Geotagging Social Media Content with a Refined Language Modelling Approach

Giorgos Kordopatis-Zilos, Symeon Papadopoulos(✉), and Yiannis Kompatsiaris

Information Technologies Institute, CERTH, Thessaloniki, Greece
papadop@iti.gr

Abstract. The problem of content geotagging, i.e. estimating the geographic position of a piece of content (text message, tagged image, etc.) when this is not explicitly available, has attracted increasing interest as the large volumes of user-generated content posted through social media platforms such as Twitter and Instagram form nowadays a key element in the coverage of news stories and events. In particular, in large-scale incidents, where location is an important factor, such as natural disasters and terrorist attacks, a large number of people around the globe post comments and content to social media. Yet, the large majority of content lacks proper geographic information (in the form of latitude and longitude coordinates) and hence cannot be utilized to the full extent (e.g., by viewing citizens reports on a map). To this end, we present a new geotagging approach that can estimate the location of a post based on its text using refined language models that are learned from massive corpora of social media content. Using a large benchmark collection, we demonstrate the improvements in geotagging accuracy as a result of the proposed refinements.

Keywords: Geotagging · Social media · Language models · Similarity search · Spatial entropy · Location detection

1 Introduction

The pervasive use of mobile media capturing equipment (smartphones, cameras) and the increased adoption of online social networking and media sharing services have disrupted the way news stories and real-world events are captured and disseminated. Shortly after the occurrence of an event, such as a natural disaster or a riot, social media platforms such as Twitter, Facebook and Instagram are flooded with data about the event, much of which comes directly from bystanders and witnesses. More often than not, information and media content from people directly involved or attending an event would be highly valuable to decision makers (e.g., reporters, emergency response teams), for assessing the situation and planning the next steps. Yet, it is often extremely challenging to find and make use of such content due to the inherent properties of social media content, namely the large volume of content, the lack of structured information and the reduced trust on the quality and veracity of posted information.

M. Chau et al. (Eds.): PAISI 2015, LNCS 9074, pp. 21–40, 2015.
DOI: 10.1007/978-3-319-18455-5_2

An important element of user-generated content is the geographic location, where a post refers to, and where it was generated. In a few cases, precise location information is made available from the media sharing platform (e.g., geotagged tweets carry the latitude and longitude of the location where the respective tweet was composed). This is extremely helpful, since it is possible for an analyst to view the content on a map (and thus better understand the location context of an event) and to establish its reliability (e.g. when a user posts far away from an event, their posts cannot be considered equally reliable to the ones of direct witnesses). Yet, the vast majority of content in social media is not accompanied by explicit geographic information (for instance, in the case of Twitter less than 1% of content is geotagged). To this end, a number of geotagging approaches have been recently proposed that analyse the posted content, typically the text of the post or the title/description and tags of a posted image, in order to estimate the location where the content refers to.

In this paper, we present a number of refinements over a popular language model-based approach [16], which has been recently demonstrated to have highly competitive performance [14]. With the help of a thorough experimental study on a widely used benchmark dataset (MediaEval 2014 Placing Task), we demonstrate that the proposed refinements result in significant improvements regarding the geotagging accuracy and the reliability of the geotagging output. Furthermore, we present an in-depth exploration of the performance of the proposed approach, including the contribution of each of the different proposed refinements, and the role of increasing the size of the training dataset. To further drive research in the area, we also make publicly available the implementation of the proposed approach as an open-source project[1].

2 Related Work

Geotagging is a very challenging task, which has attracted increasing research interest in recent years. Luo et al. [12] and Zheng et al. [20] provide surveys with detailed overviews of the geotagging research problem and a number of recent approaches. In the following paragraphs, we are briefly presenting a number of representative approaches: in particular gazetteer-based methods, language models and multimodal methods. Moreover, we present the MediaEval 2014 Placing Task, an international benchmarking activity, where the proposed approach was submitted and compared with a number of competing approaches.

2.1 Gazetteer-Based Methods

Gazetteers are essentially large dictionaries or directories that contain comprehensive lists of geographic places. These places are described by various features, such as geographic location, toponyms and alternate names (when available). The gazetteer databases typically contain high quality and precise information

[1] https://github.com/socialsensor/multimedia-geotagging

for the contained places. However, many gazetteers have limited world coverage, which makes them impractical as the sole basis for a global geotagging solution. The most well-known gazetteer databases are Geonames[2] and Yahoo! GeoPlanet[3], with the former being a free public resource with over 10 million geographical names and over 9 million unique features, of which 2.8 million populated places and 5.5 million alternate names.

Several geotagging approaches are based on gazetteers. Kessler et al. [7] combine the existing standards to realize a gazetteer infrastructure allowing for bottom-up contribution as well as information exchange between different gazetteers. They ensure the quality of user-contributed information and improve querying and navigation using a semantics-based information retrieval approach. Smart et al. [17] present a framework that accesses multiple gazetteers and digital maps in a mediation architecture for a meta-gazetteer service using similarity matching methods to conflate the multiple sources of place data in real-time. Lieberman et al. [11] introduce a heuristic method to recognize toponyms and merging list of toponyms, referring to them as *comma groups*. Toponyms in comma groups share a common geographic attribute and determine the correct interpretation of the place name.

2.2 Language Models and Multimodal Methods

In recent years, several researchers have developed data-driven techniques in order to connect the textual metadata of user-generated geotagged images to specific locations or areas with the goal of building large-scale geographical *language models*. In a typical language model-based approach, there is a large collection of geotagged textual content, composing a training set, which is clustered in discrete areas or assigned in regular cells on a virtual grid covering the surface of the earth. This process gives the opportunity to calculate useful keyword/tag statistics for each cluster or cell across the globe. One of the earliest works is [16], where Serdyukov et al. used a predefined grid of cells and calculated the prior probabilities for image tags based on the neighbourhood of the cells that they appeared. More recently, Hauff et al. [6] attempted to overcome the limitation of the fixed grid introducing disjoint dynamically sized cells. O'Hare and Murdock [13] proposed a statistical grid-based language modelling approach, which makes use of the Word-Document model, and they investigated several ways to estimate the models, based on the term frequency and the user frequency. Another approach that uses language models was described in [19], where Van Laere et al. cluster the training set images and then use the χ^2 feature selection criterion to create a vocabulary for every cluster. They also introduced a more aggressive technique, in which they calculate the most similar images, for a query image, using Jaccard similarity (on the respective sets of tags).

Other researchers have proposed multimodal methods that use visual features of images in addition to the text metadata. For instance, Crandall et al. [4] combine image content and textual metadata at two levels of granularity, at a city

[2] http://www.geonames.org/

[3] https://developer.yahoo.com/geo/geoplanet/

level (approximately 100km) and at individual landmark level (approximately 100m). They train classifiers in a relatively small set of landmarks and for a fixed set of cities. Trevisiol et al. [18] process the textual data in order to determine their geo-relevance and find the frequent matching items. Also, they build a user model using the user's upload history, social network data and user's hometown. When there is lack of such information, they use visual features for the prediction of the location. Kelm et al. [8] present a hierarchical approach, making use of external resources to identify toponyms in the metadata, and of visual and textual features to identify similar content.

2.3 The MediaEval 2014 Placing Task

Task and Dataset Description. MediaEval is an international benchmarking initiative that includes a number of tasks in the area of multimedia analysis and retrieval. The Placing Task is dedicated to the geo-localization of images [2]. Participants are challenged to determine an estimated location (in terms of latitude and longitude) of the images that are contained in a test set using another set of images for training. In MediaEval 2014, the training dataset included more than 5M images, and test datasets of different sizes, between 5K and 510K images, were distributed, with the smaller sets being subsets of the larger ones. All the datasets were subsets of the recently released Yahoo! 100M Flickr CC dataset[4] (YFCC100M).

The task participants where asked to submit up to five runs, among which one (`run1`) had to be purely text-based, meaning that only textual information was eligible, and a second (`run2`) had to be purely visual-based, i.e. using only the pixel content of images. For the other three runs, participants were allowed to utilize gazetteers, external data or any additional information, but not crawl the test images. In terms of evaluation, the submitted runs where benchmarked based on their accuracy in different ranges. The estimated location for a test (query) image was compared to its correct location; if it was located inside the circle with a radius equal to some predefined range (the used circular ranges were 10m, 100m, 1km, 10km, 100km, and 1000km) from the centre of the correct location, the estimate was considered correct for the respective range, hence resulting in the computation of *Precision at range X* (e.g., $P@1km$). Additionally, the median error was calculated, i.e. the median of the estimation errors across all test images in terms of the Haversine distance between the predicted and the actual location across all the images in the test set.

Overview of Competing Approaches. Six teams participated in the MediaEval 2014 Placing Task, including one based on the proposed approach [9]. Popescu et al. [15] used a grid-based language modelling approach, in which they divided the earth surface in a rectangular grid, and constructed a probabilistic location model based on the users that use a tag in each cell of the grid.

[4] http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67

They also built a Flickr-specific *machine tag* model, which recognizes the machine tags that contain the geographic location of the associated image. Based on users' information, they also built a user model. In [1], Cao et al. employ the Ripley's K-statistic algorithm in order to weight the tags included in the training set. They built a language model framework, utilizing the tag weights and Bayesian Smoothing with Dirichlet priors. Ferrés et al. [5], built two basic systems. The first is the Geographical Knowledge Base (GeoKB), which uses Geoname's gazetteer and an English Dictionary, refining the results through logical assumptions. The second one is Hiemstra's Language Model (HLM) with re-ranking, which combined the Terrier[5] Information Retrieval (IR) engine with the HLM weighting model. In [10], Li et al. applied a combination of textual, visual and audio analysis in order to geocode the given image/videos. Further, they re-ranked items using the RL-Sim algorithm and predicted the location of the images by clustering the top-rated results. Finally, Coi et al. [3] developed a spatial variance approach targeted to recognize the toponyms that are contained in the images and a graphical model-based approach. For the visual analysis they developed the Geo-Visual Ranking (GVR) approach, which processes the most similar training images to the query image to make an estimate based on their locations. The results of the aforementioned approaches serve as a benchmark to the results we present in Section 4 using the same experimental setting.

3 Approach Overview

The objective of the proposed system is to calculate the geographical location of social media items using text analysis on their content (e.g. tweets, image tags, etc.). Based on a pre-calculated probabilistic language model, which is derived from processing a massive amount of filtered data, an actual location is derived for a query item, in terms of latitude and longitude coordinates. To simplify the presentation of the approach, one should bear in mind that two sets of items are involved. The first, typically a massive corpus of geotagged textual items, is used for creating (training) the language model, and the second for testing the geotagging accuracy of the constructed model. The two sets will be denoted as D_{tr} and D_{ts}, respectively. In the case of Flickr images (which are used as an experimental test bed in this work), the image metadata that are used are the tags, title, user id, image id and description. In particular, the metadata of images from D_{tr} are analysed to create a probabilistic language model that is then used for predicting the location of the images (based on their metadata) from set D_{ts}. The language model is built based on the tags and titles of the images in D_{tr}. Afterwards, the model tags are processed in order to select and weight those that have the greater contribution to the location prediction problem. Finally, the approach employs some additional techniques to further refine the location predictions for the images of D_{ts}. A high-level view of the proposed geotagging approach is illustrated in Figure 1.

[5] http://terrier.org/

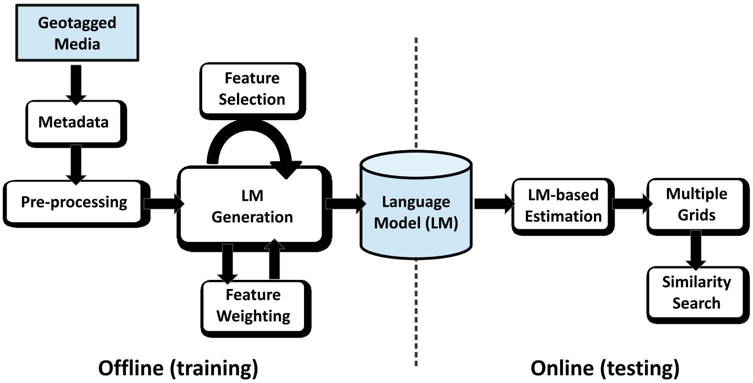


Fig. 1. Overview of proposed geotagging approach

3.1 Language Model

The construction of the language model relies on an offline processing step, in which a complex geographical-tag model is built from the tags, titles and locations of the images contained in D_{tr}. For estimating the location of a query image, the description of the images is also used in case no geographic information can be gleaned from its tags and title. A pre-processing step is first applied: all punctuation and symbols are removed (except from the add symbol "+", because in the particular dataset, it is used to link the keywords of multi-keyword toponyms, e.g., `new+york`. Also, all characters are transformed to lowercase and all tags composed of numerics are removed. Finally, phrases that contain the add symbol are split into single tags (e.g., the single tags in `new+york` are `new` and `york`). After the pre-processing, several images in D_{tr} are left with no tags and title and are hence disregarded from the remaining steps. Note that the same pre-processing is applied on the test images before the actual location estimation process. For ease of reference, we will refer to the keywords of an arbitrary social media item as tags and denote their set as T.

In order to generate discrete geographical areas, the earth surface is divided in rectangular cells with a side length of 0.01° for both latitude and longitude (corresponding to a distance of approximately 1km near the equator). Therefore, a grid C of cells is created, which is used to build the language model using the approach described in [14]. More specifically, for a query image, an estimation of the most probable cell $c \in C$ takes place based on the respective tag probabilities. A tag probability in a particular cell is calculated as the total number of different Flickr users that used the tag inside the cell, divided with the total count of different users over the whole grid C. More specifically, the tag-cell probability $p(t|c)$ is calculated for every tag $t \in T$ according to Equation 1.

$$p(t|c) = \frac{N_u}{N_t} \tag{1}$$

where N_u is the number of users in D_{tr} that used the tag t inside the borders of cell c, and N_t is the total count of different users that used the tag t in all cells. Note that a user can be counted in N_t more than once. If a user u is found in multiple cells, every time he/she is found in a different cell, he/she is considered as a new user and increases the total count of users.

In order to assign a query text to a cell, the probability of each cell of C is first calculated summing up the contributions of each individual tag in T. Then, the cell with the highest probability is selected as the text cell according to Equation 2.

$$c_j = \arg\max_i \sum_{k=1}^{N} p(t_k|c_i) \tag{2}$$

where, c_j is the most likely cell for item $j \in D_{ts}$, N is the total number of tags of in T_j and $p(t_k|c_i)$ is the tag-cell probability for tag $t_k \in T_j$ in cell $c_i \in C$. Based on the base language model presented here, the location estimation for item j is considered to be the centre of the c_j. If during this process there is no outcome (i.e. the probability for all cells is zero), then the description of the query image (in case of Flickr images) is utilized. For the D_{ts} images where there is no result (e.g. complete lack of text), their location is set equal to the centre of the most populated cell, in a coarse granularity grid (100km×100km), which is a kind of maximum likelihood estimation.

3.2 Feature Selection

To increase the robustness of the model and reduce its size, we make use of a feature selection technique. The features that need to be ranked are the language model tags. For this reason, a technique is proposed based on a cross-validation scheme using the training set only. The basic idea is to rank the tags based on the accuracy they achieve for predicting the location of items in the withheld fold. First, the set D_{tr} is partitioned into p folds. The number of partitions p is empirically selected; in this implementation, it was set to 10. Subsequently, one partition D^p_{tr} at a time is withheld, and the rest $p-1$ partitions are used to build the language model. Having built the language model, the location of every item of the withheld partition is predicted using the method described in subsection 3.1. In that way, it is straightforward to determine the contribution of each tag to the prediction of the target location: a score is computed based on the ratio of the number of correctly geotagged (in range r) items where the tag appears over the total number of items where the particular tag appears.

$$tgeo(t) = \frac{N_r}{N_t} \tag{3}$$

where, $tgeo(t)$ is the score of each tag t of the language model (essentially its *geographicity*), N_r is the total number of correctly geotagged items in D^p_{tr} where t appears, and N_t is the total number of items in D^p_{tr} where it appears. The feature selection step is carried out using a threshold, to be denoted as θ_{tgeo}, and only those tags that surpass it are selected, provided they are used by a minimum number of unique users in the whole training set (this second threshold is denoted as θ_u).

3.3 Feature Weighting Using Spatial Entropy

In order to adjust the original language model tag probabilities for each cell, we build a Gaussian weight function based on the values of the spatial tag entropy. First, for each tag in the model, its spatial entropy value is calculated. In order to calculate the entropy values, the Shannon entropy formula is applied in the tag-cell probabilities. This is a measure of the stochasticity of the tag's appearance in the cells of C and is expressed by Equation 4.

$$e(t) = -\sum_{i=1}^{M} p(t|c_i) \log p(t|c_i) \tag{4}$$

where $e(t)$ is the spatial entropy value of tag t, $p(t|c_i)$ is the tag-cell probability of t in cell $c_i \in C$ and M is the total number of cells.

Once the entropy values are computed, a Gaussian normalization is applied because the tags with either too high or too low entropy values typically carry no geographic information, and therefore their influence on the location estimation process needs to be suppressed. Tags with too low entropy values tend to be user-specific. A typical example is a tag that is only used by a single user in a single cell. This will have a zero entropy value and it is not considered useful in the location detection process. In the same way, very high entropy values indicate tag appearance that is widely spread across the globe. As a result, such a tag would carry no geographical interest (e.g., `baby` and `fun` are tags with very high spatial entropy values). Due to this fact, a Gaussian normalization is used for the re-weighting of the tag-cell probabilities. The Gaussian function is specified in Equation 5.

$$N(e(t), \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{e(t)-\mu}{2\sigma}\right)^2} \tag{5}$$

where N is the Gaussian function, and parameters μ, σ are the mean value and the variance of the entropy distribution, respectively, and are estimated on D_{tr}. Based on the Gaussian normalization, Equation 2 is adapted to Equation 6.

$$c_j = \arg\max_i \sum_{k=1}^{N} p(t_k|c_i) \cdot N(e(t_k), \mu, \sigma) \tag{6}$$

where N is the number of tags for image j, $p(t_k|c_i)$ is the probability of tag t_k for cell c_i and $e(t_k)$ is the spatial entropy of tag t_k. Figure 2 illustrates the histogram of entropy values on the training set and the respective weights.

3.4 Similarity Search

Having assigned a query item to a cell, a further location refinement is conducted using the technique of [19]. First, the k most similar images in D_{tr} that fall inside

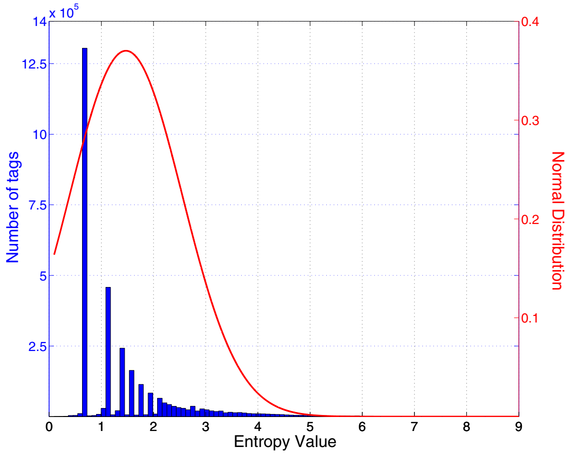


Fig. 2. Histogram of entropy values and Gaussian weighting for each range

cell c_j, are determined using Jaccard similarity on the corresponding sets of tags. For images x and y, their Jaccard similarity is defined by Equation 7.

$$J(x,y) = \frac{|T_x \cap T_y|}{|T_x \cup T_y|} \tag{7}$$

The final estimation is the centre-of-gravity of the k most similar images, weighted by the similarity values as the location estimate for the test image.

$$loc(x) = \frac{1}{k}\sum_{i=1}^{k} J(x|y_i)^a loc(y_i) \tag{8}$$

where parameter $\alpha \in [0, +\infty]$ determines how strongly the result is influenced by the most similar items. In order to perform the calculation, the location coordinates are first transformed to the Cartesian system and are then transformed back to spherical (latitude and longitude). In case that less than k similar items are found in c_j, then the centre-of-gravity is calculated by only those which are similar to x. If no similar items are found, then the centre of c_j is output as the estimated location.

3.5 Multiple Resolution Grids

In order to ensure more reliable prediction in finer granularities, we built an additional language model using a finer grid (cell side length of 0.001° for both latitude and longitude, corresponding to a square of ≈100m×100m near the equator). The grids for the coarser and finer grids are denoted as C_c, and C_f, respectively. Having computed the estimated location for both the coarse and fine granularity, we use the following refinement: if the estimated cell c^f based on the finer granularity falls within the borders of the estimated cell of the coarser granularity c^c, then the prediction is based on the fine granularity and similarity search is applied on cell c^f. Otherwise, similarity search is performed on the cell of coarser granularity c^c, since coarser granularity language models are considered more reliable by default, given that more data per cell are used for their creation, and hence the resulting probabilistic analysis is more robust.

4 Evaluation

We first evaluate the effectiveness of the proposed approach in comparison to competing approaches in the MediaEval 2014 Placing task, and next we explore in detail several performance aspects of the approach using the same reference dataset. The dataset used in this evaluation was already described in Section 2. All experiments described here were performed on the largest test set (510K).

4.1 MediaEval 2014

For the participation in the task, we submitted five runs, three of them based on text, using variants of the presented approach, and two based on the visual content of images, which are not discussed in this paper.

In the submitted text-based runs, our goal was to demonstrate the improvement of the results by applying the proposed refinements of subsections 3.3-3.5 on the base approach that relies on the language model (subsection 3.1). Hence, `run1` corresponds to using the language model, spatial entropy, similarity search, and multiple grid, `run4`, using the language model only (base approach), and `run5`, using the language model and similarity search. For all three runs the parameters used for the similarity search were $\alpha = 1$ (empirical tests indicated that the effect of this parameter on the geotagging accuracy was marginal), and $k = 4$ (which led to optimal results on the training set). The feature selection technique described in subsection 3.2 had not been developed and implemented at that time, so it was not included in these tests (but is assessed later).

Table 1 contains the result of the text-based runs for the various accuracy ranges. The best performance in terms of both median error and accuracy in all ranges was attained by `run1`. Comparing `run4` and `run5`, it appears that similarity search had considerable impact on the low range accuracy results. Also the combination of all features in `run1` further improved the overall performance (reaching a 5.85% accuracy for the 100m range, which was the second best performance in the contest), but the median error was still relatively high (230km), which means further improvements are possible.

Table 2 contains the results of the best textual runs of all participants in the MediaEval 2014 Placing Task. Note that for all methods we compare the runs that used only the training data released by the organizers (the set of ≈5M Flickr images). As can be seen, the variant of our proposed approach is positioned in the second or third place for the lower accuracy ranges, which are more important

Table 1. Geotagging precision (%) for six ranges and median geotagging error (km)

	P@10m	P@100m	P@1km	P@10km	P@100km	P@1000km	m. error
run1	0.50	5.85	23.02	39.92	46.87	60.11	230
run4	0.03	0.65	21.87	38.96	46.13	59.87	258
run5	0.31	4.36	22.24	38.98	46.13	59.87	259

Table 2. Geotagging precision (%) for five ranges and median geotagging error (km) of the best textual runs for all six participants in the MediaEval Placing Task 2014

	P@10m	P@100m	P@1km	P@10km	P@100km	m. error
SocialSensor [9]	0.50	5.85	23.02	39.92	46.87	230
USEMP [15]	0.70	1.60	23.50	40.80	48.10	N/A
UQ-DKE [1]	1.07	4.98	19.57	41.71	52.46	51
TALP-UPC [5]	0.29	4.12	16.54	34.34	51.06	84
RECOD [10]	0.55	6.06	21.04	37.59	46.14	N/A
ICSI/TU Delft [3]	0.24	3.15	16.65	34.70	45.58	N/A

for practical applications. This is also illustrated in Figure 3, where the deep blue line, that represents the team SocialSensor, lays above the other teams' lines at the leftmost part of the diagram, but it increases at a lower rate than competing approaches.

Figure 4 illustrates the median geotagging error in terms of the number of tags per test image. Obviously, run1 achieves the best result, since it is the most accurate of the three runs, achieving a median error just below 20km for images with 16-20 tags, and clearly outperforming run4 and run5 for images with more than 10 tags. The performance of run4 and run5 is very similar, with the only exception the images that contain 10-20 tags, where run5 appears to perform slightly better. It is noteworthy that images with more than 20 tags appear harder to geotag (for all runs), potentially corresponding to spammy

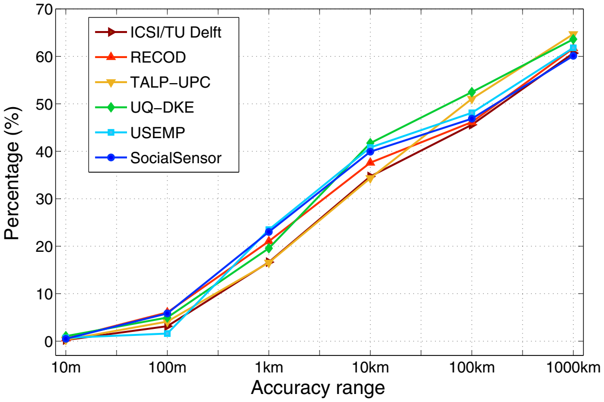


Fig. 3. Geotagging precision (%) of the best textual runs of all the participants in the MediaEval 2014 Placing Task for different accuracy ranges

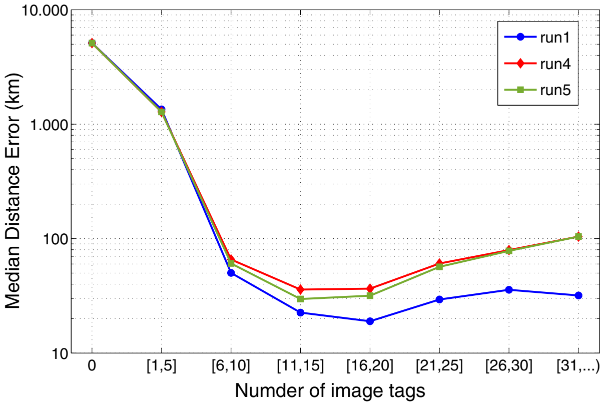


Fig. 4. Median geotagging error (km) in comparison to the number of tags that are contained in images for the submitted textual runs in MediaEval 2014 Placing Task

or very noisy metadata. Yet, it appears that the proposed extensions (feature reweighting with spatial entropy, multiple grids) are highly effective in dealing with such tags.

4.2 Further Performance Analysis

Additionally, beyond the scope of the MediaEval benchmark, in order to improve further the geotagging accuracy of the proposed approach and to explore its performance,

we made use of the full set of geotagged metadata included in the YFCC100M dataset. Excluding all the images that do not contain geo-locations and after the pre-processing step, a total set of approximately 48 million images was used for creating the language model. On this set, the feature selection method of subsection 3.2 was applied by partitioning the set in folds of 4.8 million images each. Calculating the tag geographicities according to Equation 3 for a 1km geotagging range, and filtering those tags with $tgeo > \theta_{tgeo} = 0$ and $N_t > \theta_u = 1$, we ended up with a tag model of 4,547,803 tags.

Using the language model as baseline, we tested the effect of the different refinements resulting in various configurations of the proposed approach. We denote those with FS (Feature Selection), SE (Spatial Entropy re-weighting), MG (Multiple Grid refinement), and SS (Similarity Search). The results of these experiments are presented in Table 3. We also group the experiments in two settings. In the first (the so-called `FAIR` setting), the users that appear in the test set are completely excluded from the training set, while in the second setting (`OVERFIT`), those users are not removed from the training set. The results of the latter setting are considered as overly optimistic and not transferable to different datasets, since the inclusion of tags from the same users in the training set is bound to have a very positive effect for estimating the location of those images whose owners (users) are included in the training set. The results using the

Table 3. Geotagging precision (%) for five ranges and median geotagging error (km) for different configurations of the proposed approach

	FS	SE	MG	SS	P@10m	P@100m	P@1km	P@10km	P@100km	m. error
FAIR	✓				0.00	0.74	24.44	41.37	48.29	162
	✓	✓			0.00	0.75	24.83	41.54	47.65	181
	✓		✓		0.17	6.67	24.69	41.37	48.29	162
	✓		✓	✓	0.66	7.58	24.93	41.38	48.29	162
	✓	✓	✓	✓	0.67	7.65	25.90	41.54	48.29	160
OVERFIT					0.04	1.37	40.95	54.87	60.41	3.55
	✓				0.04	1.41	42.30	57.00	62.79	2.68
	✓	✓			0.04	1.39	41.94	56.21	61.56	2.89
	✓	✓	✓	✓	1.57	20.05	42.67	57.20	62.79	2.66

`OVERFIT` setting are only presented in Table 3 as a kind of "Oracle" setting, and the setting is not further considered in the rest of the experiments.

The best results are achieved by the combination of all proposed refinements, which results to P@100m=7.65% and P@1km=25.90%. Hence, applying the proposed refinements appears to have a clear advantage compared to the base language model, especially in the fine estimation ranges (100m, 1km). In comparison to the run that was submitted to MediaEval, the use of more training data (48M compared to 5M) also contributed to the improvement of the performance. For instance, comparing the accuracy of `run4` with its counterpart in Table 3, which is the first row, it is evident that a gain of 2.57% (in P@1km), 24.44% versus 21.87%, is achieved. This is a 10% relative improvement, which came at the cost of increasing the training set size by almost 10 times.

Figure 5 depicts the median geotagging error (relative to the number of tags) of `run1`, `run4` and two configurations of the approach that use the full YFCC100M dataset, one combining only the language model with feature selection and the second using all of the proposed refinements. The combination of all proposed refinements appears to result in the best geotagging accuracy in almost all tag ranges, except the [6, 10] range where the base language model slightly outperforms the rest. Another noteworthy fact is that using the proposed improvements on the reduced training set (5M), i.e. `run1`, has almost an equivalent benefit on the geotagging accuracy, as the increase of the training set by almost 10 times in tandem with the base Language Model (`LM+YFCC100M`).

Figure 6 illustrates the median geotagging error per cell across the globe. The color bar presents the mapping of median error levels to colors. The cells with median error less than 150km are displayed with deep blue color, whereas those with more than 900km are displayed with brown red color. It is noteworthy that in North America and Australia the dominant color is brown (very high error), despite the availability of much more training data and the prevalence of English text (which is expected to be easier to handle). In contrast, in Europe a wide area is painted light blue, so in these areas the algorithm worked considerably better. A possible explanation for the high error levels in the US and Australia is the potential ambiguity in town and city names (e.g., many American towns are named after European ones). The two configurations that are displayed in

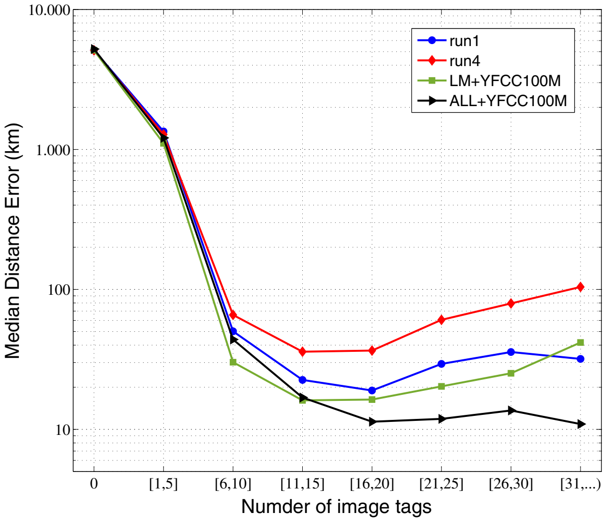
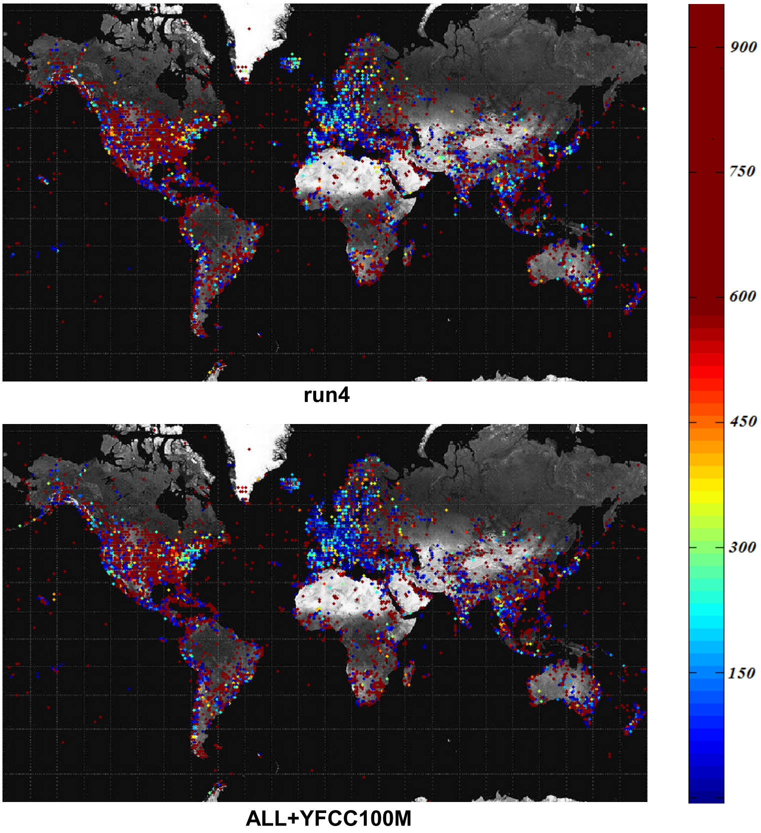


Fig. 5. Median geotagging error (km) relative to the number of tags per image for the MediaEval 2014 `run1`, `run4` and two of the runs with the extended training dataset

the figure correspond to `run4` in the MediaEval 2014, and the combination of all proposed refinements using the extended training dataset (`ALL+YFCC100M`). There appears to be an improvement in the second map in multiple locations over the globe.

As a further experiment, we were interested in investigating whether the sum of cell-tag probabilities for all the tags of an item (the sum terms in Equations 2 and 6) can be used as an indicator of the reliability of the detected location for a test item. To this end, we computed the geotagging precision only on the subset of images, for which the respective sum exceeded a user-selected threshold d, which we varied in this study. Figure 7 illustrates the obtained results. In particular, Figure 7(a) displays the geotagging precision at different ranges for the images that exceed threshold d for increasing values of the threshold. Figure 7(b) depicts the percentage of images that are placed in the range of 1km and exceed the threshold (blue line) versus the percentage of images that do not exceed the threshold (and are hence not placeable). A very important finding from this test is that the sum of cell-tag probabilities for the tags of an image is indeed a very good indicator of the location prediction reliability, at least for ranges of 1km and above. For instance, according to Figure 7(a), for images where this sum exceeds the value of $d = 0.1$, the geotagging accuracy at 1km range exceeds 70%. Figure 7(b) suggests that in that case, only 35% of images out of the original test set can be placed with such accuracy. Hence, this thresholding strategy is very practical for tuning the trade-off between geotagging accuracy and placeability.

In many cases, the location of an image may differ from the location depicted in its content, since the registered location (by use of the camera GPS sensors)

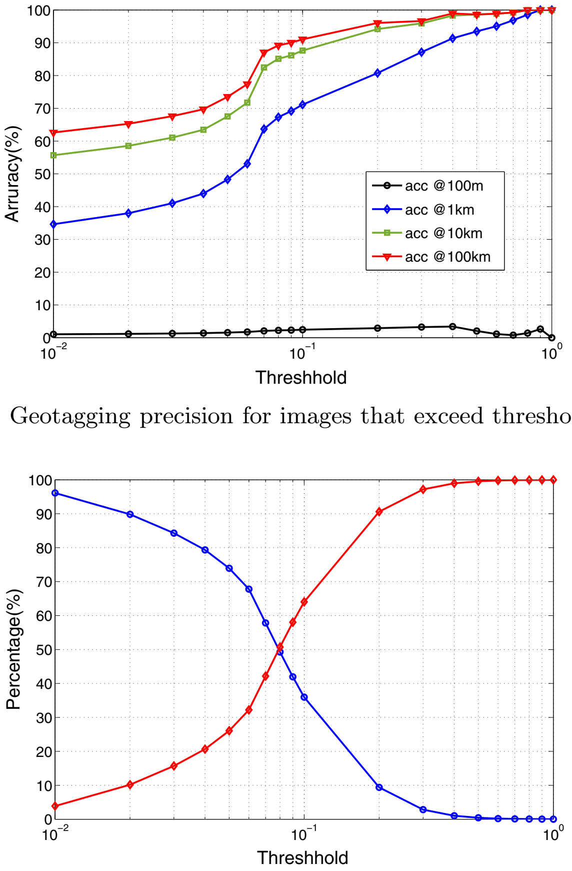

run4

ALL+YFCC100M

Fig. 6. Spatial distribution of mean geotagging error in `run4` and `ALL+YFCC100M`

typically corresponds to the location of the photographer and not of the photographed object. For example, an image of the Eiffel Tower that was taken from the opposite side of Seine is automatically located relatively far from the actual Eiffel Tower location. Combined with the fact that we used image text metadata to build our language model, one may expect that part of the geotagging errors of the proposed approach, especially in low range accuracies (<1km) could in fact be attributed to this particularity of the training dataset.

To further expose this issue with a concrete example, we collected the images from the test set that contain the exact phrase `statue of liberty` and were located close to the monument (within a square with approximately 10km sides and centred on the monument). Figure 8 depicts the images that comply with the above criteria. The points on the map correspond to the real locations of the images and are coloured based on their distance error between their estimated and real locations (green, yellow and red correspond to distance errors <0.5km, <1km and ≥1km, respectively). It is noteworthy that based on the accompanying text metadata, the photos should have been geotagged on the Statue of Liberty; yet their capture locations (which were used to generate the language model) are



(a) Geotagging precision for images that exceed threshold d

(b) Percentage of placeable images for different values of d

Fig. 7. Reliability of location predictions and placeability of social media items

the points displayed on the map. This is obviously expected to have an impact on the accuracy of the proposed approach, especially in lower ranges, since the approach tends to assign a query image to the most likely cell (given the language model) and then to the textually most similar image from the training set. In this example, there are images that are taken from the ferry in the open sea or on the coast across the statue (images 1, 2 in Figure 8). Those images were placed by the proposed approach on the actual location of the Statue of Liberty, which is the desired outcome. Yet, due to their GPS location (which was different from the statue), the resulting test distance error was relatively high, which should be interpreted with caution. In contrast, image 3 was correctly placed thanks to the accompanying tag `battery park`, which together with the phrase `statue of liberty` led the algorithm to a more accurate estimation, i.e. a location inside the mentioned park. Another challenging case is presented by image 4, where the Statue of Liberty appears far in the background, yet it appears in the textual content of the image, thus misleading the geotagging process.

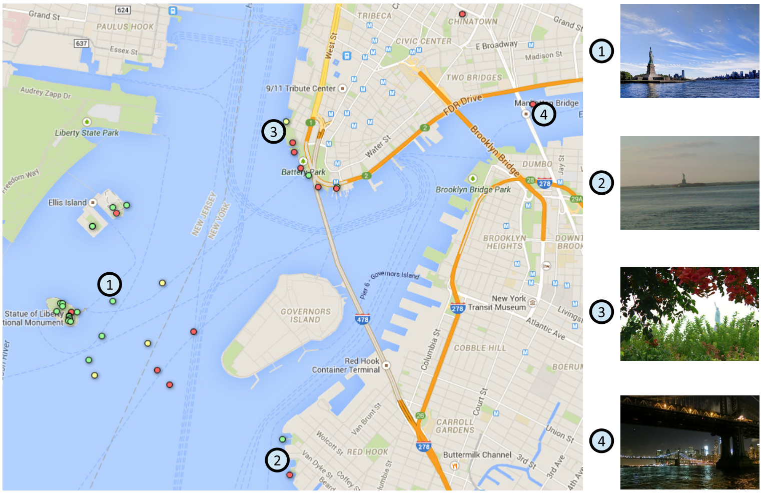
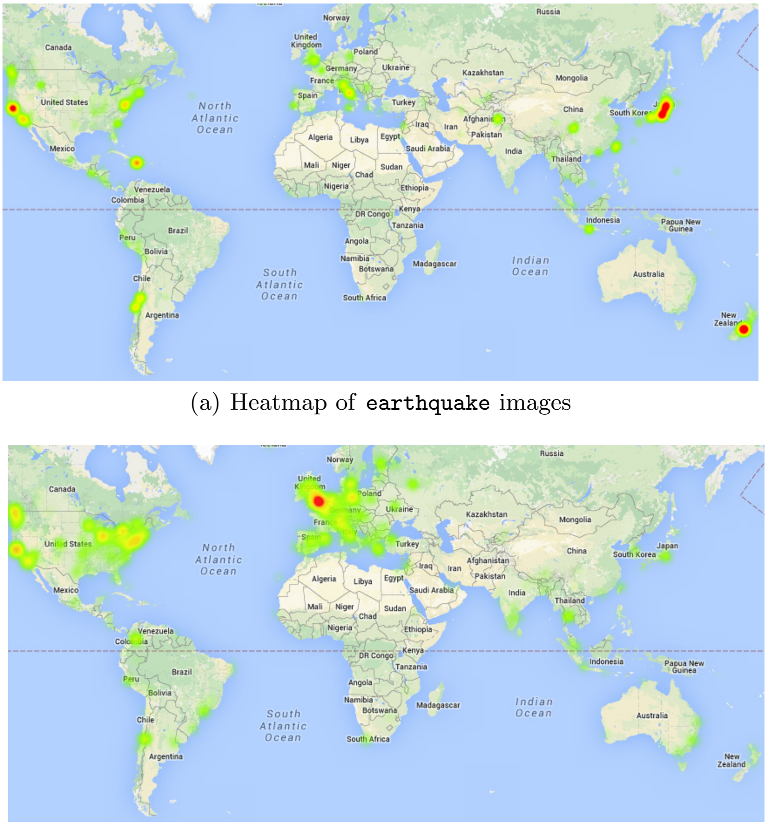


Fig. 8. Statue of Liberty example

A further issue with respect to the training data that was used for building the language model pertains to the use of textual descriptions that are related to temporary events, such as emergencies and natural disasters. Figure 9 illustrates the heatmaps generated from the locations of images carrying the tags `earthquake` and `riot` respectively. In the first case (Figure 9(a)), one may note that the most "active" regions actually correspond to seismogenic zones, such as Japan and California. We may hypothesize that taking into account this information in the location estimation process can be beneficial, since it considerably reduces the space of possible locations. In contrast, taking into account the tag `riot` in the geotagging process may actually not be particularly informative. According to the heatmap of Figure 9(b), there is a variety of locations across the globe (with higher density in London, European capitals and major US cities). One might hypothesize that when using a very large training set (such as the YFCC100M), the resulting language model would be bound to contain riot-related content in all major cities of the world. In addition, given the temporal volatility of such events (e.g., there may be periods, where more riots take place within a particular city), one should be very cautious when using such tags for building a geographic language model.

In a final experiment, we explore the performance of the approach when estimating the location of media content that is associated with emergency situations. To this end, we retrieved the test set images that contained at least one of the keywords `fire`, `flood`, `earthquake`, `hurricane`, `riot` and `demonstration`, all of which are linked to emergency situations. The total number of retrieved test images was approximately 6,000. The results of the approach on this set of images (`Emergency set`) are presented in Table 4 and compared to the results on the full MediaEval 2014 test set. We may note that the geotagging accuracy in the small accuracies (10m, 100m) is very similar (marginally lower for the emergency set). However, the geotagging accuracy in the emergency set is con-



(a) Heatmap of `earthquake` images

(b) Heatmap of `riot` images

Fig. 9. Tag-specific heatmaps based on the YFCC100M training set

Table 4. Geotagging precision on full MediaEval 2014 test set (same as fourth entry of Table 3) and reduced test set focusing on emergency-related images

	P@10m	**P@100m**	**P@1km**	**P@10km**	**P@100km**	**m. error**
`Full test set`	0.67	7.65	25.90	41.54	48.29	160
`Emergency set`	0.58	7.55	28.96	49.11	57.82	15

siderably higher for the rest of the ranges (and for the median error), which are very important for improved situational awareness and for devising appropriate emergency response strategies. An important finding from this test is that emergency-related images seem to carry text metadata that are helpful for the geotagging problem, and hence one could make use of such images even when they do not carry GPS metadata.

5 Conclusions

We presented a number of refinements over the base language model for geotagging social media content based on text. The proposed refinements included a method for performing feature selection, a feature reweighting function based on spatial entropy, similarity search, and a multiple grid technique. We presented a thorough experimental study on the MediaEval 2014 Placing Task and demonstrated the highly competitive performance of the proposed method, along with further improvements as a result of using a considerably larger training dataset, and further tuning the configuration of the refined approach. We consider that the proposed approach along with the insights gained from the conducted experimental study can lead to a reliable geotagging solution for social media settings in a variety of practical settings.

In the future, we aim at exploring the effect of using training data from different social media sources (i.e. other than Flickr) on the performance of the approach, as well as the generalization ability of the system (i.e. training using data from one social media source, and testing using data from a different one). Furthermore, we are interested in making use of the visual content of images, which is a much more challenging problem, to further improve on the geotagging accuracy and placeability.

Acknowledgement. This work was supported by the REVEAL project, partially funded by the European Commission (contract no. FP7-610928).

References

1. Cao, J., Huang, Z., Yang, Y., Shen, H.T.: UQ-DKE's Participation at MediaEval 2014 Placing Task. In: Proceedings of MediaEval 2014 Placing Task, Barcelona, Spain, October 17-18 (2014)
2. Choi, J., Thomee, B., Friedland, G., Cao, L., Ni, K., Borth, D., Elizalde, B., Gottlieb, L., Carrano, C., Pearce, R., Poland, D.: The Placing Task: A Large-scale Geo-estimation Challenge for Social-media Videos and Images. In: Proceedings of the 3rd ACM GeoMM Workshop (2014)
3. Choi, J., Li, X.: The 2014 ICSI/TU Delft Location Estimation System. In: Proceedings of MediaEval 2014 Placing Task, Barcelona, Spain, October 17-18 (2014)
4. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the World's Photos. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, pp. 761–770. ACM, New York (2009)
5. Ferrés, D., Rodríguez, H.: TALP-UPC at MediaEval 2014 Placing Task: Combining Geographical Knowledge Bases and Language Models for Large-Scale Textual Georeferencing. In: Proceedings of MediaEval 2014 Placing Task, Barcelona, Spain, October 17-18 (2014)
6. Hauff, C., Houben, G.-J.: Geo-Location Estimation of Flickr Images: Social Web Based Enrichment. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 85–96. Springer, Heidelberg (2012)

7. Kessler, C., Janowicz, K., Bishr, M.: An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 91–100. ACM (2009)
8. Kelm, P., Schmiedeke, S., Sikora, T.: A Hierarchical, Multi-modal Approach for Placing Videos on the Map using Millions of Flickr Photographs. In: Proceedings of the 2011 ACM Workshop on Social and Behavioural Networked Media Access, SBNMA 2011, pp. 15–20. ACM, New York (2011)
9. Kordopatis-Zilos, G., Orfanidis, G., Papadopoulos, S., Kompatsiaris, Y.: SocialSensor at MediaEval Placing Task 2014. In: Proceedings of MediaEval 2014 Placing Task, Barcelona, Spain, October 17-18 (2014)
10. Li, L., Penatti, O., Almeida, J., Chiachia, G., Calumby, R., Mendes, P., Pedronette, D., Torres, R.: Multimedia Geocoding: The RECOD 2014 Approach. In: Proceedings of MediaEval 2014 Placing Task, Barcelona, Spain, October 17-18 (2014)
11. Lieberman, M.D., Samet, H., Sankaranayananan, J.: Geotagging: using Proximity, Sibling, and Prominence Clues to Understand Comma Groups. In: Proceedings of the 6th Workshop on Geographic Information Retrieval, pp. 6:1–6:8 (2010)
12. Luo, J., Joshi, D., Yu, J., Gallagher, A.: Geotagging in Multimedia and Computer Vision – a Survey. MTAP 51(1), 187–211 (2011)
13. O'Hare, N., Murdock, V.: Modeling Locations with Social Media. Information Retrieval, 1–33 (2012)
14. Popescu, A.: CEA LIST's participation at MediaEval 2013 Placing Task. In: Proceedings of MediaEval 2013 Placing Task, Barcelona, Spain, October 18-19 (2013)
15. Popescu, A., Papadopoulos, S., Kompatsiaris, Y.: USEMP at MediaEval Placing Task 2014. In: Proceedings of MediaEval 2014 Placing Task, Barcelona, Spain, October 17-18 (2014)
16. Serdyukov, P., Murdock, V., Van Zwol, R.: Placing Flickr Photos on a Map. In: SIGIR 2009, pp. 484–491. ACM, New York (2009)
17. Smart, P.D., Jones, C.B., Twaroch, F.A.: Multi-source Toponym Data Integration and Mediation for a Meta-Gazetteer Service. In: Fabrikant, S.I., Reichenbacher, T., van Kreveld, M., Schlieder, C. (eds.) GIScience 2010. LNCS, vol. 6292, pp. 234–248. Springer, Heidelberg (2010)
18. Trevisiol, M., Jégou, H., Delhumeau, J., Gravier, S.: Retrieving Geo-Location of Videos with a Divide and Conquer Hierarchical Multimodal Approach. In: ICMR 2013, Dallas, United States. ACM (April 2013)
19. Van Laere, O., Schockaert, S., Dhoedt, B.: Finding Locations of Flickr Resources using Language Models and Similarity Search. In: ICMR 2011, pp. 48:1–48:8. ACM, New York (2011)
20. Zheng, Y.T., Zha, Z.J., Chua, T.S.: Research and Applications on Georeferenced Multimedia: a Survey. Multimedia Tools and Applications 51, 77–98 (2011)