Towards Automated Georeferencing of Flickr Photos

Olivier Van Laere
Department of Information Technology
Ghent University, IBBT, Gent, Belgium
olivier.vanlaere@ugent.be

Steven Schockaert*
Dept. of Applied Mathematics and Computer Science
Ghent University, Gent, Belgium
steven.schockaert@ugent.be

Bart Dhoedt
Department of Information Technology
Ghent University, IBBT, Gent, Belgium
bart.dhoedt@ugent.be

ABSTRACT

We explore the task of automatically assigning geographic coordinates to photos on Flickr. Using an approach based on k-medoids clustering and Naive Bayes classification, we demonstrate that the task is feasible, although high accuracy can only be expected for a portion of all photos. Based on this observation, we stress the importance of adaptive approaches that estimate locations at different granularities for different photos.

Categories and Subject Descriptors

I.2 [**ARTIFICIAL INTELLIGENCE**]: Miscellaneous; H.3.7 [**INFORMATION STORAGE AND RETRIEVAL**]: Digital libraries

General Terms

Algorithms, Experimentation

Keywords

Georeferencing, Web 2.0, Naive Bayes classification

1. INTRODUCTION

In recent years, tagging has emerged as one of the most prominent techniques to organize online collections of resources, such as photos, videos, bookmarks or scientific papers. Typically, users add tags (short textual descriptions) to resources they find interesting to bring structure into a collection, to facilitate retrieval, or to help others find these resources more easily, among others [2]. As a result, a rich description of the content of these resources can be obtained by statistically analyzing which tags are assigned by which users to which resources. This observation has led to techniques to automatically generate ontological knowledge [16], to improve the performance of retrieval systems [10], or to assist users with the burden of finding the right tag to add to a new resource [15].

In the case of photos, geographic location forms one of the most important forms of metadata. Accordingly, online repositories such as Flickr[1] or Panoramio[2] offer the possibility to explicitly associate geographic coordinates to photos. These coordinates are typically obtained through GPS devices that are integrated in digital cameras, or by the users indicating on a map where a photo was taken.

Analyzing the distribution of tags appearing in such large collections of georeferenced photos has been found extremely useful, for instance to model the spatial extent of vernacular places [9] or to discover toponyms [12]. Since the quality of results that are thus obtained depends crucially on the number of available georeferenced photos, the question arises of whether we can leverage the current geographic metadata to approximately localize photos in an automated way. This would allow to add approximate coordinates to the vast number of photos on Flickr that are currently not georeferenced, and would moreover allow to build more "friendly" user interfaces that attempt to zoom a map at the right place at the right level when users are trying to manually indicate where the photo was taken. In this paper, we present an approach to automatically find where a photo was taken, based on k-medoids clustering and Naive Bayes classification. We attempt to discover both in which city the photo was taken, and where it was taken within that city.

The paper is structured as follows. In the next section, we outline the procedure we have followed to construct training and test data, and explain the preprocessing steps we performed. Next, Section 3 explains how a Naive Bayes classifier could be used for the task of georeferencing Flickr photos, and contains our main experimental results. In Section 4, we discuss the implications of these experimental results and, in particular, stress the importance of adaptive techniques that make predictions at the right level of granularity, depending on how much information is available for a given photo. We sketch an approach to implement such adaptive behavior, based on Dempster-Shafer theory and possibility theory. Finally, an overview of related work is presented in Section 5, after which our conclusions are presented.

2. OBTAINING THE DATA

In our experiments, we have restricted ourselves to 55 large European cities. These cities were chosen by intersecting the set of 100 most densely populated European

*Postdoctoral Fellow of the Research Foundation – Flanders (FWO).

[1] `http://www.flickr.com/`
[2] `http://www.panoramio.com/`

GIR'10 18-19th Feb. 2010, Zurich, Switzerland

Table 1: Number of photos in the training set for each city that was considered in experiments.

name	training	name	training	name	training	name	training
Amsterdam	19601	Dusseldorf	2858	Lyon	4002	Rotterdam	5766
Antwerp	5150	Frankfurt	4589	Madrid	16540	Seville	4495
Athens	139	Genoa	2635	Malaga	1938	Skopje	143
Barcelona	29648	Glasgow	6966	Marseille	2837	Sofiya	37
Belgrade	426	Gothenburg	4143	Milan	16605	Stockholm	11196
Berlin	30695	Hamburg	10779	Moscow	5497	Stuttgart	3681
Birmingham	3707	Hanover	2514	Munich	11844	Turin	6093
Bremen	0	Helsinki	7384	Naples	3133	Valenza	5674
Bruxelles	5668	Istanbul	8390	Nuremberg	1420	Vienna	13982
Budapest	9376	Copenhagen	1837	Oslo	5941	Vilnius	747
Cologne	7957	Leipzig	2394	Palermo	1685	Warsaw	6146
Krakow	3419	Lisbon	975	Paris	72763	Zagreb	1425
Dresden	2783	Liverpool	4946	Prague	11790	Zaragoza	2568
Dublin	20449	London	188077	Rome	25120		

cities[3] with the set of 160 most important European cities for tourism[4]. Intuitively, a high population should ensure that allocating photos to locations is non-trivial (as opposed to villages where all activity is centered around a small area), while tourist activity should ensure that a sufficient number of photos is available on Flickr. For each georeferenced photo in these cities, we collected the corresponding tags and coordinates using the Flickr API, leading to a total of 3738072 photos. In addition to the coordinates themselves, Flickr provides information about the accuracy of coordinates as a number between 1 (world-level) and 16 (street level). From our initial set of photos, we removed those photos whose coordinates had an accuracy of 13 or less, to ensure that all coordinates were meaningful w.r.t. within-city location. Furthermore, we removed photos whose tag set and user name was identical to a photo that is already in our collection (to reduce the impact of bulk uploads [13]). After these two filtering steps, a set of 1029761 photos remained, which we split into 686193 photos for training ($\approx 66\%$) and 343568 photos for testing ($\approx 33\%$), such that all photos from the same user were either in the training set, or in the test set (to avoid an unfair exploitation of user-specific tags). Table 1 displays the 55 cities we considered, as well as the number of photos in the training set, for each city. Note that no photos were kept for Bremen; as it turns out, the coordinates for all photos in Bremen have an accuracy of less than 13.

To interpret the process of georeferencing photos as a classification task, we first divided the 55 cities into a set of disjoint areas that will serve as classification labels. These areas were obtained by clustering the locations of the photos in the training set using the k-medoids algorithm with geodesic distance. The k-medoids algorithm was preferred over the k-means algorithm as it handles the occurence of outliers better. We have experimented with three different values of the total number of areas k: 250, 500 and 1000. In each case, we imposed that the number of cluster centra per city was proportional to the number of georeferenced photos we had available for that city, while ensuring that every city still contained at least one cluster centre. As a result, cities for which we had only few georeferenced photos were divided in areas of a larger scale. This conforms to our intuition that we should try to be precise in estimating the location of a photo only when sufficient information is available for making that decision. In addition, whenever the number of photos in a given cluster dropped below 50, after an iteration of the k-medoids algorithm, that cluster was eliminated and the associated photos added to the nearest remaining cluster. The actual number of areas after the clustering algorithm had converged was 217, 401 and 677. The actual clusters are visualized in Figure 1 for the case of London. In this figure, cluster centra are plotted as black dots, while all photos belonging to a cluster are connected to the centre of its cluster by means of a line. The panel on which the figure is plotted is a bounding box for the actual coordinates of the photos of London. This bounding box is scaled down and distances between photos in the figure thus represent relative (geodesic) distances.

For efficiency, and to increase the robustness of the approach, we removed all tags that were used by 2 users or less. Next, we applied χ^2 feature selection to eliminate tags that are not indicative of a particular area. Let $\mathcal{A}$ be the set of areas that is obtained after clustering. Then for each area a in $\mathcal{A}$ and each tag t occurring in photos from a, the χ^2 statistic was calculated as follows:

$$\chi^2(a,t) = \frac{(O_{ta} - E_{ta})^2}{E_{ta}} + \frac{(O_{t\bar{a}} - E_{t\bar{a}})^2}{E_{t\bar{a}}} + \frac{(O_{\bar{t}a} - E_{\bar{t}a})^2}{E_{\bar{t}a}} + \frac{(O_{\bar{t}\bar{a}} - E_{\bar{t}\bar{a}})^2}{E_{\bar{t}\bar{a}}}$$

where O_{ta} is the number of photos in area a where tag t occurs, $O_{t\bar{a}}$ is the number of photos outside area a where tag t occurs, $O_{\bar{t}a}$ is the number of photos in area a where tag t does not occur, and $O_{\bar{t}\bar{a}}$ is the number of photos outside area a where tag t does not occur. Furthermore, E_{ta} is the number of occurrences of tag t in photos of area a that could be expected when occurrence of t were independent of the location in area a, i.e. $E_{ta} = N \cdot P(t) \cdot P(a)$ with N the total number of photos, $P(t)$ the percentage of photos containing tag t and $P(a)$ the percentage of photos that are located in area a; similarly, $E_{t\bar{a}} = N \cdot P(t) \cdot (1 - P(a))$, $E_{\bar{t}a} = N \cdot (1 - P(t)) \cdot P(a)$, $E_{\bar{t}\bar{a}} = N \cdot (1 - P(t)) \cdot (1 - P(a))$. The vocabulary V that was used for classification was then obtained by taking for each area a the 25 tags with highest χ^2 value. This led to a total number of 1269,

[3] `http://www.nga.mil`

[4] `http://www.visiteuropeancities.info`

Table 2: Most informative tags according to the χ^2 statistic for the area in which resp. the Sagrada Familia and Eiffel tower are located. Results are shown for the clusterings in 250, 500 and 1000 areas, as well as for the case where areas are simply the 55 cities.

	Sagrada Familia			
	city	250	500	1000
1	barcelona	sagradafamilia	sagradafamilia	sagradafamilia
2	catalunya	sagrada	sagrada	sagrada
3	spain	familia	gaudi	gaudi
4	catalonia	gaudi	familia	sagradafam??lia
5	gaudi	barcelona	sagradafam??lia	familia
6	catalu?śa	lasagradafamilia	lasagradafamilia	lasagradafamilia
7	bcn	sagradafam??lia	barcelona	barcelona
8	sagradafamilia	spain	abstraccion	templeexpiatoridelasagradafam??lia
9	barcellona	gaud??	formas	gaud??
10	espa?śa	abstraccion	gaud??	spain

	Eiffel tower			
	city	250	500	1000
1	paris	eiffeltower	eiffeltower	eiffeltower
2	france	eiffel	eiffel	eiffel
3	louvre	toureiffel	toureiffel	toureiffel
4	eiffel	paris	paris	paris
5	eiffeltower	france	champdemars	tower
6	francia	champdemars	france	france
7	parigi	tower	tower	tour
8	seine	tour	tour	torreeiffel
9	london	torreeiffel	torreeiffel	champdemars
10	notredame	champsdemars	champsdemars	latoureiffel

4701, 8452 and 13727 distinct tags, respectively in the case where k was 55 (i.e. taking the cities as areas), 250, 500 and 1000. In Table 2, the 10 highest scoring tags are shown for two well known tourist sites: the Sagrada Familia in Barcelona, Spain, and the Eiffel tower in Paris, France. In both cases, the first column contains the tags that allow to predict the corresponding cities, i.e. not the actual area around the landmarks. The right-most column contains the tags that allow to predict the immediate area around the landmarks. Unsurprisingly, most of the tags that are found correspond to toponyms and tags that directly refer to the names of the actual landmarks, although there are some notable exceptions (e.g. gaudi, abstraccion). Finally, it is interesting to note that preliminary experiments suggested a clear supriority for χ^2 over mutual information for this task [11], another well known technique for feature selection.

3. ANALYZING TAG DISTRIBUTIONS

3.1 Naive Bayes

Let $\mathcal{A}$ be a set of (disjoint) areas, and for each area $a \in \mathcal{A}$, let X_a be a set of images that were taken in that area. Given a previously unseen image x, we may then try to determine in which area x was most likely taken. In this paper, we use a (multinomial) Naive Bayes classifier to this end, which has the advantage of being simple, efficient, and robust. Initial results in [11] have shown good results for this multinomial classifier. An additional advantage, which will be exploited in the discussion below, is the fact that the output of this classifier, in contrast to e.g. support vector machines, can be interpreted as probabilities. Specifically, we assume that an image x is represented as its set of tags. Using Bayes' rule, we know that the probability $P(a|x)$ that image x was taken in area a is given by

$$P(a|x) = \frac{P(a) \cdot P(x|a)}{P(x)}$$

Using the fact that the probability $P(x)$ of observing the tags associated with image x is fixed among all areas a, we find

$$P(a|x) \propto P(a) \cdot P(x|a)$$

Characteristic of Naive Bayes is the assumption that all features are independent. Translated to our context, this means that the presence of a given tag does not influence the presence or absence of other tags. Writing $P(t|a)$ for the probability of a tag t being associated to an image in area a, we find

$$P(a|x) \propto P(a) \cdot \prod_{t \in x} P(t|a)$$

Using a multinomial language model and Laplace smoothing, the probability $P(t|a)$ is estimated as

$$P(t|a) = \frac{N_t + 1}{\left(\sum_{y \in X_a} |y|\right) + |V|}$$

where N_t is the number of images in area a containing tag t, $\sum_{y \in X_a} |y|$ is the total number of tag occurrences over all images in area a, and V is the vocabulary, as before. For the prior probability $P(a)$ of area a, the maximum likelihood estimate can be used:

$$P(a) = \frac{|X_a|}{\sum_{b \in \mathcal{A}} |X_b|}$$

Table 3: Evaluation results for the four classifiers for the entire test data.

	Acc	MRR	Dist
C_{city}	86.9%	0.89	2.60 km
C_{250}	51.4%	0.61	1.55 km
C_{500}	46.3%	0.56	1.39 km
C_{1000}	41.2%	0.51	1.29 km

Typically, as result of the classification, the most likely area is chosen. After moving to log-space to avoid numerical underflow, this leads to:

$$a^* = \arg\max_{a \in \mathcal{A}} (\log P(a) + \sum_{t \in x} \log P(t|a))$$

3.2 Experimental Results

To evaluate the performance of the approach outlined above, we determined the most plausible area for each photo in the test set for four different classifiers, viz. the Naive Bayes classifiers trained at the city level, and at the sub-city level using the 250-area, 500-area and 1000-area clusterings. Let us call these classifiers C_{city}, C_{250}, C_{500} and C_{1000}. To evaluate how good each of these classifiers is at predicting the right location of photos, we use the following measures

Acc (accuracy): the percentage of photos for which the most plausible area (according to Naive Bayes) was the correct one, i.e. the area in which the photo was actually taken.

MRR (mean reciprocal rank): the average of $\frac{1}{R}$ where for each photo, R is the position at which the correct area is found when ranking all areas according to the probability predicted by Naive Bayes. For instance, a MRR of 0.5 means that on average, the correct area is the second most plausible area according to Naive Bayes.

Dist (median distance): the median of the distance between the predicted location and the actual place where the photo was taken.

Note that the last measure does not compare the area that was predicted with the correct area, but actual locations. As predicted location, we take the medoid of the area (or cluster) that was considered most plausible by Naive Bayes. Accuracy and mean reciprocal rank are well-known evaluation measures from the fields of machine learning and information retrieval. The reason we also need to look at the median distance is because this is what really matters in most applications, and this is also the only measure that allows us to compare the results of classifiers that work at different granularity levels. Indeed, by decreasing the number of classes, it is clear that accuracy will most probably increase. Also note that the median distance is more informative than average distance here, because of its robustness to outliers.

The main results of our evaluation are presented in Table 3. Interestingly, for almost 87% of the photos in the test set, the city could be determined correctly. When considering finer granularity levels, this accuracy decreases to slightly more than 41% for the 1000-area clustering. This decrease in accuracy is what could be expected. What is important, however, is that the median distance between the predicted location and the correct location is smallest when the granularity level is finest. The reason for the poorer performance at the city-level, for instance, is because the distance between the centre of the city and the place where the photo was taken can be quite large, even when the predicted city is correct. Although the predicted area at the finest granularity level is incorrect more often than not, the small median distance suggests that the predicted area still tends to be in the vicinity of the correct area. Similar conclusions can be drawn from the high values of the MRR measure: even at the finest granularity level, the correct area is usually among the top ranked areas (around the second position in the ranking, on average).

Table 4: Evaluation results for the four classifiers, when restricted to photos having at least one tag associated to them from the vocabulary V.

	Acc	MRR	Dist
C_{city}	98.3%	0.99	1.89 km
C_{250}	59.6%	0.70	1.17 km
C_{500}	51.9%	0.63	1.07 km
C_{1000}	45.1%	0.56	1.04 km

Table 5: Evaluation results for the four classifiers, when restricted to photos having t least 6 distinct tags associated to them from the vocabulary V.

	Acc	MRR	Dist
C_{city}	99.1%	0.99	1.54 km
C_{250}	76.1%	0.84	0.66 km
C_{500}	66.4%	0.76	0.6 km
C_{1000}	56.4%	0.67	0.6 km

When analyzing the photos for which the prediction was wrong, we found that a large part of these photos did not have any tags from the vocabulary V. Thus, presence or absence of terms from the vocabulary provides useful information on whether or not we can localize a photo. In Table 4, the results are shown when restricted to those photos that have at least one tag from the vocabulary. Surprisingly, accuracy increases to 98%. This means that, in general, either we know that insufficient information is available to localize the photo, or we can reliably find the correct city. Taking this idea one step further, Table 5 contains the results when we restrict ourselves to photos that have at least 6 distinct tags from the vocabulary. Again a substantial improvement is witnessed; e.g. when using the finest granularity, half of the photos is localized with an error of less than 0.6 km. This suggests that the number of tags from the vocabulary that are associated to a photo can be a useful indication to assess the confidence we should put in a prediction.

4. DISCUSSION

The experimental results obtained above are encouraging on one hand, as it turns out that a substantial number of photos can be localized with a reasonable precision. On the other hand, there is still a large number of photos for which the predicted location is wrong. Although it may be possible to improve performance by using more advanced techniques, such as smoothing, the main conclusions will

most likely remain. For example, [13] proposes to smooth $P(a|x)$ as follows (using a uniform prior):

$$P^*(a|x) \propto \alpha P(x|a) + (1-\alpha) \cdot \sum_{b \in neigh_d(a)} \frac{P(x|b)}{(2d+1)^2 - 1}$$

where $d > 0$ and $neigh_d(a)$ is the set of all areas that are within distance d of a. Although experimental results confirm an improvement using such techniques, the gain in accuracy is not substantial. The main reason is that available tags for many photos are simply not informative enough to allow a location to be found, not even for human readers.

Clearly, if automated georeferencing of photos is to be used in practical applications, it becomes paramount that not all photos are treated in the same way. When clustering the cities into areas, we already emphasized the importance of having larger areas (clusters) in parts of cities for which less information is available. Similarly, when the tags associated to a photo are not sufficient to accurately predict its location, we should not attempt to "guess" a location anyway. Thus, in addition to finding the most likely location of a photo, we face the challenge of determining the appropriate granularity of the prediction that is reported to the user. If we are only certain about the city where the photo was taken, then this is all that should be concluded. If we are not even certain about the city, then perhaps nothing should be concluded. Below we outline two orthogonal ways to make the georeferencing process more adaptive in this sense.

4.1 Multi-scale Classification

Deciding when a city-level prediction is most appropriate and when a smaller-scale prediction is most appropriate is not an easy task. An interesting strategy, however, might be to make use of the fact that classifiers can be trained at different resolutions. Specifically, let $\{\mathcal{A}_1, ..., \mathcal{A}_k\}$ be different clusterings of the cities of interest into disjoint areas, where $\mathcal{A}_1$ corresponds to the finest clustering and $\mathcal{A}_k$ corresponds to the coarsest clustering, i.e. $|\mathcal{A}_1| > |\mathcal{A}_2| > ... > |\mathcal{A}_k|$. Then every area a from $\mathcal{A}_i$, $i \geq 2$, naturally correspond to a set $areas(a)$ of areas from the finest clustering $\mathcal{A}_1$. If the classifier that was trained on $\mathcal{A}_i$ then suggests area a as most likely location, with probability p, we can take this as evidence that the correct location, at the finest level, is among the areas in $areas(a)$. Thus the results of all classifiers can be interpreted as probability distributions on sets of areas from $\mathcal{A}_1$. Such a probability distribution on the powerset of a universe is usually called a belief function or mass assignment, and forms the basis of the evidence theory of Dempster and Shafer [14].

Using Dempster-Shafer theory, the results from the classifiers at all levels of granularity can be fused into one belief function. Since predictions are now sets of areas, this approach has the potential of justifying when enough evidence is available to conclude that the photo was taken in a specific area, and when evidence only allows to conclude in which city the photo was taken (or not even to predict the correct city). There are a wide number of ways in which this idea can be implemented, and a detailed study of using Dempster-Shafer theory in this context is left for future work. In each case, however, the intuition will be that when a is the area predicted at the finest level, and the city in which a is located is found to be unlikely by a classifier working at the city level, then the plausibility of a significantly decreases. In other words, the central idea is that agreement between different classifiers, trained with different features and operating at different levels of granularity, is a strong indication for the precision at which the right location can be predicted, in addition of course to the confidence each classifier on its own has in the prediction it makes.

4.2 Possibilistic Predictions

A second idea, when trying to make the result of the classification more adaptive, is to use a richer representation than (sets of) areas to encode the result. When reporting back to end-users, a probability distribution can easily be displayed as a heat map, for instance. Even when the prediction is used as input to other techniques, e.g. learning the spatial extent of vernacular places, probability distributions (or belief functions) could be more useful than just knowing the most plausible location. However, using probability distributions in this way has at least three disadvantages:

1. The amount of space needed for storing an entire probability distribution for every image of interest is prohibitively high, being linear in both the number of areas considered and the number of images that are analyzed. This is especially true when the total number of images is large, resolution is fine-grained and when any part of the globe is considered a priori (rather than a restricted set of cities or countries).

2. It is well-known that Naive Bayes does not produce well-calibrated probability estimates [3]. On the other hand, there is empirical and theoretical evidence that Naive Bayes can successfully be used to find the most probable outcome [5], or to rank different outcomes according to their degree of plausibility [17]. Thus it makes sense to convert the probabilities obtained using Naive Bayes to a weaker model of uncertainty, focusing on the qualitative ordering that is obtained, rather than on the actual values of the probabilities.

3. The probability distribution obtained by Naive Bayes will typically be too chaotic to adequately visualize the uncertainty associated to the prediction of a photo's location. Ideally, the visualization should be simple enough for the user the grasp immediately how much is known about the location of a photo, and where it was most likely taken.

As an alternative, we propose to approximate the mass assignment obtained from the Dempster-Shafer approach that was outlined above (or even the probability distributions obtained from a single classifier), as a weighted collection of nested areas. Formally, let $X_1, ..., X_n$ be areas of different granularity, represented as sets of areas at the finest level, i.e. $X_i \subseteq \mathcal{A}_1$. Assume furthermore that $X_1 \subseteq ... \subseteq X_n = \mathcal{A}_1$, and let w_i be a weight attached to X_i, where $0 < w_1 \leq ... \leq w_n = 1$. Intuitively, w_i reflects our certainty that the correct location is in X_i. Since X_n covers all areas, we are completely certain that it also contains the correct area, i.e. $w_n = 1$. In general, w_i can be calculated from a mass assignment m as

$$w_1 = \sum_{Y \subseteq X_1} m(Y)$$

and for $i \geq 2$,

$$w_i = \sum_{\substack{Y \subseteq X_i \\ Y \not\subseteq X_{i-1}}} m(Y)$$

Note that $\{X_1/w_1, ..., X_n/w_n\}$ is a mass assignment in which all the sets with a non-zero mass (i.e. the focal elements) are nested sets. Such mass assignments are called consonant, and can be represented as a $\mathcal{A}_1 - [0, 1]$ mapping π, i.e. a possibility distribution in $\mathcal{A}_1$, without loss of information. For each $a \in \mathcal{A}_1$, we define

$$\pi(a) = \begin{cases} 1 & \text{if } a \in X_1 \\ 1 - w_1 & \text{if } a \in X_2 \setminus X_1 \\ ... & \\ 1 - \sum_{i=1}^{l-1} w_i & \text{if } a \in X_l \setminus X_{l-1} \\ ... & \\ w_n & \text{if } a \notin \setminus X_{n-1} \end{cases}$$

This technique of approximating mass assignments was proposed in [6], where π is called the outer approximation of m. The advantage now is that π can easily be visualized as a heat map. Finding good visualizations, and indeed, good approximations of our beliefs, thus boils down to choosing good definitions of the sets $X_1, X_2, ..., X_n$. These are definitions for which the resulting possibility distribution π is maximally informative, i.e. for which the weight w_i of small-scale regions X_i (i.e. for small values of i) is as high as possible. Various options exist to measure the degree of informativeness of a possibility distribution [7]. One idea might be to take $n = k$, and for each i, choose X_i among the focal elements of m from $\mathcal{A}_i$, such that the following expression is minimized:

$$\sum_{a \in \mathcal{A}_1} area(a) \cdot \pi(a)$$

where $area(a)$ denotes the area of the spatial extent of a.

5. RELATED WORK

By far the most similar work to ours is [13], where a language modeling approach is used to predict the location of Flickr images based on their tags. To discretize the earth's surface, rectangles are used, each of which is defined as a set of locations with identical coordinates up to a fixed number of decimals. The length of the sides of the rectangles, in different experiments, ranges from about 1 km to more than 100 km, hence a substantially coarser granularity is assumed than in our experiments. Experiments are performed on a random sample of about 400 000 Flickr images, 140 000 of which are actually considered after removing images from the same user with identical tag sets. To improve the standard multinomial language model, the spatial distribution of the areas is used for smoothing, and boosting is applied to tags that are known to be (unambiguous) toponyms. These techniques are shown to lead to a significant performance increase, although accuracy (and related performance metrics) remain in the same order of magnitude.

Another related approach is presented in [4], where target locations are determined using mean shift clustering, a non-parametric clustering technique from the field of image segmentation. The advantage of this method is that an optimal number of clusters is determined automatically, requiring only an estimate of the scale of interest. Specifically, to find good locations, the difference is calculated between the density of photos at a given location and a weighted mean of the densities in the area surrounding that location. To assign locations to new images, both visual (keypoints) and textual (tags) features were used. Experiments were carried out on a sample of over 30 million images, using both Bayesian classifiers and linear support vector machines, with slightly better results for the latter. Two different resolutions were considered corresponding to approximately 100 km (finding the correct metropolitan area) and 100 m (finding the correct landmark). It was found that visual features, when combined with textual features, substantially improve accuracy in the case of landmarks. In [8], an approach is presented which is based purely on visual features. For each new photo, the 120 most similar photos with known coordinates are determined. This weighted set of 120 locations is then interpreted as an estimate of a probability distribution, whose mode is determined using mean-shift clustering. The resulting value is used as prediction of the image's location.

Using k-means to spatially cluster geotagged Flickr images has been proposed in [1], where the clusters are used to find representative textual descriptions of each area. The goal is to visualize these textual descriptions on a map, to assist users in finding images of interest.

6. CONCLUSIONS

We have proposed an approach to find the approximate location of photos on Flickr. After clustering the regions of interest into disjoint areas, a vocabulary of relevant features is compiled using the well-known χ^2 statistic. This vocabulary is then used to train a Naive Bayes classifier that can be used to predict the area in which previously unseen photos were taken. The results that were obtained in this way are encouraging, suggesting among others that predicting the city where a photo was taken can be done very reliably. Even at finer resolutions, the correct area is almost always among the areas that are considered most plausible by the Naive Bayes classifier. On the other hand, we have stressed the importance of adaptive techniques that are aware of the spatial granularity that is meaningful for a given photo. As a first step towards such techniques, we have found that the number of tags for a given photo that appear in the vocabulary is a useful indicator of how likely the predicted location will be (approximately correct). Finally, we have outlined a more advanced technique to make the overall approach more adaptive, based on Dempster-Shafer theory and possibility theory.

7. ACKNOWLEDGMENTS

We are grateful to Koen Michiels for his help with the implementation.

8. REFERENCES

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

[2] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human*

factors in computing systems, pages 971–980, New York, NY, USA, 2007. ACM.

[3] P. Bennett. Assessing the calibration of naive bayes' posterior estimates. Technical Report CMU-CS00-155, Carnegie Mellon, 2000.

[4] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 761–770, New York, NY, USA, 2009. ACM.

[5] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple Bayesian classier. *Machine Learning*, 29(2/3):103–130, 1997.

[6] D. Dubois and H. Prade. Fuzzy sets, probability and measurement. *European Journal of Operational Research*, 40(2):135–154, 1989.

[7] D. Dubois and H. Prade. Properties of measures of information in evidence and possibility theories. *Fuzzy Sets and Systems*, 100(Supplement 1):35 – 49, 1999.

[8] J. H. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[9] L. Hollenstein. Capturing vernacular geography from georeferenced tags. Master's thesis, University of Zurich, 2008.

[10] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Folkrank : A ranking algorithm for folksonomies. In *LWA*, pages 111–114, 2006.

[11] K. Michiels. Geografisch geïnformeerde zoeksystemen voor foto's. Master's thesis, Ghent University, 2009.

[12] T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Trans. Web*, 3(1):1–30, 2009.

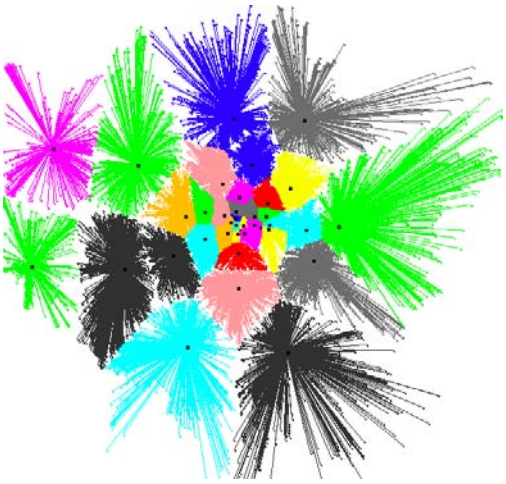
[13] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491, New York, NY, USA, 2009. ACM.

[14] G. Shafer. *A mathematical theory of evidence*. Princeton university press Princeton, NJ, 1976.

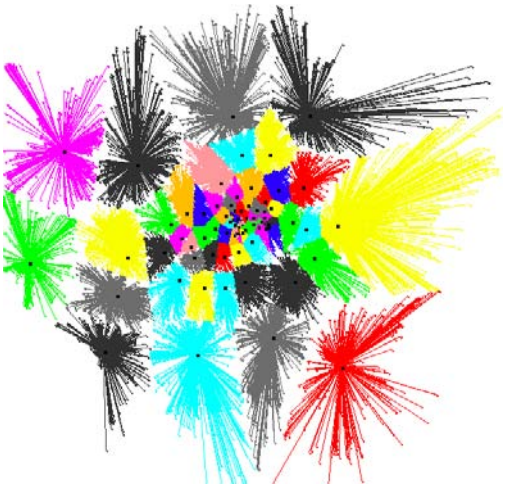
[15] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 327–336, New York, NY, USA, 2008. ACM.

[16] J. Tang, H.-f. Leung, Q. Luo, D. Chen, and J. Gong. Towards ontology learning from folksonomies. In *IJCAI'09: Proceedings of the 21st international jont conference on Artifical intelligence*, pages 2089–2094, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

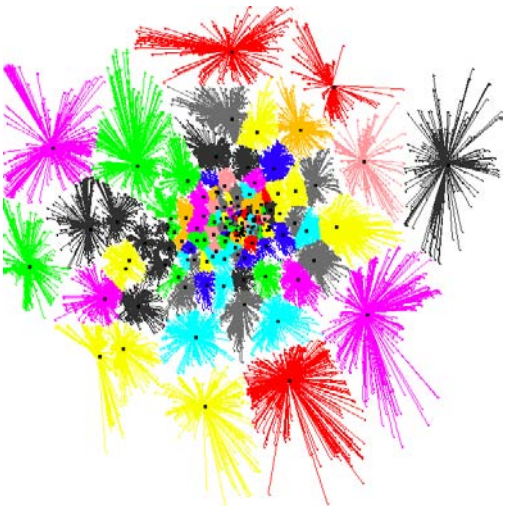
[17] H. Zhang and J. Su. Naive Bayesian classifiers for ranking. *Lecture Notes in Computer Science*, 3201:501–512, 2004.



(a) 250 area clustering



(b) 500 area clustering



(c) 1000 area clustering

Figure 1: Clustering obtained for London when the total number of clusters (over all cities) was 250, 500 and 1000.