WISTUD at MediaEval 2011: Placing Task

Claudia Hauff
Web Information Systems
Delft University of Technology
Delft, The Netherlands
c.hauff@tudelft.nl

Geert-Jan Houben
Web Information Systems
Delft University of Technology
Delft, The Netherlands
g.j.p.m.houben@tudelft.nl

ABSTRACT

In this paper, we describe our approach to estimating the geographic location of videos. Our system relies on textual meta-data and includes two basic term filtering strategies: filtering according to the general use of terms and filtering according to the geographic spread. Combining both filtering steps yields 50% accuracy within a 10km range.

1. INTRODUCTION

The goal of the Placing Task [4] is to assign geographic locations (latitude and longitude values) to test videos, using textual meta-data, audio and visual features as well as social information that is available in the training and test data. The training corpus consists of approximately three million Flickr images and ten thousand videos. All but one of our experiments rely exclusively on the textual meta-data.

2. SYSTEM AND RUNS

In the first year of participating at MediaEval, we focused our efforts on building a system that predicts a video's location based on the textual meta-data assigned to it, in particular the tags and the title terms. We follow the approach described in [5] and divide the world map into a number of cells with varying latitude/longitude ranges and assign all items of the training data to their respective cells. All available images from the development set with an accuracy of 11 or higher as well as all training videos were used for training ($2,974,635$ items in total). Since we rely on textual meta-data, we can treat images and videos in the same manner. Then, for each cell, a language model [7] is derived from the items' textual meta-data. Predicting the location of a test video is a two-step process: first, the cell C_{max} is identified whose language model generates the test video's bag-of-words (tags and title terms) T_{test} with the highest probability. In a second step, the same process is repeated within C_{max} to find the most closely matching training item I_{max}. The latitude/longitude of I_{max} is returned as the estimated location of the test video.

In contrast to [5], the grid cells in our approach are of varying size: starting with a grid cell that spans the entire world map (if viewed as a graph, this cell is the root node), the training items are added to the cell one at a time. Once the number of items in a cell exceeds the set limit ℓ_{split}, the cell is split into four equally sized cells (four children nodes are added) and the training items are re-distributed to these cells. To avoid too many splits in areas where large amounts of training data are available, a cell may not be split any further if its latitude/longitude range reaches a lower limit ℓ_{lat_lng}. This process yields cells of small size for areas where the training data is dense, and cells of large size in areas where the training data is sparse.

If a test video contains no tags or title terms (or all terms are filtered out as described below), the terms in the user location are used instead, a fall-back strategy inspired by [2]: if a user does not tag a video with its location, it is likely to be taken at the user's home location. In contrast to [2], we add the user location terms to T_{test}, instead of relying on an external resource to convert the user location to latitude/longitude coordinates. Finally, if the user location yields no usable terms, a latitude/longitude of 0/0 is assigned to the test video.

2.1 Term Filtering

We experiment with two basic term filters. Filtered out from T_{test} are (i) terms that are used by less than U users in the training data, and, (ii) terms with a geographic spread score greater than threshold θ_{geo}. Excluding terms that are used by very few users is hypothesized to improve the robustness of the approach.

Geographic spread filtering is applied for a similar reason: a video may be tagged with a number of non-geographic terms such as "wedding" or "bowling" in addition to tags that are likely to refer to locations such as "london" or "sydney". Whether a term is likely to have a geographic scope can either be determined by matching the term against a geographical database (such as GeoNames[1]) or by considering how localized the term is in the training data. We follow the latter approach here as it does not require any external resources. While in the development data the term "sydney" occurs primarily in one particular grid cell (as expected the cell containing the location of Sydney, Australia), the term "bowling" is spread considerably wider, mainly across North America. This observation leads to a simple but effective geographic spread score: a grid is placed over the world map (1 degree latitude/longitude range per cell) and the number of training items in the cell that contain the term are recorded. Neighbouring grid cells with a non-zero count are merged (in order to avoid penalizing geographic terms that cover a wide area) and the number of non-zero connected components are determined. This score is normalized by

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

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

Term	Geographic Spread Score
bowling	3.237
baby	1.809
valley	1.512
british	0.363
lakepukaki	0.049
españa	0.021
thenetherlands	0.011
london	0.010
sydney	0.007

Table 1: Examples of geographic spread scores. In our experiments, we use a threshold of $\theta_{geo} = 0.1$.

the maximum count. Thus, the smaller the score, the more localized the term occurs in the training data. Our approach is simpler than the χ^2 feature selection based geo-term filtering [6], which determines the geographic score for the tags in each cell separately. Examples of terms and their geographic spread score are shown in Table 1. While the scores of most terms appear reasonable, "british" is incorrectly identified as non-geographic (if we assume a threshold of $\theta_{geo} = 0.1$) as it is not only used to tag pictures taken in the United Kingdom. In the development data it is also used to describe British Columbia (Canada), the British Virgin Islands (Caribbean), British restaurants (mainly in the USA) and placed where historical battles against the British took places (mainly in the USA).

2.2 Run Descriptions

Based on the results of preliminary experiments, we fixed a number of parameters across all submitted runs: language modeling with Dirichlet smoothing ($\mu = 5000$), $\ell_{split} = 5000$ and $\ell_{lat\_lng} = 0.01$. These settings result in a total of 1786 non-empty cells. The maximum extent in terms of latitude and longitude are 22.5 and 45.0 in areas of the world map where the development data is sparse. Listed below are the details of the submitted runs:

Basic: baseline run without term filtering.

Gen: run with general term filtering applied, $U = 2$.

GeoGen: run with geographic and general term filtering applied, $U = 2$ and $\theta_{geo} = 0.1$.

UserSpecific: run with geographic and general term filtering applied, $U = 2$ and $\theta_{geo} = 0.1$. If the user who uploaded the test video has contributed at least one item to the training data set, only the user's training items are utilized to create the grid cells and language models (similar to [2]).

Visual: run which is based on the provided visual features. The partition of the training data is the same as in the text-based approaches, though for performance reason only 10% of the training data was used. The Naive-Bayes nearest neighbour approach [1] with all visual features was implemented.

3. RESULTS

The results of the listed runs are shown in Table 2. Reported is the accuracy within $\{1, 10, 50, 1000\}$km of the ground truth location.

	1 km	10 km	50 km	1000 km
Basic	20.3%	38.2%	49.2%	66.4%
Gen	21.5%	40.5%	51.2%	67.8%
GeoGen	17.2%	50.8%	70.0%	82.6%
UserSpecific	17.8%	38.0%	52.1%	72.7%
Visual	0.0%	0.1%	0.7%	10.9%

Table 2: Prediction accuracy of the runs for a number of distance cutoffs.

4. DISCUSSION

The biggest improvements over the baseline run are achieved by filtering out terms that have a large geographic spread. The only exception is the 1km cutoff, where **Basic** outperforms **GeoGen**. We hypothesize that once the correct cell C_{max} is identified in the first step of the estimation process, finding the closest match within the training documents of C_{max} may be more robust if all terms of T_{test} are used. Although more than 80% of the test set users also contributed items to the training set (on average 582 items), relying on only the user's contributed items for training did not yield improvements over relying on all available training items.

Our implementation of the visual features based nearest neighbour approach did not result in a usable location estimator. Future work will focus on a failure analysis of this sub-system. Exploiting weather and daylight information to place outdoor images on a map, e.g., [3], will also be investigated. Finally, we plan to research to what extent social network information (such as the home location of the user's contacts, the locations of the images the user comments on, etc.) can improve the text-based location estimation of images that are geographically underspecified.

Acknowledgments

This research has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no ICT 257831 (ImREAL project).

5. REFERENCES

[1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR '08*, pages 1–8, 2008.

[2] J. Choi, A. Janin, and G. Friedland. The 2010 ICSI Video Location Estimation System. In *MediaEval 2010 Workshop*, 2010.

[3] N. Jacobs, K. Miskell, and R. Pless. Webcam geo-localization using aggregate light levels. In *WACV '11*, pages 132–138, 2011.

[4] A. Rae, V. Murdock, P. Serdyukov, and P. Kelm. Working Notes for the Placing Task at MediaEval 2011. In *MediaEval 2011 Workshop*, 2011.

[5] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *SIGIR '09*, pages 484–491, 2009.

[6] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of flickr resources using language models and similarity search. In *ICMR '11*, pages 48:1–48:8, 2011.

[7] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, pages 334–342, 2001.