Towards Automatic Extraction of Event and Place Semantics from Flickr Tags

Tye Rattenbury*, Nathaniel Good† and Mor Naaman
Yahoo! Research Berkeley
Berkeley, CA, USA
(tye, ngood, mor)@yahoo-inc.com

ABSTRACT

We describe an approach for extracting semantics of 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 and event semantics for tags that are assigned to photos on Flickr, a popular photo sharing website that supports time and location (latitude/longitude) metadata. We analyze two methods inspired by well-known burst-analysis techniques and one novel method: Scale-structure Identification. We evaluate the methods on a subset of Flickr data, and show that our Scale-structure Identification method outperforms the existing techniques. 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

Keywords: tagging systems, event identification, place identification, tag semantics, word semantics

1. INTRODUCTION

User-supplied "tags", textual labels assigned to content, have been a powerful and useful feature in many social media and Web applications (e.g. Flickr, del.icio.us, Technorati). Tags usually manifest in the form of a freely-chosen, short list of keyword associated by a user with a resource such as a photo, web page, or blog entry. Unlike category- or ontology-based systems, tags result in unstructured knowledge – they have no a-priori semantics. However, it is precisely the unstructured nature of tags that enables their utility. For example, tags are probably easier to enter than picking categories from an ontology; tags allow for greater flexibility and variation; and tags may naturally evolve to reflect emergent properties of the data.

The information challenge facing tagging systems is to extract structured knowledge from the unstructured set of tags. Despite the lack of ontology and semantics, patterns and trends emerge that could allow some structured information to be extracted from tag-based systems [11, 17, 23]. 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 and temporal patterns. Specifically, we are looking at tags on Flickr [10], a popular photo-sharing web site that supports user-contributed tags and geo-referenced (or, *geotagged*) photos. Based on the temporal and spatial distributions of each tag's usage, we attempt to automatically determine whether a tag corresponds to a "place" and/or "event" (see Section 3 for definitions). For example, the tag `Bay Bridge`[1] should be identified as a place, and `SIGIR2007` should be identified as an event. Tag usage distributions are derived from associated photos' metadata. While the correctness of the time and location metadata for each individual photo is suspect [5], in large numbers, trends and patterns can be reliably extracted and used [9, 14].

Extraction of event and 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 and event gazetteer data that can be used, for example, to improve web search by identifying relevant spatial regions and time spans for particular keywords;
- generation of photo collection visualizations by location and/or event/time;
- support for tag suggestions for photos (or other resources) based on location and time of capture;
- automated association of missing location/time metadata to photos, or other resources, based on tags or caption text.

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

*Also affiliated with UC Berkeley, Computer Science Dept.

†Also affiliated with UC Berkeley School of Information.

[1]We use this format to represent tags in the text.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

This paper represents, to our knowledge, the first attempt to extract place and event semantics for tags. Accordingly, we are exploring a number of possible methods. We introduce a new method tailored to event and place identification, *Scale-structure Identification*, and demonstrate how this method outperforms methods borrowed from other domains.

Furthermore, 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 temporal and spatial encodings from which we can extract textual terms – like GeoRSS blog data and geo-annotated web pages or Wikipedia articles. Additionally, the general approach of analyzing a distribution of occurrences over a domain (in our case space and time) to infer semantics could be extended to other metadata domains like color (hue/saturation), visual features, audio features, and text/semantic features.

To summarize, the contributions of this paper 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 event and place identification for tag data;
- Scale-structure Identification – a new method for extracting patterns from usage data;
- a practical application of these methods to extract event and place semantics from tags associated with geotagged images on Flickr.

We formally define our problem in Section 3. Then we describe the methods (Section 4) and report on our evaluation (Section 5). We begin by reviewing the related work.

2. RELATED WORK

We address related work from a number of relevant research areas, including: event detection in time-stamped data such as web queries and personal photo collections; location-based analysis of spatially distributed data such as GPS positions, demographics information, or even information on the web; 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 [4, 18] 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. In particular, our Naïve Scan methods (see Section 4) are similar to previous work on global event detection in web query logs [25] and access logs [13] where events are semantically defined as "bursts" (cf. [15]).

More germane to this paper is the problem of event identification in personal photo collections [12, 20, 24]. A key characteristic of the personal photo collection domain is the general assumption of "a single camera", which reduces event identification to a problem of temporal segmentation. Events are considered to be a single segment of time over which a single activity was taking place, providing a coherent, unifying context. Prior work on this problem has applied a number of techniques: some rely primarily on time [12], others use both locations and times [20, 21], and another looks at the text annotation associated with photos [24]. This type of event-identification is different than ours since (1) we consider multi-person collections of photos and (2) we are interested in whether tags describe events, not whether a segment of time refers to a specific event for a specific person.

Related to event identification is the extraction of meaningful information from location-based data. Recent efforts in ubiquitous computing systems identify meaningful locations and places for GPS and other location tracking technologies [1]. In epidemiology, efforts to identify and localize disease outbreaks [16] are closely related to the place identification problem we address in this paper. Specifically, we borrow some techniques from the disease/outbreak analysis, where data is sparse and dependent on the underlying population statistics, as these two properties are echoed in our data for each tag.

More semantically-rich location analysis problems have been studied in the domain of web-based information retrieval. Specifically, the field of "GeoIR" has had two thrusts relevant to this paper. First, attempts were made (e.g. [2, 6, 8]) 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 terms and defining their spatial scope. The second related research effort in GeoIR focused on extracting the scope of geographic terms or entities based on co-occurring text and derived latitude-longitude information [3, 22]. 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.

Tagging systems in general have been of increasing research interest. Most of the prior research has looked at describing tagging systems [17], or studying trends and properties of various systems [11]. Some efforts have looked at extracting ontologies (or, structured knowledge) from tags [23] – a similar goal to ours, yet using co-occurrence and other text-based tools that could augment the methods analyzed in this paper.

More directly related to this paper are research efforts that analyzed Flickr tags (and other term associated with Flickr photos) together with photo location and time metadata [9, 14]. These projects applied ad-hoc approaches to determine "important" tags within a given region of time [9] or space [14] based on inter-tag frequencies. However, no determination of the properties or semantics of specific tags was provided. Naaman et al. created spatial models for terms appearing in geo-referenced photograph labels [19], but did not detect the location properties of specific terms.

3. 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. Each geotagged photo has, in addition to other metadata, an associated location and time. The location, ℓ_p, (consisting of latitude-longitude coordinates) associated with photo p generally marks where the photo was taken; but sometimes marks the location of the photographed object. The time, t_p, associated with photo p generally marks the photo capture time; but occasionally refers to the time the photo was uploaded to Flickr. Both location and time are recorded at high resolution (microseconds of degrees for location, seconds for time).

Tags are the second basic element type in our dataset. We use the variable x to denote a tag. Note that each photo can have multiple tags associated with it, and each tag is often associated with many photos. Based on the locations and times associated with photos, we can define the location and time usage distributions for each tag x: $\mathcal{L}_x \triangleq \{\ell_p \mid p \text{ is associated with } x\}$ and $\mathcal{T}_x \triangleq \{t_p \mid p \text{ is associated with } x\}$.

Using this data we address the following problem:

> *Can time and place semantics for a tag x be derived from the tag's location, $\mathcal{L}_x$, and time, $\mathcal{T}_x$, usage distributions?*

Example place tags are `Delhi`, `Logan Airport` and `Notre Dame`. Similarly, example event tags are `Thanksgiving`, `World Cup`, `AIDS Walk 2006`, and `New York Marathon` (interestingly, `New York Marathon` represents both an event and a place). Examples of tags not expected to represent events or locations are `dog`, `party`, `food` and `blue`.

The first step in determining whether a tag refers to an "event" or "place" is to define these terms. We aimed for definitions that address both general human perception and the generic (i.e. socially common) semantics of "event" and "place" [27]. We propose that:

Event Tags are expected to exhibit *significant temporal patterns.*

Place Tags are expected to exhibit *significant spatial patterns.*

The term "significant" in these definitions is intentionally vague – designed to capture the idea that "event" and "place" are socially defined (as illustrated by the examples above). More concretely, the definition refers to the fact that a person can expect `New York Marathon` to appear significantly more often every year around November and in New York City; whereas `dog` should appear at almost any time and in almost any location. We expect a reasonable human judge to be able to determine, for any tag and the set of photos associated with that tag, whether or not the tag represents an event and/or a place.

It is important to consider both event and place tags relative to some pre-defined geographic region. For example, `carnival` may not exhibit any patterns world wide, but does have temporal patterns if we are only considering major cities in Brazil. Similarly, `Palace` may have distinct location-based patterns in certain regions (say, London) but no significant patterns world wide. For simplicity, we do not introduce notation to handle the specification of geographic regions – 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 the given region.

Related to regions is the concept of "scale". The basic idea is that tags may exhibit significant temporal or 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 region is limited to San Francisco. Similarly, conferences lasting multiple days (e.g. SIGIR 2007) and even holidays with significant activity prior to a specific date (e.g. Christmas), do not appear to be events at the hour or single day scale, but do exhibit distinctive time patterns relative to longer time scales. Accordingly, the methods described below search for significant patterns at multiple spatial and temporal scales.

4. EVENT/PLACE IDENTIFICATION

The goal of our analysis is to determine, for each tag in the dataset, whether the tag represents an event, and whether the tag represents a place. The intuition behind the various methods we present is that an event (or place) refers to a specific segment of time (or region in space). So, the "significant patterns" for event and place tags should be manifested as bursts over small parts of time or space. More specifically, the number of usage occurrences for an event tag should be much higher in a small segment of time than the number of usage occurrences of that tag outside the segment. The scale of the segment is one factor that these methods must address; the other factor is calculating whether the number of usage occurrences within the segment is significantly different from the number outside the segment.

To simplify the discussion, we describe the methods as they pertain to event identification. The notions of segments and scales are not domain specific – i.e. both time and space can be divided into segments (perhaps overlapping) of various scale. Any place/space specific issues are addressed, but otherwise the translation to place-based analysis is left to the reader.

In the remainder of this section, we describe the methods in detail. We first present adaptations of two well-known techniques to the problem at hand. Then we present a new method for event and place identification: Scale-structure Identification.

4.1 Borrowed Methods

At a high level, the steps for the modified, burst-detection methods are the following:

1. **Scale Specification** – Choose an ordered set of scale values, $\mathcal{R} = \{r_k \mid k = 1 \ldots K, r_{k_1} > r_{k_2} \iff k_1 > k_2\}$. We generally choose an exponentially increasing set of scales (e.g., $r_k = \alpha^k$ for some $\alpha > 1.0$).
2. **Segment Specification** – For each scale r_k define a finite set of time segments to search over, say $\mathcal{Y}_{r_k}$. We use a regularly spaced grid where the grid size is based on the scale; but overlapping or arbitrary segments are possible.
3. **Partial Computation** – For each scale r_k and each time segment in $\mathcal{Y}_{r_k}$, compute a statistic on $\mathcal{T}_x$ that captures some aspect of the tag's usage pattern in time (likely, although not necessarily, based on some relationship between the usage occurrences within a time segment versus outside of the segment).
4. **Significance Test** – Aggregate the partial computation statistics for each time segment at each scale to determine whether tag x is an event.
5. **Identify Significant Segments** – Provided a significant pattern for x is found, determine which scales and time segments correspond to the event.

Before describing each method, we introduce some necessary notation. First, we use $T_r(x, i)$ to denote how many times tag x was used in time segment i (the subscript indicates that the segment is defined in relation to scale r). The maximum value for $T_r(x, i)$ is the number of photos in the segment – we use $N_r(i)$ to denote the number of photos taken during time segment i. Some of the methods below also require the total number of tag usage occurrences in a segment – we denote this as $T_r(i) = \sum_x T_r(x, i)$.

4.1.1 Naïve Scan Methods

The Naïve Scan methods are an application of a standard burst detection method used in signal processing [25]. The method computes the frequency of usage for each time segment at each scale. The method identifies a "burst" when the frequency of data in a single time segment is larger than the average frequency of the data over all segments plus two times the standard deviation of segment frequencies.

The clear 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 suffers from 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. We further relax the condition that the number of tag occurrences be larger than the mean 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 I*, the partial computation (Step 3) for each tag x is specified by: $\frac{T_r(x,i)}{\mu_N + 2\sigma_N}$, where μ_N is the mean of $\{N_r(i)|i = 1\ldots\}$ and σ_N is the standard deviation of $\{N_r(i)|i = 1\ldots\}$. We use a variable threshold of this statistic in identifying events (Step 4).

To identify the segments of time corresponding to an event for a tag (Step 5 above), we simply record the segments that pass the significance test (Step 4 above). Specifically, we record the values of i and r where the partial computation statistic is larger than the threshold.

We omit the details of Steps 1 and 2 (how to search over i and r) since any brute force search method applies. We also remind the reader that the formulation for location-based burst detection is analogous.

An alternative approach, which we refer to as *Naïve Scan II*, 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 that photo will require few tags in addition to x.

The partial computation statistic is $\frac{T_r(x,i)}{\mu_T + 2\sigma_T}$, where μ_T is the mean of $\{T_r(i)|i = 1\ldots\}$ and σ_T is the standard deviation of $\{T_r(i)|i = 1\ldots\}$. If every photo had the same number of tags, these results would be identical to those produced by Naïve Scan I. However, as photos can have an arbitrary number of tags, with some photos using far more tags than others, the Naïve Scan II method does produce (slightly) different results.

4.1.2 Spatial Scan Methods

The Spatial Scan methods are a standard application of the Spatial Scan statistic [16], 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 segment of the domain (e.g. segment of time) is abnormal relative to the underlying probability model. This abnormality test is performed for each segment.

To illustrate how the Spatial Scan methods work, we describe an example from our data. Consider `eatbrains`, which refers to a slightly obscure event that took place in San Francisco (where people dressed up as zombies and walked around certain neighborhoods). Suppose: (1) over the 2+ years covered by our data, q denotes the global probability of this tag being applied to any photo; (2) all M photos tagged with `eatbrains` occur within a single two hour segment, and (3) there are a total of N photos taken during this same two hours. If `eatbrains` refers to an event of any significance, M should be quite a bit larger than qN. The 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, it is derived from the data.)

The expression for the partial computation statistic for *Spatial Scan I* is:

$$\max_{i,r} \Big(\frac{T_r(x,i)}{N_r(i)}\Big)^{T_r(x,i)} \cdot \Big(\frac{\Sigma_{\hat{\imath}\neq i} T_r(x,\hat{\imath})}{\Sigma_{\hat{\imath}\neq i} N_r(\hat{\imath})}\Big)^{(\Sigma_{\hat{\imath}\neq i} T_r(x,\hat{\imath}))} \cdot \Big(\frac{\Sigma_{\hat{\imath}} T_r(x,\hat{\imath})}{\Sigma_{\hat{\imath}} N_r(\hat{\imath})}\Big)^{-(\Sigma_{\hat{\imath}} T_r(x,\hat{\imath}))} \cdot I\Big(\big(\tfrac{T_r(x,i)}{N_r(i)}\big) > \big(\tfrac{\Sigma_{\hat{\imath}\neq i} T_r(x,\hat{\imath})}{\Sigma_{\hat{\imath}\neq i} N_r(\hat{\imath})}\big)\Big)$$

where $I(\cdot)$ is the indicator function. For details on the derivation of this expression see Kulldorff [16].

As in the Naïve Scan methods, the significance test (Step 4) uses a single, variable threshold value – tags whose partial computation statistic exceeds this value are identified as events. Also, by storing the values of i and r where the partial computation statistic is larger than the threshold we can identify the segments in time when events occur (Step 5). Finally, details of Steps 1 and 2, how to search over i and r, are likewise omitted since brute force search methods are sufficient.

Similar to the Naïve Scan II modification, we developed *Spatial Scan II* using the total number of tags that occur inside segments. We omit the partial computation expression for Spatial Scan II – it can be produced by simply replacing occurrences of $N_r(i)$, the number of photos in each segment, with $T_r(i)$, the number of tags in each segment, in the partial computation expression of Spatial Scan I (above).

In the four methods described above, we determine the segments of time for each scale independent from the actual usage distributions of the tags. Additionally, these methods can only propose a-priori time segments as the times of events. In the worst case, these segments might hide the actual time of an event by splitting the usage occurrences into adjacent segments, none of which are above the significance test threshold. The next method we describe addresses the issue of a-priori defined time segments.

4.2 Scale-structure Identification

The Scale-structure Identification method performs a significance test (Step 4 above) that depends on multiple scales simultaneously and does not rely on a-priori defined time segments. Accordingly, the Scale-structure Identification method performs all the steps listed above except the Segment Specification step (Step 2).

The key intuition behind Scale-structure Identification is the following: if tag x is an event then the points in $\mathcal{T}_x$, the time usage distribution, should appear as a single cluster at many scales. The clustering mechanism used in Scale-structure Identification is similar to the clustering mechanism in the scale-space method developed by Witkin [26]. However, whereas Witkin was interested in any structure that exhibited robustness over a range of scales, we are in-

terested in the robustness of a single type of structure – a single cluster.

Consider the graph over $\mathcal{T}_x$ where edges between points exist if and only if the points are closer together than r (recall that r is the scale variable). Let $\mathcal{Y}_r$ be the set of connected subcomponents of this graph. The Partial Computation step (Step 3 above) computes the entropy of $\mathcal{Y}_r$ for each scale r. Specifically, the partial computation statistic is defined as: $E_r \triangleq \sum_{Y \in \mathcal{Y}_r} (|Y|/|\mathcal{T}_x|) \log_2(|\mathcal{T}_x|/|Y|)$. 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, $\mathcal{Y}_r$ (note that $E_r = 0$ when the usage distribution is a single cluster, i.e. $|\mathcal{Y}_r| = 1$).

For place identification we simply replace $\mathcal{T}_x$ with $\mathcal{L}_x$ in the calculation of E_r (we compute the distance between points in $\mathcal{L}_x$ as the L_2 distance between the points as they lie on a sphere).

A caveat to the partial computation statistic concerns periodic events. Periodic events have strong clusters, at multiple scales, that are evenly spaced apart in time. Practically, because tags occur in bursts, we also require that a periodic tag exhibit at least three strong clusters (to rule out tags that just happened to occur in two strong temporal clusters but are not truly periodic). Of course, this assumption could result in some false negatives (e.g. recurring events that only appear twice in our dataset), but it is necessary due to the sparse nature of our data (to mitigate these false negatives we could check whether the two strong clusters were spaced apart at some culturally meaningful distance like one month, one year, etc.).

We check for periodic events by: (1) identifying "strong" clusters (i.e. clusters that contain at least 2% of the data), (2) measuring how far apart the strong clusters are, (3) making sure the cluster variances are not too big relative to the distances between clusters (i.e. the standard deviations of the usage distributions for each cluster should be, on average, smaller than 10% of the average inter-cluster distance), and (4) making sure the distances between clusters are "even" (i.e. the standard deviation of inter-cluster distances is smaller than 10% of the average inter-cluster distance). If a tag's temporal distribution passes all of these tests[2], we re-compute the scale structure for this tag by treating time as modulo μ, the average inter-cluster distance. Specifically, we re-compute $\mathcal{Y}_r$ from $\mathcal{T}'_x \triangleq \{t \text{ modulo } \mu \mid t \in \mathcal{T}_x\}$ using the distance metric $\|t_1 - t_2\| \triangleq \min(|t_1 - t_2|, |\mu + t_1 - t_2|, |\mu + t_2 - t_1|)$. Intuitively, this modulo adjustment to the time dimension aligns the "strong" clusters so that they will be treated as a single cluster. For example, if a tag's temporal distribution has 3 strong clusters that are on average 365 days apart, the modulo adjustment to time corresponds to the cyclical calendar year.

Finally, the significance test calculation (Step 4) aggregates the partial computation statistics simply by summing them over the set of scales: $\sum_{k=1}^{K} E_{r_k}$. This summed value is tested against a threshold to determine if the tag is an event. By recording the scale structures at each scale, we can determine which time segments strongly characterize an event tag (Step 5). In fact, we can then characterize the tag, or rather the event it refers to, at multiple scales.

[2]The percentage thresholds used in these tests were set empirically.

5. EVALUATION

We implemented the methods described above, and performed a direct evaluation of the methods' performances over part of the Flickr dataset. The goals of the evaluation were to establish whether any of the methods can reliably identify events/places in the tag data; 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, and the results of the evaluation.

5.1 The Flickr Dataset

The data we use in this study consists of geotagged photos from Flickr and the associated tags. Location and time metadata is available for roughly fourteen million public Flickr photos (at the time this paper was written). The capture time is usually available from data embedded in the photo file by most digital cameras. 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. In this paper 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 1. In Figure 2, we plot the location and time usage distributions for the tag `Hardly Strictly Bluegrass`.

The San Francisco Bay area is currently one of the best-represented geographic regions in Flickr, increasing the likelihood of finding significant patterns at sub-city and sub-region scales. Furthermore, San Francisco is the only such region for which we could reliably generate the ground truth (see below). We note, however, that restricting the dataset to a specific geographic region did not require any alterations to the methods or the evaluation computations.

In addition to the regional specification, we applied filters to improve the time/location metadata correctness and to ensure sufficient data for the analysis. The two filters to

Figure 1: Spatial distribution of all San Francisco geotagged photos in our dataset (white markers).

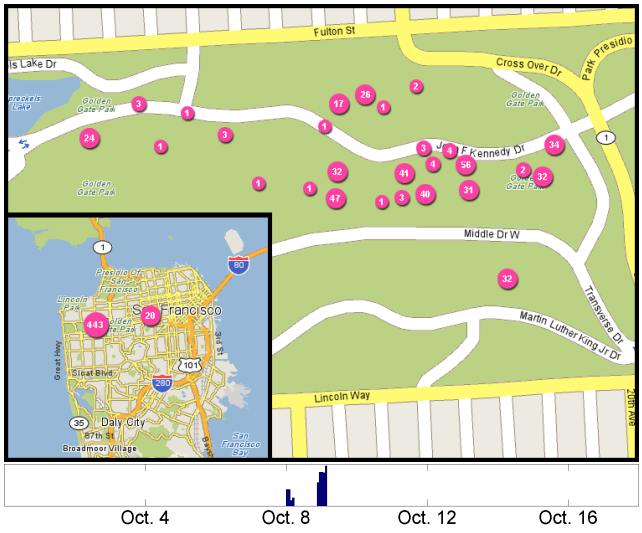

Figure 2: Location (top) and time (bottom) usage distributions for the tag `Hardly Strictly Bluegrass` in the San Francisco Bay Area. The zoomed in map view shows the details of the larger location cluster from the zoomed out view.

improve correctness were: (1) if the photo's capture time was (a) prior than 2004 or (b) not later than the upload time, the photo was removed; and (2) if the photo's location resolution was too inaccurate (i.e. anything other than the two most accurate levels in the scale from 1–16), the photo was removed. The filter to ensure sufficient data applied to tags. Any tag that was used less than 25 times or by only one user was removed.

Our final dataset consists of 49897 photos with an average of 3.74 tags per photo (s.d. 2.62). These photos cover a total temporal range of 1015 days, starting from January 1, 2004. The average number of photos per day was 49.16 (s.d. 89.89), with a minimum of zero and a maximum of 643.

From these photos we extracted 803 unique tags. As expected, and similar to previous work [9, 11], tag usage was Zipf-distributed. The maximum number of photos associated with a single tag was 34325 (for `San Francisco`), and the mean was 232.26 (s.d. 1305.40).

The Flickr dataset is rather new, and presents a number of additional challenges. While Flickr popularity is rising, the number of geo-referenced photos is still relatively low. We see sparse activity within every group of photos – for example, Flickr does not contain photos tagged `Golden Gate Bridge` for every day since January 1, 2004. 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. In the time dimension, because of the growing active community on Flickr, an order of magnitude more photos were taken and uploaded during 2006 than during 2005, complicating time-based analyses.

5.2 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 corresponds to an event, and whether the tag corresponds to a place. This in-depth analysis was needed to eliminate errors that arise from obscure tags (e.g., `eatbrains` that described a relatively unknown San Francisco event), and by issues of 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 generalize, correct, and interpolate inaccurate or sparse data.

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 people vote, without access to the photos or their metadata, on whether the tag referred to an event, a place or both. This vote-based data exhibited systematic errors relative to the ground truth data: (1) obscure or un-popular events and places were often false negatives (i.e. incorrectly labeled as not being events or places), (2) generic tags like `anniversary` and `park` were often false positives (while they have clear event and place semantics within a limited scope, over the whole data set they did not refer to specific time segments or regions of space), and (3) event tags like `Future of Web Apps` were often not labeled as places even though many events also occur in specific regions of space. For these reasons, we omit comparison of the place and event identification methods to the vote-based data.

5.3 Results

Since all of the methods produce ranked results, we can use standard IR metrics to evaluate performance. For each tag, the methods produces a number that indicates how likely this tag is to be an event (or place). Rather than choosing a single threshold for each method to categorize the tags, we can vary the threshold dynamically and examine the tradeoff in terms of recall and precision for each method.

Plots of the recall vs. precision curves are shown in figure 3 for places (top) and events (bottom). The X-axis represents a recall value – the percentage of actual events (or places) that are identified as events (or places). The thresholds for each method were adjusted to produce this recall value. The Y-axis shows the precision – the percentage of tags identified as events (or places) that are actually events (or places). For example, when the threshold for Scale-structure Identification for events is set so that recall is 50%, the precision is 82%. We can see in both figures

	P-R area	Max F1	Min CE	
Naïve Scan I	0.4455	0.5279	0.2914	PLACE
Naïve Scan II	0.4458	0.5279	0.2914	
Spatial Scan I	0.6028	0.5907	0.2441	
Spatial Scan II	0.6134	0.5955	0.2416	
Scale-structure	0.7034	0.6655	0.1930	
Naïve Scan I	0.3291	0.3636	0.1009	EVENT
Naïve Scan II	0.3320	0.3590	0.0996	
Spatial Scan I	0.4130	0.4811	0.1034	
Spatial Scan II	0.4146	0.4785	0.1046	
Scale-structure	0.6420	0.6533	0.0648	

Table 1: Precision-Recall Area, Maximum F1, and Minimum CE values for the various methods.

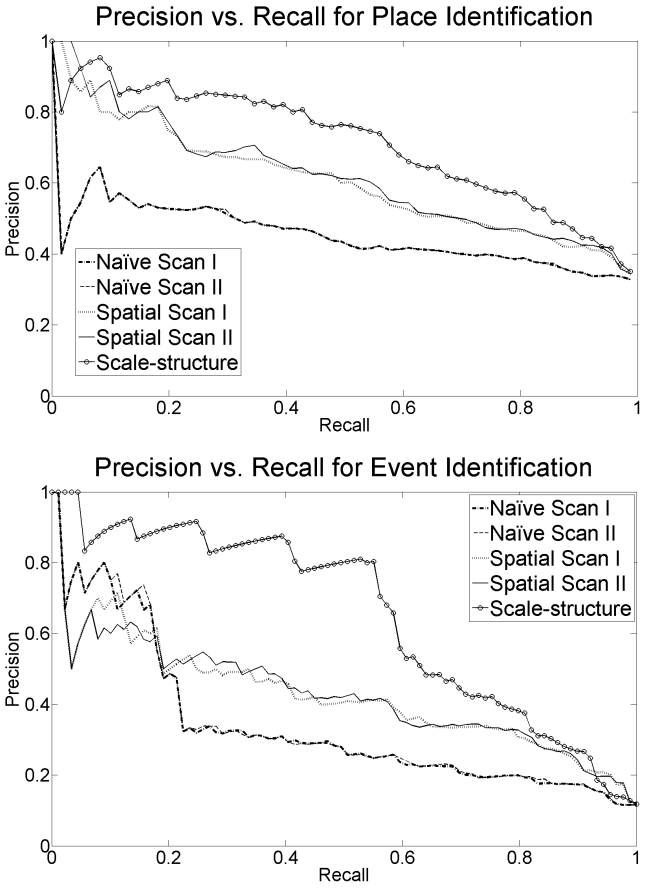

Figure 3: Precision vs. recall for the place (top) and event (bottom) identification tasks.

that Scale-structure Identification performs better than the traditional methods, for almost all recall values and for both event and place identification.

From these curves we computed (1) the area under the precision-recall curve (P-R area), (2) the maximum value of the F1 statistic for each method (MAX F1), a metric that balances precision and recall values, and (3) the minimum total classification error (Min CE) (cf. [7]). The results are shown in Table 1, again indicating that Scale-structure Identification performs better than the other methods.

As an alternative to searching for optimal threshold values for the methods, one can simply take the top N results from the ordered lists produced by the methods (where N is variable). Table 2 shows precision and recall values for $N = 50, 100$, and 200. Table 3 lists the top 10 tags for each method for both place and event detection. Again, Scale-structure Identification performs better than the other methods.

We also studied the sensitivity of the Scale-structure Identification method to the Scale Specification step (Step 1 in Section 4.1). We varied the exponential base in the scale sampling scheme from 1.1 to 5.0 (the exponents were positive integers, marking the sequential position of the scale value). The results were robust to these changes. One point to note, however, is that performance slightly, but consistently, improved as the exponential base decreased. In other words, the Scale-structure Identification method performed

	top 50	top 100	top 200	
Naïve Scan I	0.58, 0.12	0.52, 0.21	0.47, 0.39	PLACE
Naïve Scan II	0.58, 0.12	0.52, 0.21	0.47, 0.39	
Spatial Scan I	0.82, 0.17	0.68, 0.28	0.60, 0.49	
Spatial Scan II	0.80, 0.16	0.69, 0.28	0.61, 0.50	
Scale-structure	0.88, 0.18	0.83, 0.34	0.70, 0.58	
Naïve Scan I	0.38, 0.21	0.31, 0.35	0.25, 0.56	EVENT
Naïve Scan II	0.38, 0.21	0.31, 0.35	0.25, 0.56	
Spatial Scan I	0.50, 0.28	0.40, 0.45	0.33, 0.74	
Spatial Scan II	0.52, 0.29	0.41, 0.46	0.33, 0.74	
Scale-structure	0.78, 0.44	0.53, 0.60	0.36, 0.81	

Table 2: Values for (precision, recall) for different numbers of returned tags.

better with denser samplings of the space of scale values, but only slightly. Accordingly, we recommend, for computational gains, to sample the scale space fairly sparsely as the results are not strongly affected.

Due to space constraints, we do not include detailed results from our analysis of the segment identification step (Step 5 in Section 4.1). Briefly, the parts of time and space that were associated with identified events and places were mostly accurate. The only systematic errors found were due to sparse, wrong, or missing data. For example, tags like `October` and `summer` had temporal distributions that were not representative of the true duration of these events. While more data will help for some of these tags (e.g. conference names), some events like seasons and months are not likely have uniform distributions over their true durations.

In terms of error analysis, we identified several classes of common errors with the Scale-structure Identification method. First, the majority of false positives and false negatives for place identification were the result of sparse data. Likewise the false positives for event identification were often due to sparse data. False negative event tags were also caused by bad data (e.g. incorrect capture time). Additional sampling and filtering techniques could potentially alleviate some of these problems.

Overall, the performance of the Scale-structure Identification method holds promise for automatic extraction of place and event semantics. The Scale-structure Identification method clearly outperforms the methods borrowed from other domains. While the difference is significant, we believe that one reason for the gain is the "single-cluster-like" filtering; the borrowed methods simply look for outlier/bursty segments without measuring or enforcing the uniqueness of the segments.

6. CONCLUSIONS AND FUTURE WORK

We have taken a first step in showing that semantics can be assigned to free-form tags using the usage distribution 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 that outside the scope of the specific system.

In particular, we have shown that location and time metadata associated with photos and their tags enables the extraction of "place" and "event" semantics. This mapping of tags to events and locations could improve image search, serve as a basis for collection visualization, and assist in

Naïve Scan I	Alcatraz, *friends*, *event*, PFA, BB, *2006*, China Town, Lombard Street, *trip*, *August*	PLACE
Naïve Scan II	Alcatraz, *friends*, *event*, PFA, BB, *2006*, China Town, Lombard Street, *trip*, *August*	
Spatial Scan I	GG Park, Alcatraz, Baseball, Giants, HSB, de Young, *event*, SF Giants, PFA, GG Bridge	
Spatial Scan II	Alcatraz, GG Park, Baseball, HSB, Giants, PFA, de Young, SF Giants, *event*, GG Bridge	
Scale-structure	pet cemetary, Revision3, Ruby Red, Dahlias, *MashPitSF2*, VSHD, Red Devil Lounge, Club Neon, FWA, BH	
Naïve Scan I	BB, eatbrains, eatbrains2006, zombies, *wedding*, zombie, zombiemob2006, byobw, BBW, *Lombard*	EVENT
Naïve Scan II	BB, eatbrains, eatbrains2006, zombies, *wedding*, zombie, zombiemob2006, byobw, BBW, *Lombard*	
Spatial Scan I	BB, *wedding*, HSB, *event*, byobw, pride, The FWA, BBW, bigwheel, *PFA*	
Spatial Scan II	BB, *wedding*, HSB, *event*, byobw, The FWA, pride, 2004, *PFA*, *anniversary*	
Scale-structure	zombiemob, BB 2006, valleyschwag, zombie, zombiemob2006, eatbrains, VSHD, eatbrains2006, zombies, *air race*	

Table 3: Top 10 scored tags for place (top) and event (bottom) identification for each method. Higher scoring tags appear further to the left. At the optimal F1 threshold, all of the tags were labeled as places (or events). Italicized tags are false positive errors. Note, we use the following abbreviations: Palace of Fine Arts → PFA, Bay to Breakers → BB, Golden Gate → GG, Hardly Strictly Bluegrass → HSB, VS Hoe Down → VSHD, Future of Web Apps → FWA, Bottom of the Hill → BH, and Bring your own Big Wheel → BBW.

other photo-related tasks. This type of knowledge can also help create an ad-hoc gazetteer for events and locations that could be used for various tasks beyond photo management. We plan to revisit the image search, visualization and gazetteer deployment in future work.

We would also like to extend the system to handle multi-regional problems. As mentioned above, the tag `carnival` may be event-like only in major cities of Brazil; we note above that the data analysis should be limited to specific geographic regions. Ideally, we could simultaneously generate, store, and disambiguate tag semantics for different regions throughout the world.

Finally, we plan to deploy our methods to other temporally and spatially encoded data, as they become exceedingly available on the web. We also look at extending the metadata features used, beyond location and time, to verify that our methods are still effective in extracting other semantics beyond place and event.

7. ACKNOWLEDGMENTS

We would like to thank Shane Ahern, Simon King, Rahul Nair and Malcolm Slaney for their valueable insights, comments and assistance.

8. REFERENCES

[1] R. Aipperspach, T. Rattenbury, A. Woodruff, and J. Canny. A quantitative method for revealing and comparing places in the home. In *Proc. Ubicomp*. Springer, 2006.

[2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proc. SIGIR*, p 273–280. ACM, 2004.

[3] A. Arampatzis, M. van Kreveld, I. Reinbacher, P. Clough, H. Joho, M. Sanderson, C. B. Jones, S. Vaid, M. Benkert, and A. Wolff. Web-based delineation of imprecise regions. In *Proc. of the workshop on Geographic Information Retrieval* at *SIGIR2004*, 2004.

[4] G. Box and G. Jenkins. *Time series analysis : forecasting and control*. Cambridge Univ. Press, 1976.

[5] D. C. Bulterman. Is it time for a moratorium on metadata? *IEEE MultiMedia*, 11(4):10–17, 2004.

[6] O. Buyukokkten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting geographical location information of web pages. In *WebDB'99*, 1999.

[7] L. Cai and T. Hofmann. Text categorization by boosting automatically extracted concepts. In *Proc. SIGIR*, p 182–189. ACM, 2003.

[8] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proc. Very Large Databases*, p 545–556. Morgan Kaufmann, 2000.

[9] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proc. WWW*, p 193–202. ACM, 2006.

[10] Flickr.com. http://www.flickr.com.

[11] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, 2006.

[12] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd. Time as essence for photo browsing through personal digital libraries. In *Proc. of ACM/IEEE-CS Joint Conf. on Digital Libraries*, 2002.

[13] V. Guralnik and J. Srivastava. Event detection from time series data. In *Proc. SIGKDD*, p 33–42. ACM, 1999.

[14] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proc. Multimedia Information Retrieval*, p 89–98. ACM, 2006.

[15] J. Kleinberg. Bursty and hierarchical structure in streams. *Knowledge discovery and data mining*, 7(4):373–397, 2003.

[16] M. Kulldorff. Spatial scan statistics: models, calculations, and applications. In *Scan Statistics and Applications*, p 303–322, 1999.

[17] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. HYPERTEXT*, p 31–40. ACM, 2006.

[18] D. McDowall, R. McCleary, E. E. Meidinger, and R. A. H. Jr. *Interrupted Time Series Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 1980.

[19] M. Naaman, A. Paepcke, and H. Garcia-Molina. From where to what: Metadata sharing for digital photographs with geographic coordinates. In *Proc. CoopIS*, 2003.

[20] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proc. JDCL*, 2004.

[21] A. Pigeau and M. Gelgon. Organizing a personal image collection with statistical model-based ICL clustering on spatio-temporal camera phone meta-data. *J. of Visual Comm. and Image Rep.*, 15(3):425–445, 2004.

[22] R. Purves, P. Clough, and H. Joho. Identifying imprecise regions for geographic information retrieval using the web. In *GISRUK*, 2005.

[23] P. Schmitz. Inducing ontology from flickr tags. In *Proc. of the workshop on Collaborative Web Tagging* at *WWW2006*, 2006.

[24] A. Stent and A. Loui. Using event segmentation to improve indexing of consumer photographs. In *Proc. SIGIR*, p 59–65. ACM, 2001.

[25] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *Proc. SIGMOD*, p 131–142. ACM, 2004.

[26] A. Witkin. Scale space filtering. In *Proc. Int'l Joint Conf. Artificial Intelligence*, p 1019–1022, 1983.

[27] J. Zacks and B. Tversky. Event structure in perception and conception. In *Psychological Bulletin*, 127(1):3–21, 2001.