# Discovering Relationships among Tags and Geotags

**Sang Su Lee, Dongwoo Won, and Dennis McLeod**

Computer Science Department
University of Southern California
Los Angeles, CA 90089
{sangsl, dwon, mcleod}@usc.edu

**Abstract**

This paper presents an analysis of the correlation of annotated information unit (textual) tags and geographical identification metadata geotags. In this paper, to make it possible for geotagging to be used in analysis with tagging, we prove that there is a strong correlation between tagging and geotagging information. Our approach uses tag similarity and newly employed geographical distribution similarity to determine inter-relationships among tags and geotags. From our initial experiments, we show that the power law is established between tag similarity and geographical distribution similarity; they are strongly correlated and the correlation can be used to find more relevant tags in the tag space. The power law, which is any polynomial relationship that exhibits the property of scale invariance, confirms that there is the relationship between tagging and geotagging and the relationship is scalable in size of tags and geotags.

## Introduction

The use of user-generated tags on the unit of information is very popular. The popularity is based on the fact that there is no restriction on the format of the tag. This characteristic, however, produces inconsistencies like polysemy, synonyms, and word inflections. These inconsistencies hinder users in searching for appropriate resources from tag spaces. To overcome this drawback, researchers are trying to find the relations among tags. Tag relations, however, still have a deficiency. Although tagging systems are evolving, tag relation does not reflect the dynamics of tag space, especially for the new function called geotagging. Geotagging is the process of adding geographical identification metadata to various resources. However, geotagging information has not been included in the analysis to improve tag relations. We believe that adding geotagging information to retrieve new relation among tags enables the current tag relation to be more precise and relevant. To support this, our paper focuses on discovering the relationship between tagging and geotagging.

To find tag relationships, we present the tag similarity based on the numbers of photos annotated by each tag in Flickr.com. Then we calculate the geographical clusters for each tag and calculate geographical distribution similarities for clusters. The derived similarities are used for retrieving the relationship between tagging and geotagging.



## Approach

### Tag Similarity Calculation

Each photo in the tag space such as Flickr.com has related tags that describe the characteristic of the photo and are attached by tagging users. From photo-tag information, we create the feature vector for each tag to calculate similarity among tags. If tag A is co-annotated with other tag B, A was considered feature of B and vice versa. Following previous work (Pantel & Lin 2002), the value of the feature vector is the point-wise mutual information between tag and its each feature (co-occurring tags). Point-wise mutual information between the tag and co-occurring tag was used as feature weight.

$$mi_{w,c} = \frac{p(w,c)}{p(w) \times p(c)} \quad \ldots (1)$$

In equation (1), $c$ is the co-occurring tags, $w$ is the tag, and $p(w, c)$ is the frequency count of a tag $w$ occurring in co-occurring tags $c$. Again, these point-wise mutual information values were multiplied with a discounting factor to mitigate bias towards infrequent words. Once feature vectors are created, simple cosine similarity was used to calculate similarity between all tags.

### Geographical Cluster Calculation

After calculating tag similarities, we create the geographical clusters for each tag using the coordinate (latitude and longitude) information of the photos. First, we retrieve all relevant location information for each tag. From that information, we use k-means and k-means++ algorithms to generate geographical clusters for tags. The k-means algorithm is efficient, but it comes with the low accuracy. The accuracy of the result largely depends on the initial set of clusters. To find the best possible initial set of seed points, the k-means++ algorithm (Authur & Vassilvitskii, 2007) is adapted. The idea of the k-means++ algorithm is to maintain the distances among the seed points as much as possible. By the k-means++, we can generate clusters with better accuracy. Every generated cluster has three attributes: name of the tag, coordinate of

the centroid, and radius of the cluster. The radius of the center is the average distance from the centroid to its member points and is calculated by the Euclidean distance. Clusters are defined as the shape of a circle.

### Geographical Distribution Similarity (GDS) Calculation

The next step is to calculate how geographically similar two tags are. In the previous section, the first output is the circle-shape clusters held by each tag on the coordinate system. For two arbitrary tags, corresponding clusters are retrieved and the similarity of the clusters from the two tags is calculated. To calculate the similarity of two clusters, we find the size of the overlapped regions in clusters of two different tags and calculate the total size of the clusters from two tags. Then the size of overlapped regions is divided by the total size of the clusters and the result of division becomes the geographical distribution similarity (GDS).

## Experiment and Evaluation

We have collected raw data from a photo-sharing web site, Flickr.com. We have randomly selected approximately 340 tags and retrieved 5000 photos data per tag using Flickr API. For our experiment, 729,948 photos are collected as an initial dataset. The dataset includes 12,545 distinct tags and 54,811 users. 89,855 photos are retrieved with geotagging information and 50,262 tags are associated with geotagging information.

We now consider how to find the relation between tag similarity and geographical distribution similarity of tags. To discover the relation between two different similarities, we need additional factors. One of them is the photo frequency of tag x, which we denote as *pf(x)*. Each tag has different number of annotations for photos and thus has different popularities. To distinguish the popularity of tags, we introduce the photo frequency of tags, which is the percentage of photos that use the specific tag over all photos. Another is the number of users using this tag. Each tag has a different number of users who use the tag and needs to be dealt differently based on the popularity of users. For this reason, the idea of user frequency is introduced. The user frequency *uf(x)*, where x is the tag, is the percentage of users that use the certain tag over all users in the tag space.

To determine the relationships between tag similarities and GDS, following factors such as *sim(x, y)*, *geo_sim(x, y)*, *pf(x)*, *pf(y)*, *uf(x)*, and *uf(y)* are employed. The equations for similarity *SIM(x, y)* and the weighted geographical distribution similarity *GEO_SIM(x, y)* are as follows.

$$SIM(x,y) = sim(x,y) * pf(x) * pf(y) * uf(x) * uf(y) \quad \text{... (2)}$$

$$GEO\_SIM(x,y) = geo\_sim(x,y) * pf(x) * pf(y) * uf(x) * uf(y) \quad \text{... (3)}$$

For a visualization of the relation, we first provide the log-log plot for two weighted similarities. The left graph in Figure 1 shows the distribution of linear regression in the log-log space, which is denoted as $\log y = \alpha \log x + \log c$ , where *y* is *GEO_SIM* and *x* is *SIM*.

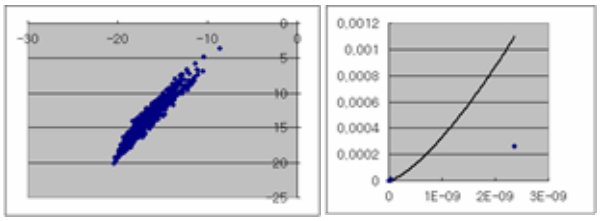


**Figure 1. Linear Distribution of SIM(x,y) and GEO_SIM(x,y) in log-log space(left) and Power Law Distribution of SIM(x,y) and GEO_SIM(x,y)**

Usually the linear regression in the log-log space has a common meaning for following the power law, which is the relationship between two scalar quantities *x* and *y* in the form of $y = cx^{\alpha}$ . The linearity in the left graph in Figure 1 may be the evidence of the power law. In the power law equation, *c* is $9.04690438 \times 10^{-10}$ and $\alpha$ is 1.3914. The right graph in Figure 1 shows the power law distribution. But, linear regression in the log-log space can cause a bias in the value of the power law exponent. Hence, we need to refine the result by the power law validation suggested from Newman (2005).

As a result of our evaluation, the following two interpretations can be drawn from this distribution. First, the result shows that the relationship between *SIM* and *GEO_SIM* follow the power law distribution with the high probability. This reveals that geotagging and tagging are closely related to each other in terms of tag similarity and GDS. This evidence helps us to draw the conclusion that both geotagging and tagging information can be integrated into the tag search problem, allowing users to get more refined and relevant tag search results. Second, our approach assures the scalability. Our analysis is supported by the scale-free characteristic of power law. Scale invariance is a feature of objects or laws that do not change when length scales are multiplied by a common factor. Thus, this relationship is maintained regardless of the size of tag pair examples.

## Acknowledgements

This work was funded in part by a grant from the Department of Homeland Security, ONR Grant number N00014-07-1-0149.

## References

Arthur, D and Vassilvitskii, S. 2007. k-means++: The Advantages of Careful Seeing. In *Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027-1035.

Pantel, P. and Lin, D. 2002. Discovering word sense from text, In *Proc. of the 8th ACM SIGKDD*, 613-619.

Newman, M.E.J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*. 323–351.