Beyond Co-occurrence: Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities

Haipeng Zhang, Mohammed Korayem, Erkang You, David J. Crandall
School of Informatics & Computing
Indiana University
Bloomington, IN
{zhanhaip,mkorayem,erkyou,djcran}@indiana.edu

ABSTRACT

Studying relationships between keyword tags on social sharing websites has become a popular topic of research, both to improve tag suggestion systems and to discover connections between the concepts that the tags represent. Existing approaches have largely relied on tag co-occurrences. In this paper, we show how to find connections between tags by comparing their distributions over time and space, discovering tags with similar geographic and temporal patterns of use. Geo-spatial, temporal and geo-temporal distributions of tags are extracted and represented as vectors which can then be compared and clustered. Using a dataset of tens of millions of geo-tagged Flickr photos, we show that we can cluster Flickr photo tags based on their geographic and temporal patterns, and we evaluate the results both qualitatively and quantitatively using a panel of human judges. We also develop visualizations of temporal and geographic tag distributions, and show that they help humans recognize semantic relationships between tags. This approach to finding and visualizing similar tags is potentially useful for exploring any data having geographic and temporal annotations.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining

General Terms

Measurement, Theory.

Keywords

tag semantics and visualization, Flickr, geo-spatial and temporal clustering

1. INTRODUCTION

Online photo sharing is booming: as of late 2010, Flickr hosted over 5 billion photos and users were uploading more than 3,000 new images every minute [29]. In addition to simply hosting images, these sites include social features that allow users to annotate photos in a variety of ways, including adding descriptive keywords (called *tags*), titles, captions, quality scores, and free-form comments. In the particular case of tags, empirical studies of Flickr have shown that users add tags for a variety of reasons, including to indicate geographic locations, descriptions of actions and events, identities of objects, people, and groups, and so on [30]. Thus tags provide a rich (albeit noisy, incomplete and inconsistent) source of information about the semantic content of photos. Tags have been used in the data mining community to study the properties of online photo collections, including identifying temporal bursts of photographic activity corresponding to important events [26], finding geo-spatial peaks of activity corresponding to important landmarks [22], selecting iconic images to represent particular places [6], and even predicting product adoption rates by monitoring the popularity of product photos [14].

A major theme of this existing work has been to study *tag co-occurrences*, finding tags that frequently occur with one another on the same photograph as a means of identifying semantically-related tags. Co-occurrence has been particularly helpful for suggesting new tags based on the existing tags of a photo [11, 18, 30]. Other work has applied clustering algorithms to find semantically-related concepts by looking for groups of tags that frequently co-occur [3, 28]. While these approaches often yield reasonable results, they make the assumption that tags are related only if they co-occur often on the same photographs. But some related tags may seldom co-occur: the Statue of Liberty and Central Park are clearly related – both are major New York City landmarks – but the tags *statueofliberty* and *centralpark* do not have many co-occurrences because it is nearly impossible to take a photo that includes both. Moreover, co-occurrences give little information about the nature of the relationship – i.e. *why* two tags are related.

In this paper, we explore other more specific types of connections between tags – and, by extension, between the concepts that the tags represent – by comparing the spatial and temporal distributions of tags instead of their co-occurrences. The intuition here is that related concepts have similar geo-temporal distributions because they occur at about the same times and places, even if they rarely co-occur on photos. Our work takes advantage of the rich metadata available on photo-sharing sites like Flickr, including the timestamps recorded by modern digital cameras, and the geo-tags specifying the latitude and longitude of where a photo was taken. Geo-spatial and temporal properties of tags have been studied in existing work (e.g. [14, 22, 26]), but we are not aware of work that has used these properties to quantify connections between tags. We also present methods for visualizing the relationships between tags by comparing their geo-temporal distributions. These techniques complement other methods like tag clouds [16] and temporal tag evolution [8] to help people find and visualize the relationships between tags.

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.

In the remainder of this paper, we describe an approach that finds relationships between tags from geo-spatial and temporal similarities, and we apply this approach on a large dataset of tens of millions of photos downloaded from Flickr. After surveying related work in Section 2, we show how to define feature vectors to compactly represent geo-temporal distributions and cluster tags based on these features (Section 3). We then present qualitative visualizations of the resulting clusters as well as a more quantitative evaluation using a panel of human judges in Section 4.

2. RELATED WORK

Tags on social sharing systems have been studied extensively. Here we review the work most relevant to studying tag semantics and relationships in photo collections.

Clustering tags based on co-occurrences. Much work has been based on photo tag co-occurrences, mostly in the context of tag suggestion systems. Garg and Weber [11] use tag co-occurrences to suggest additional tags for a new image. Sigurbjörnsson and Van Zwol [30] take a similar approach but also conduct a study of tagging behavior on Flickr. Liu *et al* [18] rank tags by performing Pagerank-like random walks on tag graphs where the edge weights are frequency of co-occurrence. Other work has clustered tags using co-occurrence features in order to find groups of semantically-related tags. Shepitsen *et al* [28] use TF-IDF to build trees of del.icio.us and Last.Fm tags, and then partition them into homogeneous clusters. Begelman *et al* [3] partition tag graphs using spectral clustering. These papers inspired our idea of clustering tags, but our work differs in that we use features other than co-occurrence, instead looking for tags with second-order connections like similarities in spatial and/or temporal distributions.

Temporal and geo-spatial properties of tags. Timestamps and geo-tags of photos have been used to study temporal and geo-spatial distributions of individual tags, often in order to identify peaks in the temporal distribution (corresponding to events) and/or in the geo-spatial distribution (corresponding to cities and popular landmarks). Rattenbury *et al* [26] use burst detection techniques to find tags with significant peaks in time and space, while Moxley *et al* [22] build on this work by using entropy analysis on a quadtree data structure to improve performance. Chen and Roy [4] model tag occurrences as points in a 3D geo-temporal space, use wavelet transform-based techniques to find tags with bursts in both temporal and spatial distributions, and then cluster these tags using DBSCAN. Ahern *et al* [1] cluster photos to find dense areas based on geo-tags and find representative tags for the areas using TF-IDF, while Moxley *et al* [21] use a similar approach to rank local tags. Crandall *et al* [6] and Kennedy *et al* [17] find both distinctive tags and images for clusters of photos found based on geo-tags. While these papers analyze geo-spatial and temporal attributes of photos, they are primarily focused on finding and studying "event" tags corresponding to peaks in the geo-temporal tag usage distributions, whereas we are interested in comparing and grouping larger and more general sets of tags.

Visualizing tag clusters. The usual method to visualize tags is to draw tag clouds [16], while Dubinko *et al* [8] visualize interesting Flickr tags that evolve over time through animations. In this paper we propose visualizations of the temporal semantics of tag clusters by plotting time series representing their usage along time, and the geo-spatial semantics of tag clusters in a 3-D space over a map to represent their usage across space. We also show that the visualizations can help humans understand subtle semantic relationships.

Spatial clustering and co-location pattern mining. Our work is reminiscent of spatial clustering [24, 27] which groups together similar spatial data points based on their locations, but differs from our problem in that it clusters spatial data points while we cluster spatial distributions. Our goal is more related to co-location pattern mining [9, 10, 13, 35], in which the goal is to identify features that are often located near one another. Clustering is not used by Xiao *et al* [35], while Huang and Zhang [13] and Estivill-Castro *et al* [9,10] take different clustering approaches. We are not aware of work that applies these techniques to vast datasets of user-generated social media content as we do here.

Studies of query logs, tweets and news articles. Temporal and geo-spatial patterns have been studied to discover concept relationships in other domains, including tweets, news articles, and queries in search engine logs. Radinsky *et al* [25] extend Explicit Semantic Analysis to represent concepts as time series of word occurrences in the New York Times archive. Vlachos *et al* [33] use frequency-space analysis of search query time series to identify bursts and semantically-similar queries. Chien and Immorlica [5] use similar techniques but perform an experimental evaluation, and find that for 70% of the queries, at least three of the top ten keywords identified by temporal similarity are semantically related. The geo-spatial distributions of search engine queries were studied by Backstrom *et al* [2], who estimate geographic centers and dispersions of queries. Vadrevu *et al* [32] use the co-occurrence of a query term with place names in a region to determine whether the query is related to this region. Perhaps most similar to ours is very recent work in the domains of search engine queries and Twitter hashtags. Mohebbi *et al* [20] developed a tool which takes a search engine query and finds other queries with similar temporal or spatial distributions. They quantize the weekly time series data and state-by-state usage data of individual queries into vectors to represent temporal and spatial distributions, and then apply K-means clustering and an approximate nearest neighbor alogorithm to look up similar vectors efficiently. Meanwhile, Yang and Leskovec [36] cluster Twitter hashtags and short phrases in news documents to identify their temporal patterns and the dynamics of human attention they receive. Neither of these latter two papers evaluates the semantic quality of the resulting clusters and connections, whereas we use panels of human judges to evaluate our results.

Finally, our work is related in spirit to that of Wu *et al* [34], who compute the "Flickr distance" between a pair of tags by computing the visual similarity of photos having those tags, and then cluster tags based on this score. Our work is similar in that it defines connections between tags using a property other than co-occurrence, but is complementary in that we define similarity metrics based on metadata like timestamps and geo-tags instead of the visual content of images. This both allows us to discover connections that may not be apparent from visual features, and also allows our techniques to scale to much larger datasets (having tens of millions of photographs) because processing metadata is much more efficient than visual analysis.

3. DISCOVERING TAG RELATIONSHIPS

We assume that we have a dataset of online objects (e.g. photos), each of which has a user id of the person generating the object (e.g. the photographer), a timestamp specifying when the object was created, a geo-tag specifying latitude-longitude coordinates for the object, and a set of zero or more text tags. To define this formally, it is useful to think of tagging *events* – individual acts of a user tagging a photo with a text tag. The information associated with a particular event includes the tag that was applied, the photo to which the

tag was applied, the user who uploaded the tagged photo, the geographic location of the tagged photo, and the timestamp indicating when the photo was taken. Letting $\mathcal{T}$ be the set of all possible text tags, then the set of tagging actions $\mathcal{A} = \{a_1, a_2, ..., a_q\}$ can be defined as a set of tuples of the form $a_i = (u_i, t_i, \tau_i, g_i)$, where u_i is a user, $t_i \in \mathcal{T}$ is a tag, τ_i is a timestamp, and $g_i \in \mathcal{R} \times \mathcal{R}$ is a geo-tag (latitude-longitude coordinate).

Given a collection of tagged objects, our goal is to cluster the tags based on their geo-spatial and temporal properties. To do this, we first extract a geo or temporal signature for each tag and represent it with a corresponding feature vector, and then we cluster the feature vectors using an unsupervised algorithm like k-means [19]. In addition to finding the geo-spatial and temporal distributions of each tag, we also compute a feature vector based on the cross-product of these two attributes, which allows us to represent a tag's joint geo-temporal distribution – i.e. the "motion" of how a tag's spatial distribution changes over time. The following three subsections explain the geo, temporal, and motion feature vectors in turn.

3.1 Geo-spatial feature vectors

Since different types of tags have different geographical distributions, our first feature aims to characterize the geographical distribution of a tag. Because the distribution of photographs over the world is highly non-uniform, with most of the photographic activity concentrated in cities, many areas of the world have very few photos. It is thus useful to aggregate photos together into coarse geo-spatial buckets instead of clustering using raw geo-tags. (Quantization of geo-tags also reduces the impact of noise in the geo-tags.) To do this, we divide the world into n bins, each with s degrees of latitude by s degrees of longitude. We assign each of the bins a unique index number in the range $[1, n]$, and define a quantization function $q_G(g)$ that maps a latitude-longitude coordinate g into the index number of the corresponding geo bin. In the results presented in this paper, we use $s = 1$ degree, which corresponds to grid cells of roughly 100 km $\times$ 100 km at the middle latitudes. Note that the bins do not have the same surface area because degrees of longitude became closer together near the poles; this tends not to be a problem in practice because the vast majority of photos are taken near the middle latitudes. An equal-area partitioning of the globe [12] would address this issue and is a direction for future work.

To compute a geo feature vector for tag t, we first count the number of unique users who have used that tag in each geo bin g,

$$U_G(g, t) = || \{u_i | (u_i, t_i, \tau_i, g_i) \in \mathcal{A}, t_i = t, g = q_G(g_i)\} ||.$$

We count the number of *users* who applied a tag within a geographic area instead of the number of photos in order to prevent high-activity users from biasing the distribution [1]. (This can be thought of as giving each user a single "vote" for whether or not a tag applies to a given geographic area.) Then we normalize the vector to get the geo feature $v^G(t)$ of tag t,

$$v_i^G(t) = \frac{U_G(i, t)}{\sqrt{\sum_{j=1}^{n} U_G^2(j, t)}}.$$

Normalization is necessary since tags sharing similar geo distributions might have different overall frequencies of occurrence. While other work has found that L1 normalization works better in high dimensional spaces [7, 23], we found that L2 normalization works better in our context. (For example, for the clustering results presented in Section 4, we found that L2 norm generates clusters that have more uniform and moderate sizes. When clustering 2000 tags into 50 clusters, L1 norm produces 17 singletons (clusters that contain only one tag) while L2 norm produces no singletons. L1 norm cluster sizes also have much greater variation: their standard deviation is 127.8 while the standard deviation of L2 norm is 54.1.)

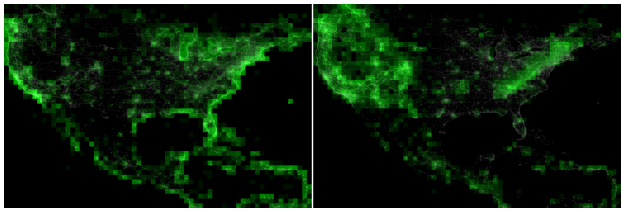


Figure 1: Geographic distributions for tag "beach" (left) and "mountains" (right).

As an example, Figures 1 visualizes the normalized matrices for tags "beach" and "mountains" over North America, in which greater intensity indicates that more users applied the tag to photos in a given geo-bin. Notice that the Appalachian and Rocky Mountain ranges are immediately apparent in the "mountains" map, while the "beach" map highlights the coastline of North America.

3.2 Temporal feature vectors

Tags also have different temporal distributions, because some tags (and semantic concepts) are much more popular at certain times than others — for example, we might expect "beach" to be used more often during the summer than in the winter, while "restaurant" might occur more often during the meal times of the day. As with the geographic feature vectors described above, with temporal features it is also useful to aggregate photos together into coarse temporal bins. Let $q_T(\tau)$ be a quantization function that maps a timestamp τ into one of m temporal bins, returning a bin index in the range $[1, m]$. This quantization function could be designed to operate at different levels of granularity, for example mapping timestamps to hours of the day, days of the week, months of the year, etc. For any tag t, we then build an m-dimensional vector, again counting the number of unique users who have used the tag in each temporal period p,

$$U_T(p, t) = || \{u_i | (u_i, t_i, \tau_i, g_i) \in \mathcal{A}, t_i = t, p = q_T(\tau_i)\} ||,$$

and then normalize to produce an m-dimensional temporal feature vector $v^T(t)$,

$$v_i^T(t) = \frac{U_T(i, t)}{\sqrt{\sum_{j=1}^{m} U_T^2(j, t)}}.$$

In this paper we primarily use a quantization function that maps timestamps into one of 26 two-week periods of the year: January 1-14, January 15-28, etc. We disregard the specific year and as a result, all the data is merged together into a single year – for example, photos taken on January 1, 2008 and January 1, 2009 will be mapped to the same cell. In addition to the 26-dimensional 2-week vectors, we also create 7-dimensional day-of-week vectors and 24-dimensional hour-of-day vectors.

Flickr users are significantly more active at certain times of the year than others, as illustrated in Figure 2: note that nearly twice as many users take photos during the first two weeks of July (period 14) than in early February (period 3). To correct for this effect, in practice we normalize the temporal counts for tag t and period p, $U_T(p, t)$, by the total number of photos taken in North America during p, before applying L2 normalization to produce v^T. Figure 3 shows the process of obtaining temporal feature vectors.

3.3 Geo-temporal (motion) features

Finally, we also want to produce a signature for a tag based on

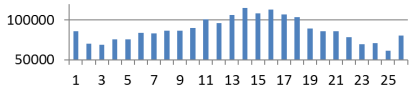


Figure 2: Number of unique Flickr users active in North America, for each 2-week period of the year.

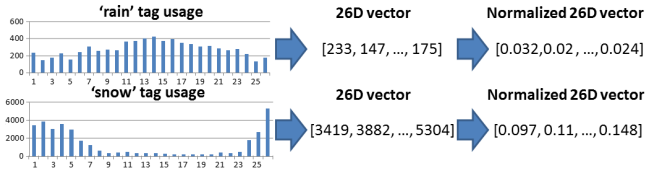


Figure 3: Computing temporal feature vectors.

its joint geo-temporal distribution – that is, how the geo-spatial distribution of the tag varies over the course of a year (or equivalently, how the temporal distribution varies with spatial location). We call these geo-temporal signatures our "motion" features. Given geo and temporal quantization functions q_G and q_T (described above) that map geo-tags into one of n bins and timestamps into one of m bins, a motion feature vector has one bin per entry in the cross product of these two sets of indices, or mn dimensions total. More precisely, we define a motion quantization function that maps a geo-tag g_i and timestamp τ_i to a bin index in $[1, mn]$,

$$q_M(g_i, \tau_i) = m \times (q_G(g_i) - 1) + q_T(\tau_i),$$

then count the number of unique users who used a given tag t in each geo-temporal bin,

$$U_M(m, t) = || \{u_i | (u_i, t_i, \tau_i, g_i) \in \mathcal{A}, t_i = t, m = q_M(g_i, \tau_i)\} ||,$$

and take the L2 norm (as above) to define a final motion feature vector, v^M. For the experiments in this paper, this vector has $mn = 124,800$ dimensions. As with the geo feature vectors, we remove empty dimensions (geo-temporal cells having no photos) as an optimization.

3.4 Co-occurrence features

For comparison purposes, we also define similarity metrics using using two more traditional techniques. First, the pairwise co-occurrence between two tags t_1 and t_2, co_occur(t_1, t_2), is computed by simply counting the number of photos that are tagged with both t_1 and t_2. A disadvantage of this simple co-occurrence measure is that it favors pairs of tags that occur very often, since very frequent tags will co-occur more often than infrequent tags even if they are unrelated. Thus we include a second baseline feature, mutual information, which overcomes this problem by normalizing the co-occurrence measures by the overall frequency of the tags [3],

$$\text{mutual\_info}(t_1, t_2) = \frac{\text{co\_occur}(t_1, t_2)}{K} \log\left(\frac{\text{co\_occur}(t_1, t_2) K}{\text{occur}(t_1)\text{occur}(t_2)}\right)$$

where occur(t) is the total number of photos having tag t and K is the total number of photos. This score can be thought of as a measure of the independence of the two tags: it is minimized if the two tags are completely independent (never co-occur), and is maximized if the tags are strongly correlated (always co-occur).

4. EXPERIMENTS AND VISUALIZATIONS

To test our techniques for characterizing tags based on geo and temporal signatures, we used a dataset of geo-tagged, time-stamped photos downloaded from Flickr through the site's public API interface, using a crawling technique similar to that described in [6]. We collected the following information for each photo: the geo-tag (latitude and longitude) of where the photo was taken, the timestamp of when it was taken, and the set of textual tags (if any) associated with the photo. From this collection of nearly 80 million photos, we selected only the photos in North America (which we defined to be a rectangular region spanning from 10 degrees north, -130 degrees west to 70 degrees north, -50 degrees west).

We then computed the top 2000 most frequent text tags (ranked by the number of unique users applying the tag) in North America. (We chose this relatively small number of tags so that our geo-temporal distributions would have substantial mass (each of these tags has been used by at least 1,100 unique Flickr users), and to make human evaluation tractable. Note that the majority (66.7%) of photos on Flickr are tagged with at least one of these 2,000 tags since Flickr tag frequency follows a long tailed distribution [30].) For each of these tags, we extracted the geo feature vectors, temporal feature vectors and motion vectors described in Section 3. In preparation for geo feature vector extraction, we filtered the data by removing photos with geotag precision less than about city-scale (according to the precision reported by Flickr), resulting in a dataset with about 44 million photos. For the temporal feature vectors, we removed photos with inaccurate or suspicious timestamps (including photos supposedly taken in the future or distant past), resulting in about 41 million photos; for the motion feature vectors, both of these filters were applied, yielding about 39 million photos.

4.1 Tag relationships

We can use the similarity metrics defined in Section 3 to find pairs of similar tags. Given a tag t', we can find a list of related tags using each of the distances defined above, including geo-spatial, temporal, and geo-temporal. To do this, we compute the feature vectors $v^G(t)$, $v^T(t)$, and $v^M(t)$ for each tag t, and then compute the pairwise Euclidean distances between these vectors and those of t'. The tags are ranked according to their distances to t' in ascending order, and the k tags with lowest distance are found. For the co-occurrence and mutual information features, we compute the similarity for each tag t using the co_occur(t, t') and mutual_info(t, t') functions, and then rank the tags in increasing order of similarity.

As an example, Table 1 lists the tags that are most similar to the tag "cherryblossoms" under the various measures of similarity. The first column shows the 20 most similar tags according to the geo-spatial similarity metric. Most of these tags are strongly related to Washington, DC (which is of course famous for its annual cherry blossom festival in April), including "president", "whitehouse", "smithsonian" and "lincolnmemorial" among others. The second column shows tags having high temporal similarity, including "easter", "spring", "april" and "magnolia". The list of tags under motion similarity appear to be a mixture of geographically similar tags and temporally similar tags. In contrast, the co-occurrence list has arguably much lower quality: "canon", "water" and "usa" are popular tags that also co-occur with many other tags, and are not particularly relevant to "cherryblossoms". Mutual information gives more meaningful results compared to raw co-occurrence, but it missed tags like "whitehouse", "tulips" and "kite" which were picked up by the temporal and geo-spatial analyses. These tags do not frequently co-occur with "cherryblossoms" on the same photos, but do share similar geo and/or temporal patterns.

4.2 Clustering tags

The analysis in the last section can be used to compute the similarity between any arbitrary pair of tags, but it is difficult to visualize or quantify the performance of these results directly because there are so many possible pairs. In past work, tag similarity re-

Table 1: Top 20 most similar tags to "cherryblossoms" using different similarity metrics. The columns rank the tags according to (from left): geo-spatial, temporal, motion (geo-temporal), co-occurrence, and mutual information.

	Geo-spatial		Temporal		Motion		Co-occurrence		Mutual information	
1.	president	0.321	cherry	0.451	cherry	0.291	washingtondc	4568	washingtondc	2.41e-4
2.	whitehouse	0.321	blossoms	0.505	blossoms	0.417	dc	3443	dc	1.70e-4
3.	monument	0.323	blossom	0.612	blossom	0.546	spring	2319	spring	1.54e-4
4.	smithsonian	0.324	easter	0.636	jefferson	0.673	washington	2089	blossoms	1.37e-4
5.	memorial	0.324	spring	0.638	kite	0.868	flowers	1969	flowers	1.18e-4
6.	georgetown	0.325	april	0.654	washingtonmonument	0.908	blossoms	1367	pink	9.22e-5
7.	washingtonmonument	0.327	magnolia	0.735	monument	0.990	pink	1007	washington	8.37e-5
8.	dc	0.327	buds	0.758	magnolia	0.998	trees	979	cherry	7.88e-5
9.	lincolnmemorial	0.328	washingtonmonument	0.813	spring	1.026	canon	754	washingtonmonument	5.55e-5
10.	wwii	0.331	tulip	0.822	bloom	1.042	cherry	753	trees	5.21e-5
11.	washingtondc	0.332	jefferson	0.837	memorial	1.057	tree	703	tree	4.85e-5
12.	jefferson	0.367	egg	0.858	lincolnmemorial	1.085	usa	610	flower	3.00e-5
13.	arlington	0.370	tulips	0.862	washingtondc	1.086	flower	560	blossom	2.04e-5
14.	lincoln	0.372	bloom	0.868	dc	1.091	washingtonmonument	552	bloom	1.94e-5
15.	mall	0.392	break	0.869	whitehouse	1.092	water	498	april	1.42e-5
16.	capitol	0.401	poppy	0.870	mall	1.096	2007	415	water	1.31e-5
17.	soldier	0.407	eggs	0.883	festival	1.103	unitedstates	412	white	1.29e-5
18.	war	0.421	bud	0.895	tulip	1.106	festival	379	sky	1.26e-5
19.	cherry	0.429	kite	0.922	government	1.121	nature	377	blue	1.25e-5
20.	capital	0.448	olympics	0.924	capital	1.122	brooklyn	347	nature	1.23e-5

sults have been summarized by grouping tags into a small number of similar clusters, typically using co-occurrence information (e.g. [5]). We follow a similar strategy and cluster Flickr tags according to each of our three types of distance metrics (temporal, geo-spatial, geo-temporal) as well as traditional co-occurrence and mutual information measures.

For each feature vector type, we clustered the 2000 tags using k-means [19]. Squared Euclidean distance was used to measure distances between vectors. Since k-means clustering is sensitive to the initial choice of centroids, we ran k-means five times with different random initial cluster centers and chose the best result (in this case, choosing the clustering with the minimum total vector-to-centroid distance). Of course, this clustering algorithm requires an *a priori* choice of the number of clusters (k). For the purposes of this paper, where our primary focus is on presenting techniques for comparing tags based on their geo and temporal distributions and not on presenting an end-to-end system for tag analysis, we simply set k to a value (50) that gave reasonable results in our subjective judgment. Various techniques exist for selecting k automatically based on properties of a dataset (see e.g. [31]) and these techniques could easily be applied to our work.

As we show in the next few sections, the overall "shape" of the geo-spatial and temporal distributions varies dramatically from tag to tag and cluster to cluster: some clusters contain tags that are diffuse across space and time (like "canon", "geotagged", "blackandwhite", etc.), while other distributions are very "peaky" ("newyorkcity", "washingtondc", etc.), and others are somewhere in between ("usa", "newengland", etc.). It is thus useful to compute a statistical measure of the peakiness of a distribution, in order to compactly characterize its overall "shape". We measure the peakiness of a vector v by computing its second moment,

$$\text{second\_moment}(v) = v \cdot v = \sum_{i=1}^{n} v_i^2,$$

and measure the peakiness of a cluster of tags C as the average second moment of the vectors that it contains,

$$\frac{\sum_{v \in C} \text{second\_moment}(v)}{|C|}.$$

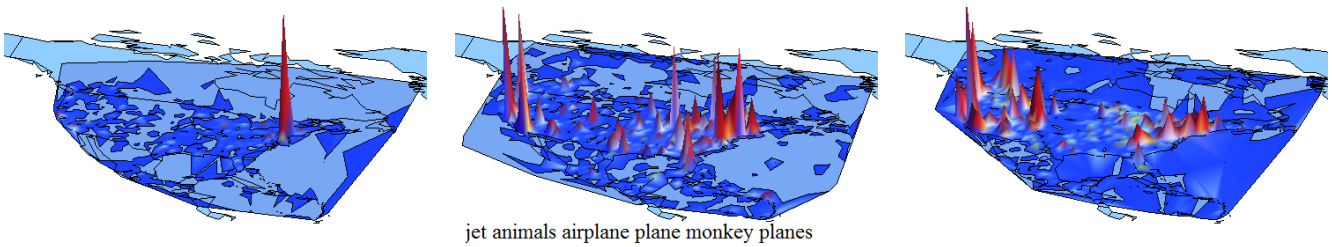
Peaky distributions will have higher second moment values, while distributions close to uniform will have low second moment. (Note that the second moment of a discrete probability distribution is the likelihood of sampling twice from the distribution and drawing the same value both times.) The second moment gives a statistic by which to rank clusters, as clusters with high average peakiness usually have bursts in temporal distributions or geo-spatial distributions which indicate particularly interesting clusters.

We present sample results and visualizations for several sample tag clusters in the following sections. Due to space limitations we do not show all the tag clusters generated from the three different perspectives, but these and other detailed results are available at `http://vision.soic.indiana.edu/tagclusters`.

4.2.1 Geo-spatial clusters

Figure 4 shows visualizations of several tag clusters produced by analyzing geo-spatial distributions. The visualizations were created by taking a cluster centroid and converting it back to a two-dimensional matrix: i.e. for each geo bin, we find the corresponding latitude-longitude coordinates for the bin center and plot them together with their values in a 3D space over a schematic map of North America. The result is a topographical visual effect with the heights of the peaks as well as the intensity of red color indicating the usage of the tags in the cluster at corresponding locations underneath. The figure shows three sample clusters. Figure 4(a) consists of tags from the New York City area. Some not very obvious tags are: "soho" which is a shopping area in New York City, "bull" which is the Wall Street Bull and "ground" which relates to Ground Zero referring to World Trade Center site. This cluster is ranked 10th out of 50 clusters by second moment. Figure 4(b) visualizes the cluster in which most tags are related to zoos and animals and others are related to airports. As a result, the visualization peaks at major US cities with famous zoos and airports. It is ranked 37th. Figure 4(c) displays the cluster of tags that occur predominantly in national parks. This cluster is ranked 27th.

Images that record the visualizations of all the 50 geo clusters are available at the above website. The top ranked clusters by second moment are more concentrated geographically. From these top clusters, we see state clusters, city clusters, zoo clusters, park clus-



911 liberty chelsea wtc worldtradecenter statueofliberty jersey newjersey nj hudson hudsonriver soho bull ground

(a)

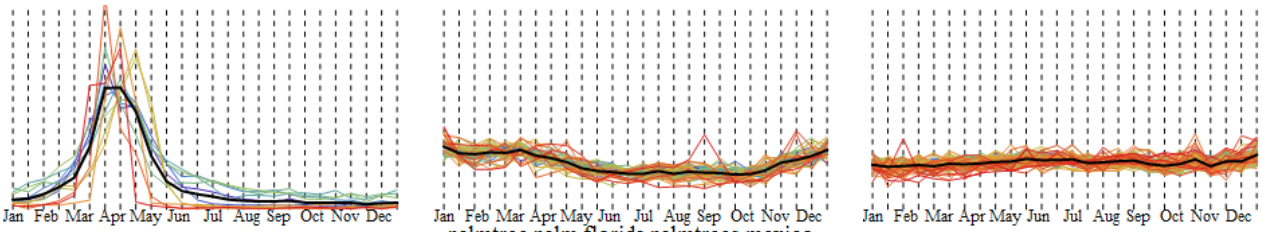
jet animals airplane plane monkey planes aircraft zebra f18 ray tiger international landing bear elephant airport gorilla aviation zoo giraffe starwars peacock orchid penguin convention bears navy airshow

(b)

mountain wilderness meadow mountains summit peak lakes lodge chipmunk granite mount mt glacier yosemite

(c)

Figure 4: Visualizations of sample clusters produced by analyzing similarity of geo-spatial distributions. The clusters seem to correspond to (a) tags related to New York City, (b) tags related to cities with popular zoos and airports, and (c) tags related to national parks and outdoor areas. Best viewed in color.



blossoms spring blossom magnolia cherry buds tulip tulips april cherryblossoms easter

(a)

palmtree palm florida palmtrees mexico warm miami bowling keys pelican palms key impressedbeauty ruins fl arena azul caribbean tea tropical phoenix sol tulum iguana méxico tucson cuba puertorico...

(b)

john amy friends funny happy smile child jason tom eric matt michael son mike jeff 2007 2008 david tongue chris adam eating andy alex ben us bell ryan dave dan ...

(c)

Figure 5: Visualizations of three clusters produced by analyzing similarity of temporal distributions: (a) tags related to spring, (b) tags related to winter, (c) tags related to gatherings of friends.

ters, northern city clusters, coastal area clusters and so on. For lower ranked clusters, the tags are more geographically distributed, such as a cluster of rural regions and a cluster of urban regions.

4.2.2 Temporal clusters

Figure 5 shows visualizations of three of the clusters produced by the temporal similarity metric. For each temporal cluster, we plot the cluster centroid as well as the distributions of the individual tags within the cluster, with each point representing the usage in the corresponding two-week period. Figure 5(a) displays the cluster with a strong peak during spring. The thick black curve corresponds to the cluster centroid while the other curves correspond to the signals for individual tags. This cluster is ranked 11th out of 50 clusters by second moment. Figure 5(b) visualizes a cluster with a shallow peak during the winter season. Most tags are related to winter vacations in warm locales. This cluster is ranked 34th. The cluster in Figure 5(c) (ranked 48th by second moment) seems to correspond to family gatherings, with slight temporal peaks around Thanksgiving and Christmas. There are also some year tags (e.g. "2008", "2009", etc.) which appear frequently around New Year's Day. Visualizations of all 50 temporal clusters are available at the website above. We see that top ranked clusters of tags have sharp bursts in smaller time windows and lower ranked clusters have more general seasonal patterns.

We also clustered the data according to temporal features at other time scales, including 7-dimensional day-of-week vectors and 24-dimensional hours-of-day vectors. Due to space constraints we do not present detailed results, but instead mention a few interesting findings. For day-of-week vectors, we are able to see weekday clusters such as "work office desk students commute" and weekend clusters such as "live sushi gallery concert macys highschool moma". For hours-of-day clusters, we see clusters peak at different time of the day, such as a morning cluster, "early sunrise dawn morning," and a nighttime cluster, "lightning concert longexposure campfire nighttime nightphotography exposure live".

4.2.3 Geo-temporal clusters

Tags within a motion cluster typically share either geo and temporal similarities, or both. Cluster "vegas lasvegas las bellagio strip paris casino nevada fountains flamingo" captures Las Vegas and its hotels and casinos which has a counterpart in geo clusters. It is ranked 5th out of 50. Cluster "christmas holiday xmas holidays christmastree christmaslights december decorations ornament decoration cookies gift santa fireplace" captures Christmas which has a counterpart in temporal clusters and is ranked 31st. We observe that top ranked clusters are more likely to have obvious geographic connections while the clusters having temporal patterns are ranked in the middle. All the motion clusters can also be found at the website above.

4.3 Evaluation

We evaluated the idea of finding similar tags using geo-spatial

Top 10 temporal clusters

	Tags in cluster	# tags	2nd moment
1	**4th fourthofjuly 4thofjuly independenceday july4th**	7	0.578
2	**january newyearseve**	2	0.5
3	**turkey thanksgiving november**	3	0.355
4	**august**	1	0.2785
5	iris may dandelion graduation memorialday	5	0.269
6	**costume costumes halloween**	3	0.223
7	**christmastree christmaslights christmas ornament holidays**	9	0.215
8	pride june	2	0.206
9	**fallcolors pumpkins autumn fall foliage**	7	0.190
10	irish march	2	0.144

Top 10 geo-spatial clusters

	Tags in cluster	# tags	2nd moment
1	***toronto niagara niagarafalls cntower falls***	9	0.415
2	***golden cablecar francisco sanfrancisco sf***	27	0.402
3	***los angeles santamonica la losangeles***	8	0.397
4	***broadway brooklyn empire cab empirestatebuilding***	34	0.394
5	***strip paris vegas las lasvegas***	10	0.379
6	***seattle needle pugetsound spaceneedle wa***	8	0.374
7	***chicago bean searstower illinois il***	7	0.366
8	***ma massachusetts boston cambridge newengland***	6	0.332
9	prairie pennsylvania pa philadelphia philly	58	0.287
10	***911 liberty chelsea wtc worldtradecenter***	14	0.276

Top 10 motion (geo-temporal) clusters

	Tags in cluster	# tags	2nd moment
1	***losangeles angeles santamonica los la***	7	0.021
2	***taxi broadway empirestatebuilding brooklyn empire***	38	0.01693
3	***tx texas austin houston dallas***	5	0.01691
4	***chicago searstower bean illinois il***	7	0.01679
5	***vegas lasvegas las bellagio strip***	10	0.01671
6	***alberta calgary banff***	3	0.01606
7	***francisco sanfrancisco goldengatebridge goldengate berkeley***	30	0.0158
8	***pa philadelphia philly pennsylvania***	4	0.0157
9	***statueofliberty liberty newjersey jersey nj***	13	0.0148
10	ski skiing snowboarding tahoe fdsflickrtoys	73	0.0138

10 randomly-chosen co-occurrence clusters

	Tags in cluster	# tags
1	***sea ocean beach boat island***	41
2	***coast waves sun shore pier***	41
3	***washington statue museum sculpture washingtondc***	41
4	people model female face hair	43
5	***vacation travel trip desert arizona***	41
6	***water canada winter sky nature***	41
7	***trees mountains mountain hiking hike***	41
8	party wedding friends love dance	40
9	light city night bed sleep	39
10	geotagged us building architecture canon	41

10 randomly-chosen mutual information clusters

	Tags in cluster	# tags
1	rails rail railway train railroad	36
2	**independenceday july4th fourthofjuly 4th weird**	39
3	***marriage rings groom love couple***	38
4	plane jet aviation aircraft planes	38
5	furry sleepy pet kitten fur	38
6	jeans jacket socks shoes feet	39
7	furniture toilet sink seat couch	34
8	***tide waves surf ocean wave***	41
9	***rockies rockymountains peak glacier summit***	38
10	***sail port harbor docks sailing***	37

Figure 6: Comparison of clusters produced by different similarity metrics: temporal (top left), geo-spatial (top right), and geo-temporal (center). Clusters judged to be temporally significant by human judges are printed in blue boldface, while clusters judged to be geographically related are printed in *red boldface italics*. Clusters are sorted in decreasing order of second moment. For comparison, also shown are clusters produced by co-occurrence (bottom left) and mutual information (bottom right). For each cluster, up to 5 top ranked tags are displayed. Relevancy was judged by users without visualizations being shown.

and temporal features by comparing the clustering results from our proposed methods with those of the co-occurrence based techniques. Because it is difficult to define the quality of a tag cluster objectively, we involved humans in our experiments to judge the geo-spatial and temporal relevance of the clusters we found. We then used the human judgment as ground truth to compute the precisions and recalls when the task is to retrieve semantically meaningful clusters in time and/or space by thresholding the average second moment values. The goal of this evaluation was to test whether our techniques produce coherent tag clusters that correspond to intuitive geo-spatial and temporal concepts, and how these clusters compare to traditional techniques that use co-occurrence. We also showed our visualizations of geo-spatial and temporal tag distributions to a subset of the human judges, to try to measure how well these visualizations could help people appreciate subtle semantic connections between tags.

4.3.1 *Clustering based on tag co-occurrences*

As a baseline we used the method based on tag co-occurrence described in [3] to cluster the top 2000 tags into 50 clusters for a fair comparison. In particular, we constructed an undirected graph of tags, weighted the edges between tags by metrics of their co-occurrences and removed weak edges by thresholding the weights. We then applied a graph partitioning program, KMETIS [15], to partition the graph into clusters. We tried two different methods to weight the edges, one by raw co-occurrence counts and the other by mutual information (defined in Section 3). As a result, we generated two sets of 50 clusters: co-occurrence clusters and mutual information clusters. For each cluster, we ranked its tags by the numbers of edges inside the cluster: tags with more edges are considered to be more representative.

4.3.2 *Ground truth from human judgment*

To conduct the human judgment study at a large scale, we used Amazon's Mechanical Turk service, asking users to judge the geo coherence and temporal coherence of clusters produced by the various similarity metrics that we propose. To improve the quality of the human judgment, we required users to be in the United States (so that they would be familiar with North American geography and cultural events) and have a good (>=95%) historical approval

rate. For each cluster discovered by our methods, we selected its ten top ranked tags to present to the user.

For each geo, motion, co-occurrence and mutual information cluster, we asked the users to judge its geo relevance among the following three options: "more than 50% of its tag representatives represents a specific geographic area such as NYC", "more than 50% of its tag representatives represents an abstract geographic concept such as ocean," or "not geo relevant." Similarly, for each cluster we also asked the users to judge its temporal relevance according to the scale: "more than 50% of its tag representatives represents a specific temporal event such as Thanksgiving", "more than 50% of its tag representatives represents a broad temporal indication such as spring", "not temporally relevant". For both the geo and temporal relevance questions, if at least 80% of the users choose the first or second option for a cluster, we consider it to be geo or temporally relevant respectively. We put the clusters in 16 batches of 25 clusters per assignment and each batch was assigned to up to 20 Mechanical Turk users. On average, each cluster was judged by 19.9 users. Users were shown only the top ten tags associated with each cluster. We conducted two independent sets of experiments, one in which the 20 users were not shown the visualization graphs described above, and another in which a separate group of 20 users were shown the visualizations, to quantify how useful the visualizations might be in practice.

4.3.3 *Evaluation results*

The evaluation found that geo-spatial and motion clustering were more effective in finding geo relevant clusters than the other techniques: 29 (58%) of the geo clusters were found to be geo relevant by the human judges who were not shown the visualizations, compared to 30 (60%) of motion clusters, 11 (22%) of the co-occurrence clusters, and 11 (22%) of the mutual information clusters. A case study of some geo-temporal clusters judged not to be geo or temporally relevant showed the visualizations gave hints to people and helped understand the not-so-obvious semantics behind the clusters. For the geo clusters, among the top 31 clusters ranked by average second moment, 28 were judged to be geo relevant. We examined the 3 clusters that were judged to be "not geo relevant," and even these appeared to have interesting geographical semantics that were likely not obvious to the human judges. Visualizations for these 3 clusters are shown in Figure 7. The cluster in Figure 7(a) has an obvious peak in San Diego. The terms "polarbear" and "border" may not be immediately associated with San Diego, but in fact they refer to the San Diego Zoo's famous polar bears while the tag "border" refers to the Mexican border which is just a few miles away. In Figure 7(b), most tags are state or city names. They are in one cluster as their geographical distributions are very concentrated resulting in very peaky geo vectors which are not far from each other as measured by Euclidean distances. In Figure 7(c), there is a peak in Northern California and the tags are related to wine. Northern California is famous for its wine industry (Wine Country). Some other lower ranked clusters judged to be "not geo relevant" also show some geographical signal, such as the cluster displayed in Figure 4(b), which highlights zoos and airports.

Temporal clustering also found more temporally-relevant clusters than other techniques: 13 (26%) of the clusters produced by the temporal similarity metric were found to be temporally relevant, versus 5 (10%) of motion clusters and only 1 (2%) of the co-occurrence clusters and 6 (12%) of the mutual information clusters. We examined the temporal clusters judged not to be temporally relevant and found that some did have temporal patterns that were hard to observe when only the text tags were presented to users. We present three such clusters in Figure 8. The cluster in Figure 8(a)

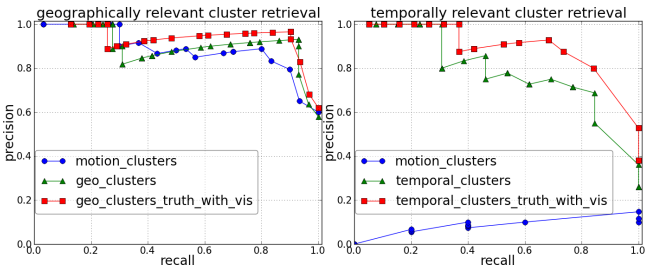


Figure 9: Precision-recall curves for retrieving geographically (left) and temporally (right) relevant clusters.

has bursts around important dates for the Presidential election. The cluster in Figure 8(b) has an autumn pattern and includes mostly insects and plants that are active or increasing in autumn. Finally, the cluster in Figure 8(c) has a January and February pattern, with tags related to the Chinese New Year and the Super Bowl. Some other such examples can be found in Figure 5(b) and (c) which were also judged to be "not temporally relevant."

In all of these cases, the visualizations of geo and temporal clusters were helpful for us to discover the hidden semantics behind the tag clusters. To try to measure the effectiveness of these visualizations, we conducted two separate evaluations on Mechanical Turk, one in which the visualizations were shown and another in which they were suppressed. The results suggested that the visualizations were helpful; the results only differed in that some of the clusters that were judged to be "not geo relevant" or "not temporally relevant" by the group who did not see visualizations were judged to be "geo relevant" or "temporally relevant" by the group that did. For geo clusters, 2 more high-ranked clusters (ranked 11 and 31 by second moment) mentioned above and visualized in Figure 7(a) and (c) were judged to be "geo relevant," which gave in total 62% of the 50 clusters judged to be "geo relevant." For temporal clusters, 6 more clusters (ranked in top 21) were judged to be "temporally relevant," which gave in total 38% of the clusters judged to be "temporally relevant" and 18 out the 22 top ranked clustered were "temporally relevant." Clusters displayed in Figures 8(a) and (b) and Figure 5(b) and (c) are examples of the 6 clusters. Though the cluster in Figure 8(c) was still judged to be "not temporally relevant," 65% of the users who saw the visualizations judged it to be "temporally relevant" versus the 40% who did not. On average, for each geo cluster, 66.7% of the users who saw visualizations judged it to be "geo relevant," compared with 64.4%; for each temporal cluster, 56.9% of the users who saw visualizations judged it to be "temporally relevant", comparing with 49.7%.

Motion clustering found both geographically and temporally relevant clusters. However, no motion clusters were judged to be both geo *and* temporally relevant. Mutual information clustering and co-occurrence clustering found the same number of geo relevant clusters, but mutual information clustering found 5 more temporally relevant clusters. Using this measurement, mutual information clustering performed better in finding temporally relevant clusters than co-occurrence clustering.

A subset of the evaluation results are shown in more detail in Figure 6. The figure shows the top 10 clusters produced by the geo-spatial, temporal, motion, co-occurrence, and mutual information analyses, and indicates which of these clusters were found to be geo or temporally significant by the panel of human judges which did not see the visualizations.

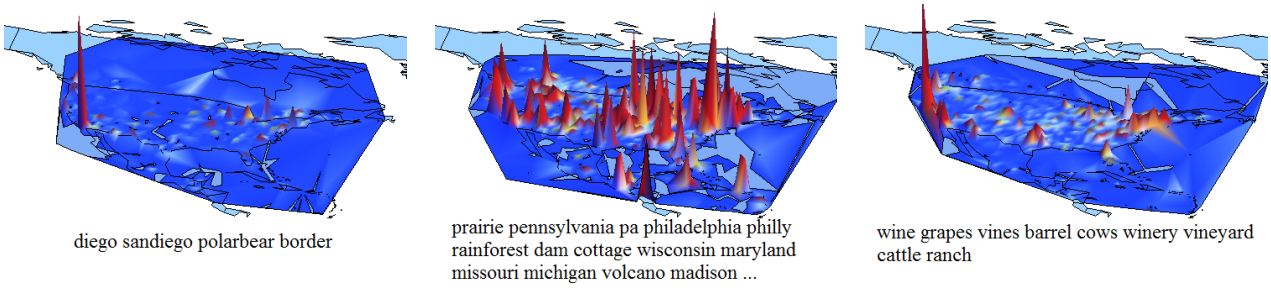


Figure 7: Sample geo clusters judged to be not geo relevant with high average second moment by users who were not shown the visualizations. The clusters correspond to (a) tags related to San Diego, (b) tags related to cities and states, and (c) tags related to Northern California Wine Country. Best viewed in color.

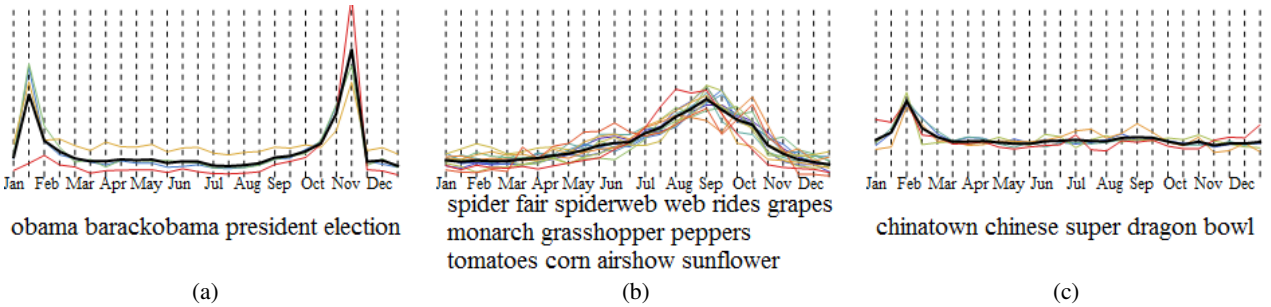


Figure 8: Sample temporal clusters judged to be not temporal relevant by users who were not shown the visualizations. The clusters correspond to (a) tags related to the Presidential election, (b) tags related to autumn, (c) tags related to Jan and Feb.

4.3.4 Geo and temporally relevant cluster retrieval

We observed that the average second moment of the geo-spatial, temporal, and motion clusters appears to be a good indicator of whether a cluster will be judged to be geo or temporally relevant. We quantified this relevance by studying a retrieval problem, in which the task is to find relevant clusters using different average second moment thresholds. Clusters are considered to be retrieved if their average second moment is equal to or above a certain threshold. We can then summarize the results in terms of standard precision and recall statistics. The precision and recall for geographically relevant cluster retrieval is computed as:

$$\text{precision} = \frac{|R \cap G|}{|R|} \qquad \text{recall} = \frac{|R \cap G|}{|G|}$$

where R is the set of retrieved clusters and G is the set of clusters judged to be geographically relevant. The precision and recall for temporally relevant cluster retrieval are computed in a similar way.

Figure 9(left) shows the precision-recall curves for retrieving geo relevant clusters, in which the average second moment threshold decreases from left to right on each curve. For example, for geo clusters in geo relevant cluster retrieval, when the average second moment threshold is 0.04, both the precision and recall are 93.1%. Motion clusters performed slightly worse at high recalls.

Figure 9(right) shows the precision-recall curves for retrieving temporally relevant clusters. The precisions and recalls for temporal clusters are worse than geo relevant cluster retrieval. When the average second moment threshold is 0.07, the precision is 71.4% and recall is 76.9%. Motion clusters performed much worse, because (as we discussed above) for motion clusters the average second moment does not have strong correlation with temporal relevance. In the ground truth, only 5 clusters were judged to be temporally relevant and their average second moment ranks ranged from 13 to 33. As future work, it would be interesting to study alternative statistics other than second moment that may perform better for motion clusters.

We found that the retrieval statistics were best when ground truth was defined by the group of users who were shown our visualizations; when the recall is 90.3%, the precision reaches 96.6% for the geo-spatial clusters, and reached 68.4% recall and 92.9% precision for the temporal features. This result suggests that the visualizations helped users see subtle geospatial or temporal connections between tags in a cluster.

5. CONCLUSION

In this paper, we proposed techniques to measure the semantic similarity of tags by comparing geo-spatial, temporal, and geo-temporal patterns of use. We used these techniques on a large dataset from Flickr to cluster tags using geo-temporal distributions and proposed novel methods to visualize the resulting clusters. An evaluation and case study showed the overall high quality of the semantics mined by our approach, and that the second moment served as a simple filtering measurement that achieved promising performance in selecting geographically- and temporally-relevant clusters. A case study suggests that our visualizations of tag semantics can help people understand subtle geo-temporal relationships between tags.

There are many possible improvements and future directions for this research. Currently, we are using only North American data and clustering the top 2000 most used tags into 50 clusters. It would be interesting to apply our approach within a more flexible framework, deciding the number of tags and the number of clusters in an automatic way. It would also be interesting to build a tag recommendation system that integrates our techniques, using multiple kinds of tag similarity metrics to improve results and give corresponding visualizations to the user. Finally, our approach could be

applied to other collections of objects with geographical and temporal attributes, such as data from Wikipedia or Twitter.

6. ACKNOWLEDGEMENTS

We thank Prof. Andrew Hanson for discussions and advice on visualization, and the anonymous reviewers for their helpful comments. This work was supported in part by a grant from the Lilly Endowment Inc. and by the Data to Insight Center at Indiana University.

7. REFERENCES

[1] S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL*, 2007.

[2] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *WWW*, 2008.

[3] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop*, 2006.

[4] L. Chen and A. Roy. Event detection from flickr data through wavelet-based spatial analysis. In *CIKM*, 2009.

[5] S. Chien and N. Immorlica. Semantic similarity between search engine queries using temporal correlation. In *WWW*, 2005.

[6] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW*, 2009.

[7] C. Ding, D. Zhou, X. He, and H. Zha. R1-PCA: Rotational invariant L1-norm Principal Component Analysis for robust subspace factorization. In *ICML*, 2006.

[8] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. *ACM Transactions on the Web*, 1(2), 2007.

[9] V. Estivill-Castro and I. Lee. Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In *Intl. Conf. on Geocomputation*, 2001.

[10] V. Estivill-Castrol and A. Murray. Discovering associations in spatial data – an efficient medoid based approach. In *PAKDD*, 1998.

[11] N. Garg and I. Weber. Personalized tag suggestion for Flickr. In *WWW*, 2008.

[12] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2), 2005.

[13] Y. Huang and P. Zhang. On the relationships between clustering and spatial co-location pattern mining. In *ICTAI*, 2006.

[14] X. Jin, A. C. Gallagher, L. Cao, J. Luo, and J. Han. The wisdom of social multimedia: using Flickr for prediction and forecast. In *ACM MM*, 2010.

[15] G. Karypis and V. Kumar. Parallel multilevel k-way partitioning scheme for irregular graphs. In *Proc. Supercomputing*, 1996.

[16] O. Kaser and D. Lemire. Tag-cloud drawing: Algorithms for cloud visualization. In *WWW Workshop on Tagging and Metadata for Social Information Organization*, 2007.

[17] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *ACM MM*, 2007.

[18] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *WWW*, 2009.

[19] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[20] M. Mohebbi, D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi, and S. Kumar. Google correlate whitepaper. http://www.google.com/trends/correlate/whitepaper.pdf, 2011.

[21] E. Moxley, J. Kleban, and B. S. Manjunath. Spirittagger: a geo-aware tag suggestion tool mined from flickr. In *MIR*, 2008.

[22] E. Moxley, J. Kleban, J. Xu, and B. S. Manjunath. Not all tags are created equal: Learning Flickr tag semantics for global annotation. In *ICME*, 2009.

[23] A. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *ICML*, 2004.

[24] R. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *VLDB*, 1994.

[25] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *WWW*, 2011.

[26] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *SIGIR*, 2007.

[27] J. Sander, M. Ester, and H. P. Kriegel. Density-based clustering in spatial databases: A new algorithm and its applications. *Data Mining and Knowledge Discovery*, 1998.

[28] A. Shepitsen, J. Gemmell, B. Mobasher, and R. D. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proc. ACM Conf. on Recommender Systems*, 2008.

[29] Z. Sheppard. Flickr blog: 5,000,000,000. http://blog.flickr.net/en/2010/09/19/5000000000/, 2010.

[30] B. Sigurbjörnsson and R. V. Zwol. Flickr tag recommendation based on collective knowledge. In *WWW*, 2008.

[31] C. A. Sugar and G. M. James. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 98, 2003.

[32] S. Vadrevu, Y. Zhang, B. Tseng, G. Sun, and X. Li. Identifying regional sensitive queries in web search. In *WWW*, 2008.

[33] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *SIGMOD*, 2004.

[34] L. Wu, X. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *ACM MM*, 2008.

[35] X. Xiao, X. Xie, Q. Luo, and W.-Y. Ma. Density based co-location pattern discovery. In *Intl. Conf. on Advances in Geographic Information Systems*, 2008.

[36] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, 2011.