On Contextual Photo Tag Recommendation

Philip J. McParlane*
School of Computing Science
University of Glasgow
Glasgow, UK
p.mcparlane.1@research.gla.ac.uk

Yashar Moshfeghi
School of Computing Science
University of Glasgow
Glasgow, UK
Yashar.Moshfeghi@glasgow.ac.uk

Joemon M. Jose
School of Computing Science
University of Glasgow
Glasgow, UK
Joemon.Jose@glasgow.ac.uk

ABSTRACT

Image tagging is a growing application on social media websites, however, the performance of many auto-tagging methods are often poor. Recent work has exploited an image's context (e.g. time and location) in the tag recommendation process, where tags which co-occur highly within a given time interval or geographical area are promoted. These models, however, fail to address *how* and *when* different image contexts can be *combined*. In this paper, we propose a weighted tag recommendation model, building on an existing state-of-the-art, which varies the importance of time and location in the recommendation process, based on a given set of input tags. By retrieving more temporally and geographically relevant tags, we achieve statistically significant improvements to recommendation accuracy when testing on 519k images collected from Flickr. The result of this paper is an important step towards more effective image annotation and retrieval systems.

Categories and Subject Descriptors: H.3.1 Information Storage and Retrieval - *Content Analysis and Indexing*; I.2.10 AI - *Vision and Scene Understanding*
General Terms: Performance, Experimentation
Keywords: Photo Tag Recommendation, Temporal, Geolocation

1. INTRODUCTION

With the amount of multimedia data rapidly increasing, it becomes important to organize this content effectively. Photographs uploaded to image sharing websites such as Flickr[1] often contain few or no tags, making effective retrieval difficult. Photo tag recommendation has offered one such solution where new, additional tags are offered based on those already assigned to an image. Much research has been taken out in this area where tag recommendation models have exploited the co-occurrence of tags [7], user tendencies [2] and the image context [8, 6] in the recommendation process. To be able to facilitate efficient multimedia retrieval, in this paper we propose to annotate these images with keywords[2] by exploiting its context.

The time and location an image is taken in, has been seen to be a reliable source of evidence for tag recommendation, achieving significant improvements over a baseline which ignores the image context [8, 6]. These works, however, fail to *combine* the time and location contexts of images, as well as capture the varying levels of association different keywords have with different temporal windows (e.g. the time of day) and geographical areas (e.g. continents). For example, Figure 1 demonstrates this by showing four tags, `sunrise`, `sunset`, `autumn` and `leaves`. As can be seen, `sunrise/sunset` have strong, recurring *hourly* trends, whereas on the contrary they have noisy, *monthly* temporal distributions. The opposite effect is observed on the tags `autumn` and `leaves`. Therefore, in this paper, we consider a tag's *association* with the different image contexts, and their *combination*. For time intervals we consider the *time of day, the season and the day of the week*, and for geographical areas we consider the *continent*[3] an image is taken in, in our tag recommendation process.

The rest of this paper is organised as follows. In Section 2, we present an overview of work in image tag recommendation. Section 3 describes how we exploit an image's context in the tag recommendation process. Section 4 details our evaluation procedure, the results of which are detailed in Section 5. Finally, we conclude in Section 6.

2. RELATED WORK

The automatic process of annotating images with tags takes two forms: *automatic image annotation*, which looks to identify tags based solely on the image contents, and *tag recommendation* which takes the tags already present in an image's tag list as a query, in order to offer new tag suggestions to the user. *Automatic image annotation* has been a widely researched area over the last decade with a large number of works attempting to bridge the *semantic gap* between low level image features and high level concepts [1, 3, 4, 5]. Although these works have made significant progress in achieving this end goal, the semantic gap still largely exists, and annotations are often very unreliable. As a result,

*This research was supported by the the European Community's FP7 Programme under grant agreements nr 288024 (LiMoSINe)

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

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

[2] In this paper we refer to tags and keywords synonymously

[3] For the remainder of this paper we refer to these as *dimensions*

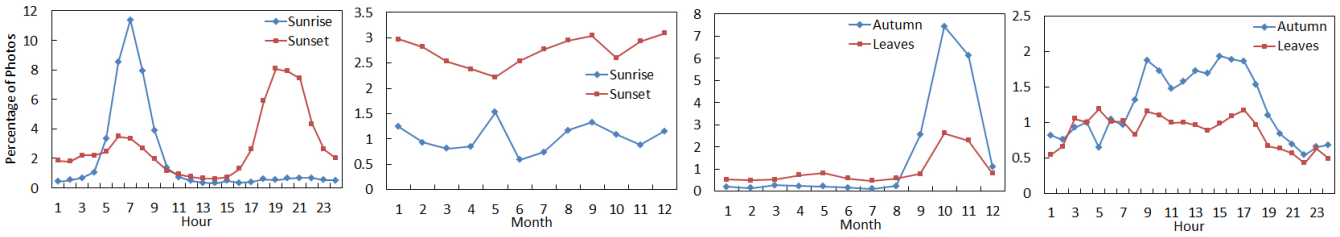

Figure 1: Differing Levels of Temporal Association for four tags: `sunrise`, `sunset`, `autumn` and `leaves`.

social image tagging websites, such as Flickr, depend solely on user tagging to allow for image retrieval.

For *tag recommendation*, a number of works have been proposed in recent years which attempt to offer new tags to the user, based on the already existing tags in the collection. For example, Sigurbjornsson *et al.* proposed a tag recommendation strategy to support users annotating photos on Flickr [7]. The relationships between tags were exploited to suggest highly co-occurring tags. Garg *et al.* offered personalised tag recommendations [2] in their *Hybrid* approach which combined suggestions made from personalised and global tag co-occurrence matrices. These works, however, ignore the time and place an image is taken in the recommendation process.

Zhang *et al.* looked to go beyond tag co-occurrence by clustering tags based on geolocation and temporal trends [8]. McParlane *et al.* exploited the daily, monthly and yearly trends of tags are exploited for the task of image tag recommendation [6]. These works which consider image context, however, fail to address the way in which these time and location dimensions differ from tag to tag, and their combination.

In this paper we consider the exploitation of time and location in the recommendation process by offering a weighted model which considers a tag relationship with each of the temporal and geographical dimensions.

3. CONTEXTUAL RECOMMENDATION

In the following sections, we introduce and formulate the problem of image tag recommendation and detail our weighted model which attempts to improve recommendation accuracy by considering an image's context.

Problem Statement Let m denote an image in our collection, containing a number of tags, d, assigned by the user. The overall goal in tag recommendation is therefore to recommend a set of tags p, given a subset of tags, q, from d ($q \subset d$), so that is maximizes $p \cap (d - q)$.

In order to make tag recommendations, our state-of-the-art (SOTA) model, as defined in Section 3.2, calls upon a tag co-occurrence matrix. In the following section we introduce a number of co-occurrence matrices for this purpose.

3.1 Contextual Tag Co-occurrence Space

If we assume that in total k unique tags represent the images in a collection of size n, the tag co-occurrence matrix would be a square matrix C_k where the value of the element $c_{t_{ij}}$ represents the number of images that contain both t_i and t_j tags. We define the representation of a tag t_i as a vector $\underline{t'_j} = (c_{t_{i1}}, c_{t_{i2}}, \ldots, c_{t_{ik}})$ where each dimension corresponds $\underline{t_i}$'s co-occurrence value with another tag.

In this work, we compute a number of temporal and geographical matrices. These capture the time and location an image is taken in the co-occurrence measures (i.e. $c_{t_{ij}}$), by computing the number of images two tags coexist in for a given time interval, or geographical location. Firstly, let us introduce a number of image sets used to create the tag co-occurrence matrices.

Sets Let S^x denote a set of images in our training set, where x can be either *Global* (i.e. a set containing all images in our training set) or *Context*, where the context can be one of the following values:

- *Time(y)*: a subset of images taken within a given timespan. y takes four different values: morning (06:00 to 11:59), afternoon (12:00 to 17:59), evening (18:00 to 23:59) and night (00:00 to 05:59).
- *Day(y)*: a subset of images taken on a particular day of the week. y takes two different values: weekday and weekend.
- *Season(y)*: a subset of images taken in a particular season. y takes four different values: winter, spring, summer and autumn.
- *Cont(y)*: a subset of images taken in a particular continent. y takes seven different values: Africa, Antarctica, Asia, Europe, North America, Oceania and South America.

For example, $S^{Cont(Africa)}$ denotes the subset of images taken in Africa. In our approach, we consider the *time* an image is taken from its *exchangeable image file (exif) data*, and the *continent* from its *GPS location*.

Matrices The definition of these sets, allows us to construct different co-occurrence matrices C^x, where x, takes the same values as defined above. Each co-occurrence matrix is built using the images corresponding to its set (S^x). For example, $c_{t_{ij}}^{Cont(Africa)}$ is the number of images, taken in Africa, in which tags t_i and t_j coexist.

In our baseline approach, our tag recommendation model uses the *Global* co-occurrence matrix (C^{Global}), whereas in our experimental approaches, we take co-occurrence values from our *contextual* matrices ($C^{Time(y)}$, $C^{Day(y)}$, $C^{Season(y)}$ and $C^{Cont(y)}$), thus offering temporally and geographically significant suggestions.

3.2 Tag Recommender (TR) Model

Given a co-occurrence matrix C^x, a number of existing algorithms can be used to generate tag recommendations. We choose to adopt a tf-idf approach proposed in Algorithm 2 in [2] due to its *simplicity* in implementation and *performance* in comparison to other tag recommendation baselines. It begins by computing a new matrix $\hat{C^x}$ from C^x in two stages. Firstly, all diagonal values of C^x are set to zero. Secondly, each column of this new matrix is scaled, so that the maximum value in each column is 1. The output from this model (a vector of scores where each element refers to a tag), denoted as O_q, is then computed as: $O_q = (\hat{C^x} \times \underline{q}) \cdot \underline{\text{idf}}$. We define $\underline{\text{idf}}$ to be the vector of *inverse document frequencies*, where each element computes the *idf* score, $\log(n/n^{(t_j)})$,

for each tag in the collection, where $n^{(t_j)}$ is the number of images containing tag t_j. $\underline{q}$ is the binary vector of tags used as a query from the image's tag list. The "$\cdot$" is the component-wise product of the output vector from $\hat{C^x} \times \underline{q}$ and $\underline{\text{idf}}$. For multiple tags, the corresponding contributions are added. In our approach, it is the C^x matrix which is changed (and combined) between the various co-occurrence matrices defined in Section 3.1.

3.3 Multi-Context TR Model

The main novelty of this paper considers the combination of co-occurrence measures based on the association of tags, with each of the different time and location dimensions. Therefore two combination approaches are introduced:

Option (1) As our first approach we take the average co-occurrence score, from the four *contextual* co-occurrence matrices (i.e. time, day, season and continent). This model is therefore a *non-weighted* approach.

Option (2) As a more elaborated approach we consider the relationship between the input tags and the different dimensions using a weighted combination approach. Therefore, the overall weighting for a tag, for each *contextual* co-occurrence matrix, can be expressed as follows: $\lambda P(C^x|t_j, m)+(1-\lambda)P(C^x)$. In the computation of co-occurrence vectors for a tag t_j, we consider (a) the association $P(C^x|t_j, m)$ of the given input tag, for image m, to each matrix C^x and (b) the global perceived effectiveness $P(C^x)$ of each co-occurrence matrix.

$P(C^x|t_j, m)$ is computed as the likelihood of the tag occurring within a given set of images, normalised by the sum of probabilities for the tag existing in *all* of the contextual sets i.e. $P(C^x|t_j, m) = P(t_j|S^x)/\sum_{x \in Context} P(t_j|S^x)$, where $P(t_j|S^x)$ is the fraction of images containing t_j in S^x.

$P(C^x)$ is a weighting, computed on a validation set, for each type of *contextual* co-occurrence matrix. Each weight, is computed as the proportional improvement (using *precision at five* as a measure) of the given matrix over the baseline (C^{Global}). These weights are normalised and summed to 1. In our experiments these weights were computed as 0.32, 0.20, 0.24, 0.24 for $C^{Cont(y)}$, $C^{Day(y)}$, $C^{Season(y)}$ and $C^{Time(y)}$, respectively.

In our experiments, in order to select λ we tested on our validation set, optimising for precision at five. We varied λ between $[0, 1]$ with a step of 0.1. Best results were achieved when $\lambda = 0.2$, showing that weighting the most effective co-occurrence dimensions higher is most important.

4. EXPERIMENTAL SETUP

In this section, we describe the experimental setup that supports the evaluation of our proposed framework. In particular, our experiments aim to answer two main research questions: (i) can the association between different tags and time and location dimensions be exploited to better model the use of time and place in photo tag recommendation? (ii) can image meta-data, such as time and location, be *combined* to improve tag recommendation accuracy?

In the following, we detail the training and test collections, the parameter tuning and the metrics used in our evaluation.

Data Collection For our experiments we collect an image dataset from Flickr[4] consisting of 2M images. The dataset was collected by querying Flickr for 2000 popular nouns extracted from WordNet (categorised as `animal, artifact, body, food, plant, substance`). The top 2000 images (containing GPS co-ordinates) from the results page, for each search were then considered for use in our collection.

Pre-processing A number of pre-processing stages were then taken out on the dataset to make it useful for tag recommendation. Due to the large amount of noise present in online image collections [7], we first had to remove a number of tags deemed irrelevant for tag recommendation. We therefore manually removed (using three assessors) those tags which fell into the following categories: camera meta data (e.g. `d60`), Flickr awards (e.g. `excellentphotographeraward`) and Flickr groups (e.g. `5photosaday`), from the top 1000 most frequently occurring keywords. Additionally, those tags which were used by less than 20 users were also removed. This approach has been used by previous work [6].

Training and Test Set After this pre-processing stage, there existed 517k images in the *training set*, uploaded by 77k users, using 20k tags. Each image contained on average 15.4 tags and over 99% of the images were taken between January 1999 and October 2012. The tags within these images were used to build the various co-occurrence matrices as described in Section 3.1.

We created two further collections (external from the training set), the first of which was used for *parameter tuning*, and the other for *performance testing*. These collections were collected using the same procedure as described in the previous section and comprised of 1000 images each. In each of these, the user tags were used as the *ground truth*.

Evaluation Metrics To evaluate our methods, we use four standard metrics, all of which are commonly used and are adopted by previous work in image tag recommendation [2]. We evaluate performance for a recommended tag list by comparing those suggested tags, with those already provided by the user. The metrics are as follows: (i) *Precision at One (P@1)*: The percentage of runs where the top tag is relevant (equal to S@1). (ii) *Precision at Five (P@5)*: The percentage of relevant tags amongst the top five, averaged over all runs. (iii) *Success at Five (S@5)*: The percentage of runs, where there exists at least one relevant tag amongst the top five returned. (iv) *Mean Reciprocal Rank (MRR)*: Computed as $1/r$ where r is the rank of the first relevant tag returned, averaged over all runs.

Systems We denote R^x a system which takes co occurrence measures from a *single* matrix, where x denotes the different types of co-occurrence matrices defined in Section 3.1. For example, $R^{Cont(y)}$, is a system where co-occurrence values are taken from $C^{Cont(y)}$, where y, is the continent, the given image is taken in. R^{Global} is our baseline. We introduce two further systems, $R^{Option(1)}$ and $R^{Option(2)}$, which take co-occurrence values as a combination of the four *contextual* matrices. These approaches compute these values using the two combinations options described in Section 3.3 e.g. $R^{Option(1)}$, is the system where co-occurrence values are combined using the non-weighted method i.e. *Option 1*.

Evaluation Procedure We selected N *random* tags (with $N = [1, 2, 3]$) from an image's ground truth which were used to query the recommendation model. The top five tags were retrieved and used as recommendations. This evaluation procedure has been used by previous work [2]. Finally, we compute paired t-test statistical significance tests comparing our the experimental approaches against our baseline (R^{Global}).

[4]Available for download at `http://dcs.gla.ac.uk/~philip/`

Table 1: Overall Results for N = {1, 3} ($N = 2$ is not shown due to space limitations). The best performing models for each measure are displayed in bold. The statistical significance results against the baseline (R^{Global}) are denoted as * being $p < 0.05$ and ** being $p < 0.001$.

		Static	*Temporal*			*Geographical*	*Combined*	
		$\mathbf{R}^{Global}$	$\mathbf{R}^{Time}$	$\mathbf{R}^{Day}$	$\mathbf{R}^{Season}$	$\mathbf{R}^{Cont}$	$\mathbf{R}^{Option(1)}$	$\mathbf{R}^{Option(2)}$
1 Input	**S@5**	0.566	0.559** (-1.2%)	0.564 (-0.3%)	0.568 (+0.3%)	0.596** (+5.3%)	0.613** (+8.3%)	**0.622** (+9.8%)**
	P@1	0.247	0.244* (-1.2%)	0.247 (+0.0%)	0.264** (+6.8%)	0.276** (+11.7%)	**0.291** (+17.8%)**	0.288** (+16.5%)
	P@5	0.240	0.254** (+5.8%)	0.251** (+4.5%)	0.268** (+11.6%)	0.280** (+16.6%)	0.286** (+19.1%)	**0.290** (+20.8%)**
	MRR	0.358	0.357 (-0.2%)	0.359 (+0.2%)	0.372* (+3.9%)	0.389** (+8.6%)	0.403** (+12.5%)	**0.404** (+12.8%)**
3 Input	**S@5**	0.722	0.729 (+1%)	0.743 (+2.9%)	0.734 (+1.6%)	0.743* (+2.9%)	**0.799** (+10.7%)**	0.793** (+9.8%)
	P@1	0.337	0.367* (+8.9%)	0.361 (+7.1%)	0.376** (+11.6%)	0.378** (+12.2%)	**0.400** (+18.7%)**	**0.400** (+18.7%)**
	P@5	0.342	0.362** (+5.8%)	0.362** (+5.8%)	0.369** (+7.9%)	0.382** (+11.7%)	0.414** (+21.1%)	**0.415** (+21.3%)**
	MRR	0.480	0.500* (+4.2%)	0.499* (+4%)	0.507* (+5.6%)	0.515** (+7.3%)	**0.549** (+14.4%)**	**0.549** (+14.4%)**

5. RESULTS

As can be seen from Table 1, exploiting time and location in the photo tag recommendation process can have dramatic increase in effectiveness, where up to 21% (for P@5) statistically significant improvements are achieved over our baseline. An image's location gives larger increases to recommendation accuracy when compared to the temporal dimensions. This is possibly due to the large number of location type tags used by users on Flickr [7]; this is an interesting feature which needs further investigation. For the temporal dimensions, the season is the most effective dimension, achieving larger increases to accuracy than the day and time based approaches. Encouragingly, all of the temporal and location based dimensions increase recommendation accuracy, across different metrics and N (except for R^{Time} & R^{Day} where $N = 1$).

The findings of our results show that a combination of temporal and geographical dimensions are complimentary in the photo tag recommendation process. Of our two combined approaches, using the weighted model ($R^{Option(2)}$) gives better results than simply averaging the co-occurrence measures. This therefore supports our hypothesis that by combining different evidences, given the associations between the input tags and each of the contextual evidences, recommendation accuracy improves.

Finally, recommendation accuracy increases as the number of inputs increases (the same trend exists for $N = 2$). Further, our combined approaches, are able to maintain the same level of performance over our baseline as N increases.

To further investigate the effectiveness of our approach, we computed the top tags for each of the contextual approaches by computing $P(t|S^x) - P(t|S^{Global})$ (where $x \neq Global$), for each dimension. As can be seen in Table 2, there exists a strong temporal and geographical link between tags and each subset; the tags are highly relevant each of the given temporal and geographical dimensions.

Table 2: The top tags for each dimension

		Top 5 (adjusted) Tags
Time	**Morning**	morning, nature, sunrise, bird, birds
	Afternoon	uk, england, autumn, green, museum
	Evening	night, sunset, light, music, concert
	Night	film, night, berlin, smile, yoko
Season	**Spring**	spring, may, april, 2012, march
	Summer	summer, august, july, germany, june
	Autumn	autumn, fall, september, october, november
	Winter	winter, snow, christmas, cold, ice
Day	**Weekend**	2010, race, racing, car, festival
	Weekday	night architecture, travel, light, art
Continents	**Africa**	africa, southafrica, egypt, kenya, morocco
	Antarctica	ice, antarctica, mountains, snow, wildlife
	Asia	asia, japan, india, china, travel
	Europe	uk, england, europe, london, germany
	N America	usa, california, canada, newyork, nyc
	Oceania	australia, victoria, nsw, sydney, newzealand
	S America	brasil, brazil, southamerica, argentina, rio

6. CONCLUSION AND FUTURE WORK

In this paper we exploited the effect of the contextual aspects of an image in the tag photo recommendation process. In particular, we constructed several *contextual* co-occurrence matrices built only on images taken within three time intervals (i.e. the time of day, the day of the week, the season) and one geographical area (i.e. continent). By offering temporally and geographically significant tag recommendations, statistically significant improvements of up to 21% (for P@5), in the best case, were achieved when testing on a Flickr image collection.

Our results demonstrated that, firstly, the exploitation of time and place can improve tag recommendation accuracy. Second of all, combing these evidences can further improve the accuracy of the model if the association between tags and each of the dimensions is taken into account.

This work opens up a number of interesting questions, such as why is the continent an image is taken in considered a more reliable source of evidence than the time it is taken? Future work will look to answer this question as well as the exploration of a number of new image dimensions and more elaborate methods of contextual combination.

7. REFERENCES

[1] P. Duygulu et al. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. ECCV, 97-112, 2002.

[2] N. Garg et al. Personalized, Interactive Tag Recommendation for Flickr. RecSys, 67-74, 2008.

[3] J. Jeon et al. Automatic Image Annotation and Retrieval using Cross-media Relevance Models. SIGIR, 119-126, 2003.

[4] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual Relevance Models. SIGIR, 175-182, 2002.

[5] A. Makadia, V. Pavlovic, and S. Kumar. Baselines for Image Annotation. Int. J. Comput. Vision, 88-105, 2010.

[6] Philip J. McParlane, and Joemon M. Jose Exploiting Time in Automatic Image Tagging. ECIR, 2013.

[7] B. Sigurbjörnsson et al. Flickr Tag Recommendation Based on Collective Knowledge. WWW, 327-336, 2008.

[8] H. Zhang et al. Beyond Co-occurrence: Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities. WSDM, 33-42, 2012.