A MACHINE LEARNING APPROACH TO DETERMINING TAG RELEVANCE IN GEOTAGGED FLICKR IMAGERY

Mark Hughes, Noel E. O'Connor

CLARITY Centre for Sensor Research
Dublin City University
Ireland

Gareth J. F. Jones

Centre for Next Generation Localisation
School of Computing
Dublin City University
Ireland

ABSTRACT

We present a novel machine learning based approach to determining the semantic relevance of community contributed image annotations for the purposes of image retrieval. Current large scale community image retrieval systems typically rely on human annotated tags which are subjectively assigned and may not provide useful or semantically meaningful labels to the images. Homogeneous tags which fail to distinguish between are a common occurrence, which can lead to poor search effectiveness on this data. We described a method to improve text based image retrieval systems by eliminating generic or non relevant image tags. To classify tag relevance, we propose a novel feature set based on statistical information available for each tag within a collection of geotagged images harvested from Flickr. Using this feature set machine learning models are trained to classify the relevance of each tag to its associated image. The goal of this process is to allow for rich and accurate captioning of these images, with the objective of improving the accuracy of text based image retrieval systems. A thorough evaluation is carried out using a human annotated benchmark collection of Flickr tags.

1. INTRODUCTION

In recent years, there has been unprecedented growth in the number of images being stored in online image repositories such as Flickr. The unparalleled scale of these repositories means that efficient methods are required to store and retrieve images based on user queries. The constraints of computer vision technologies means it remains difficult to automatically infer high level semantic meaning from image content alone. The majority of image repositories therefore rely on short human entered annotations (referred to as "tags"), which are created at upload time, and to provide text-based image search for relevant images of interest to a user.

Many of the users who upload images to Flickr are unaware of the potential value of their annotations for organisation and search of the images in the collection. Even those who are, frequently leave this as an activity to be done "later"

Fig. 1. *Example of the problems associated with homogeneous tags in text based image retrieval systems. Pictured are the top three ranked results returned from Flickr (24-Jan-2012) for the query text: 'statue of liberty'.*

and never actually get around to adding high quality detailed annotations. Image tags therefore, generally, lack more than rudimentary detail, are informal and are often not objective. The poor quality of annotation limits the potential effectiveness of image retrieval systems. An example of these problems can be seen in Figure 1. To help overcome this problem and improve the performance of image retrieval systems, a method is desired that can automatically differentiate between semantically relevant and non-relevant tags. We propose a machine learning method to solve this problem.

In this study we focus on classifying tags describing commonly photographed landmarks. Landmarks are selected for our investigation due to the significant proportion of the images in large scale public photo repositories such as Flickr which focus on Landscapes. For example, a search on Flickr "Eiffel Tower" returns over 450,000 images, and a Flickr search for "Empire State" returns over 370,000 images (June 2011)). It should be noted however, that our approach is relatively generic and is not limited to clusters of landmark images, and we anticipate that this method could be used for other datasets such as images containing events etc.

In this paper we propose a novel method to enable the precise classification of Flickr tags associated with landmark images into groups of semantically relevant and non semantically relevant. We use an SVM combined with a novel feature set as our classification tools. The current state of the art in Flickr tag selection approaches is based on the tf-idf metric.

In this paper we demonstrate that our approach outperforms this standard approach by a significant margin for this task. The paper is split into two main sections: the first describes the image corpus used in our evaluation; the second section describes our machine learning approach to tag classification in each of these clusters. The paper concludes with a thorough evaluation of each of these approaches, including an evaluation against the state of the art tf-idf method.

2. PREVIOUS WORK

In previous years, several different approaches to selecting representative tags from clusters of Flickr images have been suggested. The majority of these focus on the tf-idf metric and slight variants of tf-idf.

Kennedy et al. [1] explored different methods to structure Flickr data, and to extract meaningful patterns from this data. Specifically, they were interested in selecting metadata from image collections that might best describe a geographical region. Similar work by Kennedy and Naaman [2], focused on extracting textual descriptions of geographical features, specifically landmarks, from large collections of Flickr metadata. Tags are clustered based on location, and tags are selected using a tf-idf approach, so as to correlate with nearby landmarks. Ahern et al. [3] also employed a tf-idf approach on sets of Flickr tags, in this work it was used to create a visualisation of representative tags overlaid on a geographical map. Their "World Explorer" system enables users to view unstructured textual tags in a geographically structured manner.

Xirong et al. [4] combine visual information with a tf-idf scoring metric to estimate tag relevance within a dataset of Flickr images. For each test image, they carried out a visual search procedure to find its nearest visual neighbours within the dataset. They showed that by calculating co-occurrences of tags within visually similar images, it is possible to estimate relevant tags for a query image over using text based methods alone with a higher probability.

Most of the approaches to date have focused on variations of text-retrieval based models using a tf-idf scoring approach to choose relevant representative tags from a cluster of metadata [5]. In this paper, we improve on this existing work by adopting a machine learning based approach.

3. LANDMARK IMAGE CORPUS

Our tag classification method requires a corpus of geo-tagged images containing landmarks to train the classifier. For this investigation, it was decided to focus on landmark images photographed within a large city. The city of Paris was chosen, mainly because there is a high distribution of landmarks which are commonly photographed.

Our training corpus of geo-tagged images was harvested using the publicly available Flickr API. To return possible landmark images, the Flickr system was queried with a list of generic words that may indicate a landmark is present in an image, such as landmark, church, bridge, building, facade etc. These queries retrieve images based on the presence of these labels among manual image tags.

To filter out non-landmark images from the corpus, an approach based on the use of stop words was adopted. To build a list of stop words, an image set collected from Flickr consisting of 1000 images was manually inspected and classified as containing a large landmark. This set was labelled as S_1. A further set of 1000 images that did not contain a large landmark, but rather depicted an event or different types of objects, people, and animals was also collected and denoted S_2.

For the set S_1 a list of all associated tags was extracted and denoted as T_1. A second list of tags T_2 was created containing all the tags associated with images in S_2. All tags contained in $T_2 \backslash T_1$ were considered possible candidate tags, however the presence of a tag in $T_2 \backslash T_1$ alone is not enough to indicate that the tag would suggest a non-landmark image. It was decided therefore, to selected the tags that occurred the highest number of times in T_2 but not T_1. The final set of stop words was selected based on the tag frequency of each possible candidate tag from $T_2 \backslash T_1$. The frequency was calculated using the following formula:

$$tf_i = \frac{t_i}{|T_2 \backslash T_1|} \tag{1}$$

where t_i is the number of occurrences of the tag i in the list $T_2 \backslash T_1$. If the term frequency was above a threshold of .005 (roughly translating to a frequency of 10), the tag was marked as a candidate tag.

Any image within our corpus containing one of these candidate tags was filtered out. In total, we downloaded just under 200,000 geo-tagged images from Flickr in the Paris region. Over 100,000 were filtered out using this approach, leaving a final training corpus consisting of 90,968 images. From informal empirical inspection this tag filtering approach was observed to be generally effective.

4. TAG RELEVANCE CLASSIFICATION

The approach that we propose to classify tags as semantically relevant is based a novel set of features that we calculate based on statistical information. A training collection of tags was randomly selected from our main corpus that were not contained within the test set of tags used in our evaluation. Each of these tags were manually classified (using criteria described in Section 5) as either having a high(denoted as positive) or low(denoted as negative) semantic value. In total our training collection consisted of 120 positive tags and 157 negative tags. We trained classification models using an SVM based on combinations of our proposed feature set and used these models to classify a manually annotated evaluation set

of tags. Additionally, we trained sets of models using three different SVM kernel functions, linear, polynomial and RBF.

4.1. Features

4.1.1. Geographical Distribution

Combining the geographical and textual metadata of each image within the training corpus has the potential to improve tag relevance classification, since not only does a geo-tag have a semantic relationship with an image, it also has a semantic relationship with the associated textual tags. It is logical to assume that a text tag that re-occurs within a small geographical area is more likely to be describing a geographical feature or object within that area, whereas a generic text tag with low semantic relevance is less likely to relate to a specific region.

To measure the geographical distribution of a text tag, we extract a feature based on geographical variance. A metric calculating the standard deviation was utilised, as shown in equation 2:

$$dev_i = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(x_i - \bar{x})^2} \tag{2}$$

where x_i is the geographical location for an i^{th} instance of a tag and $\bar{x}$ is the mean geographical location of the tag. This standard deviation correlates with distance of each tag from the mean tag location. To calculate our feature we create a histogram H for each tag where the value of H_t is equal to the number of tag occurrences where the geographical variance falls into the distance threshold t (in metres) where $t \in \{0-99, 100-249, 250-499, 500-999, 1000-1999, >2000\}$

4.1.2. Tag Rankings

The Flicker interface prompts users uploading images to add tags which describe the image content We assume that they will enter the tags that they deem most relevant to the image in descending order of significance. This order is preserved within the data, and therefore can be considered as a list ranked by the importance of the tag.

To create a metric to measure the rankings, we use a histogram based approach. A histogram H is calculated for each tag where the value of H_r is equal to the number of tag occurrences within the corpus where the tag has a ranking position of r and $r \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, >9\}$. The hypothesis behind this metric is that a tag with a high semantic relevance is more likely to have peaks in the lower entries of the histogram (i.e. users will enter them first), whereas a more generic tag will have peaks in the higher entries or perhaps a more balanced distribution.

4.1.3. Inverse Document Frequency

One important measurement in determining the relevance of a tag is its level of 'uniqueness' or 'specificity' across the entire corpus. The inverse document frequency (idf) is a metric that calculates the frequency of a tag across the entire image corpus. This method assigns a higher score to tags that have a lower frequency across the entire corpus. Additionally, it will assign a low score to any tag that occurs regularly across the corpus.

To calculate the idf metric we use the following formula:

$$idf_t = \log \frac{N}{df_i} \tag{3}$$

where df_i is the document frequency of the tag i and N is the total number of images within the corpus.

5. TAG SELECTION EVALUATION

To evaluate the proposed tag classification method, we created a benchmark set of tags extracted from our corpus, consisting of a total of 3444 tags. Each tag was manually analysed and deemed semantically relevant or irrelevant to it's associated image using the following protocol:

- **Relevant**: A classification of relevant is given to a tag that contains a high level semantic description. It must contain the name of the main landmark or surrounding geographical area (localised, not on a city wide scale), such as 'Notre Dame Cathedral' or 'Place de la Concorde'. A tag was also deemed relevant if it contains a mid-level semantic description of the content within an image. For example, if a tag describes the type of landmark or location depicted, it is deemed relevant. Some examples are: 'Cathedral', 'Facade' or 'Fountain'.

- **Non-Relevant**: A classification of non relevant is given to a tag that contains temporal information or a low-level semantic description of an image. Examples of a low-level semantic tag might be 'outdoor', 'sky', 'night', 'river' or 'park'. Tags that contain vague geographical descriptions such as 'Europe','city' or 'continent' provide little discrimination value and were deemed irrelevant along with heterogenous tags such as 'vacation', 'honeymoon' and 'trip'.

We then processed each of these benchmark tags through our SVMs and the outputs from this were evaluated against the human defined benchmark. The results of this evaluation can be seen in Tables 1 and 2. Additionally we implemented the current state of the art, tf-idf process and classified all of the benchmark tags and compared our SVM method against it, the results of which can be seen in Figure 2. To calculate the tf-idf metric, we randomly selected 100 images from the corpus and retrieved visually near identical images from the corpus based on the matching of SURF image features [6], to form 100 clusters of semantically related images. The tf score for each tag was extracted from each of these clusters whereas the idf score was calculated from the entire corpus.

All Features	Precision	Recall	F-Score
GeoVariance	.54	.65	.58
Tag Ranking	.07	.96	.13
Ranking + Geo	.82	.13	.22
All Features	**.64**	**.60**	**.61**

Table 1. *Precision results of the different evaluated combinations of features using the RBF SVM kernel function.*

SVM Kernel	Linear	Polynomial	RBF
Precision	.655	.777	.640
Recall	.536	.415	.606
F-Score	.58	.53	**.61**

Table 2. *Precision of a combination of all features using three different kernels with the SVM.*

As can be seen from the results, our best performing approach is that using all three proposed features. It would seem from Table 1 that the tag ranking approach is not as precise as the geographical variance feature. It can be seen in Figure 2 that our best performing approach, achieves a precision increase of over .21 above the current optimal approach to this task, the tf-idf metric.

Our hypothesis is that the tf-idf metric performs poorly on this dataset due to the high distribution of landmarks. A commonly photographed landmark such as 'The Eiffel Tower' will have a high distribution within the dataset, and therefore will have a low idf score. This will bias the metric against commonly occurring, but semantically relevant tags. tf-idf is designed to reward words with high selectivity of relevant documents in information retrieval, this is less appropriate for our task, since we are looking for tags which are strongly correlated with the image.

6. CONCLUSIONS

In this paper we proposed a novel approach to classifying the semantic relevance of a text tag associated with a Flickr image. We proposed a machine learning methodology and described a detailed evaluation which demonstrates that this methodology outperforms the current state of the art solution to this problem (tf-idf).

It should be noted that the implementation of this approach as described in this paper, does not take multilingual text into account. Thus manually contributed image labels on language other than English are ignored. Performance could thus be improved by utilising machine translation of non-English languages to include this information in the dataset. Additionally, synonyms of equivalent labels are treated as separate labels. We aim to continue work to exploit resources such as

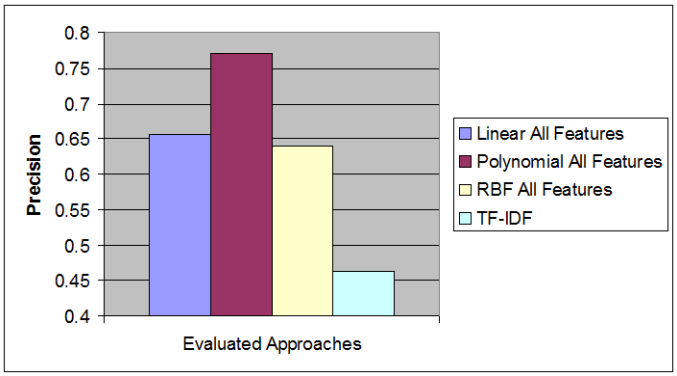

Fig. 2. *Diagram showing the precision of each evaluation approach compared against tf-idf.*

Wordnet [1] to help address this issue by combinining equivalent labels to improve the input features for the SVM.

7. REFERENCES

[1] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How Flickr helps us make sense of the world: context and content in community-contributed media collections," in *Proceedings of MULTIMEDIA 2007*. 2007, pp. 631–640, ACM.

[2] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proceeding of the 17th international conference on World Wide Web (WWW 2008)*, 2008, pp. 297–306.

[3] S. Ahern, M. Naaman, R. Nair, and J. Yang, "World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections," in *In Proceedings of the Seventh ACM/IEEE-CS Joint Conference on Digital Libraries*. 2007, pp. 1–10, ACM.

[4] L. Xirong, C. G. M. Snoek, and M. Worring, "Annotating images by harnessing worldwide user-tagged photos," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, ICASSP '09, pp. 3717–3720.

[5] A. Mahapatra, X. Wan, Y. Tian, and Jaideep J. Srivastava, "Augmenting image processing with social tag mining for landmark recognition," in *Proceedings of the 17th international conference on Advances in multimedia modeling*.

[6] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *9th European Conference on Computer Vision*, pp. 404–417, 2006.

[1] Wordnet:http://wordnet.princeton.edu/