Resolving Tag Ambiguity

Kilian Weinberger
Yahoo! Research
Santa Clara, CA USA
kilian@yahoo-inc.com

Malcolm Slaney
Yahoo! Research
Santa Clara, CA USA
malcolm@ieee.org

Roelof van Zwol
Yahoo! Research
Barcelona, Spain
roelof@yahoo-inc.com

ABSTRACT

Tagging is an important way for users to succinctly describe the content they upload to the Internet. However, most tag-suggestion systems recommend words that are highly correlated with the existing tag set, and thus add little information to a user's contribution. This paper describes a means to determine the ambiguity of a set of (user-contributed) tags and suggests new tags that disambiguate the original tags. We introduce a probabilistic framework that allows us to find two tags that appear in different contexts but are both likely to co-occur with the original tag set. If such tags can be found, the current description is considered "ambiguous" and the two tags are recommended to the user for further clarification. In contrast to previous work, we only query the user when information is most needed and good suggestions are available. We verify the efficacy of our approach using geographical, temporal and semantic metadata, and a user study. We built our system using statistics from a large (100M) database of images and their tags.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.5 [**Pattern Recognition**]: Models

General Terms

Algorithms

Keywords

tagging, photos, query expansion, ambiguity

1. INTRODUCTION

Tags are an important part of today's multimedia databases. They are often contributed by users when they submit an image or video and form a key part of the search experience. Content-based multimedia search remains out of reach, and a simple tag like "Tokyo" provides more information than we can possibly glean from content-based algorithms. Thus, making it as easy as possible for users to enter tags alongside multimedia content is important. This work addresses the problem of eliciting high-quality tags from users.

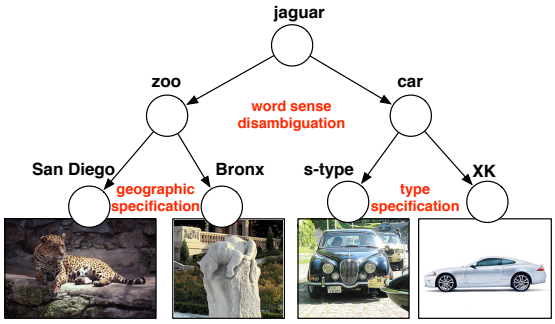

Figure 1: An example of how different expansion of one tag ("jaguar") can lead to very different image descriptions.

There have been numerous efforts to suggest tags to users [4, 15, 17, 19]. A common method is to suggest the most likely co-occuring tags. However, in many cases, the most likely tag is also the most obvious and least informative. For example, given the tag "goldengate," another common tag is "sanfrancisco," but this tag does not add any new information. Instead, we want to ask the user if they want to add "night," "sunny" or "fog."

There are two scenarios for which one would want to suggest a new tag to the user. The first scenario is if the current tag set has more than one meaning. Resolving this type of ambiguity is non-trivial, as there exist many different ways a tag set can appear ambiguous. Examples of ambiguity are word-sense ambiguity (e.g. "jaguar" can be a car or an animal), geographic ambiguity (e.g. "Cambridge" as in MA or UK), temporal ambiguity (e.g. "Superbowl" from 2006 or 2005), language ambiguity (e.g. "mist" means dung in German and fog in English), etc. Ideally, we would like to have one algorithm to handle all of these cases, without resorting to different additional side information every time (e.g. time or location analysis). The second scenario is if the current tag set is not sufficiently specific. For example, the tag "Asia" is not technically ambiguous; however, the images accompanied with this tag are very diverse and can range from busy street scenes in Tokyo to panorama shots of the Himilayas.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

We unify both scenarios into a single framework that allows us to find additional tags and quantify the potential benefit from adding them to the tag set. We will refer to both cases as *ambiguity*. Figure 1 illustrates how expanding a tag set can lead to very different image descriptions.

This work makes an important contribution to the literature. Most importantly, we propose in Section 3 a statistical means to suggest the tags that best reduce the ambiguity of the starting tag(s). As part of this statistical model, we measure the ambiguity of tag sets within the context of the data. This is important because we only want to interrupt the user to ask for more information when it is truly necessary and when additional tags can significantly reduce the measure of ambiguity. After tuning and verifying our algorithm with a user study (Section 5), we further test our approach by measuring the resolution of ambiguity using alternate channels of metadata, such as time and location (Section 6).

2. RELATED WORK

Tagging is a popular means of annotating objects on the web [10, 14]. Tagging allows people to describe and find interesting objects [4] and organize them by position [2]. Tagging is an important form of user-generated content and our work aims to improve the quality of tags.

There are several approaches for suggesting tags to users. Both Mishne [15] and Xu [19] propose systems that make suggestions by aggregating tags from similar textual content. Ames and Naaman propose a system called ZoneTag to make it easier for mobile-phone users to tag the photos they upload based on location and previous tags [4]. Finally, Sigurbjörnsson proposes a system based on a probabilistic model of tag usage across all users [17]. Each of these systems is looking for the most likely tags to describe content.

Similarly, there is an extensive literature for describing images based on image content and the words that surround an image reference. Heesch notes the importance of automated annotation for searching and browsing large image collections [12]. Wang selects candidate tags and determines which tags are suitable using the visual content of the image [18]. Zhou uses a heuristic greedy, iterative algorithm to estimate the probability that words are in the caption of an image by examining the text surrounding the images [20]. Compared to this work, our approach follows a different philosophy. We want to recommend clarifying tags when the current tags are not specific enough to describe an object.

Our work is perhaps most similar to the work that is done on query performance and query expansion. Cronen-Townsend suggest that query performance is correlated with the clarity of a query [8]. Clarity measures the ambiguity of a query with respect to a collection of textual documents. Amati proposed the notion of query difficulty to predict query performance [3]. Their basic idea is that the term weight, which is given by the divergence of the query terms' distribution in the top-retrieved documents from their distribution in the whole collection, provides evidence of the query performance. A similar approach is suggested by He [11]. Carpineto proposes an information-theoretic approach that expands a search query based on the top results of an initial ranking. The words of the top-retrieved documents are added to the search query and weighted based on the Kullback–Leibler (KL) divergence from the overall distribution of the entire corpus [5]. Though highly related to the detection of ambiguity of tags, the set of metrics for search query expansion and predicting query difficulty, such as clarity, is targeted for rich-text documents. We have a more difficult problem because our data is extremely sparse.

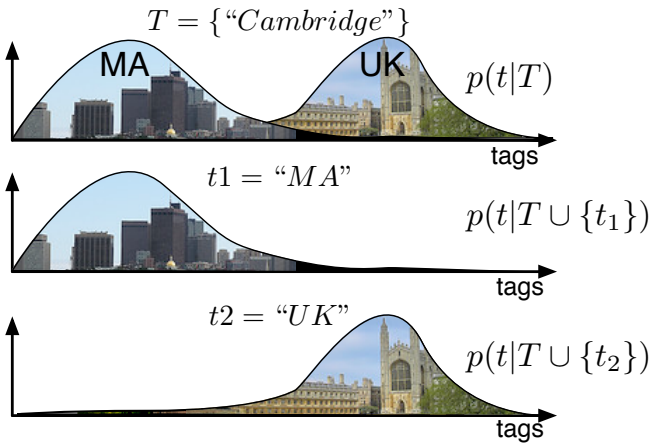

Figure 2: Three probability distributions over co-occuring tags, illustrating the ambiguity of a tag like “Cambridge” and the new distributions with additional tags such as “MA” or “UK”.

3. METHOD

In this paper, we propose a method that measures the level of ambiguity of a tag set T and selects two additional tags that can be proposed to a user to best disambiguate it. Our method underlies the intuition that a tag set is *ambiguous* if it can appear in at least two different tag contexts. These could be defined by geographic locations, word senses, languages or temporal events, etc. We measure the tag contexts by the distribution over all tag co-occurrences.

A good example of an ambiguous tag is the word “Cambridge,” since there are well-known examples of Cambridge in both Massachusetts and England. Suggesting a tag such as “university” is very likely in both contexts, but does little to resolve the ambiguity. Thus given the tag “Cambridge,” we want to note that this is an ambiguous tag, and suggest either “MA” or “UK” because these words do the most to remove the ambiguity. We assume that the tag set {“Cambridge”,“MA”} co-occurs with different tags than {“Cambridge”,“UK”}. These additional tags are defined by locations and events that differ strongly between the two very distant cities.

First, we will introduce a probabilistic framework that provides us with a probability $p(t|T)$ that a tag t co-occurs with the set T. Instead of suggesting the tags that are most likely within this framework, we suggest the two tags t_i, t_j that, once added to T, give rise to maximally different probability distributions $p(t|\{T \cup t_i\})$ and $p(t|\{T \cup t_j\})$. The level of ambiguity of a set T is measured by a weighted KL divergence of these two probability distributions. This idea is illustrated in Figure 2.

Probabilistic Framework

We propose a probabilistic framework to model tag co-occurrences and measure ambiguity. We assume an image

is labeled with a set of tags $T = \{t_a, t_b, ...\}$. The expression $I(T)$ is the number of images that contain the tag set T. For any pair of tags t_i, t_j, we denote the number of image co-occurrences by $I(t_i \cap t_j)$, and then we form an estimate of the probability that a co-occuring tag, t_i, appears in another tag's presence, t_j, by calculating

$$p(t_i|t_j) = \frac{I(t_i \cap t_j)}{\sum_k I(t_k \cap t_j)}. \quad (1)$$

Here we normalize by the sum of co-occurrences for all other possible contexts (t_k). We further sum over all contexts to calculate the prior probability that any one tag is used on any image to find

$$p(t_i) = \frac{\sum_j I(t_i \cap t_j)}{\sum_{j,k} I(t_k \cap t_j)}. \quad (2)$$

We base all of our models on these two probability distributions, which we calculate from pair-wise co-occurrence data.

Tags do not appear only in pairs. We really want to know the probability of a tag in any context, but we can not store this quantity for all tag sets, T. To simplify our computation, we assume that conditional co-occurences are independent. This leads to:

$$p(T|t_i) = \prod_{t \in T} p(t|t_i). \quad (3)$$

Using this assumption, we can write the probability of a tag given any context using Bayes' rule

$$p(t_i|T) = \frac{p(T|t_i)p(t_i)}{p(T)} = \frac{p(t_i)\prod_{t \in T} p(t|t_i)}{\sum_j p(t_j)\prod_{t \in T} p(t|t_j)}. \quad (4)$$

Pairwise Disambiguation

We make a basic assumption about the meaning of ambiguity: A set of labels T is ambiguous if there exist two labels t_i and t_j such that adding one or the other gives rise to very different distributions over the remaining labels. Thus, given the tag "Cambridge," adding the tags "MA" or "UK" leads to very different locations; and the other tags we see in this context are likely to change (stores, people, etc.). We will measure the deviation between two posterior distributions with the KL divergence [13]. Let T denote the current set of tags, and let t_i, t_j be two additional tags. To abbreviate notation, let us denote the corresponding conditional distribution, after t_i has been added to T, as $p_i(t) = p(t|\{T \cup t_i\})$. The KL divergence between these two distributions is:

$$KL(p_i||p_j) = \sum_t p_i(t) \log\left(\frac{p_i(t)}{p_j(t)}\right). \quad (5)$$

This equation integrates the amount of disagreement between the two distributions over all tags t, weighted by the probability $p(t|\{T \cup t_i\})$. It is strictly non-negative but not necessarily symmetric. As in our case, there is no meaningful notion of order for the tags t_i, t_j, we use a commonly used symmetric variation

$$\bar{KL}(p_i, p_j) = KL(p_i||p_j) + KL(p_j||p_i). \quad (6)$$

Given that our data base is limited, it is always possible to find tags t_i, t_j with maximal disagreement by selecting two terms that only appear in very different contexts and are unrelated to the set T. For example, for the tag set $T = \{$"*Cambridge*"$\}$, we could add $t_1 =$ "*fridge*" and $t_2 =$ "*mercedes*" and the KL divergence between the two posterior distributions would presumably be very high. To prevent this, we weight Eq. 5 by the conditional probabilities of the two terms and therefore discount additional tags that have no real relation with the original tag set. We define the weighted divergence as

$$div(p_i, p_j) = p(t_i|T)p(t_j|T)g(\bar{KL}(p_i||p_j)), \quad (7)$$

where $g()$ is some monotonically-increasing function that trades off the impact of the KL divergence with the conditional probabilities.

We define the measure of ambiguity of a tag set T as the maximum divergence between two potential posterior distributions:

$$f(T) = \max_{i,j} div(p_i, p_j). \quad (8)$$

The function $g(x)$ can be any monotonic function that influences the impact of the KL divergence on the output. We experimented with $g(x) = x^e$ for a range of values of e. If the value of $f(T)$ is above a certain threshold, we suggest the labels t_i and t_j, because they represent the "direction" of greatest ambiguity, $f(T)$, to the system. Eq. 8 can be adapted to more than two tag suggestions in several ways, e.g. greedily, given t_i, t_j, by letting the value of a third tag t_k be the minimum divergence from the top two tags $\min(div(p_k, p_i), div(p_k, p_j))$.

Large-Scale Implementation

A naïve implementation of Eq. 8 results in a computational complexity of $O(n^3)$, where n denotes the number of terms in the database. Clearly, this is impractical. However, for any given tag set T, almost all tags t_i have a very small conditional probability $p(t_i|T)$. As we are only interested in the two terms with *maximum* disambiguation value, it is generally sufficient to restrict the search over the top N most common terms, where N is some small number. For our experiments, we found $N = 25$ to be sufficient, under which 97.5% of all computations resulted in exact results. (Note: this approximation mostly affects the value of Eq. 8 and not the suggested tags nor the ordering from most to least ambiguous tags.) A more detailed discussion of how to set the value N is in Section 5. Even finding the top N tags can be safely approximated, as the majority of all tags are never likely in any context. (See Figure 3 for a plot of the distribution of co-occurences.)

For a very large scale implementation, one can take advantage of the fact that Eq. 8 is truly parallelizable. In a map-reduce framework [9], the mapper implements the $div()$ operator defined in Eq. 7 and the reduce phase calculates the $max()$ operator.

4. DATA

In Section 3, we described a generic approach for detecting the ambiguity of a given set of tags. To evaluate our method, we generated a data corpus consisting of image tags from publicly available Flickr™ images. We also used these photos and their tags in a small user study described later.

Photo Data

Flickr™is an online photo-sharing service that contains more than 2 billion photos that are uploaded, tagged and organized by more than 8.5 million registered users. For the

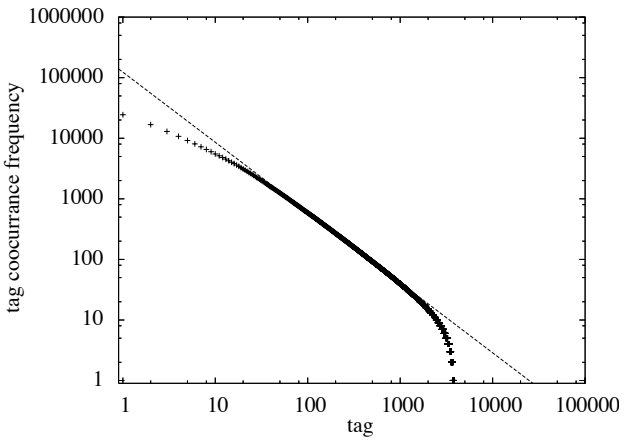

Figure 3: Frequency distribution of the co-occuring tags in Flickr. The tags are sorted according to co-occurrence frequency along the horizontal axis.

research described in this paper, we used a random snapshot from Flickr of 102 million publicly-available photos with annotations. The photos were uploaded between February 2004 and December 2007 and each photo has at least one user-defined tag. The collection we use in this paper contains about 407 million tags in total, and about 6.2 million unique tags.

Figure 3 shows the tag co-occurance distribution on a log-log scale. The horizontal axis represents the 950,000 unique tags that are contained in the co-occurance matrix, ordered by the co-occurance frequency in descending order. The vertical axis refers to the number of tags co-occuring with a given tag. We can model the distribution accurately with a power law. The head of the distribution contains tags that are generic and co-occur with many other tags. For example, the five most frequent co-occurring tags are: *2006*, *2005*, *wedding*, *party* and *2007*. Normally, the tail of the distribution contains infrequent tags that typically can be categorized as incidentally occurring words, such as mis-spellings, and complex phrases. Due to their infrequent nature, we expect that these highly-specific tags are not useful for disambiguation. Our pre-processing steps remove these tags.

We train and test the proposed tag suggestion method with a subset of these photo annotations. We first removed photos from the collection that only have one tag, as these do not contribute to the tag co-occurance matrix. Second, we removed tags used less than five times or by only one user. The final data set contained 950,000 unique tags.

We report our results using the top 235 most frequently occurring tags with distributions ranging over all tags (after removing the year-tags 2004–2008, since their co-occurrence patterns are artificial). For the user study described next, we chose 50 tags within the top 235, with scores that are uniformly distributed through the range of ambiguity scores.

User Data

We collected data for a user experiment using a web survey that showed a tag and two sets of photos that our ambiguity measure predicted would help disambiguate the original tag. Subjects were told, "We are interested in measuring the ambiguity of different tags." We then asked them, "Do you expect the photos with the tag 'Jaguar' to need further clarification?" Users were asked to respond on a four-point scale, from "no further tags are necessary," to "yes, additional tags are necessary to clarify." We collected data from 12 of our colleagues (who didn't know our hypothesis) using 50 tags with a wide range of ambiguities. There was a wide variance in their scores because of the types of data we examined. For example, for most people, the tag "Athens" is highly specific (and Greek), unless you happen to live near Athens, Georgia or Athens, Ohio. This ambiguity is seen in our photo database. There are 68,000 entries for Athens, of which 35,000 are labeled "Greece" and 6,000 are labeled either "Ohio" or "Georgia." A user's ability to know of these ambiguities is limited by his/her personal experience.

5. USER STUDY

We performed a user study to tune the parameters of our model, and to verify that our ambiguity measure correlates with human expectations. Our algorithm has two different parameters. One is an exponent that controls the relative importance of tag frequency versus the KL divergence, and this affects the behavior of our algorithm. The other is a parameter that specifies how deep we look for tag recommendations, and is important because of computational concerns.

The more important parameter controls the behavior of the algorithm. We would like to adjust the trade-off between tag probabilities and the KL divergence by finding the function, g, that gives the best fit to human judgments of ambiguity. We restrict ourselves to functions of the form, $g(x) = x^p$, for some constant p, so the entire function is multiplicative. Thus, we only have to consider different values of the exponent.

Based on experimental data described in Section 4, we have human data that measures the degree of (human) ambiguity for different terms in our database. We use this data and evaluate its correlation with different versions of our ambiguity function. We assume a simple, linear model and use Pearson's correlation [7] to measure performance, but non-linear measures based on ranking can also be used.

We evaluate the correlation with human data using 12 different exponents between 0 and 6. Figure 4 shows the fraction of variance explained by our model for different versions of the ambiguity function and our human data. We find a broad peak for an exponent between 2 and 4. In the rest of this work, we used the simplest result, setting $g(x) = x^2$.

In a second experiment, we measured the number of tags we need to check to find the best reduction in ambiguity. The expression in Eq. 8 is a search over all possible i and j for the tags that maximize the measure. This has complexity $O(N^2)$, where N is the number of tags we consider and thus the computational complexity quickly grows unwieldy. But because of the probabilities at the front of the equation, there is no need to examine terms that are rare. Thus, we can cut the search after we have seen the most popular sub-tags for each search. We tuned this parameter by numerically evaluating, using the 235 most popular tags, the effect of varying a cutoff. We computed the ambiguity for each tag with N varying from $N = 2$ to $N = 40$, and counted how many tags were necessary to obtain the exact score (at $N = 40$). Figure 5 shows the results of this study. As expected, the ambiguity scores grow rapidly and by the time we get to 25 tags, the curve has almost entirely flat-

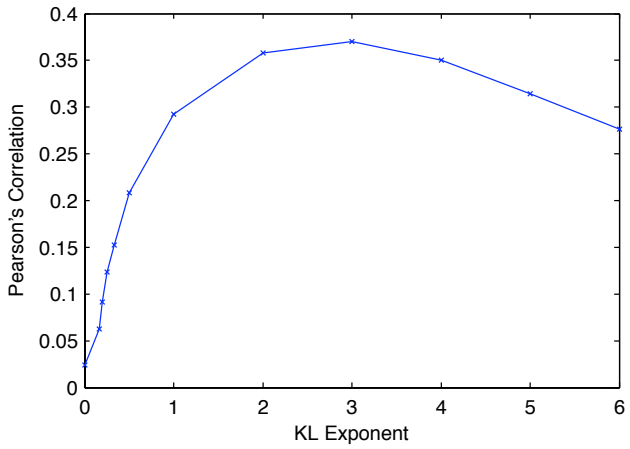

Figure 4: Pearson's correlation between different versions of our ambiguity function and human judgments of ambiguity. We limited our study to different exponents. Higher correlations indicate better fits between the ambiguity function and our human raters.

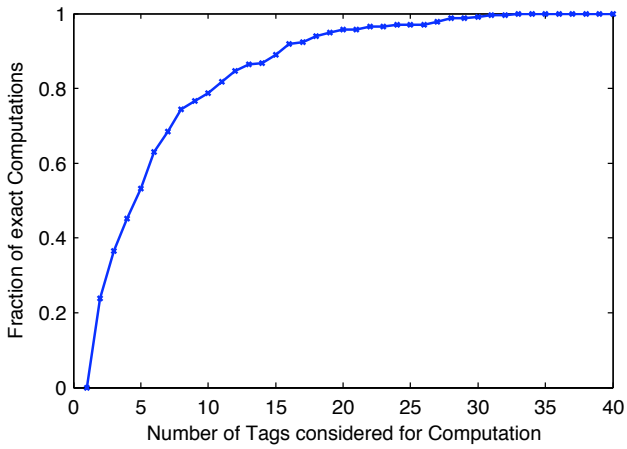

Figure 5: The graph shows the fraction of images whose tag suggestion computation lead to exact results given a cutoff parameter N.

tened. If we cut off the search after 25 tags, then we only lose about 3% of the final ambiguity score.

The correlation between the (averaged) outcomes of the user study and our algorithm is shown in Figure 4. Setting the exponent in the ambiguity measure Eq. 8 to zero causes the measure to ignore the KL divergence, and for our system to behave much like a normal tag-suggestion system. In effect, we rate tags highly if there are two additional tags that both account for much of the additional information (because the probability of both t_i and t_j are high). This forms a rudimentary strawman—we are picking tags that are popular. Tags that had such popular suggestions performed very poorly, and explained only 3% of the human variance. With the right exponent (e.g. $g(x) = x^2$), even with the "ambiguity" of our user's task, our measure of ambiguity was able to explain more than 35% of the variance in the user survey.

6. METADATA RESULTS

In this section we show that for 57% of the most ambiguous tags, our algorithm suggests additional tags that are minimally correlated along one of three measurable dimensions: temporal, geographic and semantic. In these cases, the resulting tag sets resolve the ambiguity. (We expect the explanation is not so simple for the remaining 43%.)

Many of the images in Flickr come with metadata we can use to explain our results. For this paper, we look at three different types of data: geographic, temporal and semantic. We describe each type of data in turn and describe how they can be used for testing the disambiguation algorithm. In each case, the goal is to measure the degree of correlation, or more importantly for our task, the degree of overlap between two sets of data. We want to see if the tags we suggest to resolve ambiguity separate the original photos in one of these three dimensions. At its heart, this is a correlation operation.

Temporal Comparison:

Photos uploaded to Flickr often come with automatically annotated time-stamps that specify when the photo was taken. This data is not always correct, but this noise establishes a lower bound on our algorithm. To measure the temporal ambiguity, we collect all photos with tags $\{T \cup t_2\}$ or $\{T \cup t_3\}$ for single-tag sets $T = \{t_1\}$.

For each of these collections, we form a histogram, using one-month wide bins between 2004 and the end of 2007, and normalize these to create a temporal probability model for each tag t_i. Given a photo with all tags in T and also tag t_i, the empirical probability of it being uploaded in month m is $t_i(m)$. Figure 6 shows one example of the temporally-ambiguous tag set $T = \{$"*holiday*"$\}$ with co-occuring tags "Christmas" and "vacation". Although both tags co-occur frequently with T, their temporal distributions differ dramatically.

We wish to measure the degree to which two tags, t_i, t_j, lead to different temporal distributions, provided both co-occur frequently with T. To measure the degree of similarity between tag sets $\{T \cup t_i\}$ and $\{T \cup t_j\}$, we then compute the cross-correlation

$$R_g(t_i, t_j) = \sum_m c_i(m) c_j(m) \tag{9}$$

where $c_i(m)$ is the number of photos with tags $\{T \cup t_i\}$ in month m. Sets of photos and times with no overlap have a temporal correlation of 0, while photos with similar temporal distributions have a correlation of 1.

Geographic Comparison:

Flickr allows users to tag their photos with a geographic location. This can be done automatically based on the camera's GPS information, or entered manually using a map-based user interface. We use this longitude and latitude data to study the geographic ambiguity of our photo collections.

The geographic data is both two-dimensional and is distributed over widely different scales. Thus, we don't have enough data for a straightforward calculation of the his-

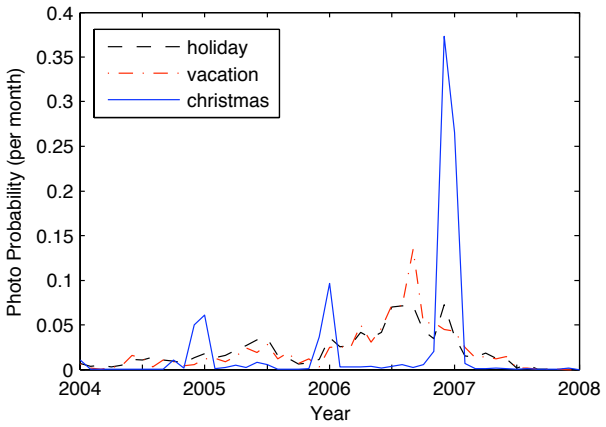

Figure 6: Probability distributions showing the distribution of photos for the tag sets $\{holiday\}$, $\{holiday, vacation\}$ and $\{holiday, christmas\}$ showing these two subtags are used at different times. Holiday–christmas pictures are more common at the end of the year, while holiday–vacation pictures are more common during the Northern-hemisphere summer.

togram and correlation. Instead, given a set of location data $l(x, y)$, we measure the statistics of the data and fit a full-covariance Gaussian model to the geographic data for each set of photos—this is a form of smoothing. Thus, we model the geographic spread of a set of photos with the Gaussian

$$G_i(\vec{x}) = \frac{1}{2\pi|\Sigma|^{1/2}} exp[-1/2(\vec{x} - \vec{\mu})\Sigma^{-1}(\vec{x} - \vec{\mu})] \quad (10)$$

where $\vec{x} = (x, y)$, $\vec{\mu}$ is the mean of the data and Σ is a 2×2 covariance matrix. We fit two Gaussians to the locations of the photos tagged $\{T \cup t_j\}$ versus $\{T \cup t_j\}$. We analytically compute the cross-correlation of the two Gaussians and then calculate the geographic correlation using

$$R_g(t_i, t_j) = \frac{1}{Z} \iint G_i(\vec{x}) G_j(\vec{x}) d\vec{x}, \quad (11)$$

where Z is a normalization constant. This gives us a (smoothed) estimate of the geographic cross-correlation of the photos with the two sets of tags.

Figure 7 shows an example of geographic independence. In this case the tag "elephant" is often seen with the words "Africa" or "Thailand" accompanying it. Evidently from the other tags that co-occur, there is enough difference in their contexts so that geographic ambiguity is present.

Semantic Comparison:

Finally, we also look for information about the semantic meaning of each tag set. Like previous work [6], we use the number of web pages returned in a web search as a rough measure of semantic content. In general, adding a tag's synonym does not change the number of results, but adding a semantically-orthogonal term dramatically reduces the number of results. Thus, we compare the results for $\{T \cup t_i\}$ or $\{T \cup t_j\}$ versus $\{T \cup t_i \cup t_j\}$ to measure semantic ambiguity.

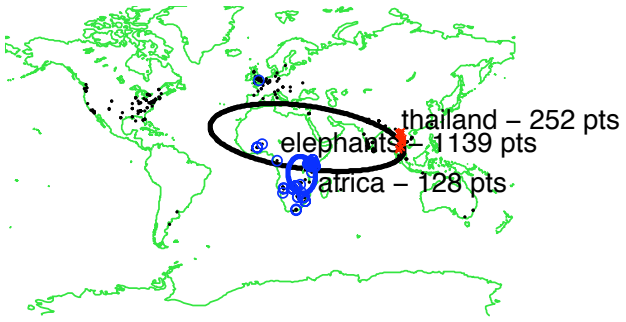

Figure 7: A scatter plot showing the distribution of photos for the tag sets $\{elephant\}$ (shown with black dots), $\{elephant, thailand\}$ (red 'x') and $\{elephant, africa\}$ (blue 'o'). We have smoothed the data and represented each set with a two-dimensional Gaussian as shown above. (The ellipse for Thailand's photos is too small to see here.) There are clusters of data in North America and Europe that are not matched by either submodel.

If t_i and t_j are semantically related, then we do not expect to see much overlap in the results.

We do this by querying the web and looking at the number of web pages that satisfy our queries. We use the Jaccard index to compare two sets of web pages and measure their correlation [16]. This is done by dividing the size of the intersection of the two sets by their union. If the two sets have no overlap, then the Jaccard index is 0, while if there is perfect overlap, then the measure is 1.

Let $web(T)$ be the set of web-search results returned by a query for all the tags in set T. Thus, the correlation in semantic space between two queries is equal to

$$R_s(t_i, t_j) = \frac{|web(T \cup t_i \cup t_j)|}{|web(T \cup t_i) \cup web(T \cup t_j)|} \quad (12)$$

and we calculate the union of web pages in the denominator using: $|web(a) \cup web(b)| = |web(a)| + |web(b)| - |web(a \cup b)|$.

The Jaccard index is calculated by counting the number of web pages returned from the public Yahoo search API.[1] Figure 8 shows the distribution of the correlations. In many cases, the suggested tags produce distinct distributions, which shows up as different semantic ideas and thus a low correlation. More importantly, there are a number of tags that have zero semantic correlation, suggesting that these tags are semantically ambiguous and we have resolved the ambiguity by suggesting two new tags that are semantically diverse.

Meta Explanation:

We do not expect the metadata analysis described in this section to explain all types of ambiguity. But it is promising that we can identify many of the issues that make a tag ambiguous from the data associated with each image. This gives us confidence that we are measuring ambiguity in a useful way.

The metadata can only explain how we resolve some ambiguous data, but not give a measure that is correlated with ambiguity. For example, "June" or "July" resolve a tempo-

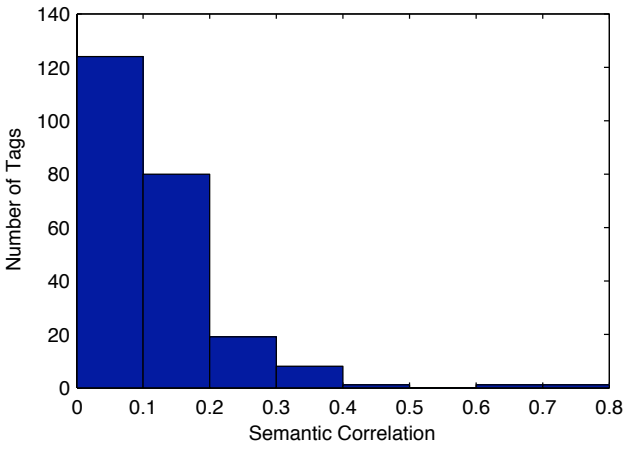

Figure 8: Histogram of semantic correlations. The graph shows how many of the original tags have disambiguating tags that fall into each of correlations bins in 0.1 increments.

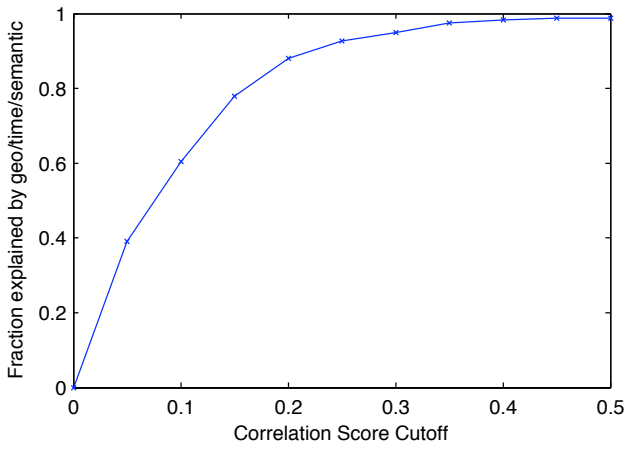

Figure 9: The fraction of tags for which we can explain the source of the ambiguity, as a function of the correlation threshold used to fix an explanation.

ral ambiguity, however our system does not suggest them as the photos taken in June and July are usually not significantly different. Similarly, the images taken in Sunnyvale and Santa Clara, two neighboring towns in Silicon Valley, are not that different—people are taking the same kinds of photos and are probably labeling them with the same kinds of tags. On the other hand, North and South Korea, even though they are adjacent, will have very different kinds of photographs.

Figure 9 shows the fraction of photographs for which we can explain how we resolve the ambiguity using any of the three metadata dimensions: geographic, temporal and semantic. A metadata dimension is said to explain the ambiguity of a photo, and resolve it, if the correlation of the two suggested tags in that dimension is low enough. Figure 9 shows the fraction of tags that are explained as we vary the explanation threshold from 0 to 1. More than 40% of the photos are explained at even the lowest thresholds (0.05).

Figure 10 shows how all photos are explained by different measures using a correlation threshold cutoff of 0.1. At all levels, the semantic measure, even though it was measured using text-web data, explains the majority of the images.

Table 1 shows the 50 most ambiguous examples (from the 250 most common tags) together with the tag suggestions, the ambiguity score and the KL-divergence value. The table also shows the metadata correlation values. Correlation values below 0.1 are highlighted in bold. Not all ambiguous cases can be explained through metadata, in particular none of the terms in the top 50 can be explained through time. Due to the nature of image tags, the top 50 ambiguous terms contain many geographically ambiguous tags (eg "washington"), but also word-sense disambiguations ("football") and terms from composite expressions ("world" or "spring").

7. CONCLUSION

In this paper, we describe a new means to suggest tags based on ambiguity. We do not want to suggest just the most common tags, but instead we want to suggest tags that allow people to better describe their content.

We define a novel measure of tag ambiguity, based on a weighted KL divergence of tag distributions. Our approach both measures the ambiguity of the existing tags, and suggests new tags that best reduce this ambiguity. This is important because asking for new tags is a user-intensive activity and we want to ask when the benefits are significant.

We tested our approach with a user study that asked users to evaluate the ambiguity of 50 different tags. We used this data to validate and tune the parameters of our algorithm. We further showed that we can explain more than half of the found ambiguous tags along one of three dimensions: temporal, geographic or semantic. Based on our user test and the metadata, we explain more than 35% of the variance of our ambiguity measure, and infer the reason for the ambiguity in more than 50% of the cases.

In the future, we wish to examine how these tag suggestions impact search, both for multimedia and for text-based web search.

8. ACKNOWLEDGMENTS

The authors would like to thank Börkur Sigurbjörsnsson, Alex Berg and Serguei Mourachov for their tremendous help in creating the metadata collection.

9. REFERENCES

[1] Yahoo search API. http://developer.yahoo.com/search/.

[2] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. *JCDL '07: Proceedings of the 7th ACM/IEEE joint conference on Digital libraries*, pages 1–10, 2007.

[3] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. *Advances in Information Retrieval: 26th European Conference on IR Research, ECIR 2004, Sunderland, UK*, pages 127–137, 2004.

[4] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. *Proceedings*

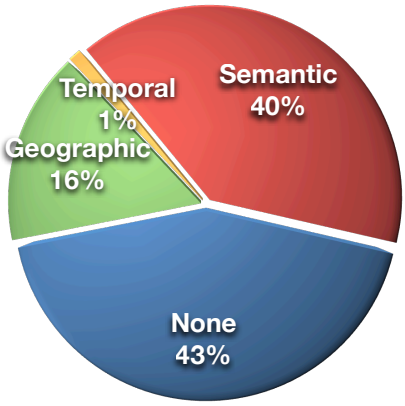

Figure 10: A pie chart showing which metadata feature explains a tag's ambiguity. We used a correlation threshold of 0.1 to assign an explanation. Some photos are explained by more than one type of data and they thus contribute a uniform fraction to each successful category.

of the SIGCHI conference on Human factors in computing systems, pages 971–980, 2007.

[5] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, 2001.

[6] R. L. Cilibrasi and P. M. B. Vitanyi. The Google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, 2007.

[7] J. Cohen. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2003.

[8] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306, New York, NY, USA, 2002. ACM.

[9] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107, 2008.

[10] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[11] B. He and I. Ounis. Query performance prediction. *Information Systems*, 31(7):585–594, November 2006.

[12] D. Heesch, A. Yavlinsky, and S. Rüger. Nnk networks and automated annotation for browsing large image collections from the world wide web. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 493–494, New York, NY, USA, 2006. ACM.

[13] S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[14] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, toread. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM Press.

[15] G. Mishne. AutoTag: A collaborative approach to automated tag assignment for weblog posts. *Proceedings of the 15th international conference on World Wide Web*, pages 953–954, 2006.

[16] T. Pang-Ning, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 1st edition, May 2005.

[17] B. Sigurbjörsnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th International World Wide Web Conference (WWW2008)*, Beijing, China, April 2008.

[18] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Image annotation refinement using random walk with restarts. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 647–650, New York, NY, USA, 2006. ACM.

[19] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May*, 2006.

[20] X. Zhou, M. Wang, Q. Zhang, J. Zhang, and B. Shi. Automatic image annotation by an iterative approach: Incorporating keyword correlations and region matching. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 25–32, New York, NY, USA, 2007. ACM.

T	t1	t2	ambiguity	KL	time	geo	semantic
washington	dc	seattle	3.94	84.14	0.905	**0.000**	0.120
san	francisco	diego	3.05	60.69	0.947	0.507	0.161
world	cup	disney	2.1	151.96	0.332	**0.068**	**0.020**
new	york	zealand	1.84	80.23	0.896	**0.000**	**0.049**
la	losangeles	de	1.79	72.52	0.954	**0.009**	**0.045**
australia	sydney	melbourne	1.29	53.49	0.976	0.587	0.234
d50	nikon	horses	1.24	106.18	0.666	0.965	**0.004**
cameraphone	moblog	zonetag	1.22	90.82	0.491	0.888	**0.002**
texas	austin	dallas	1.18	64.12	0.803	0.348	0.193
south	africa	korea	0.97	115.12	0.803	**0.000**	0.250
canada	ontario	vancouver	0.92	71.36	0.968	**0.000**	0.162
japan	tokyo	kyoto	0.91	27.69	0.979	0.501	**0.078**
temple	japan	cambodia	0.88	121.57	0.837	**0.000**	0.101
oregon	portland	coast	0.75	37.84	0.876	0.404	0.113
china	beijing	shanghai	0.73	31.07	0.971	**0.051**	0.187
baseball	okinawa	mlb	0.72	157.91	0.774	**0.000**	**0.001**
asia	china	thailand	0.69	87.22	0.812	**0.082**	0.219
africa	south	tanzania	0.68	66.93	0.828	**0.003**	0.193
sports	action	athletes	0.68	137.2	0.639	0.524	**0.029**
canon	eos	powershot	0.66	86.02	0.967	0.966	**0.097**
hawaii	maui	oahu	0.65	38.39	0.914	0.944	0.291
seattle	washington	photobooth	0.64	92.68	0.246	**0.001**	**0.000**
zoo	animals	sandiego	0.63	42.79	0.925	**0.028**	**0.055**
florida	orlando	beach	0.58	54.77	0.941	0.114	0.157
france	paris	provence	0.58	48.22	0.907	**0.010**	**0.078**
race	bike	car	0.58	50.88	0.880	0.916	**0.097**
scotland	edinburgh	glasgow	0.58	37.39	0.947	0.510	0.348
spring	flowers	break	0.58	59.19	0.563	0.707	**0.069**
usa	california	newyork	0.58	50.47	0.962	**0.013**	0.190
man	burning	male	0.57	71.75	0.340	0.123	**0.050**
show	music	car	0.56	58.23	0.928	0.841	0.157
dance	party	bellydance	0.55	93.03	0.392	0.898	**0.004**
animals	zoo	pets	0.54	49.4	0.960	0.998	**0.049**
newzealand	wellington	southisland	0.54	45.38	0.797	0.648	**0.000**
town	city	cape	0.54	42.66	0.782	**0.000**	**0.072**
america	usa	south	0.53	33.94	0.846	0.034	0.172
california	sanfrancisco	losangeles	0.53	52.31	0.956	**0.090**	**0.000**
deutschland	germany	horse	0.52	144.66	0.216	0.593	**0.013**
nikon	d50	d200	0.52	35.99	0.939	0.971	0.167
football	soccer	nfl	0.51	67.02	0.414	0.185	0.223
halloween	party	pumpkin	0.51	30.35	0.970	0.993	**0.095**
bike	bicycle	motorcycle	0.5	50.01	0.888	0.971	**0.042**
festival	music	japan	0.5	47.35	0.865	0.281	**0.081**
germany	deutschland	berlin	0.5	21.17	0.986	0.818	0.106
thailand	bangkok	chiangmai	0.5	34.63	0.918	0.464	**0.022**
ireland	dublin	travel	0.49	30.82	0.826	0.463	0.115
europa	europe	oldenburg	0.48	88.59	1.000	**0.013**	**0.006**
ice	snow	hockey	0.47	74.02	0.817	0.946	**0.054**
uk	england	scotland	0.47	39.05	0.948	0.447	0.233
boston	massachusetts	beantownsoftballleague	0.46	77.12	0.626	1.000	**0.000**

Table 1: Examples of disambiguating suggestions for the 50 most ambiguous tags within the 250 most common tags. The table shows the original tag T, the two suggested tags $t1$ and $t2$, the ambiguity score, the value of the KL divergence, and the meta data correlations: time, geo and semantic. A low correlation score indicates that the ambiguity might be explained through the respective meta-data context (bold). Some co-occurrences like "beantownsoftballleague" and "boston" or "deutschland" and "horse" are over represented due to single power-users whose uploading coincided with our data set generation. We expect these anomalies to disappear with more data and better uniform sampling.