Exploiting Twitter and Wikipedia for the Annotation of Event Images

BLIND FOR REVIEW

ABSTRACT

With the rise in popularity of smart phones, there has been a recent increase in the number of images taken at large social (e.g. festivals) and world (e.g. natural disasters) events which are uploaded to image sharing websites such as Flickr. As with all online images, they are often poorly annotated, resulting in a difficult retrieval scenario. To overcome this problem, many photo tag recommendation methods have been introduced, however, these methods all rely on *historical* Flickr data which is often problematic for a number of reasons, including *the time lag problem* (i.e. in our collection, users upload images on average 50 days after taking them, meaning "training data" is often out of date). In this paper, we develop an image annotation model which exploits textual content from related Twitter and Wikipedia data which aims to overcome the discussed problems. The results of our experiments show and highlight the merits of exploiting social media data for annotating event images, where we are able to achieve recommendation accuracy comparable with a state-of-the-art model.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

Keywords: Twitter, Wikipedia, Tag recommendation

1. INTRODUCTION

Today, taking and sharing images is a far easier and accessible process than it once was; with the advancement of smart phone camera technology, photographers no longer need expensive equipment. This change has opened up photography to a much wider audience, increasing the amount of visual content uploaded to image sharing websites such as Flickr[1]. Of this content, an ever increasing number of users are uploading photographs taken at large social (e.g. London 2012 Olympics) and world (e.g. Philippines Typhoon) events, where the user acts the role of the amateur photo journalist. As a result, photo sharing websites such as YouthMedia[2] have been setup targeted at these photo journalists.

Organising these images is difficult, however, as a result of the semantic gap [5] and lack of annotations provided by users [11]. An entire field of work has focused on the automatic annotation of images [5, 3]. Despite the progress made in the last two decades, fully automatic methods still perform lower than what is required for industry and therefore real life applications have instead adopted semi-automatic tag recommendation approaches, allowing users to annotate their images from a list of suggested tags. Aside from Flickr's recommendation approach, there have been many photo tag recommendation methods proposed in recent years [11, 2]. These approaches all recommend tags based on *historical* Flickr data which introduces a number of problems, however, as discussed in the following section. We conjecture that many of these problems can be alleviated by exploiting social media and encyclopaedic data.

By recommending on historical data, if a new social/world event occurs, the recommendation model will be slow in considering the new evidence, due to the delay between users *taking* and *uploading* photographs, hence reducing the training set size and quality of suggested tags. Figure 1 highlights this "time lag" problem for our collection covering the Austin City Limits (ACL) 2012 music festival. This difference is clearly observed where images are uploaded, on average, around 50 days later than they are taken. Not only are images uploaded much later, but the volume of images is many magnitude smaller than the volume of tweets posted for a given event, as shown in Figure 2. Twitter is both faster and offers wider and denser coverage (more users) for a given event than Flickr. Further, other issues exist with relying on Flickr data for tag recommendation purposes; batch tagging functionality (where users can tag multiple images with a single tag set) allows a single user to have overriding influence over the content of a tag co-occurrence matrix and ultimately the tag recommendations computed from this. Additionally, as users tend to annotate Flickr images with a small number of popular tags [11], there is a lower topic

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ACM SIGIR 2014
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$10.00.

[1] http://www.flickr.com
[2] http://www.youthmedia.eu

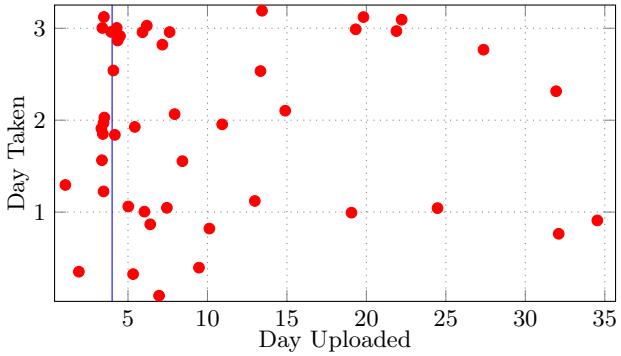

Figure 1: Comparison of an image's taken vs upload time. Vertical line indicates the *end* of the festival.

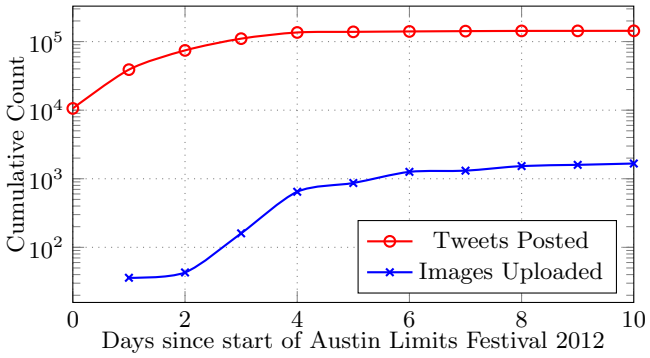

Figure 2: Volume of ACL Tweets vs Flickr Images

coverage on Flickr in comparison to the vast amount of content and coverage of tweets and users on Twitter. In this paper, we propose to use social media streams to annotate Flickr images corresponding to events.

However, Flickr and Twitter both contain noisy user data; in comparison, the collaborative crowdsourced approach of Wikipedia offers a structured data source containing less irrelevant content (i.e. noise), whilst maintaining fast update speeds [7]. Due to the curated nature of Wikipedia, in this paper we propose to use its content as a more reliable source of information in order to "counter" the noisy nature of Flickr/Twitter for photo tag recommendation purposes.

This paper attempts to address the following questions:

1. *RQ1:* Can noisy social media streams, such as Twitter, be exploited in order to annotate images online? How can we effectively address these noise issues?
2. *RQ2:* Can Wikipedia content also be exploited in order to offer reliable photo tag recommendations? Does a combination of social media and Wikipedia sources improve photo tag recommendation accuracy further?

The rest of this paper is as follows: In Section 2 we discuss existing image annotation and photo tag recommendation techniques. Section 3 details our recommendation methodology whilst we discuss our collection and evaluation procedure in Section 4. Section 5 details the findings of our results before concluding 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 *photo tag recommendation* which considers the tags already assigned to an image in order to offer new suggestions.

Automatic Image Annotation (AIA): 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 [4, 5]. Lavrenko *et al.* [4] introduced the Cross-Media Relevance Models (CMRM) to predict the probability of generating a word given visual "blobs". Makadia *et al.* [5] showed that many existing "state-of-the-arts" could be outperformed by adopting a K-nearest neighbour approach trained on global colour and texture features.

As these works consider only visual appearance, their performance is often unsatisfactory due to the presence of the semantic gap [5]. In this paper, we therefore focus on the semi-automatic process of tag recommendation.

Photo Tag Recommendation (PTR): Sigurbjornsson *et al.* [11] proposed a tag co-occurrence recommendation strategy to support users annotating photos on Flickr. Garg *et al.* [2] offered personalised tag recommendations by combining suggestions made from personalised and global tag co-occurrence matrices. Rae *et al.* [10] exploited a user's social context on Flickr in the recommendation process by combining different contexts, such as a user's tagging history, their social circles and groups they are members of.

These works, however, make suggestions based only on the tags available in historical images, often resulting in outdated and unsatisfactory tag recommendations. In our work, we address this problem by exploiting instantaneous text streams from Twitter and Wikipedia for PTR.

Social Media for Annotation: Recently a number of works have also considered the vast amount of social media content for multimedia purposes. Picault *et al.* [9] presented a framework for the indexing and retrieval of video segments by employing text mining and topic modelling techniques in order to collate related tweets. Shamma *et al.* [1] exploited microblog posts for the segmentation and summarization of broadcast media events e.g. 2008 presidential debate.

Despite the number of works exploiting social media for multimedia applications, the potential of social media data for photo tag recommendation purposes has not yet been explored. In this work, we propose a photo tag recommendation model which draws evidences from both social media streams and Wikipedia, presenting preliminary results for this application.

3. METHODOLOGY

In the following section we first formulate the problem of tag recommendation, before describing how we exploit Twitter and Wikipedia data for the purposes of photo tag recommendation.

Problem Statement: Let m denote an image in our collection, containing a set of tags, d, assigned by the user. The overall goals 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 it maximizes $P \cap (d - q)$.

Annotating Event Images: Photographs taken at social and world events present an interesting challenge for annotation models as there exists much evidence from many disparate sources (i.e. Tweets, Flickr images and Wikipedia articles). Given the amount, varying quality and types of data present, there are many challenges regarding its exploitation for PTR purposes. The following sections detail the challenges and exploitation of each data source.

Twitter Data: Using Twitter data presents a number of challenges for tag recommendation; the largest problem being that of noise where tweets are short, contain misspelt words, colloquial expressions and often irrelevant information. In order to overcome the problem of irrelevant data, we consider only those tweets containing predefined hashtags[3] which refer to the event in question. We address the identification of hashtags for an event manually as this is not the purpose of this paper; however, in a real world scenario, we would rely on an event detection model [6].

Using this approach we are able to address noise from a tweet *topic* relevance perspective, but not from a tweet *content* perspective. In order to remove irrelevant terms from tweets, we use the popular Stanford Parser [6] to conduct part-of-speech (POS) tagging on each tweet. Table 1 summarises the term types for each data type. As can be observed, Flickr images are mostly annotated with nouns and entities; therefore a successful recommendation strategy should also suggest mostly nouns. In our approach, we therefore suggest only nouns/entities, thus ignoring many irrelevant terms present in tweets (e.g. punctuation, stopwords, foreign words *etc*) which are not useful for PTR purposes.

Wikipedia Data: Our approach assumes that we are able to identify the relevant Wikipedia article for the event in question. We achieve this process automatically by exploiting a database of Wikipedia URL redirects (containing over 5M url {extension, article} pairs); specifically, we match the event hashtags against this database. As before, we classify each term within the Wikipedia article using the described POS tagger.

	Noun	*Entity*	*Verb*	*Adjective*
Flickr	**0.566**	**0.204**	0.0677	**0.091**
Twitter	0.135	0.098	**0.109**	0.055
Wikipedia	0.198	0.143	0.083	0.006

Table 1: Fraction of Term Types per collection. Highest fraction per type are shown in bold.

4. EXPERIMENTATION

In the following section we first detail our collection before discussing our systems and evaluation procedure.

Evaluation Collection: we collect tweets, Flickr images and Wikipedia content related to the ACL 2012 music festival. We selected the ACL music festival for experimentation purposes as (i) there exists much related Flickr, Twitter and Wikipedia content (ii) there exist many sub-events within this overall event e.g. bands playing *etc* (iii) the event contains temporally and geographical diverse content. The collection is as follows:

1. *Images:* we searched Flickr using the standard search API[4] for images annotated with one of the discussed event tags[3] taken between 11-15 Oct 2012. In total we collect 2,750 images taken by 68 users, annotated with 732 different tags, with each image containing on average 10.6 tags.
2. *Tweets:* we used a subset of a well known public Twitter event collection [6], selecting tweets containing one of the predefined hashtags[3] which are posted between 11-15 Oct 2012. In total we collect 1,507 tweets, containing around 14,570 different terms, posted by 1,309 users.
3. *Wikipedia:* Finally, as previously discussed, we also consider the Wikipedia page. From this document, which describes the history of the festival and not the 2012 festival in isolation, we extract 949 different terms.

Systems: we compare recommendations computed from Twitter and Wikipedia as well as a combination, against a naïve and an industry strength baseline. Firstly, we introduce our systems which offer suggestions from a *single* source:

1. *Flickr(F):* firstly, we compare against a naïve baseline which suggests the most popular tags on Flickr. We propose this baseline to replicate the cold start scenario i.e. where an image contains no tags to suggest upon.
2. *Flickr(FR):* secondly, we use an industry strength baseline by using those tag recommendations made on the Flickr website. Specifically, we consider the top tags as suggested from the getRelated API method[5], for the input tag, `acl`.
3. *Twitter(T/TP):* in our first Twitter approach, we suggest the most frequent terms within the related stream of Tweets (T). In our second approach, we suggest only the most frequent extracted nouns and entities (TP), thus reducing noise.
4. *Twitter(TR(N)):* inspired by [2], we use a tf-idf based tag recommendation approach based on N input tags (randomly extracted from an image) which computes recommendations from on a stream of n tweets. In this approach, we model tf as a normalised co-occurrence vector for a given term, where each position counts the number of tweets the given term co-exists in; *idf* is the vector of *inverse document frequencies* computed as $\log(n/n^{(t_j)})$, where $n^{(t_j)}$ is the number of tweets containing term t_j. Recommendations are computed as the dot product of these vectors with contributions added for multiple input tags.
5. *Wikipedia(W/WP):* in our first Wikipedia approach, we suggest the most frequent terms within the related Wikipedia article (W). In our second approach, we suggest only the most frequent extracted nouns and entities (WP), thus reducing noise.

Secondly, we combine recommendations from Twitter and Wikipedia using the following methods:

[3] `acl, acl2012, acl2012acl, aclfest, aclfest2012, aclfestival, aclfestival2012, aclmusicfest, aclmusicfestival`

[4] `www.flickr.com/services/api/flickr.photos.search.html`

[5] `www.flickr.com/services/api/flickr.tags.getRelated.html`

	Baselines		*Individual*							*Combination*		
	F	*FR*	*T*	*W*	*TP*	*WP*	*TR(1)*	*TR(2)*	*TR(3)*	TP ∪ WP	TP ∩ WP	TR(1) ∩ WP
P@5	0.045	0.525	0.003	0.105*	0.231*	0.101*	0.271*	0.313*	0.312*	0.101*	0.333*	**0.469***
MRR	0.075	0.690	0.004	0.510*	0.609*	0.507*	0.521*	0.452*	0.389*	0.507*	0.573*	**0.693***

Table 2: Recommendation Performance; statistical significance results against the F are denoted as * $p < 0.05$

1. *Intersection (∩):* We combine into one list containing only the *intersecting* tags weighted by a given tag's position in the original lists. This weighting scheme is computed as $1/p$, where p is the tag's position in a list, thus giving precedence to those in higher ranks. The weights from each list for each tag are summed. The top tags ordered by decreasing weight are returned.
2. *Union (∪):* We combine by considering the union of the lists. The same weighting scheme is used as before, with the top tags returned in decreasing order.

Metrics: We compute performance metrics by comparing those recommendations against those provided by the user. Using these user tags, we compute metrics used by previous work in image tag recommendation [2]:

1. *Precision at Five (P@5)*: The percentage of relevant tags amongst the top five, averaged over all runs.
2. *Mean Reciprocal Rank (MRR)*: $1/r$ where r is the rank of the first relevant tag returned, averaged over all runs.

5. RESULTS

Firstly, from Table 2, we observe that by suggesting frequent nouns and entities from a related stream of tweets (TP) we are able to significantly outperform our naïve baseline, supporting our hypothesis (RQ1) that social media can be exploited for PTR purposes. We improve upon this technique in a more elaborate tf-idf model (T(N)) which makes suggestions based on N input tags, achieving up to 31% recommendation accuracy for P@5. Further, we observe the importance of using part-of-speech tagging methods as a technique to address noise (RQ1) in Twitter for PTR by comparing the large difference in accuracies between the systems which suggest only nouns/entities (TP) against the system which suggests all terms (T).

Secondly, from Table 2, we observe that Wikipedia can also be exploited for tag recommendation purposes, however its application is not as effective as when recommending on Twitter data, perhaps due the narrower coverage of Wikipedia articles. The most effective recommendation strategy combines suggestions based on both Twitter and Wikipedia data (TR(1) ∩ WP) highlighting the complementary nature of these evidences and supporting our initial hypothesis (RQ2). Specifically, using an intersecting combination approach for both sources (TR(1)∩WP), recommendation accuracy which is almost comparable with state-of-the-art techniques (FR) is achieved. Therefore, for images taken at new events, which lack sufficient training data on Flickr (due to the time lag problem), tag recommendation approaches can instead make suggestions based on instantaneous social media and Wikipedia streams with high accuracy.

6. CONCLUSION AND FUTURE WORK

In this paper we developed an automatic approach for annotating event images by exploiting relevant social media and Wikipedia data. Specifically, we proposed photo tag recommendations based on significant nouns and entities present in tweets and Wikipedia data related to the Austin City Limits 2012 music festival. In this work, we highlighted the merit of computing recommendations based on these streams as an alternative to recommending based on Flickr data (which is often sparse and out-of-date due to users uploading images long after they are taken). In order to address noise present in social media streams, we applied natural language processing techniques and combined recommendations made with those computed from structured Wikipedia data. This work proposes a new area for image annotation research, and for this purpose we will release our test collection on acceptance. In future work, we plan to evaluate our approach for more, and varying types of events (e.g. natural disasters), as well as employ more elaborate techniques (e.g. topic modelling, clustering *etc*) for recommendation purposes.

7. REFERENCES

[1] D. A. Shamma, L. Kennedy and E. F. Churchill. Statler: Summarizing media through short-message services. In *ACM CSCW 2010*.

[2] N. Garg and I. Weber. Personalized, interactive tag recommendation for flickr. In *ACM RecSys 2008*.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS 25, 2012*.

[4] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *ACM SIGIR 2002*.

[5] A. Makadia, V. Pavlovic, and S. Kumar. Baselines for image annotation. *IJCV 90, 2010*.

[6] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. Building a large-scale corpus for evaluating event detection on twitter. In *ACM CIKM, 2013*.

[7] M. Osborne et al. Bieber no more: First story detection using twitter and Wikipedia. In *SIGIR '12*.

[8] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *HLT 2010*.

[9] J. Picault, M. Ribiere, and Y. Gaste. Indexing video segments using microblogs. In *IEEE CBMI, 2013*.

[10] A. Rae, B. Sigurbjörnsson, and R. van Zwol. Improving tag recommendation using social networks. In *RIAO, 2010*.

[11] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *ACM WWW, 2008*.