Tag-Geotag Correlation in Social Networks

Sang Su Lee
Computer Science Department
University of Southern California
Los Angeles, CA 90089
1-213-740-3696
sangsl@usc.edu

Dongwoo Won
Computer Science Department
University of Southern California
Los Angeles, CA 90089
1-213-740-4521
dwon@usc.edu

Dennis McLeod
Computer Science Department
University of Southern California
Los Angeles, CA 90089
1-213-740-4504
mcleod@usc.edu

ABSTRACT

This paper presents an analysis of the correlation of annotated information unit (textual) tags and geographical identification metadata geotags. Despite the increased usage of geotagging in collaborative tagging systems, most current research focuses on textual tagging alone in solving the tag search problem. This may result in difficulties to search for precise and relevant information within the given tag space. For example, inconsistencies like polysemy, synonyms, and word inflections with plural forms complicate the tag search problem. Therefore, more work needs to be done to include geotag information with existing tagging information for analysis. In this paper, to make geotagging possible 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 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: this means that tag similarity and geographical distribution similarity has a strong correlation and the correlation can be used to find more relevant tags in the tag space. The power law confirms that there is an increased relationship between tagging and geotagging and the increased relationship is scalable in size of tags and geotags. Also, using both geotagging and tagging information instead of only tagging, we show that the uncertainty between derived and actual similarities among tags is reduced.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Content Analysis and Indexing; H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Measurement

Keywords

tagging, geotagging, clustering, power law for correlation among tagging and geotagging

1. INTRODUCTION

Uses of user-generated tags are increasingly popular. A tag is the relevant keyword or term that is associated with or assigned to a unit of information. A tag describes the item and enables keyword-based classification of information that the tag is related to. Tagging is acknowledged as a useful way to accumulate and categorize information (e.g. bookmarks, blog posts, articles, photos and videos). Often, users can annotate resources without restriction in format and without the limitation on number of tags per each resource. These characteristics allow regular users to facilitate tagging. In addition, another benefit of tagging systems is the vocabulary enhancement [12]. It is aided with shared tag data set generated by numerous users. It can also reduce the burden of building comprehensive and correct metadata. In spite of these benefits, tagging systems have a critical limitation. One characteristic of tagging system is that there is no rigid format. This may produce the following inconsistencies:

1. polysemy, words with multiple related meanings (eg. a window can be a operating system or a sheet of glass)

2. synonyms, multiple words with the same or similar meanings (eg. tv and television, Netherlands/Holland/Dutch)

3. word inflections with plural forms (eg. "cat" versus "cats")

The inconsistencies impede users from finding appropriate resources by keywords. To overcome the drawback, researchers are trying to find the relations among tags. The relation among tags can bridge the links between synonyms and provide a standard to classify polysemy into several subgroups. One of the researches is found in the work of Flickr [5], where it attempts to cluster tags for user convenience.

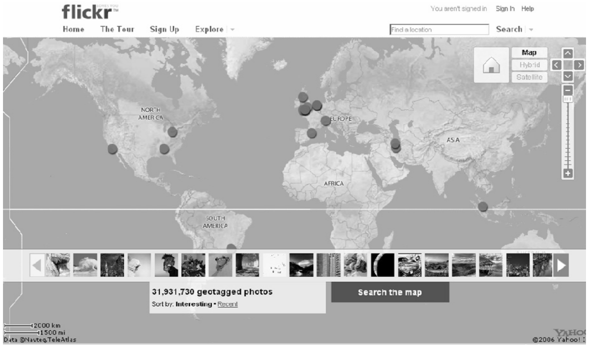

Figure 1. Geotagged Photos on Flickr

SSM'08, October 30, 2008, Napa Valley, California, USA.

Tag relation, however, still has a deficiency. Even though tagging systems are evolving, tag relation does not reflect the change, especially the new function called geotagging. Geotagging is the process of adding geographical identification metadata to various media such as websites, RSS feeds, or images. Recently, geotagging has been used widely from users in collaborative tagging systems. Figure 1 shows an example of a geotagging photo page in Flickr. But, geotagging information has not been included for 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 finding out strong relationship between tagging and geotagging.

In this paper, we show three steps to confirm our approach. This approach has been first introduced in [10], but here, we further elaborate the detail of our approach by providing examples and a new evaluation method. We present the tag similarity based on cosine similarity and point-wise mutual information in order to articulate similarities among tag pairs. Then, we calculate the geographical clusters for each tag based on k-means and k-means++ algorithm [1] for lowering squared sum of errors in cluster creation. After the creation of the geographical clusters, we calculate geographical distribution similarity for clusters. The remainder of our paper is the following: we discuss related work in section 2. We describe the algorithm for tag similarity, the algorithm for generating the geographical clusters for each tag, and the algorithm to calculate the geographical distribution similarity of tags from clusters which are created by the algorithm in Section 3. Section 4 evaluates our approach for finding the correlation between tag similarity and geographical distribution similarity. Section 4 also shows the reduction of entropy when tagging and geotagging are used together. We conclude our paper with future work in section 5.

2. RELATED WORK

We have investigated related works to propose our approach. At first, papers about the structure of the collaborative tagging system can be found. A paper regarding social analysis for tagging behavior is discussed next. Then, special attention is paid to papers focusing on the tag similarity using various techniques. Lastly, a paper about geotagging in collaborative tagging system is mentioned.

There are intriguing papers about the analysis for current tagging framework [6, 12]. Both papers build the criteria to classify tags and investigate social aspects of tagging. [6] asserts that tagging is a kind of social activities, because tag usages are stabilized by imitation and shared knowledge. For examples, users from a social bookmark website called del.icio.us [4] can imitate other users' tag choices and illustrate tagging activity as a kind of social activities. Also, [12] refers to social incentives that express the communicative nature of tagging. Authors show the increase in number of tags is proportional to the increase in number of contacts. Furthermore, they have made known the relationship between affiliation and tag vocabulary formation by showing that users linked by the contacts use similar tag vocabularies, i.e. tagging activity can be related to social activity as authors in [6] have pointed out. Both papers indicate that it is necessary for us to take users into consideration in analyzing tagging systems.

The social aspect of tagging systems is further investigated in [9]. This work focuses on social psychological aspects of tagging behavior in del.icio.us. Here, it articulates the relation between the user's annotative tendency and the degree of perceived social presence, which is the key concept of this approach. Many human social activities are carried out due to social position and association. This point of view can be applied in the analysis of collaborative tagging systems. Users who recognize other users in online communities have higher chance to tag resources more precisely and actively.

There are some interesting works about finding tag relations to solve the search problem in the tagging system [2, 3, 7, 15]. Brooks and Montanez [2] induce a hierarchy of tags by utilizing data from Technorati [18]. They have used agglomerative clustering technique to iteratively cluster similar blog articles using cosine similarity metric. Belelman *et al.* [3] propose a technique similar to spectral clustering to generate tag clusters in del.icio.us. Several small graphs of tag relation are resulted from clustering a big graph using tag similarity. Heyman and Garcia-Molina [7] suggest creating hierarchical taxonomies of tags that are aggregated into tag vectors using cosine similarity metric. In the work of Schmitz [16], he has generated an ontology of tags in Flickr. A subsumption-based statistical model is adapted to generate a graph of possible parent-child relationships. All papers above are using different ways to find tag similarity, but there is one thing in common. They have tried to find tag similarity based on co-occurrence of tags from resources.

[8] is one of few papers available regarding geotagging in collaborative tagging systems. It employs disparate information - tags, the location information of photos, and photos themselves - to generate knowledge like the representative photos in certain areas. Authors use location-driven approach to generate aggregate knowledge in the form of representative tags for arbitrary areas in the world. They also use a tag-driven approach to automatically extract place and event semantics for Flickr tags, based on each tag's metadata patterns. Based on the extracted patterns, vision algorithms are employed with greater precision. The significance of this paper is that it has been the first approach to create knowledge from tagging, geotagging, and photos. This work, however, extracts knowledge separately, and therefore lacks in expressing compound information like tagging and geotagging information together.

3. APPROACH

The approach taken in this research consists of three parts. The first part is calculating tag similarity to discover tag relations. The second part is building geographical clusters with tags. The last part is calculating geographical distribution similarity for the geographical clusters of each tag.

3.1 Tag Similarity Calculation

Each photo has related tags which are used to describe the characteristic of the photo by the user of tagging systems. 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 [11, 14], the value of feature vector is point-wise mutual information between tag and its each feature (co-occurring tags). Point-wise mutual information between the tag and co-occurring tag is used as feature weight.

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

, where *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, following the work of [14], 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.

3.2 Geographical Cluster Calculation

In order to calculate the similarity of the geographical distribution between tags, at first we create the geographical clusters for each tag using the coordinate (latitude and longitude) information of photos. A photo has coordinate and annotated tags. We organize the data in order to observe which tags are annotated in which places. Based on geotagging data and annotated tags from photos, we assign the latitude and longitude information for each tag. Then, a tag which holds several related coordinate information is used to generate geographical clusters. Figure 2 below is the example of the tag-location assignment.

Photo	Tag
P_1	california
P_1	angels
P_1	football
P_2	football
P_2	seattle
P_3	football

Photo	Coordinate
P_1	34.013752,-118.288593
P_2	47.738169, - 122.359886
P_3	34.192778, - 118.174782

→

Tag	Coordinate
football	34.013752,-118.288593
football	47.738169, -122.359886
football	34.192778, -118.174782
angels	34.013752,-118.288593
california	34.013752,-118.288593
seattle	47.738169, -122.359886

Figure 2. Tag-Location Assignment

From tag-location information, we use k-means algorithm to generate geographical clusters for tags. The k-means algorithm is widely used in cluster generation because of its efficiency. In short, k-means algorithm is to cluster objects based on those attributes into k groups. The objective of k-means is to minimize the total intra cluster variance, or the sum of squared errors. Usually k-means works as follows:

1. Select K points randomly as the initial centroids
2. Form K clusters by assigning all points to the closest centroid
3. Recompute the centroid of each cluster
4. Repeat Step 2 and 3 until centroids does not change any more

But, the efficiency of k-means comes with the low accuracy. There is no guarantee that k-means algorithm finds a global optimum. On the contrary, there are many examples that k-means generates bad clusters in terms of the accuracy. The accuracy of the result largely depends on the initial set of clusters. Another disadvantage is that the number of k should be specified prior to executing the algorithm. We propose a way to overcome those disadvantages of k-means as follows.

To improve the accuracy, we need to find the best possible initial set of seed points. The k-means++ algorithm [1] is adapted to find appropriate seed points. From the experiment of [1], the authors show that k-means++ improves the accuracy of k-means algorithm while maintaining the speed and simplicity of the algorithm. The idea of k-means++ algorithm is to maintain the distances among the seed points as farther as possible. The k-means++ selects initial centers in a way that they are already initially close to large quantities of points. After that, *D(x)*, which is the shortest distance from a data point x to the closest center already chosen, is calculated. Using *D(x)*, the probability named D^2 weighting is calculated and is employed to choose the next center. The k-means++ algorithm works like below.

1. Take one center c_1, chosen uniformly at random from X
2. Take a new center c_i choosing $c_i = x' \in X$ with probability $D(x')^2 \Big/ \sum_{x\in X} D(x)^2$
3. Repeat Step 2 until we have taken k
4. Proceed as with the standard k-means

Then we use heuristics in choosing the appropriate number of k, which enables to maintain the sum of squared errors as small as possible. We assume the k-means++ algorithm can give the lowest possible sum of squared errors for the arbitrary number of k. Based on this assumption, we start to find the location of the initial seed point, which holds the lowest squared sum of errors for k=1. Then we increase the number of k gradually and execute k-means++ until we find the lowest squared sum of errors. As a result, we are able to find the number of k and the locations of k initial seeding points that give the lowest sum of squared errors from all possible numbers of k. The whole procedure works as follows.

1. Find locations of initial seeding points by k-means++
2. For calculated initial seeding points, execute k-means
3. Increment the number of k
4. Repeat above steps until the sum of squared errors is the smallest

Based on k-means and k-means++ algorithm, we generate clusters for tags. Every cluster has three attributes: name of the tag, coordinate of the centroid, and radius of the cluster. 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 a circle shape.

3.3 Geographical Distribution Similarity (GDS) Calculation

The next step is to calculate how geographically similar two tags are. To find the geographical similarity of two tags, we exploit the geographical aspect of tags. In the previous section, the first output has been the circle-shape clusters held by each tag on the coordinate system. These clusters are resources in articulating the geographical distribution similarities of different tags. For two arbitrary tags, corresponding clusters are retrieved and the similarity of clusters from two tags is calculated. This similarity of two clusters indicates how two tags are similar in the geographical locations. To calculate the similarity of two clusters, we find the size of overlapped regions in clusters of two different tags. Then, we calculate the total size of clusters from two tags. Figure 3 shows geographical clusters and overlapped regions of different clusters.

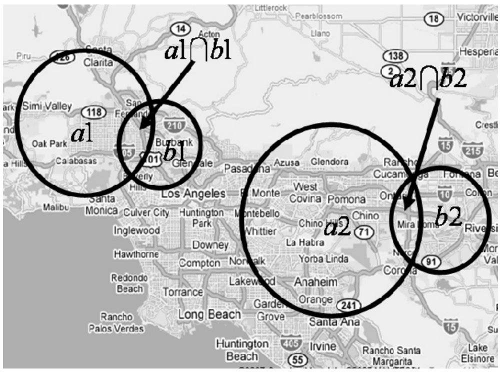

Figure 3. Overlapped Regions among Two Different Clusters

There are $\{a1, a2\} \in A$ and $\{b1, b2\} \in B$, where *A* and *B* are sets of clusters for different tags. *a1, a2* and *b1, b2* are geographical clusters for tag A and B respectively. As shown in Figure 3, $a1 \cap b1$ and $a2 \cap b2$ mean the overlapped regions from A and B tags. The regions that is proportional to whole regions from A and B is referred to as the geographical similarity of two tags. The equation to find the geographical distribution similarity is shown below.

$$geo\_sim = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} a_i \cap b_j}{\sum_{i=1}^{m} a_i + \sum_{j=1}^{n} b_j - \sum_{i=1}^{m}\sum_{j=1}^{n} a_i \cap b_j} \quad \ldots (2)$$

, where $\{a_1, ..., a_m\} \in A$ and $\{b_1, ..., b_n\} \in B$. Prior to using Eq. (2), we need to check whether at least one overlapped region exists or not for the efficiency. To do this, cluster pairs from two different tags are retrieved and overlapped regions are uncovered if exist. If there is at least one overlapped region, the similarity is calculated. Next, we calculate the similarities regarding all possible tags pairs. The whole procedure works as follows:

1. For each tag T_i, retrieve all relevant geographical clusters, a_1, …, a_m
2. For each tag T_j, retrieve all relevant geographical clusters, b_1, …, b_n
3. If T_i and T_j have overlapped regions, calculate overlapped regions and retrieve geographical distribution similarity(*geo_sim*) by Eq. (2)
4. Repeat above steps until there is no overlapped regions for tag pair T_i and T_j

So far, we have shown the steps for calculating GDS (Geographical Distribution Similarity). In the next section, we are going to reveal the relation between the tag similarity and GDS.

4. EXPERIMENT AND ANALYSIS

4.1 Experiment

The machine we have used for this experiment has Pentium 4 2.4GHZ CPU and 1GB memory. The operating system is Windows XP. Our approach is implemented by Python 2.5 and Java J2SE 1.5. At the beginning, we have collected raw data from a photo-sharing web site, Flickr.com. The data from Flickr.com consists of four elements: the owner information of the photo, the tags attached to the photo, geotag, and the photo itself. We have randomly selected approximately 340 tags and retrieved 5000 photos data per tag. The raw data is retrieved 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.

Table 1. Example Raw Data from Flickr

Photo ID	User ID	Tags	Lng/Lat
138602759	1277457 4@N00	audience, baseball, nyc, ny, yanks,bronx, newyork, newyorkcity, stadium, yank, yankee, yankees, yankeestadium	-74.157715/ 40.797176
143045374	5426641 9@N00	swedenborgian, seder,church, newyork, maundythursday holyweek, newchurch	-73.980354/ 40.747257
149113015	4320966 5@N00	yankeestadium, newyork,	-73.92859/ 40.82696

Table 1 shows a partial example of our raw dataset. Using data, we first calculate the tag similarity. From the data, photo IDs and relevant tags are retrieved. Then the feature vector for each tag is calculated and those vectors are used to calculate the cosine similarity between two tags. Table 2 shows the partial result of tag similarity calculation for a tag named "newyork".

Table 2. Relevant Tags for "newyork"

Tag 1	Tag 2	Similarity
newyork	newyorkcity	0.2591218990962
newyork	gothamist	0.2255912261205
newyork	bronx	0.1284546640474
newyork	aia150	0.1207746002249
newyork	nycpb	0.1152881157987
newyork	podcast	0.1092873413209

newyork	yankeestadium	0.1039956324392

Next, again from the raw data, we generate geographical clusters for each tag. As we have already shown in Figure 2, Tag-Location assignment is performed. After that, tags and coordinates from Tag-Location assignment are applied to the algorithm (see Section 3.2) to generate geographical clusters for each tag. Table 3 is the partial result of geographical clusters for tag "newyork" and "newyorkcity". A cluster which Cluster ID starts with ny is a cluster of "newyork" and a cluster which Cluster ID starts with nyc is a cluster of "newyorkcity".

Table 3. Geographical Clusters for "newyork" and "newyorkcity"

Cluster ID	Longitude of Centroid	Latitude of Centroid	Cluster Radius
ny1	40.8275305	-73.9265935	0.009425306704
ny2	40.730645142	-73.990243428	0.018889052616
ny3	40.77757425	-73.970645	0.016690971574
ny4	40.826908947	-73.928367578	0.000606728735
nyc1	40.76105525	-73.9758085	0.0093464984609
nyc1	40.82771825	-73.92622025	0.0007698066395
nyc1	40.703349666666	-73.99447833333	0.0223051413547
nyc1	40.827328090909	-73.92839290909	0.0008228534878

Once we have generated the geographical clusters, we can calculate the geographical distribution similarities for arbitrary tag pairs. For example, suppose we are to find GDS for tags "newyork" and "newyorkcity". According to Eq. (2), the overlapped regions of clusters and the total size for all of clusters from two tags "newyork" and "newyorkcity" need to be derived. Derived overlapped regions are these:

$$ny1 \cap nyc2 = 1.5542242964121083\text{E-}6$$

$$ny2 \cap nyc3 = 2.8555575041052224\text{E-}4$$

$$ny3 \cap nyc1 = 1.0982379886217187\text{E-}4$$

$$ny4 \cap nyc2 = 9.65614433130694\text{E-}7$$

All above values are added into total overlapped size:

$$total\_overlapped\_size = 3.978993880022369\text{E-}4$$

Then, we calculate the total size of all clusters by adding each cluster's area which is same as $\pi * (cluster\_radius)^2$.

$$total\_size = 0.0038395523703844159$$

By applying total overlapped size and total size to Eq. (2), the GDS for "newyork" and "newyorkcity" can be calculated:

$$geo\_sim(\text{"}newyork\text{"},\text{"}newyorkcity\text{"})$$
$$= \frac{3.978993880022369\text{E-}4}{0.0038395523703844159 - 3.978993880022369\text{E-}4}$$
$$= 0.11561287266295696$$

GDS for other tag pairs can be calculated exactly as the procedure we have just mentioned. Table 4 is a list of related tags to the tag "newyork" with tag similarity and GDS.

Table 4. Tag Similarity-GDS List for "newyork"

Tag 1	Tag 2	Similarity	GDS
newyork	newyorkcity	0.2591218990962	0.1156128726629
newyork	gothamist	0.2255912261205	0.0047746618822
newyork	bronx	0.1284546640474	3.926857876e-06
newyork	aia150	0.1207746002249	0.0002197668497
newyork	nycpb	0.1152881157987	0.0010182582011
newyork	podcast	0.1092873413209	1.139363029e-06
newyork	yankeestadium	0.1039956324392	0.0010063674171

4.2 Analysis

In this section, we find the relation between tag similarity and geographical distribution similarity of tags. To do this, we have first calculated the tag similarities and geographical distribution similarities. Then, we introduce other factors to discover the relation between two different similarities. The one thing is the photo frequency *pf(x)*, where *x* is a tag. It means how many photos use this tag. The other thing is the user frequency *uf(x)*, where *x* is a tag. It means how many users use this tag. If there is a tag which is used by only one photo and a tag which is used by various photos, these two tags have different popularities and must be dealt differently. To distinguish the popularity of tags in terms of the numbers of photos that are using the tags, we introduce a term called photo frequency of tags. It is the percentage of photos that use the specific tag by all photos. The other thing we need to consider is how many users use this tag. One user can take a number of photos and he or she can use only one tag to annotate all his or her photos. In this case, even though the number of the certain tags is huge, it is only used by a single user. A tag used by only one person and a tag used by many people need to be distinguished. For that reason, the idea of user frequency is introduced. The user frequency is the percentage of users that use the certain tag.

In finding out the relation 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. We have weighted similarity *SIM(x, y)* and the weighted geographical distribution similarity *GEO_SIM(x, y)* as the equation below.

$$SIM(x,y) = sim(x,y) * pf(x) * pf(y) * uf(x) * uf(y) \ \ldots \ (3)$$

$$GEO\_SIM(x,y) = geo\_sim(x,y) * pf(x) * pf(y) * uf(x) * uf(y) \ \ldots \ (4)$$

, where *x* and *y* are tags, *sim(x, y)* is the similarity between two tags *x* and *y*, *geo_sim(x, y)* is the GDS of two tags *x* and *y*, *pf(x)* means the photo frequency of tag *x,* and *uf(x)* means the user frequency of tag *x*. For the clear visualization of the relation, we provide a log-log plot for two weighted similarities. Figure 4 reveals the relation between two similarities.

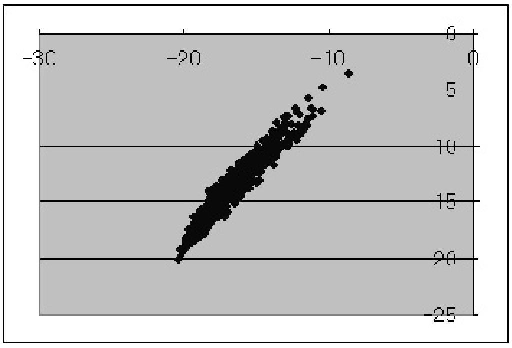

Figure 4. Distribution of log (SIM(x, y)) and log(GEO_SIM(x, y))

(*SIM(x, y)*) and Y axis is log (*GEO_SIM(x, y)*). From the graph above, the regression equation is derived as shown below.

$$\log(GEO\_SIM(x,y)) = 1.3914(\log(SIM(x,y)) - 9.0435 \ldots (5)$$

We suppose that the regression is written as:

$$\log y = \alpha \log x + \log c \ldots (6)$$

Generally linear regression in the log-log space is considered that the distribution follows the power law. The straight trend in Figure 4 can be the evidence of the power law. The power law is a relationship between two scalar quantities x and y of the form:

$$y = cx^{\alpha} \ldots (7)$$

Eq. (7) is the same form once we remove the log-log scale from two axes. If we remove log-log scale from Eq. (5), the equation can be also written as:

$$(GEO\_SIM(x,y)) = \mathrm{c} * (SIM(x,y))^{\alpha} \ldots (8)$$

In Eq. (8), c is $9.04690438 \times 10^{-10}$ and α is 1.3914. Before we investigate the meaning of this distribution regarding *SIM(x, y)* and *GEO_SIM(x, y)*, we need to validate whether this distribution follows precisely the power law or not. As mentioned earlier, the most simple and widely used way to check if a distribution follows the power law is to perform linear regression in the log-log space. However, [12] suggests that this can cause a bias in the value of the exponent. So, the following formula to determine α is proposed as one of reliable alternatives.

$$\alpha = 1 + n * \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]^{-1} \ldots (9)$$

, where x_i , i = *1…n* are the measured values of x and $x_{\min}$ that corresponds to the lowest value for which the power law holds. By applying Eq. (9), α is calculated as 1.184648305. Hence, the value of α from Eq. (8) and (9) implies that the distribution of tag similarity and GDS follows the power law with α *< 2.*

As a result of our evaluation, the following two interpretations can be drawn from this distribution. First important point is that it follows the power law distribution with increase relation. Hence, it reveals the fact that geotagging and tagging are closely related to each other in terms of tag similarity and GDS. The evidence helps us to arrive at a conclusion that both geotagging and tagging information can be integrated into the tag search problem allowing user to get more refined and relevant tag search results.

The other point is that 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, the shape of the distribution curve does not depend on the scale when we measure the quantity of the similarity. In other words, the increase relation is maintained regardless of the size of tag pair examples.

4.3 Mutual Information of Tagging and Geotagging

In this section, we try to show the effectiveness of using tagging and geotagging information together from a different point of view. If we suppose there is actual similarity which shows the exact degree of similarity between two tags, the derived tag similarity could reflect the actual similarity in some degree. However, we cannot say that the derived similarity is identical to the actual similarity. In other words, there is some uncertainty between the derived and the actual similarity among tags. Our view is that using tag similarity from 3.1 and GDS of tag from 3.3 together can reduce the uncertainty over using tag similarity from 3.1, and this reduced uncertainty shows the effectiveness of using tagging information and geotagging information together. Generally in measuring the uncertainty, Entropy [17] is useful. In addition, Mutual Information (MI) is applied in order to calculate the reduction of the uncertainty. MI is a measure of the reduction in the uncertainty about one random variable given the knowledge of another [17]. In our case, MI shows degree of the reduction in the uncertainty about tag similarity given by the knowledge of GDS of tag.

At first, tag similarity is calculated for all tags. The sum of tag similarities is a prior probability *p(TAG)*. Once tag similarity is calculated, we start to calculate GDS for tag pairs which already have tag similarities. The sum of these GDSs is set to a conditional probability *p(GEO_TAG|TAG)*. Also *p(GEO_TAG)* is calculated to retrieve conditional entropy *H(GEO_TAG|TAG)*. Using these probabilities, entropy for each probability is derived as below.

$$H(TAG) = -\sum_{i=1}^{n} p(TAG_i) * \log_2 p(TAG_i) \ldots (10)$$

$$H(GEO\_TAG) = -\sum_{i=1}^{n} p(GEO\_TAG_i) * \log_2 p(GEO\_TAG_i) \ldots (11)$$

$$\begin{aligned} &H(GEO\_TAG \mid TAG) \\ &= -\sum p(TAG) * \sum p(GEO\_TAG \mid TAG) * \log_2 p(GEO\_TAG \mid TAG) \end{aligned} \ldots (12)$$

So we can apply MI (mutual information) for measuring the reduction in the uncertainty about the tag similarity given GDS.

$$I(TAG;GEO\_TAG) = H(GEO\_TAG) - H(GEO\_TAG \mid TAG) \ldots (13)$$

For this calculation, totally 1937 flickr photos are selected and 2008 tags are retrieved from these photos. Based on different sizes of flickr photo samples, mutual information is calculated. Table 5 is derived from equation (10), (11), (12) and (13). Table 5 shows entropy values and mutual information values for each sample.

In order to compare MI from samples with different sizes, MI is normalized. One of the normalized mutual information is symmetric uncertainty coefficient [15]. Symmetric uncertainty coefficient is defined by

$$S(TAG;GEO\_TAG) = \frac{I(TAG;GEO\_TAG)}{\frac{1}{2}[H(TAG) + H(GEO\_TAG)]} \ldots (14)$$

Table 5. Entropy Values for Each Sample

	500 Samples	**1000 Samples**	**1500 Samples**	**1937 Samples**
H(TAG)	3155.86944477	5345.21053903	11168.7326426	13429.3327337
H(GEO_TAG)	256.949329838302	455.355260909928	1406.30046973579	2479.73547763694
H(GEO_TAG|TAG)	50.0350380542	92.934916633	229.015865051	254.147779714
MI	206.9142918	362.4203443	1177.284605	2225.587698

Using values from Table 5 and equation (14), Figure 5 is derived.

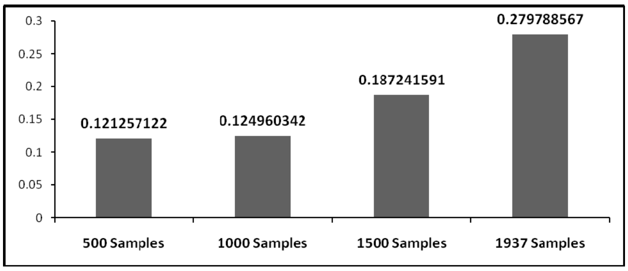

Figure 5. Symmetric Uncertainty Coefficient for Samples

Figure 5 shows symmetric uncertainty coefficient for each sample. From this graph, we can find the fact that MI between tags and geotags exists, and this MI indicates that the uncertainty between derived and actual similarities among tags is reduced by using tagging and geotagging information together rather than using tagging information only. In other words, using tags with geotags has a higher possibility to reach the actual similarity among tags than using just tags. The other information from Figure 6 is that symmetric uncertainty coefficient is increased as the size of sample data increases. It means that as the more data is retrieved, the less uncertainty between derived and actual similarities can be obtained.

5. CONCLUSION AND FUTURE WORK

We have shown how the tag similarity has strong relationships with the geographical distribution similarity. To do this, we first calculate the tag similarities from tag pairs. Then, we calculate geographical clusters for each tag. From those geographical clusters, we compute the geographical distribution similarities for tag pairs. Next, we introduce the weighted tag similarity and the weighted geographical distribution similarity for revealing the relation between the tag similarity and the geographical distribution similarity. By using those two weighted similarities, the linear regression in log-log scale is discovered. The result shows that one similarity increases as the other similarity increases. Additionally, the mutual information of tagging and geotagging is calculated and it shows using both tagging and geotagging information rather than using tagging information only can reduce the uncertainty between derived similarities and actual similarities among tags.

In the future, we plan to further explore a more appropriate metric for finding relevant tags by the association between tag similarity and geographical distribution similarity. We hope to see more refined results in searching the tag space. Next, we try to improve the geographical cluster generation. Usually, k-means clustering is weak from outliers and hence, we plan to generate good clusters in removing outliers. As well, the mutual information of tagging and geotagging is researched further. Lastly, we are working on finding relevant users in the collaborative tagging system by using the tag similarity. Users can be classified by the tags which they frequently use and users can be grouped together by the classification.

6. ACKNOWLEDGMENT

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

7. REFERENCES

[1] Arthur, D and Vassilvitskii, S. k-means++: The Advantages of Careful Seeing. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (New Orleans, Louisiana, USA, January 7-9, 2007). SODA 2007. Society for Industrial and Applied Mathematics, Philadelphia, PA. 1027-1035. DOI=http://doi.acm.org/10.1145/1283383.1283494

[2] Brooks, C.H. and Montanez, N. Improved Annotation of the Blogosphere via autotagging and hierarchical clustering. In Proceedings of the 15th international conference on World Wide Web (Edinburgh, Scotland, UK, May 23-26, 2006). WWW '06. ACM Press, New York, NY, 625-632. DOI=http://doi.acm.org/10.1145/1135777.1135869

[3] Belelman, G., Keller, P., and Smadja, F. Automated Tag Clustering: Improving search and exploration in the tag space. In Proceedings of Collaborative Web Tagging Workshop at WWW (Edinburgh, Scotland, UK, May 23-26, 2006).

[4] del.icio.us. http://del.icio.us/

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

[6] Golder, S. and Huberman, B. 2006. The structure of collaborative tagging systems. Journal of Information Science, 198-208.

[7] Heyman, P. and Garcia-Molina, H. 2006. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging System. Technical Report, Stanford University.

[8] Kennedy, L., Naaman, M., Ahern, S., Nair, R., and Rattenbury, T. 2007. How Flickr Helps us Make Sense of the World: Context and Contents in Community-Contribute Media Collections. In Proceedings of the 15th International Conference on Multimedia (Augsburg, Bavaria, Germany, September 23-28, 2007). MM '07. ACM Press, NY, 631-640. DOI=http://doi.acm.org/10.1145/1291233.1291384

[9] Lee, K.J. What goes around comes around: an analysis of del.icio.us as social space. In Proceedings of the 2006 ACM Conference on Computer Supported Cooperative Work (Banff, Alberta, Canada, November 4-8, 2006). CSCW '06. ACM Press, NY, 191-194. DOI=http://doi.acm.org/10.1145/1180875.1180905

[10] Lee, S., Won, D. and McLeod, D. 2008. Discovering Relationships among Tags and Geotags. In Proceedings of the Second International Conference on Weblogs and Social Media (Seattle, Washington, USA, March 30- April 2, 2008). ICWSM 2008.

[11] Lin, D. An Information-Theoretic Definition of Similarity. In Proceedings of the Fifteenth International Conference on Machine Learning (Madison, Wisconsin, USA, July 24-27, 1998). ICML 1988. Morgan Kaufmann, 296-304.

[12] Marlow, C., Naaman, M., Boyd, D., and Davis, M. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In Proceedings of the 17th ACM Conference on Hypertext and Hypermedia (Odense, Denmark, August 22-25, 2006). HYPERTEXT 2006. ACM Press, New York, NY, USA, 31-40. DOI=http://doi.acm.org/10.1145/1149941.1149949

[13] Newman, M.E.J. 2005. Power laws, Pareto distributions and Zipf's law. Contemporary Physics. 323–351. DOI=http://dx.doi.org/10.1080/00107510500052444.

[14] Pantel, P. and Lin, D. Discovering word sense from text, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Edmonton, Alberta, Canada, July 23-26, 2002). KDD 2002. ACM Press, New York, NY, USA, 613-619. DOI=http://doi.acm.org/10.1145/775047.775138

[15] Sarndal, C. A comparative study of association measures. 1974. Psychometrika. Vol. 39. 165-187

[16] Schmitz, P. Inducing Ontology from Flickr Tags. In Proceedings of Collaborative Web Tagging Workshop at WWW2006 (Edinburgh, Scotland, UK, May 23-26, 2006

[17] Shannon, C.E. 1948. A Mathematical Theory of Communication. Bell System Technical Journal. Vol. 27. 379-423, 623-656.

[18] Technorati. http://www.technorati.com