewhere; whereas `dog` should appear in many locations. We expect a reasonable human judge to be able to determine, for any tag and the set of photos associated with that tag, whether the tag has a coherent place semantic.

It is important to consider place tags relative to some predefined geographic area. For example, `carnival` may not exhibit any patterns worldwide, but does have spatial patterns if we are only considering the dataset of photos taken in Brazil. Similarly, `Palace` may have distinct location-based patterns in certain areas (e.g., London) but no significant patterns worldwide. For simplicity, we do not introduce notation to handle the specification of geographic areas; we generally assume that the set of photos considered by the algorithm is such that for all photos p in the set, ℓ_p is contained in a specific area of interest.

Related to areas is the concept of "scale." The basic idea is that tags may exhibit significant spatial patterns at various scales. For example, `museum` refers to specific locations within the San Francisco Bay area, while `California` is not expected to show significant patterns if our area is limited to San Francisco. Accordingly, the methods described next search for and aggregate measurements of significant patterns at multiple spatial scales.

3. PLACE SEMANTIC IDENTIFICATION

The goal of our analysis is to determine, for each tag in the dataset, whether the tag has a coherent place semantic (i.e., is a place tag). The intuition behind the various methods we present is that a coherent place semantic should reference a specific spatial region within the area of inquiry. So, the significant patterns for place tags should be manifested as a burst (like a probability distribution which is highly peaked over a small number of nearby values in its domain). It's important to note that spatial patterns are often positively autocorrelated; that is, if a region exhibits a characteristic, it is likely that nearby regions will also exhibit this characteristic. However, unlike, for example, income or disease distributions that can vary over entire areas, there is usually some boundary to a place. If we sample regions that subdivide the area of the place, it should

be possible to see fairly high autocorrelation. However, at larger granularities, namely scales, the place will be localized to a single region. It is in this latter case, when the place is localized to a single sampled region, that we apply the notion of bursts. In other words, above a certain scale, the number of usage occurrences for place tags should be much higher in a small region of the area than the number of usage occurrences of that tag outside the region. The scale of the region is one factor that these methods must address; the other factor is calculating whether the number of usage occurrences within the region is significantly higher than the number outside the region.

Another aspect of place semantics is that they should be generally first-order phenomena. In other words, the geographic location of a single tag's place semantic should be based primarily on the physical geography of the world and the socio-cultural interpretations of this geography, as opposed to the locations of other tags' place semantics. Of course, place semantics are not strictly first-order. For example, we would expect `art` and `museum` to have positively correlated place semantics. In this work, we focus on identifying place semantics using first-order data (i.e., primarily a single tag's distribution of usage). Some of the methods described in this article use the spatial distribution of all the tags in combination in addition to a single tag's distribution. (We briefly discuss how using second-order, intertag relationships should help disambiguate some tag's place semantics in the Future Work section of this article.)

In the remainder of this section, we describe the methods in detail. We first present adaptations of three existing techniques to the place semantic identification problem. Then we present the scale-structure identification methods which we developed particularly for identifying tag semantics [Rattenbury et al. 2007]. Finally, we describe a simple technique for combining the described methods into a hybrid approach. We present empirical evidence that demonstrates the superiority of hybrid methods.

All of the methods we describe perform the same generic steps.

(1) *Scale Specification*. Choose a finite set of scale values, $K = \{k_1, \ldots, k_n\}$. These scales can be specified without reference to the data, in which case we generally choose an exponentially increasing set of scales (i.e., scale k_i corresponds to a spatial range of α^i for some $\alpha > 1.0$).

(2) *Region Specification*. For each scale k define a finite set of spatial regions to search over, say $\mathbb{R}_k = \{R | \mathbb{P}_R \subseteq \mathbb{P}\}$. The simplest set of regions is a regularly spaced, square grid where grid size is based on scale. However, it is certainly feasible to use overlapping, irregularly sized, and unconnected regions.

(3) *Partial Computation*. For each scale $k \in K$ and each spatial region $R \in \mathbb{R}_k$, compute a statistic on $\mathcal{L}_x$ that captures some aspect of the tag's usage pattern in space (likely, although not necessarily, based on some relationship between the usage occurrences within R versus outside of R).

(4) *Significance Test*. Aggregate the partial computation statistics for each spatial region $R \in \mathbb{R}_k$ at each scale $k \in K$ and determine whether x is a place tag.

(5) *Identify Significant Regions*. Provided a significant pattern for x is found, determine which scales and spatial regions are referenced by x's place semantic.

Before describing each method, we introduce some additional notation. First, we use $|\mathbb{P}_{R,x}|$, the number of times tag x was used in spatial region R, as an important computational element in some of our methods. Also, while region $R \in \mathbb{R}_k$ is defined relative to some scale k, we drop the scale index for readability. Note that according to the preceding definitions, $|\mathbb{P}_{R,x}| \leq |\mathbb{P}_R|$. Finally, some of the methods that follow also require the total number of tag usage occurrences in a region: $\sum_{x \in \mathbb{X}_R} |\mathbb{P}_{R,x}|$.

3.1 Baseline Methods

At a high level, the baseline methods are based on well-known techniques for burst detection. We describe three baseline methods: naïve scan, spatial scan, and TagMaps TF-IDF.

3.1.1 *Naïve Scan Methods.* Naïve scan methods consist of an application of a standard burst detection method used in signal processing [Vlachos et al. 2004]. The method computes the frequency of usage for each region at each scale. The method identifies a burst at a specific scale when the frequency of data in a single region is larger than the average frequency of the data over all regions plus two times the standard deviation of the region frequencies.

The majority of tags in our data have sparse usage distributions, which results in low average frequencies and low standard deviations. Consequently, the standard formulation of this method generates too many false positives. To combat this problem we compute the average and standard deviation values from aggregate data: either from all of the photos or from all of the tags combined, rather than the average and standard deviation for each tag separately. We further relax the condition that the number of tag occurrences be larger than the average plus two standard deviations, instead requiring that the ratio of these values be larger than some threshold, which we can vary for optimal performance.

For Naïve Scan 1, the partial computation (step 3) for each tag x and region R (at scale k) is specified by

$$\frac{|\mathbb{P}_{R,x}|}{\mu_1 + 2\sigma_1},$$

where μ_1 is the average of $\{|\mathbb{P}_R| \mid R \in \mathbb{R}_k\}$ and σ_1 is the standard deviation of $\{|\mathbb{P}_R| \mid R \in \mathbb{R}_k\}$. To identify place tags, we compare the maximum partial computation value over all regions R and scales k to a threshold (step 4). We can vary this threshold to obtain different results, which we discuss in Section 4.3 to follow.

To identify the regions of space corresponding to a tag's place semantic (step 5 given before), we simply record the regions that pass the significance test (step 4 given before) at each scale k. Specifically, we record the region R where the partial computation statistic is larger than the threshold.

For the scale specification step (step 1), we define $K = \{k_1, \ldots, k_n\}$ such that k_i corresponds to a spatial length of 2^i (an exponential set of scales). The spatial length is used to define a square grid of regions (step 2 given earlier).

An alternative approach, which we refer to as Naïve Scan 2, compares the individual tag occurrences to the total number of tag occurrences, instead of the number of photo occurrences. The reasoning behind this modification is based on the assumption that if tag x captures the important aspects of a photo, then this photo will require few tags in addition to x.

The partial computation statistic is

$$\frac{|\mathbb{P}_{R,x}|}{\mu_2 + 2\sigma_2},$$

where μ_2 is the average, and σ_2 the standard deviation, of $\{(\sum_{x \in \mathbb{X}_R} |\mathbb{P}_{R,x}|) \mid R \in \mathbb{R}_k\}$. If every photo had the same number of tags, these results would be identical to those produced by Naïve Scan 1. However, as photos can have an arbitrary number of tags with some photos using far more tags than others, the Naïve Scan 2 method does produce (slightly) different results.

3.1.2 *Spatial Scan Methods.* Spatial scan methods comprise a standard application of the spatial scan statistic [Kulldorff 1999], a burst detection method used in epidemiology. These methods assume an underlying probability model of observing some phenomenon over some domain. The methods then test whether the number of occurrences of a phenomenon in a region of the domain (e.g., region of space) is abnormal relative to the underlying probability model. This abnormality test is performed for each region.

To illustrate how the spatial scan methods work, we describe an example from our data. Consider `Yoda`, a tag that refers to a little-known statue of the widely popular Star Wars character in the Presidio of San Francisco.[3] Suppose: (1) Over the entire San Francisco Bay area, q denotes the global probability of the tag `Yoda` being applied to any photo; (2) all M photos tagged with `Yoda` occur within a single spatial region; and (3) there are a total of N photos located within this same region. If `Yoda` is a place tag, M should be quite a bit larger than qN. Spatial scan methods are designed to test whether the value M represents a significant deviation from the global probability distribution (an important note is that q is not defined a priori, but is derived from the data.)

The expression for the partial computation statistic for Spatial Scan 1 is

$$\left(\frac{|\mathbb{P}_{R,x}|}{|\mathbb{P}_R|}\right)^{|\mathbb{P}_{R,x}|} \cdot \left(\frac{|\mathbb{P}_{R^c,x}|}{|\mathbb{P}_{R^c}|}\right)^{|\mathbb{P}_{R^c,x}|)} \cdot \left(\frac{|\mathbb{P}_x|}{|\mathbb{P}|}\right)^{-|\mathbb{P}_x|} \cdot I\left(\left(\frac{|\mathbb{P}_{R,x}|}{|\mathbb{P}_R|}\right) > \left(\frac{|\mathbb{P}_{R^c,x}|}{|\mathbb{P}_{R^c}|}\right)\right),$$

where R^c is the complement set to R (i.e., $\mathbb{P}_R \cap \mathbb{P}_{R^c} = \emptyset$ and $\mathbb{P}_R \cup \mathbb{P}_{R^c} = \mathbb{P}$) and $I(\cdot)$ is the indicator function. For details on the derivation of this expression, see Kulldorff [1999].

As in naïve scan methods, the significance test (step 4) searches for the maximum partial computation statistic value over all scales k and regions R. This

[3]We refer the reader to the Star Wars movie series by George Lucas and urge the reader to visit the statue of the Master.

maximum statistic value is tested against a threshold; tags whose maximum partial computation statistic exceeds the threshold are identified as place tags. Also, by storing those regions where the partial computation statistic is larger than the threshold, we can identify the spatial regions referred to by the tag's place semantic (step 5). Finally, as with naïve scan methods, we use an exponential set of scales and a square grid of regions (steps 1 and 2 given earlier). We describe how to set the threshold in Section 4.3.

Similar to the naïve scan 2 modification, we developed Spatial Scan 2 using the total number of tags that occur inside regions. The expression for the partial computation statistic for the Spatial Scan 2 method is

$$\left(\frac{|\mathbb{P}_{R,x}|}{\sum_{x\in\mathbb{X}_R}|\mathbb{P}_{R,x}|}\right)^{|\mathbb{P}_{R,x}|}\cdot\left(\frac{|\mathbb{P}_{R^c,x}|}{\sum_{x\in\mathbb{X}_{R^c}}|\mathbb{P}_{R^c,x}|}\right)^{|\mathbb{P}_{R^c,x}|}\cdot\left(\frac{|\mathbb{P}_x|}{|\mathbb{P}|}\right)^{-|\mathbb{P}_x|}\cdot I\left(\left(\frac{|\mathbb{P}_{R,x}|}{\sum_{x\in\mathbb{X}_R}|\mathbb{P}_{R,x}|}\right)>\left(\frac{|\mathbb{P}_{R^c,x}|}{\sum_{x\in\mathbb{X}_{R^c}}|\mathbb{P}_{R^c,x}|}\right)\right)$$

This expression differs from the expression for the Spatial Scan 1 statistic in that the number of photos in each region has been replaced with the number of tag occurrences in each region. This replacement impacts the baseline occurrence statistics used by the Spatial Scan 2 method.

In the four methods described before, we determine the regions of space for each scale independent from the actual usage distributions of the tags. It follows that these methods can only propose a priori defined regions as the locations of a place semantic. In the worst case, these regions might hide the actual location of a place semantic by splitting the usage occurrences into adjacent spatial regions, none of which are above the significance test threshold. This is a general problem known as the modifiable areal unit problem, which we discuss in Section 5. The next two methods we describe address the issue of a priori defined regions; they both generate regions based on the actual tag occurrences.

3.1.3 *TagMaps TF-IDF Method.* The TagMaps method was originally developed to automatically identify tags that are representative for each given geographical area, namely tags that uniquely define regions within the area in question. TagMaps would ideally give a high score to tags such as `Golden Gate Bridge`, `Alcatraz`, and `Yoda` which uniquely represent specific locations, landmarks, and attractions within the city. Using TF-IDF-like (term frequency, inverse document frequency) techniques, TagMaps assumes (similarly to the other methods described in this article) that tags that primarily occur in a single region and do not occur often outside this region are more representative than tags that occur diffusely over the whole area.

For example, a sample set of representative tags for San Francisco is shown in Figure 1. In Ahern et al. [2007] and Jaffe et al. [2006] we supply more details on the algorithm, and on how we extend the computation to support multiple regions and zoom levels. Using this algorithm, we had created a live visualization[4] of the world; the details and evaluation of this system can also be found in Ahern

[4]http://tagmaps.research.yahoo.com.

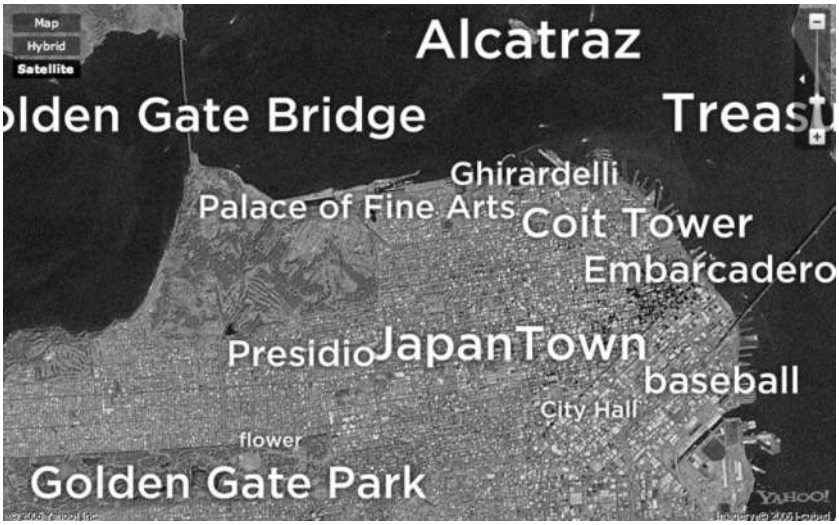

Fig. 1. Representative tags for San Francisco generated by the TagMaps method.

et al. [2007]. Here we apply TagMaps analysis to extract the spatial semantics of tags.

A key difference between the TagMaps method and the naïve scan and spatial scan methods is the use of spatial regions that are defined by the data (instead of being defined a priori). This difference also impacts how the scale specification step (step 1) happens. Whereas the naïve scan and spatial scan methods define the scales directly in the spatial domain, the TagMaps method defines its scales indirectly. Specifically, the TagMaps method starts by selecting a fixed number of photo clusters to find in the data. Generally, the fewer the number of clusters to find, the larger the associated spatial scale.

The TagMaps computation follows the computation steps listed previously. Scale specification (step 1) in this case is based on the number of clusters generated from the photo data. The region specification step (step 2) is a simple k-means clustering on the location data of all photos in $\mathbb{P}$. Geographical distance is used as the distance metric, and the stopping condition for the k-means algorithm is when every cluster's centroid moves less than 50 meters during an update. Each cluster defines a region.

Once the clusters have been determined, the system scores the tags in each cluster to identify region-specific tags. In other words, we consider each cluster R, and the set of tags $\mathbb{X}_R$ that appear with photos from the cluster. We score each tag $x \in \mathbb{X}_R$ according to the factors defined next.

The factors used for scoring a tag are modified elements of the traditional TF-IDF calculation. This method assigns a higher score to tags that have a larger frequency within a cluster compared to the rest of the area under consideration. The TF-IDF score is computed with slight deviation from its regular use in the information retrieval domain. The term frequency for a given tag x within a cluster R is the count of the number of times x was used within the cluster: $\text{tf}(R, x) \stackrel{\triangle}{=} |\mathbb{P}_{R,x}|$. The inverse document frequency for a tag

x computes the overall ratio of the tag x amongst all photos under consideration: $\text{idf}(x) \stackrel{\triangle}{=} |\mathbb{P}|/|\mathbb{P}_x|$. Note, this inverse document frequency definition is similar to the ones used in other algorithms used to score, and ultimately rank, elements [Zaragoza et al. 2007]. Recall that our total set of photos $\mathbb{P}$ is restricted to a specific area (e.g., the San Francisco Bay area). This restriction to a specific area allows us to identify local trends for individual tags, regardless of their global patterns.

The tag's score, as computed, can often be affected by a single photographer who takes a large number of photographs using the same tag. To guard against this scenario, we include a user element in our scoring that also reflects the heuristic that a tag is more valuable if a number of different photographers use it. In particular, we factor in the percentage of photographers in the cluster R that use the tag x: $\text{uf}(R, x) \stackrel{\triangle}{=} |\mathbb{U}_{R,x}|/|\mathbb{U}_R\}$.

The final score for tag x in cluster R is computed as

$$\text{Score}(R, x) = \text{tf}(R, x) \cdot \text{idf}(x) \cdot \text{uf}(R, x).$$

The higher the score and the user score, the more likely the tag is a place tag.

Again, the partial computation statistic is an aggregate value that combines the separate scores, $\text{Score}(R, x)$, from each tag and cluster. Specifically, since we are interested in identifying place semantics, we are interested in tags that are: (1) located within a single cluster (since places are spatially localized) and (2) receive a high score within the cluster. Accordingly, the partial computation statistic for the TagMaps method is

$$\sum_k \left(\sum_{R \in \mathbb{R}_k} \text{Score}(R, x) \cdot I \left(\sum_{R' \neq R} \text{Score}(R', x) == 0 \right) \right),$$

where $I(\cdot)$ is the indicator function. What this expression calculates is the sum of TF-IDF scores for a tag x at scales, k, where the tag occurred in at most one cluster. Note that both $\text{tf}(R, x)$ and $\text{uf}(R, x)$ will equal zero if tag x does not occur in region R. Of course, the larger the partial computation statistic value, the more likely that the tag is a place tag.

As with the other methods described earlier, a simple threshold can be used to determine which tags are place tags and which ones are not. And, by storing those clusters (i.e., regions) where the partial computation statistic is larger than the threshold, we can identify the spatial regions (as defined by spatial extent of the photos in the cluster) that correspond to each tag's place semantics (step 5). We describe how to set the threshold in Section 4.3.

3.2 Scale-Structure Identification Methods

Similar to the aforementioned TagMaps method, scale-structure identification methods perform a significance test (step 4) that depends on multiple scales simultaneously and does not rely on a priori defined spatial regions. However, scale-structure identification methods differ from TagMaps in the scale specification step (step 1). Whereas the scale values in the TagMaps method refer to a number of clusters, the scale values in scale-structure identification methods refer to the minimum spatial distance between clusters, as explained next.

The key intuition behind scale-structure identification methods is the following: If tag x is a place tag then the points in $\mathcal{L}_x$, the spatial usage distribution, should appear as a single cluster at many scales. The clustering mechanism used in scale-structure identification methods is proximity-based: Points that are closer together get merged before points that are further apart. This is a well-known hierarchical clustering method known as nearest neighbor or single-linkage, which is equivalent to the minimum spanning tree algorithm developed by Kruskal [1956]. It is also similar to the clustering mechanism in the scale-space method developed by Witkin [1983]. However, whereas Witkin was interested in any structure that exhibits robustness over a range of scales, we are interested in the robustness of a single type of structure: a single cluster containing a majority of the tag occurrences.

Consider the graph over the points in $\mathcal{L}_x$ where edges exist if and only if the points are closer together than d_k, where d_k is the spatial distance defined such that $k_i > k_j \iff d_{k_i} > d_{k_j}$. We compute the distance between points in $\mathcal{L}_x$ as the L_2 distance between the points as they lie on a sphere. $\mathbb{R}_k$ is then defined as the set of connected subcomponents of this graph. We can measure a number of interesting statistics on $\mathbb{R}_k$. For example, we might measure how much entropy $\mathbb{R}_k$ contains: Does it consist of a single cluster containing all the tag occurrences, or are the tag occurrences evenly split among a number of disjoint clusters/subcomponents? We could also measure the range of scales k over which $\mathbb{R}_k$ exibits a single cluster that contains the majority of tag occurrences. Yet another thing we could measure is the area delimited by the entropy of $\mathbb{R}_k$ as the scale varies (a natural combination of the two previous measurements).

Each of these measurements results in a variation of the general scale-structure identification approach. We describe each variation in more detail.

3.2.1 *Scale-Structure Identification 1.* For this method, the partial computation step (step 3 given before) computes the entropy of $\mathbb{R}_k$ at each scale k. We chose an exponential sampling method to select the scales: $d_{k_i} = \alpha^i$, $1.1 \leq \alpha \leq 5$ (see the discussion in Section 4.3). The entropy for a single scale can be written as

$$E_{k,x} \stackrel{\triangle}{=} \sum_{R \in \mathbb{R}_k} (|\mathbb{P}_{R,x}|/|\mathbb{P}_x|) \log_2(|\mathbb{P}_x|/|\mathbb{P}_{R,x}|).$$

We use the entropy value as a measurement of how similar the data is to a single cluster, since entropy increases as data becomes more distributed. We are interested in low-entropy structures, $\mathbb{R}_k$ (note that $E_{k,x} = 0$ when the usage distribution is a single cluster, i.e., $|\mathbb{R}_k| = 1$).

Accordingly, the significance test calculation (step 4) aggregates the partial computation statistics simply by summing them over the set of scales specified in the scale specification step (step 1): $\sum_k E_{k,x}$. This summed value is tested against a threshold to determine if the tag is a place tag. Unlike the naïve scan and spatial scan methods, which are interested in exceeding a threshold, the scale-structure identification 1 method is interested in tags whose summed partial computation statistics falls below a threshold; recall that low entropy corresponds to a more concentrated distribution of tag occurrences.

Finally, by recording the scale structures at each scale, we can determine which spatial regions strongly characterize a place tag (step 5). In fact, we can then characterize the tag, or rather the tag's place semantic, at multiple scales.

3.2.2 *Scale-Structure Identification 2.* For this method, the partial computation step (step 3 given previously) computes the threshold scale value, denoted by $\hat{k}$, that marks the point where the largest set of tag occurrences in $\mathbb{R}_k$ becomes "stable," that is, the set is stable for any scale value larger than $\hat{k}$. "Stable" is quantified as not changing in size by more than 10% at any point along scale dimension. This method performs the scale specification step (step 1) by finding all the spatial distances $D = \{d_1, \ldots, d_n\}$, and the corresponding scale values $K = \{k_1, \ldots, k_n\}$, such that

$$|\mathbb{R}_{k_i}| > |\mathbb{R}_{k_j}| \; \forall \; k_i < k_j;$$

in other words, all the scale values k which result in a change in the graph structure captured by $\mathbb{R}_k$ relative to smaller scale values.

To compute the threshold scale value $\hat{k}$, we first initialize its value to zero: $\hat{k} = 0$. Then, we walk through the scale values in K and assess whether each subset in $\mathbb{R}_k$ remained stable. To illustrate this step, let us assume that R_1 merges with R_2 at scale value k. Furthermore, let us assume that $|\mathbb{P}_{R_1,x}| \leq |\mathbb{P}_{R_2,x}|$. Since R_1 is not the largest existing set of tag occurrences, its stability is irrelevant (in fact, since $\frac{|\mathbb{P}_{R_2,x}|}{|\mathbb{P}_{R_1,x}|} \geq 1$, which is larger than 10%, R_1 is not stable). For R_2, if $\frac{|\mathbb{P}_{R_1,x}|}{|\mathbb{P}_{R_2,x}|} \geq 0.1$ then we would know that the threshold scale value is at least as large as k, and so we would set $\hat{k} = k$. This process continues for each $k \in K$. The final value of $\hat{k}$ is that scale value marking when the largest set of tag occurrences became stable.

The intuition behind this method is that tags whose spatial distributions tend to look like a strong single cluster (i.e., look like they reference a place from their usage) should have a core set of tag occurrences that is stable over a large range of scales. In other words, place tags should have a smaller threshold scale value, $\hat{k}$, compared to nonplace tags. Nonplace tags will likely have multiple strong clusters. When these clusters finally merge, at some scale, this merger will be significant for every cluster (assuming they are of about equal size). Hence, the threshold scale value computed by this method will be large.

For the significance test (step 4), $\hat{k}$ is tested against a threshold to determine if the tag is a place tag. Like the Scale-Structure Identification 1 method, the Scale-Structure Identification 2 method is interested in tags whose partial computation statistic value, in this case $\hat{k}$, falls below a threshold. And, by recording the scale structures at each scale, we can determine which spatial regions strongly characterize a tag's place semantic (step 5). In fact, we can then characterize the tag's place semantic at multiple scales.

3.2.3 *Scale-Structure Identification 3.* For this method, we combined the ideas behind scale-structure identification 1 and 2. The partial computation

step (step 3) for this method is based on the following calculation.

$$\int_0^\infty E_{k,x}dk = \int_0^\infty \sum_{R\in\mathbb{R}_k} \frac{|\mathbb{P}_{R,x}|}{|\mathbb{P}_x|} \log_2\left(\frac{|\mathbb{P}_x|}{|\mathbb{P}_{R,x}|}\right) dk$$

As discussed earlier, entropy is a measurement of how similar the data is to a single cluster, since it increases as the data becomes more distributed. Notice that $E_{k,x}$ is a step-wise constant function that changes value at a finite set of spatial distances $D = \{d_1, \ldots, d_n\}$, with corresponding scale values $K = \{k_1, \ldots, k_n\}$ such that

$$|\mathbb{R}_{k_i}| > |\mathbb{R}_{k_j}| \; \forall \; k_i < k_j.$$

Note: This is the same set of scales used in the aforesaid Scale-Structure Identification 2 method. If we include the extreme point $d = 0$ in D, and in the corresponding set of scales K, then we can use the equation

$$\int_0^\infty E_{k,x}dk = \sum_{i=2}^{n}(d_i - d_{i-1}) \cdot E_{k_i,x}$$

to solve for the partial computation statistic.

As with the scale-structure identification 2 method, this method simply walks through the ordered scale values in K to compute the value used in the significance test.

The intuition behind this method is that tags whose spatial distributions tend to look like a strong single cluster (i.e., look like they reference a place from their usage) should reach an entropy of zero (i.e., be merged into a single scale) fairly quickly. In other words, place tags should have a small partial computation value compared to nonplace tags. Nonplace tags will likely have multiple, nontrivial clusters. Hence, nonplace tags will have a high entropy value over a large range of scales.

Like previously described methods, the partial computation value is tested against a threshold to determine if the tag is a place tag (step 4). Like the scale-structure identification 1 and 2 methods, the scale-structure identification 3 method is interested in tags whose partial computation value falls below a threshold. And, as with all the other methods, by recording the scale structures $\mathbb{R}_k$ at each scale, we can determine which spatial regions, $R \in \mathbb{R}_k$, strongly characterize the tag's place semantic (step 5).

3.3 Hybrid Methods

Finally, we can combine the methods described previously into hybrid place semantic identification methods. The basic process for combining methods is to take a weighted sum of the normalized significant test statistics. We illustrate this process with an example.

Suppose, for each of tag $x \in \mathbb{X}$, each of the methods described earlier produces a real-valued significance test statistic: $s_m(x)$, where m references the method. Further, we assume that every method has been arranged so that a tag x is more likely to be a place tag the smaller $s_m(x)$ is, for every method m; note

that some of the aforementioned methods preferred larger values, and we can reverse this preference by multiplying each score by −1.0.

Before combining the statistics from different methods, we normalize them. Normalization is performed according to the following equation.

$$\bar{s}_m(x) \stackrel{\triangle}{=} \frac{s_m(x) - (\min_{z \in \mathbb{X}} s_m(z))}{\max_{y \in \mathbb{X}}(s_m(y) - (\min_{z \in \mathbb{X}} s_m(z)))}$$

Now, we can take the weighted average of two or more methods, and treat it as a new method. For example, given non-negative weights w_1 and w_2, we can create a new method which produces statistics, $\hat{s}(x)$, according to

$$\hat{s}(x) \stackrel{\triangle}{=} w_1 \bar{s}_1(x) + w_2 \bar{s}_2(x).$$

To simplify, we assume that the weights sum to 1.0. Note that we could take any nondecreasing transformation of each method's tag scores before combining them, and/or we could combine them nonlinearly. However, for the present article, it suffices to demonstrate a hybrid approach using a simple, linear combination of the previously described methods.

In the evaluation section, we provide results from a number of hybrid methods. Ultimately, we demonstrate that hybrid methods outperform the others methods described before.

4. EVALUATION

We implemented the methods described earlier, and performed a direct evaluation of each method's performance over part of the Flickr dataset. The goals of the evaluation were to establish whether any of the methods can reliably identify place tags, compare the performance of the different methods, and evaluate the performance with varying parameters. Finally, we seek to understand the type of errors made by the different methods.

We begin by describing the Flickr data used in our evaluation. We then provide details on how we generated the ground truth for the tags in the dataset. Finally, we discuss the results.

4.1 The Flickr Dataset

The data we use in this study consists of geotagged photos from Flickr and their associated tags. Location metadata was available for roughly fourteen million public Flickr photos at the time we collected our data. Currently, over forty million photos on Flickr now have location metadata. While the photo location could also be provided by the camera, it is more likely to be entered by the user using maps on the Flickr Web site, or possibly obtained from an external GPS device via synchronization software.

We applied several filters to improve the metadata correctness and to ensure sufficient data for the analysis. Specifically, Flickr allows an accuracy level to be assigned to the location metadata for each photo. We only used photos whose location resolution was in the two most precise levels of the 1–16 scale. In addition, to ensure sufficient tag occurrence data, we only considered tags used more than 25 times, and by more than one user. (These thresholds are

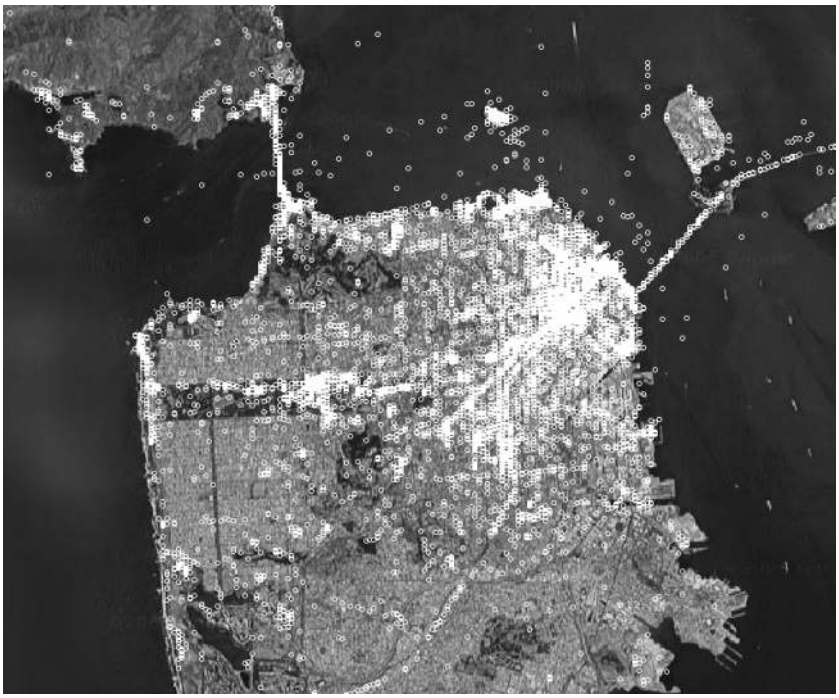

Fig. 2. Spatial distribution of all San Francisco geotagged photos in our dataset (white markers).

arbitrary and reflect simple heuristics to ensure reasonable data coverage. We did not perform any testing to assess the impact of these threshold values on the results presented in this article.)

In this work we focus our evaluation on photos from the San Francisco Bay area. We plot the location for every geotagged photo in our dataset in Figure 2. In Figure 3, we plot the location usage distribution for the tag `Hardly Strictly Bluegrass`. The San Francisco Bay area is one of the best-represented geographic regions in Flickr, increasing the likelihood of finding significant patterns at subcity and subneighborhood scales. We note, however, that restricting the dataset to a specific geographic area did not require any alterations to the methods or the evaluation computations.

After applying the filters to ensure sufficient data for each tag, our dataset consists of 49897 photos with an average of 3.74 tags per photo (standard deviation 2.62). These photos cover a total temporal range of 1015 days, starting from January 1, 2004. From these photos we extracted 803 unique tags (according to the filters described before). As expected, and similar to previous work [Dubinko et al. 2006; Golder and Huberman 2006], the number of photos for each tag ($\mathbb{P}_x$) was Zipf-distributed. The maximum number of photos associated with a single tag was 34590 (for `San Francisco`), and the mean was 232.26 (standard deviation 1305.40). Figure 4 shows the most common tags used in our dataset.

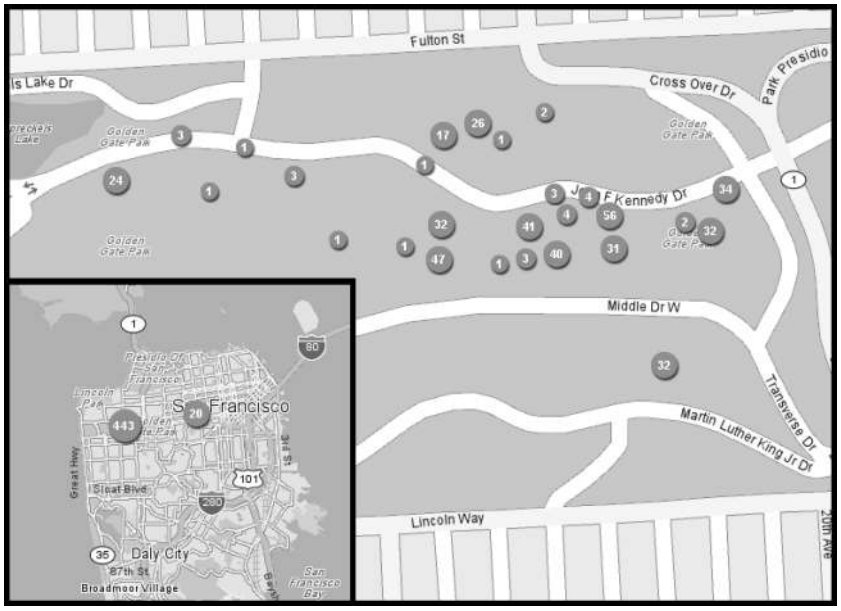

Fig. 3. Location usage distributions for the tag `Hardly Strictly Bluegrass` in the San Francisco Bay area. The zoomed-in map view shows the details of the larger location cluster from the zoomed-out view.

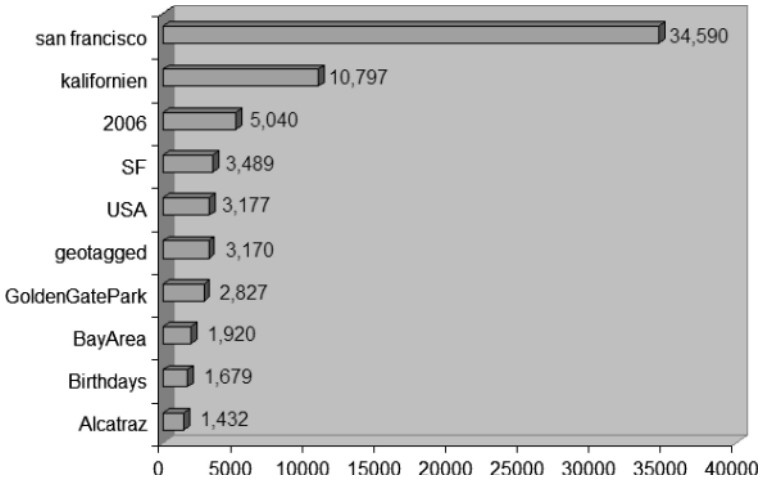

Fig. 4. Top ten tags by usage frequency in our San Francisco Bay area dataset.

The Flickr dataset is rather exciting in that it possesses many of the expected challenges for this type of computation. While Flickr popularity is rising, the number of geo-referenced photos is still relatively low. For every tag in the dataset, the spatial usage could be sparse. For example, even the set of photos tagged `Golden Gate Bridge` does not demonstrate an even distribution; some days have no `Golden Gate Bridge` photos; and there are progressively fewer

photos uploaded as you go back to the earlier data (e.g., photos from 2004). Moreover, photos with the tag `Golden Gate Bridge` tend to cluster into a number of spatially disjoint regions corresponding to popular viewing points. Another complicating factor is the fact that the data is often uneven: More photos are likely to be uploaded with the tag `Golden Gate Bridge` than `Bay Bridge`, for example. Nevertheless, we expect these types of challenges in every real-life, user-generated dataset.

In addition to sparseness and popularity, the Flickr dataset exhibits other interesting patterns. For example, local residents in a certain location, like San Francisco, are more likely to take pictures of events while tourists are more likely to take pictures of landmarks. More pertinent to the place semantic identification problem, Flickr is far more likely to contain geotagged photographs in major urban areas than in small cities or rural areas. The notable exception to this are famous landmarks like Stonehenge or Mount Rushmore.

4.2 Ground Truth

To generate the ground truth for our evaluation we manually annotated each of the 803 tags. Specifically, we looked at a sample of pictures associated with each tag in our dataset, including their locations and times of capture, to determine whether the tag is a place tag. This in-depth analysis was needed to accurately assess obscure tags (e.g., the tag `Yoda` mentioned earlier that referred to the statue in the San Francisco Presidio) and tags subject to polysemy and homonymy (e.g., `Apple` in San Francisco was mostly assigned to photos of the Apple Computer store). Examining the content of the photographs was often required; from the photo and caption content we were often able to assess the intended semantics of the tags associated with the photo, and hence whether they were associated with specific geographical locations.

To measure the discrepancy between common-sense interpretations of the tags in our dataset and the ground truth, we also collected a set of labels for the tags generated by having four judges vote, without access to the photos or their metadata, on whether the tag referred to a place. Our goal was to assess whether the tag name by itself, that is, without any details on the photos or locations where that tag was applied, was sufficient for determining if the tag was a place tag. Interestingly, the vote-based data exhibited systematic errors relative to the ground-truth data: (1) Obscure or unpopular place tags were often false negatives (i.e., incorrectly labeled as not being place tags); (2) generic tags like `park` were often false positives (while they have clear place semantics within a limited scope, over the whole dataset they did not refer to specific spatial regions); and (3) tags that superficially refer to temporal events like `Future of Web Apps` were often not labeled as place tags, even though many such events also occur in specific regions of space. In terms of interjudge agreement, 503 tags were voted as nonplace tags by all judges, 50 (6.2%) tags were voted as place tags by only one judge, 39 (4.9%) tags were voted as place tags by only two judges, 58 (7.2%) tags were voted as place tags by three judges, and the remaining 103 tags were voted as place tags by all of the judges.

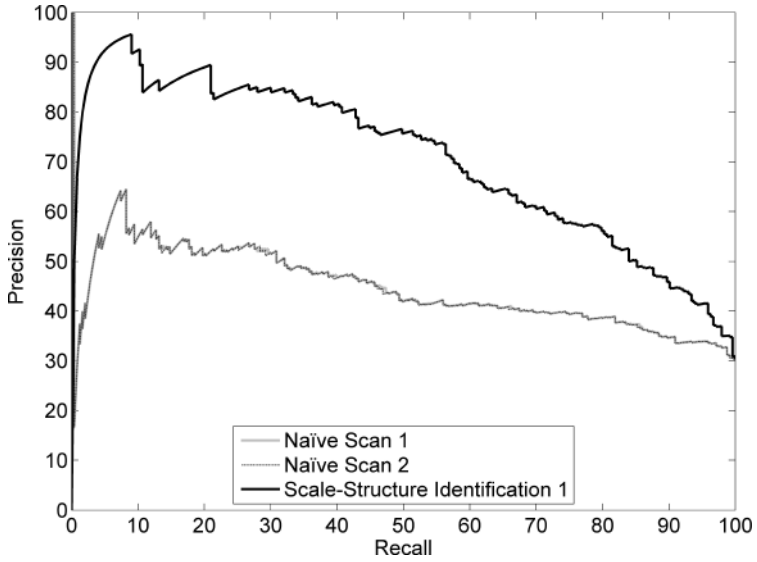

Fig. 5. Precision vs. recall for Naïve Scan 1 and 2 as well as Scale-Structure Identification 1. Scale-Structure Identification 1 is used as a baseline for comparison between the results figures. Note that the two naïve scan methods produced curves that are almost identical.

4.3 Results

Since all of the methods produce ranked results, we can use standard IR metrics to evaluate performance. For each tag, the methods produce a score that indicates how likely the tag is to be a place tag. Rather than choosing a single threshold for each method to categorize the tags, we can vary the threshold dynamically and examine the trade-off in terms of recall and precision for each method.

Plots of the precision versus recall curves are shown in Figures 5–10. The x-axis represents a recall value: the percentage of place tags (according to the ground-truth labels) that are identified as place tags by the algorithm in question. The thresholds for each method were adjusted to produce the recall values. The y-axis shows the precision: the percentage of tags identified as place tags by the algorithm that are actually place tags (according to the ground-truth labels). For example, Figure 5 demonstrates that when the threshold for scale-structure identification 1 is set so that the algorithm identifies half of the place tags (recall is 50%), then 76% of the identified place tags are correctly labeled as place tags according to our ground truth (76% precision).

In every figure, we include the curve for the Scale-Structure Identification 1 (SSI1) method as a reference for comparison, since plotting every curve in a single figure would be illegible. Figure 5 plots both naïve scan (NS1 and NS2) methods as well as the SSI1 method. Both NS1 and NS2 perform worse, in terms of precision, for every possible recall value. Figure 6 plots both spatial scan (SS1 and SS2) methods as well as the SSI1 method. While both SS1 and SS2 outperform SSI1 at low recall values (less than 4%), meaning that the top

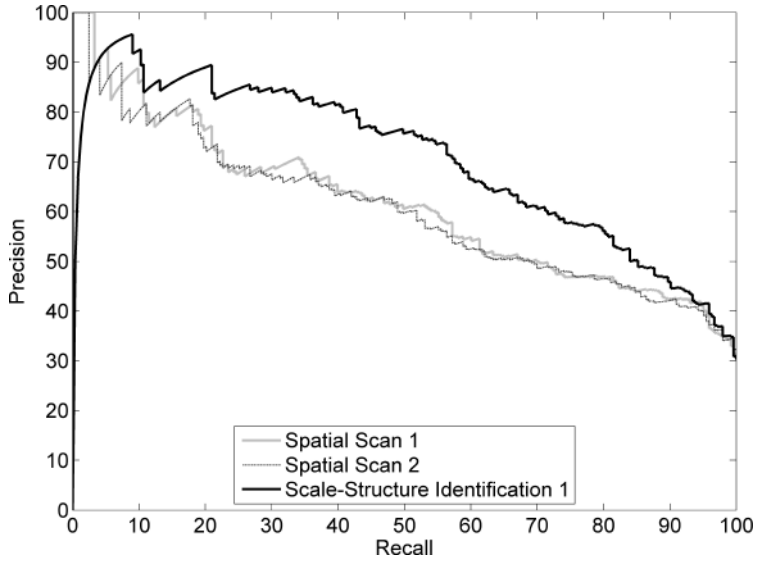

Fig. 6. Precision vs. recall for Spatial Scan 1 and 2 as well as Scale-Structure Identification 1. Scale-Structure Identification 1 is used as a baseline for comparison between the results figures.

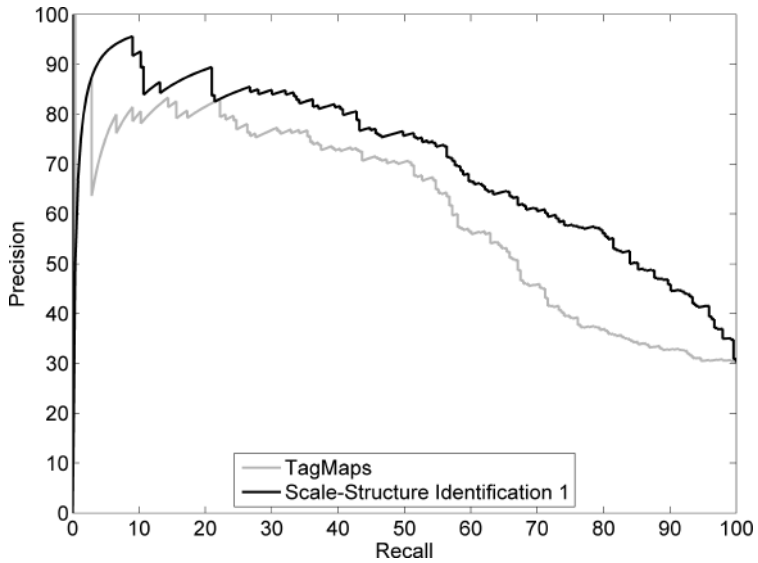

Fig. 7. Precision vs. recall for TagMaps and Scale-Structure Identification 1. Scale-Structure Identification 1 is used as a baseline for comparison between the results figures.

few tags scored by SS1 and SS2 are correctly labeled as place tags relative to the group-truth data while the top few tags of SSI1 are false positives, SSI1 outperforms both SS1 and SS2 over the remaining recall values. Figure 7 plots the TagMaps (TM) method results relative to SSI1. Like NS1 and NS2, the TM method consistently performs worse than the SSI1 over all recall values;

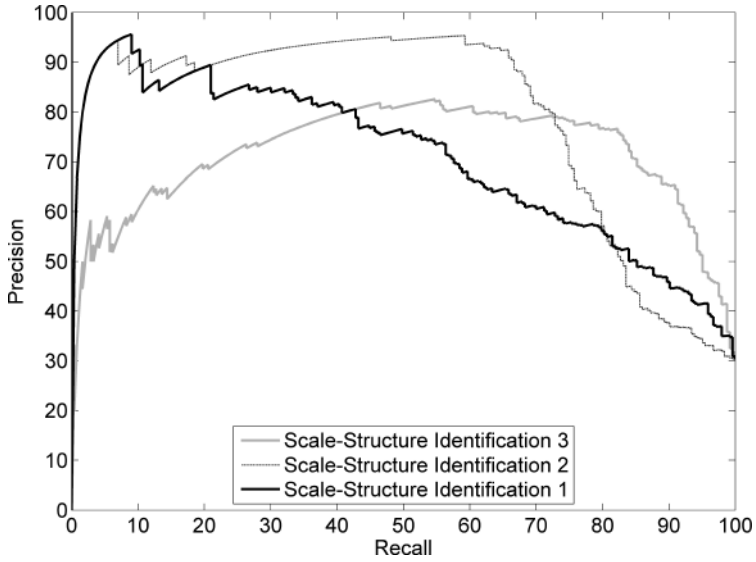

Fig. 8. Precision vs. recall for Scale-Structure Identification 1, 2, and 3. Scale-Structure Identification 1 is used as a baseline for comparison between the results figures.

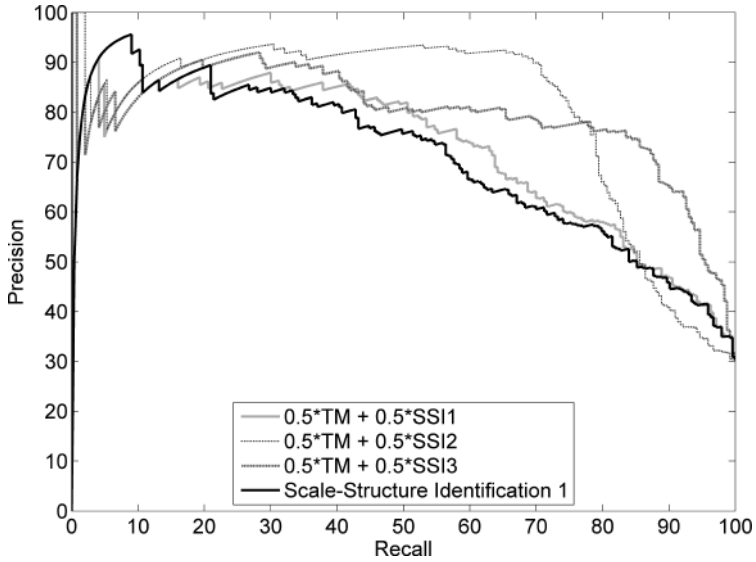

Fig. 9. Precision vs. recall for hybrid methods that combine TagMaps with Scale-Structure Identification 1, 2, and 3. Scale-Structure Identification 1 is used as a baseline for comparison between the results figures.

however, the TM method results are much closer to SSI1's than either NS1 or NS2.

Figure 8 plots the results for all three scale-structure identification methods (SSI1, SSI2, and SSI3). Interestingly, there is a large variation in the performance of these three methods. SSI2 and SSI3, like SSI1, perform rather poorly

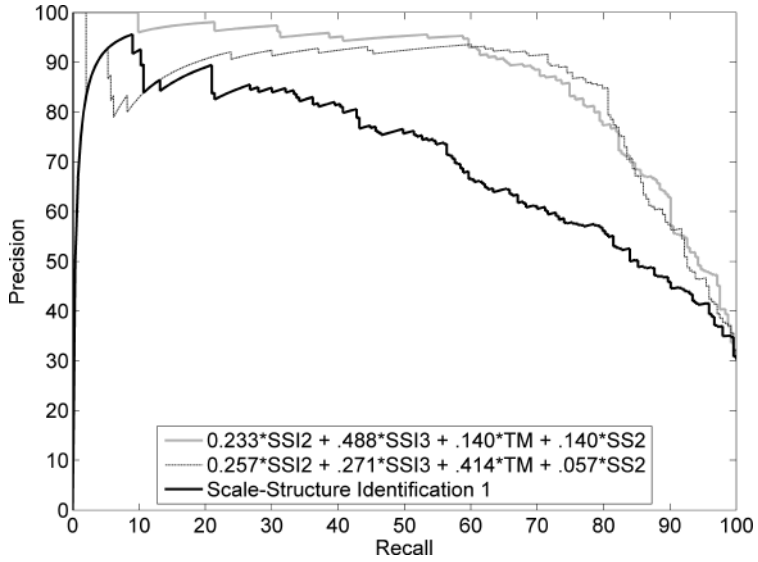

Fig. 10. Precision vs. recall for hybrid methods optimized for P-R area (the first curve in the legend) and for maximum F1 value (the second curve in the legend). The actual P-R area and F1 values are shown in Table II. Scale-Structure Identification 1 is used as a baseline for comparison between the results figures.

over small recall values, indicating that the top-scoring tags in these methods are false positives (i.e., labeled as place tags when the group-truth data indicates that they are not actually place tags). Over the middle range of recall values, 20%–70%, SSI2 performs the best; but its performance falls below that of SSI1 for high recall values. SSI3, on the other hand, performs the best over high recall values, greater than 70%. The SSI1-3 results, demonstrating variable performance over different ranges of recall values, indicate that some form of hybrid, mixture method should be superior.

Figure 9 plots the results of hybrid methods that each combine TagMaps and a scale-structure identification method. Again, SSI1 is shown for reference. Interestingly, by combining TagMaps with the scale-structure identification methods, we can consistently improve the performance of each SSI method for mid to high recall values, greater than 25%. However, over smaller recall values, the hybrid methods perform worse than the SSI methods (with the exception of SSI3, whose performance at low recall values is greatly improved).

Finally, Figure 10 plots the results of hybrid methods combining TagMaps, scale-structure identification methods 2 and 3, and Spatial Scan 2. The mixture weights of these hybrid methods were chosen to maximum the P-R curve area in one hybrid, and to maximize the F1 test statistic in the other hybrid (see the discussion to follow for definitions of P-R area and the F1 test statistic). The weights were found using randomized cross-validation. Ten training-testing datasets were created from the original dataset; each training-testing dataset was created by randomly splitting the original dataset into two equal halves. Using an iterative, refinement search, the mixture weights were chosen that

Table I. (Precision, Recall) Values for Different Numbers of Returned Tags

	top 50	top 100	top 200
Naïve Scan 1 (NS1)	0.58, 0.12	0.52, 0.21	0.47, 0.39
Naïve Scan 2 (NS2)	0.58, 0.12	0.52, 0.21	0.47, 0.39
Spatial Scan 1 (SS1)	0.82, 0.17	0.68, 0.28	0.60, 0.49
Spatial Scan 2 (SS2)	0.80, 0.16	0.69, 0.28	0.61, 0.50
TagMaps (TM)	0.80, 0.16	0.76, 0.31	0.67, 0.55
Scale-Structure Ident. 1 (SSI1)	0.88, 0.18	0.83, 0.34	0.70, 0.58
Scale-Structure Ident. 2 (SSI2)	0.90, 0.19	0.94, 0.39	0.84, 0.69
Scale-Structure Ident. 3 (SSI3)	0.64, 0.13	0.75, 0.31	0.80, 0.65
Vote-Based	0.84, 0.17	0.83, 0.34	0.70, 0.58
0.5*SSI1 + 0.5*TM	0.86, 0.18	0.85, 0.35	0.74, 0.60
0.5*SSI2 + 0.5*TM	0.90, 0.19	0.91, 0.37	0.87, 0.72
0.5*SSI3 + 0.5*TM	0.90, 0.19	0.89, 0.37	0.80, 0.65
0.233*SSI2 + .488*SSI3 + .140*TM + .140*SS2	0.94, 0.38	0.86, 0.70	0.56, 0.91
0.257*SSI2 + .271*SSI3 + .414*TM + .057*SS2	0.91, 0.38	0.89, 0.73	0.56, 0.92

The mixture weights of the last two hybrid methods were set using randomized cross-validation. The values in the table for these methods are averages over the testing portion of cross-validation datasets. The standard deviations for these values range from 0.014625–0.047378.

maximized the averaged results over all training sets. The results presented in the tables that follow are from the testing portion of the datasets.

The first hybrid demonstrates superior performance to SSI1 over the entire range of recall values, especially over low recall values. The second hybrid method, while underperforming over low recall values, achieves the highest combination of precision and recall (i.e., is the closest to the top-right corner of the precision-recall plot). The top-right corner of the precision-recall plot, corresponding to 100% precision at 100% recall, would be perfect performance for a classification method. The second hybrid achieves about 85% precision at 81% recall.

As an alternative to searching for optimal threshold values for the methods, we can simply take the top N results from the ordered lists produced by the methods (where N is variable). Table I shows precision and recall values for $N = 50$, 100, and 200.

From the precision-recall curves we computed: (1) the area under the curve (P-R Area); (2) the maximum value of the F1 statistic[5] for each method (Max F1), a metric that balances precision and recall values; and (3) the minimum total classification error[6] (Min CE) (see Cai and Hofmann [2003] for more discussion of these metrics). The results are shown in Table II. These metrics demonstrate the superior performance of the hybrid methods.

We also studied the sensitivity of the scale-structure identification 1 method to the scale specification step (step 1 in Section 3.1). We varied the exponential base in the scale sampling scheme from 1.1 to 5.0 (i.e., scale k_i corresponds to a spatial distance of α^i, where α ranged between 1.1 and 5.0). The results

[5]The F1 statistic is defined as $\frac{2pr}{p+r}$, where p is the precision and r is the recall.

[6]The total classification error is defined as $(\frac{N_p}{N}) \cdot (1 + \frac{r}{p} - 2r)$, where N_p is the total number of place tags (according to the group-truth data), N is the total number of tags, and p and r are precision and recall, respectively.

Table II. Precision-Recall Area, Maximum F1, and Minimum CE Values for the Various Methods

	P-R Area	Max F1	Min CE
Naïve Scan 1 (NS1)	0.4455	0.5279	0.2914
Naïve Scan 2 (NS2)	0.4458	0.5279	0.2914
Spatial Scan 1 (SS1)	0.6028	0.5907	0.2441
Spatial Scan 2 (SS2)	0.6134	0.5955	0.2416
TagMaps (TM)	0.6050	0.6018	0.2130
Scale-Structure Ident. 1 (SSI1)	0.7034	0.6655	0.1930
Scale-Structure Ident. 2 (SSI2)	0.7888	0.7692	0.1196
Scale-Structure Ident. 3 (SSI3)	0.7073	0.7937	0.1295
Vote-Based	0.6797	0.6586	0.1843
0.5*SSI1 + 0.5*TM	0.7229	0.6770	0.1806
0.5*SSI2 + 0.5*TM	0.8048	0.7908	0.1133
0.5*SSI3 + 0.5*TM	0.7938	0.7976	0.1283
0.233*SSI2 + .488*SSI3 + .140*TM + .140*SS2	**0.8664**	0.7998	0.1166
0.257*SSI2 + .271*SSI3 + .414*TM + .057*SS2	0.8335	**0.8258**	**0.1021**

The best values for each metric are in bold. The mixture weights of the last two hybrid methods were set using randomized cross-validation. The values in the table for these methods are averages over the testing portion of cross-validation datasets. The standard deviations for these values range from 0.009734–0.025120.

were robust to these changes. One point to note, however, is that performance slightly but consistently improved as the exponential base decreased (0.035 increase in the P-R area). In other words, the scale-structure identification 1 method performed better with denser samplings of the space of scale values, but only slightly.

Results of the region identification step (step 5 in Section 3.1) are straightforward. To summarize, the regions of space that were associated with identified place tags were accurate. The only systematic errors found were due to sparse, wrong, or missing data. For example, tags like `Bay to Breakers` had spatial usage distributions that were not representative of the true spatial expanse of the referenced social gathering: Pictures tended to be taken more at the beginning and end of the race, leaving the middle parts too sparse to accurately identify as a single, connected place.

In terms of error analysis, we identified common errors with the methods. We will discuss the false positive and false negatives of many of the methods described in this article. For the TM-SSI1 (TagMaps, Scale-Structure Identification 1) hybrid, the false positives were all due to missing or sparse data. Interestingly, this "lack of data" problem appears in two forms. First, there are tags that refer to specific events like `August 2006`, which, due to the limited amount of data, appear to occur in a specific spatial region, when in fact more data would demonstrate that this specific event actually occurs in multiple regions. Second, generic tags like `orchids` and `baby shower` are only represented in single spatial regions in our data, whereas more data would demonstrate that these tags refer to generic entities that can, and do, occur in multiple spatial regions. In terms of false negatives, the TM-SSI1 hybrid missed specific place tags (e.g., `Coit Tower` and `Chinatown`) in the San Francisco Bay area for one of two reasons. Either there was not enough data, so the spatial usage

distribution of the tag appears as a number of disjoint clusters which would be connected if more data was available, or the place referenced by the tag was visible from a number of locations (e.g., `Coit Tower` or `Golden Gate Bridge`), making it impossible to automatically identify as a single spatial region. The other hybrid methods, TM-SSI2, TM-SSI3, and TM-SSI2-SSI3-SS2s, exhibit very similar false positive and false negative error characteristics.

Interestingly, TM, NS1, NS2, SS1, and SS2 all exhibited another type of false negative error (relative to the SSI and hybrid methods). Specifically, tags whose spatial usage distributions deviated significantly from a single circular or square region (which are the types of regions that these methods search over) were incorrectly labeled as not being place tags. Examples include `Golden Gate Park`, which is basically a long rectangle, and `Embarcadero`, which is a curved street on the waterfront which many people walk along. In addition to the clustering mechanism in SSI, which extends the nearest-neighbor or single-linkage hierarchical clustering algorithm, further research has been done to handle nonconvex regions [Ng et al. 2001], and could be incorporated into new place semantic identification methods.

Additionally, TagMaps has some false negatives which result from its clustering mechanism to specify possible spatial regions of interest. The problem is that the cluster locations will be dominated by popular places. Since many clusters will have a number of tags that are specific to them, inverse document weighting will not be able to reveal small, relatively unknown tags like `Club Neon` and `Bring Your Own Big Wheel`. They will be hidden at all but the smallest scales, and hence receive relatively low overall scores. However, this problem can be overcome by mixing the TagMaps methods with other methods, as demonstrated in this article.

Overall, we believe that hybrid methods hold the most promise for automatic extraction of place semantics. The simple, linear mixtures between TagMaps, scale-structure identification 2 and 3, and Spatial Scan 2 are the best for our dataset. We speculate that nonlinear combinations of the methods described in this work could yield even better results; however, linear mixtures are sufficient to demonstrate the potential benefit of a hybrid method.

5. RELATED WORK

We address related work from a number of relevant research areas, including event detection in time-stamped data, location-based analysis of spatially distributed data (often referred to as spatial analysis), and analysis of tagging systems.

Many scientific domains have studied the general problem of time-based event detection. Time-series analysis techniques such as ARIMA [Box and Jenkins 1976; McDowall et al. 1980] analyze trends in time-series data with the goals of: (1) explaining spikes and valleys over various time windows and (2) producing future trend forecasts. While we are not addressing event detection in this article, we note that the general problem of explaining data trends in the time domain can be extended to the spatial domain: Specifically, methods that can effectively detect bursts in the temporal domain can often be extended

to detect bursts over the spatial domain [Jones et al. 2001; Kleinberg 2003]. In particular, our naïve scan methods (see Section 3) are similar to previous work on global event detection in Web query logs [Vlachos et al. 2004] and access logs [Guralnik and Srivastava 1999]. The general idea of the naïve scan methods is to assume that a coherent place semantic coincides with a burst of data in the spatial domain (e.g., many people take pictures of the Ferry Building in San Francisco, generating a concentration of photos in that particular location).

The primary issue that must be addressed in detecting bursts is how to define the spatial extent of the burst, namely its area. This issue is well known in the spatial analysis and geography literature and is referred to as the Modifiable Areal Unit Problem, or MAUP [Openshaw 1984]. The basic concern is that analytical results of spatially distributed data, in particular analyses that produce aggregate descriptive statistics, are sensitive to the definition of spatial units. For example, looking at burglary rates with counties as the spatial unit is problematic because counties with higher populations should naturally exhibit higher raw crime numbers. Common methods for dealing with this problem are to normalize the data by obvious independent variables like population or population density. Another approach is to define the spatial units to equalize these variables (e.g., breaking counties into smaller units with equal populations).

A common technique for dealing with MAUP is to define multiple regular grids over the total spatial extent of the dataset being analyzed, where each grid differs by its scale. The analysis is then performed at every grid point [Openshaw et al. 1987]. This method is known as the Geographical Analysis Machine. Our naïve scan and spatial scan methods utilize this multiscale, regular grid search method. To better deal with the MAUP problem, the spatial scan methods use the baseline data rates in each spatial unit, namely region, to calculate the significance test statistics.

In terms of the specific problem of place semantic identification, recent efforts in ubiquitous computing systems have attempted to identify meaningful locations and places from GPS and other location tracking data [Aipperspach et al. 2006]. The general approach in that work is to search for accumulations of data points in fixed locations. These accumulations can be interpreted as bursts, relative to other locations that have few or no data points. In epidemiology, efforts to identify and localize disease outbreaks [Kulldorff 1999] are closely related to the place semantic identification problem we address in this article. Our spatial scan methods described earlier borrow directly from disease/outbreak analysis, where data is sparse and dependent on the underlying population statistics (two properties exhibited by our Flickr tag data).

To specify the geographic locations of a tag's place semantic we relied on either the a priori specified regions (in the case of the naïve scan and spatial scan methods) or the clusters created by the tag occurence data points (in the TagMaps and scale-structure identification methods). An alternative approach would be to smooth the original data points to create potentially more robust place semantic descriptions. Recent methods in the geographical information systems literature have looked at methods for smoothing raw data points to

create continuous distributions, with the advantage of creating summary statistics that are less senstitive to high-frequency noise in the data [Brunsdon 1995; Brunsdon et al. 2002]. The basic idea of these methods is to replace the data points with continuous kernal functions, often Gaussian probability distributions, which are then summed to create a single distribution for the entire dataset (basically a two-dimension convolution). Choosing an appropriate kernal function radius is important because values that are too small will preserve high-frequency noise, while values that are too large will hide the spatial structure of the data [Brunsdon 1995]. More recently, researchers have studied the effects of changing the kernal function radius based on the actual data distribution to create more localized summary statistics [Brunsdon et al. 2002]. Data smoothing could be incorporated into place semantic identification methods as a preprocessing step, potentially improving problems arising from the Modifiable Areal Unit Problem.

Other geographical information systems research has looked at defining place semantics using field-of-view information in addition to the GPS location associated with photos [Epshtein et al. 2007]. The basic idea is that most photos' field of view covers some surface area heading in a specific direction from the location where the photo was taken. By overlapping these areas from multiple photos, important locations (i.e., places) can be identified.

While the aforementioned work has mostly studied place semantic identification using spatial distributions of data points, the field of GIR (Geographic Information Retrieval) has studied how to derive place semantics using only terms and place names. Two research directions from GIR are relevant to this article. First, attempts were made (e.g., Amitay et al. [2004], Buyukokkten et al. [1999], Ding et al. [2000], and Wang et al. [2007]) at extracting geographic information for a Web page, based on the page links and network properties, as well as geographic terms that appear on the page. Our system described here could potentially help these systems by identifying additional geographic/location terms. The second related research effort in GIR focuses on extracting the scope of geographic terms or entities based on co-occurring text and derived latitude-longitude information [Arampatzis et al. 2004; Purves et al. 2005; Zhou et al. 2007]. With geo-annotated photos and tags, as well as any system with direct location annotation, the potential exists not only to delineate known geographic terms, but also to identify new regions of interest based on the data.

Finally, we discuss related work on tagging systems. Most of the prior research has looked at describing tagging systems [Ames and Naaman 2007, Marlow et al. 2006], or studying trends and properties of various systems [Golder and Huberman 2006]. Some efforts have looked at extracting ontologies (or structured knowledge) from tags [Schmitz 2006]: a similar goal to ours, yet using co-occurrence and other text-based tools that could augment the methods analyzed in this article. Other tagging work has looked at semi-automatic photo annotation [Davis et al. 2004; Sarin et al. 2007].

More directly related to this work are research efforts that analyzed Flickr tags (and other terms associated with Flickr photos) together with photo

location and time metadata [Dubinko et al. 2006; Jaffe et al. 2006]. These projects applied ad hoc approaches to determine "important" tags within a given region of space based on intertag frequencies. However, no determination of the properties or semantics of specific tags was provided. Naaman et al. [2003] created spatial models for terms appearing in geo-referenced photograph labels, but did not detect the location properties of specific terms.

6. CONCLUSIONS AND FUTURE WORK

In this work, we have taken a first step in showing that some semantics can be assigned to free-form tags or text labels using the usage distributions of each tag. The ability to extract semantics can improve current tagging systems, for instance, by allowing more powerful search and disambiguation mechanisms. Additionally, the knowledge that these methods extract can help with tasks outside the scope of the specific system.

In particular, we have shown that location metadata associated with photos and their tags enables the extraction of place semantics. This mapping of tags to geographic locations could improve image search, serve as a basis for collection visualization, and assist in other photo-related tasks. This type of knowledge can also help create a gazetteer for places and landmarks that could be used for various tasks beyond photo management [Jones et al. 2001]. We plan to revisit the image search, visualization and gazetteer deployment in future work. We show in Rattenbury et al. [2007] that similar techniques can be successfully applied to extract event semantics using photos' time-stamp patterns. Using metadata distributions in other domains could potentially allow us to mine yet other types of semantics from tags.

We would also like to extend our current system to handle multi-area problems. As mentioned previously, `Palace` may exhibit place semantics in London, but perhaps not in most cities. To handle this characteristic of the data, we stated earlier that the data analysis should be limited to specific geographic areas. Ideally, we could simultaneously generate, store, and disambiguate tag semantics for different areas throughout the world. A number of arbitrary area specifications could be used: country boundaries, language geographies, city boundaries, etc. We could also consider reversing the process and choose a tag with a specific geographical place semantic and then extend the area around this location until we encounter another place semantic for the tag.

Tags with multiple place semantics (i.e., associations with different geographical locations) in different areas present interesting disambiguation problems. Clearly, `Palace` means something different in San Francisco than it does in London or Paris. We could use the co-occurrence of tags to create higher-order place semantics to disambiguate (e.g., `Fine Art`, `Buckingham`, or `Versailles`).

It is worth noting that there is likely a difference between identifying place semantics in urban areas like the San Francisco Bay area versus rural or less populated urban areas. We did not test this in the current article. However, it would be a logical step for future work.

Finally, we plan to deploy our methods to other temporally and spatially encoded data, as it becomes pervasively available on the Web.

ACKNOWLEDGMENTS

We thank S. Ahern, S. King, R. Nair, N. Good and M. Slaney for their valuable insights, comments, and assistance in the initial develpments of this work. We also thank the reviewers for their valuable comments and pointers to related work.

REFERENCES

AHERN, S., NAAMAN, M., NAIR, R., AND YANG, J. H.-I. 2007. World Explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the Conference on Digital Libraries (JCDL)*. ACM, New York, 1–10.

AIPPERSPACH, R., RATTENBURY, T., WOODRUFF, A., AND CANNY, J. 2006. A quantitative method for revealing and comparing places in the home. In *Proceedings of the International Conference on Ubiquitous Computing (Ubicomp)*. Springer.

AMES, M. AND NAAMAN, M. 2007. Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

AMITAY, E., HAR'EL, N., SIVAN, R., AND SOFFER, A. 2004. Web-a-Where: Geotagging Web content. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR)*. ACM Press, 273–280.

ARAMPATZIS, A., VAN KREVELD, M., REINBACHER, I., CLOUGH, P., JOHO, H., SANDERSON, M., JONES, C. B., VAID, S., BENKERT, M., AND WOLFF, A. 2004. Web-Based delineation of imprecise regions. In *Proceedings of the Workshop on Geographic Information Retrieval*.

BOX, G. AND JENKINS, G. 1976. *Time Series Analysis: Forecasting and Control*. Cambridge University Press.

BRUNSDON, C. 1995. Estimating probability surfaces for geographical point data: An adaptive kernel algorithm. In *Comput. Geosci. 21*, 7, 877–894.

BRUNSDON, C., FOTHERINGHAM, A., AND CHARLTON, M. 2002. Geographically weighted summary statistics: A framework for localized exploratory data analysis. In *Comput. Environm. Urban Syst. 26*, 501–524.

BULTERMAN, D. C. 2004. Is it time for a moratorium on metadata? *IEEE MultiMedia 11,* 4 (Oct.), 10–17.

BUYUKOKKTEN, O., CHO, J., GARCIA-MOLINA, H., GRAVANO, L., AND SHIVAKUMAR, N. 1999. Exploiting geographical location information of Web pages. In *Proceedings of the Workshop on Web Databases (WebDB)*. Held in conjunction with ACM SIGMOD'99. http://dbpubs.stanford.edu/pub/1999-4.

CAI, L. AND HOFMANN, T. 2003. Text categorization by boosting automatically extracted concepts. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 182–189.

DAVIS, M., KING, S., GOOD, N., AND SARVAS, R. 2004. From context to content: Leveraging context to infer media metadata. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 188–195.

DING, J., GRAVANO, L., AND SHIVAKUMAR, N. 2000. Computing geographical scopes of Web resources. In *Proceedings of the 26th International Conference on Very Large Databases*. Morgan Kaufmann, 545–556.

DUBINKO, M., KUMAR, R., MAGNANI, J., NOVAK, J., RAGHAVAN, P., AND TOMKINS, A. 2006. Visualizing tags over time. In *Proceedings of the 15th International Conference on World Wide Web (WWW)*. ACM Press, New York, 193–202.

EPSHTEIN, B., OFEK, E., WEXLER, Y., AND ZHANG, P. 2007. Hierarchical photo organization using geo-relevance. In *Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. ACM, 1–7.

GOLDER, S. A. AND HUBERMAN, B. A. 2006. Usage patterns of collaborative tagging systems. *J. Inf. Sci. 32*, 2, 198–208.

GURALNIK, V. AND SRIVASTAVA, J. 1999. Event detection from time series data. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM Press, New York, 33–42.

JAFFE, A., NAAMAN, M., TASSA, T., AND DAVIS, M. 2006. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR)*. ACM Press, New York, 89–98.

JONES, C., ALANI, H., AND TUDHOPE, D. 2001. Geographical information retrieval with ontologies of place. In *Proceedings of the Conference on Spatial Information Theory*. Vol. 2205. Springer, 322–335.

KLEINBERG, J. 2003. Bursty and hierarchical structure in streams. *Data Mining Knowl. Discov. 7,* 4, 373–397.

KRUSKAL, J. B. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. In *Proc. Amer. Math. Soc. 7*, 1, 48–50.

KULLDORFF, M. 1999. Spatial scan statistics: Models, calculations, and applications. In *Scan Statistics and Applications*, Glaz and Balakrishnan, eds., Springer, Boston, Birkhauser, 303–322.

MARLOW, C., NAAMAN, M., BOYD, D., AND DAVIS, M. 2006. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the 7th Conference on Hypertext and Hypermedia*. ACM, 31–40.

MCDOWALL, D., MCCLEARY, R., MEIDINGER, E. E., AND JR., R. A. H. 1980. *Interrupted Time Series Analysis*. Sage University PaperSeries on Quantitative Applications in the Social Sciences.

NAAMAN, M., PAEPCKE, A., AND GARCIA-MOLINA, H. 2003. From where to what: Metadata sharing for digital photographs with geographic coordinates. In *Proceedings of the 10th International Conference on Cooperative Information Systems (CoopIS)*. Springer, Berlin, 196–217.

NG, A., JORDAN, M., AND WEISS, Y. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*. Vol. 14.

OPENSHAW, S. 1984. *The Modifiable Areal Unit Problem: Concepts and Techniques in Modern Geography*. Geo Books, Norwich.

OPENSHAW, S., CHARLTON, M., WYMER, C., AND CRAFT, A. 1987. A mark 1 geographical analysis machine for the automated analysis of point data sets. *Int. J. Geograph. Inf. Syst. 1*, 4, 335–358.

PURVES, R., CLOUGH, P., AND JOHO, H. 2005. Identifying imprecise regions for geographic information retrieval using the web. In *Proceedings of the Conference GISRUK*.

RATTENBURY, T., GOOD, N., AND NAAMAN, M. 2007. Towards automatic extraction of event and place semantics from Flickr tags. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, 103–110.

SARIN, S., NAGAHASHI, T., MIYOSAWA, T., AND KAMEYAMA, W. 2007. Exploiting users' personal and public information for personal photo annotation. In *Proceedings of the IEEE International Conference on Multimedia*. IEEE, 564–567.

SCHMITZ, P. 2006. Inducing ontology from Flickr tags. In *Proceedings of the Workshop on Collaborative Web Tagging at WWW2006*.

VLACHOS, M., MEEK, C., VAGENA, Z., AND GUNOPULOS, D. 2004. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM Press, New York, 131–142.

WANG, C., WANG, J., XIE, X., AND MA, W.-Y. 2007. Mining geographic knowledge using location aware topic model. In *Proceedings of the ACM Workshop on Geographical Information Retrieval*. ACM, 65–70.

WANG, L., WANG, C., XIE, X., FORMAN, J., LU, Y., MA, W.-Y., AND LI, Y. 2005. Detecting dominant locations from search queries. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 424–431.

WITKIN, A. 1983. Scale space filtering. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

ZARAGOZA, H., RODE, H., MIKA, P., ATSERIAS, J., CIARAMITA, M., AND ATTARDI, G. 2007. Ranking very many typed entities on Wikipedia. In *Proceedings of the ACM Conference on Information and Knowledge Management*. ACM, 1015–1018.

ZHOU, C., FRANKOWSKI, D., LUDFORD, P., SHEKHAR, S., AND TERVEEN, L. 2007. Discovering personally meaningful places: An interactive clustering approach. *ACM Trans. Inf. Syst. 25*, 3, 1–31.

Received December 2007; revised June 2008; accepted August 2008