Improving tag recommendation using social networks

Conference or Workshop Item

Improving Tag Recommendation Using Social Networks

Adam Rae
Open University
Milton Keynes, UK
a.rae@open.ac.uk

Börkur Sigurbjörnsson
Yahoo! Research
Barcelona, Spain
borkur@yahoo-inc.com

Roelof van Zwol
Yahoo! Research
Barcelona, Spain
roelof@yahoo-inc.com

ABSTRACT

In this paper we address the task of recommending additional tags to partially annotated media objects, in our case images. We propose an extendable framework that can recommend tags using a combination of different personalised and collective contexts. We combine information from four contexts: (1) all the photos in the system, (2) a user's own photos, (3) the photos of a user's social contacts, and (4) the photos posted in the groups of which a user is a member. Variants of methods (1) and (2) have been proposed in previous work, but the use of (3) and (4) is novel.

For each of the contexts we use the same probabilistic model and Borda Count based aggregation approach to generate recommendations from different contexts into a unified ranking of recommended tags. We evaluate our system using a large set of real-world data from Flickr. We show that by using personalised contexts we can significantly improve tag recommendation compared to using collective knowledge alone. We also analyse our experimental results to explore the capabilities of our system with respect to a user's social behaviour.

Categories and Subject Descriptors

H.3.1 [**Information Retrieval**]: Content Analysis and Indexing; H.3.5 [**Information Retrieval**]: On-line Information Services

General Terms

Algorithms, Experimentation, Performance

Keywords

Flickr, tag recommendation, social networks, personalisation

1. INTRODUCTION

Tagging of media objects has proven to be a powerful mechanism that can improve search options for images and video in social media sharing sites such as Flickr[1] and You Tube[2]. Tags are an unstructured form of meta data where the vocabulary and reasoning behind each user's choice of tags varies. Common usage themes tend to emerge where people agree on the semantic description of a media object.

In popular social media sharing sites there are can be billions of images and videos being annotated by millions of users. This provides a wealth of information that can form the basis for tag recommender systems. We envision two tasks where recommender systems are particularly useful. In one scenario a user annotating a photo is recommended tags related to the photo that can be used to extend the existing annotation—this helps to simplify the task for the user and helps expand the coverage of the tags annotating the image. In a different scenario, the role of the recommender system is to provide search recommendations. This can be done through automated query expansion, or in an interactive process by means of search assistants that provide additional query terms.

Recommender systems based on "collective knowledge" have been proven to provide relevant suggestions [10]. Typically these systems aggregate the annotations used in a large collection of media objects independently of the users that defined the annotations. Alternatively, the recommendations can be personalised by using the annotations for the photos of a single user [1]. Both approaches come with their advantages and drawbacks. When the recommendations are based on collective knowledge the system can make good recommendations on a broad range of topics, but is likely to miss some recommendations that are particularly relevant in a personal context. Basing the recommendations on the personal context will provide good results if the user has been active, making the statistics underlying the recommendation system reliable, and if the user is conscientious while annotating.

Users participating in social media sharing sites interact with other users. For example, in Flickr users can maintain *contacts* with other users, who then can be further identified to be their *friend*, *family* member, or *other* type of contact. Additionally, a user can subscribe to *groups* in Flickr. The group membership of a user defines the explicit interest of a user in a certain topic, or community of users sharing a common interest.

In this paper we propose a personalised recommender system that aggregates and exploits the knowledge that exists at four different contextual layers in an extendable proba-

RIAO'10, 2010, Paris, France.

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

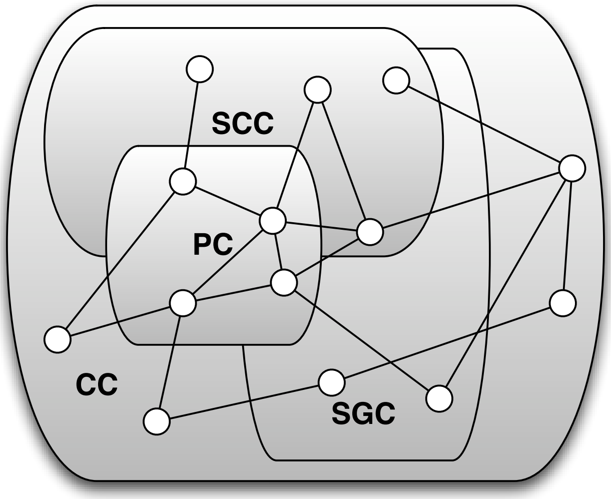


Figure 1: The contextual layers in our personalised recommender system describe overlapping subsets of the complete interconnected tag network. Here, instances of tags are represented by nodes and co-occurrences of tag values by edges.

bilistic framework. In our approach the focus on the user is central, therefore the first contextual layer is the "personal" (PC), constructed from the annotations provided by the user. Secondly, a "social contact context" (SCC) is defined by aggregating the annotations over all users that are identified as a contact of that user. Thirdly, a "social group context" (SGC) is obtained by aggregating the photo annotations of photos posted in the groups that the user is subscribed to. Finally, a "collective context" (CC) is derived by aggregating the annotations for all photos posted by all users. In Figure 1 the scope and interaction between the different contextual layers is schematically depicted.

A network of tags is derived for each context, based on co-occurrence analysis of tags used to annotate the photos within that context. Different vocabularies, and co-occurrence statistics emerge per user for each of the contexts defined. The exact formulation of these networks is discussed in more detail in Section 4.1.

The Personal Context (PC) defines the personal tag dictionary and network of related tags, which we expect to be most accurate when recommending tags using only personalised tag networks. The social activities of a user are of great influence on the size of the Social Contact and Group Contexts (SCC and SGC) and so their relative performance is dependent on the scale of these activities. The effectiveness of these two contextual layers has not been studied before in similar recommender systems and we will focus our evaluation in particular on these two contextual layers. Finally, we will use the recommendations based on the Collective Context (CC) as a baseline for measuring the effectiveness of our personalised recommender system.

The remainder of this paper is organised as follows. In Section 2 we give an overview of related work. We introduce the data collection in Section 3. The recommendation framework is described in Section 4 and and its evaluation in Section 5. Finally, we draw conclusions and discuss future work in Section 6.

2. RELATED WORK

Tag recommendation has been studied extensively in the past few years. The methods have been applied to a range of different media objects, such as blog posts [8], web bookmarks [2, 11], scientific articles [3, 11], music tracks [3], and photos [10, 1, 7]. The methods have been applied to various features of the media objects, such as content [8, 2, 11, 7], global tags [10, 1], personal tags [3, 1, 7] and social network [7]. In the remainder of this section we will focus on related work for photos and work that uses personalisation or social networks.

Sigurbjörnsson and van Zwol proposed a method of tag recommendation using the collective knowledge of a large collection of Flickr photos [10]. Their approach used global tag co-occurrence to make recommendations for partially tagged photos. They did not use user specific information in order to provide any personalisation nor did their method take into account the context of the user and their past interaction with the system. Their approach will be used as a baseline in our experimentation.

In contrast, Garg and Weber proposed a personalised approach to tag recommendation for Flickr photos [1]. They compare three methods: (i) using query-independent personal tag usage, (ii) query-dependent personal tag usage and (iii) query dependent group tag usage based on the group the photo belongs to. Our personal context approach is similar to their approach (ii) but we use significantly more data (see Section 3). Our group context approach differs from theirs (iii) in that we use the group context of the user and it can thus be applied to photos that have not been assigned to groups. Furthermore, we introduce a user social context approach which is not present in their work. They highlight the good performance of a hybrid method combining the personal and general contexts that gives improvement over either context alone. Their results also demonstrate how the balance between personal and general evidence, when combined, is influenced by the activity of the user—in that users who tag a lot tend to benefit relatively less from general evidence and more from personal evidence.

The personal context we define is similar to the user-tag matrices used by Jäschke et al. [3]—which they call *personomies*. While they made recommendations based on collaborative filtering using implicit tag usage similarity between users, we separate out the user specific and general tag data to allow us to examine them individually, before combining them. Their work also demonstrates the potential for data extracted from graph based representations of tag, resource and user interaction.

The nature of the differing graphs that make up the complete graph comprised of users, resources and their metadata is highlighted in the work of Kern et al. [4] who calculate and explore the overlap between different perspectives on the same data. The ability to avoid this overlap is important in being able to provide complementary but individually useful results from different contexts which, when combined, produce even better results than a single context in isolation. This is discussed in Section 6. Their experimental results also demonstrate the good performance of *personomy* derived recommendations.

A measure of how well a photo's tags match with a user's personomy is discussed in the work of Lerman et al. [6] along with a collaborative filtering approach based on the same 'Contact' label used in Flickr that we use to form our Social

Contact Context. Although their filtering method showed good improvement over their non-personalised baseline, it could only be used for users who had contacts and there was no discussion about how performance varied with the number of contacts (we address this with our system in Section 5.2.1). Their personomy-based personalisation is similar to our Personal Context in that it aggregates a user's tag usage from their own photos, but they use a different probabilistic model to calculate tag relevance.

The work of Lindstaedt et al. [7] also looked at making tag recommendation based on a range of complementary data sources: the text associated with image, the visual content and the user context. In particular their work on social data demonstrates the varied and diverse range of different interpersonal relations that can be used to model a user's position in a community and although their performance results were low, they showed the potential in using such data. Unlike in our work, they did not address any methods of combining result from different contexts, which is how we extend the concepts brought up in their experimentation.

3. DATA COLLECTIONS

In this section we discuss the data collection on which we base our experiment. We describe exactly the task we address in this paper and describe the setup of its evaluation.

3.1 Data Collection

Throughout our experiments we use publicly available data from Flickr, the online image sharing website where users can interact in many different ways with their media as well as each other. Users upload a diverse range of images—from diagrams to art photography—and form social connections with each other both explicitly and implicitly through interaction. Figure 2 shows the interactions that we use as social data in this work. A given user owns a set of photos which are annotated with zero or more tags. The user may also belong to zero or more groups and have zero or more contacts. Those contacts will in turn have their own photos and groups that will contain sets of photos.

Our recommendation system uses tag-based annotations from a large collection of Flickr photos. The collection contains the annotations of over 700 million public Flickr photos, uploaded before May 2008. Due to the scale of the data set used, a distributed, parallelised approach was taken to process the tag occurrence and co-occurrence values required. This was done on a Hadoop cluster using 100 nodes[3], with processing taking a matter of a few hours. This processing produced all the conditional probabilities for all users in the data set.

3.2 Evaluation Collection

We evaluate our tag recommendation system on a set of photos uploaded after May 2008. Thus we ensure that there is no overlap between the set of photos for which our co-occurrence statistics are calculated and the set of photos used for the evaluation. The evaluation collection was created using the Flickr API[4].

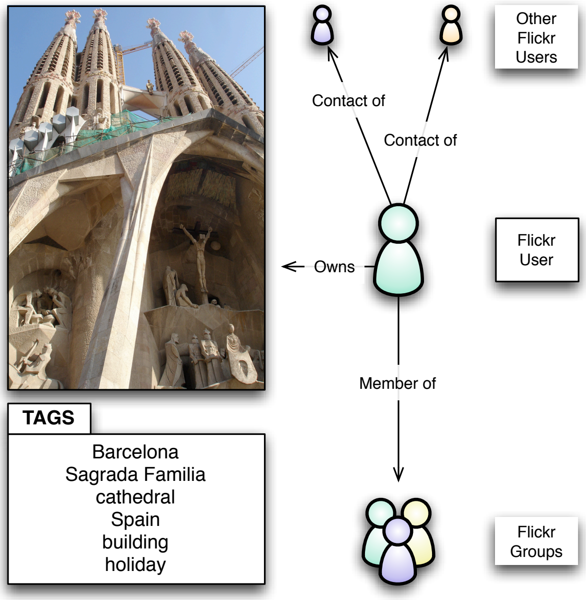


Figure 2: Overview of the relationship between Flickr users, their annotated media and their groups.

Table 1: Statistics over the 300 users in our experiment.

Statistic	Min.	Max.	Mean	StDev	Median
No. Contacts	0	1472	122.7	243.7	11.0
No. Groups	0	656	89.8	135.5	25.5
No. Photos	102	94415	1185.9	5586.9	385.0

3.2.1 Task

We evaluate the performance of our system through a "proxy task" - for a given photo with 10 tags or more we use two tags as input for our system and measure its performance in terms of how many of the photo's remaining tags it can recommend. Since this is a "simulated" evaluation of our system it may not give the right picture of the absolute performance of our system—in fact it is likely to underestimate the performance of our system. However, it is appropriate for comparing the relative performance of different tag recommendation approaches. The two photo tags are chosen at random. Our future work will involve looking at other choices of query tags and how this effects performance.

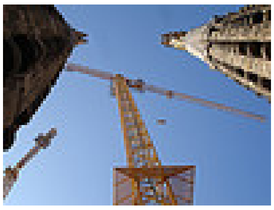
The top part of Figure 3 shows and example photo from Flickr together with its original tags added by the photo owner. The bottom part of Figure 3 shows an example of how the original tags were split into two sets: *input tags* and *target tags*. In our evaluation we pass the input tags to our algorithm and measure how well it can recommend the tags in the set of target tags.

3.2.2 Users

For the evaluation we chose to focus our attention on a selection of 300 users. This number was chosen to allow a balance between a large number of users for analysis and the time cost in computing the personalised larger tag

[3] http://hadoop.apache.org/

[4] http://www.flickr.com/services/api/



Owner tags: towers, cranes, architecture, construction, buildings, Sagrada Familia, Spain, Barcelona, Antoni Gaudi, Catalunya, blue sky.

Input tags: construction, Antoni Gaudi.

Target tags: towers, cranes, architecture, buildings, Sagrada Familia, Spain, Barcelona, Catalunya, blue sky.

Figure 3: Example of a photo in the test collection. Above, is the photo and the original annotation by the photo owner. Below is the split of the tag set into two parts, *input tags* and *target tags*.

co-occurrence networks. They were select at random from among users that represented a variety in "socialness"—i.e., the collection contained both users with few contacts and users with many contacts to better allow us to observe how this factor affects the ability of our system to make recommendations. We divide the users into buckets based on how many contacts they have:

Bucket 0: Users with zero contacts.

Bucket 1: Users with 1 or 2 contacts.

Bucket 2: Users with 3 to 10 contacts.

Bucket 3: Users with 11 to 50 contacts.

Bucket 4: Users with 51 to 250 contacts.

Bucket 5: Users with 251 contacts or more.

From each bucket we select 50 users who satisfy the following criteria:

- They have at least 100 photos in the data collection. This is because we aim our system at reasonably active users.
- They have at least 20 photos uploaded after May 2008 that satisfy the following criteria: 1) the photos need to have at least 10 tags; 2) no two photos have the same tag-set. From the resulting photos, 10 were randomly chosen for testing.

Hence our evaluation collection contains 3,000 photos from 300 different users. Table 1 shows some characteristics of our 300 users in terms of the number of contacts they have, the number of groups to which they belong and the number of photos they have in our data set.

4. RECOMMENDATION FRAMEWORK

4.1 Probabilistic Prediction Framework

For each context we derive a weighted network of tags, with nodes representing unique tags $t_i \in T$ and edges occurring when two tags have been used to annotate the same photo. Weights are defined by the number of times this happens in our data set. For all the tags of all the photos in the collection, we calculate the occurrence tally $o(t_i)$ and the co-occurrence tally $c(t_i, t_j)$. A tag 'occurs' if it is used to annotate a photo in the collection. Two tags 'co-occur' if they have been used to annotate the same photo. The probability of a tag occurring and the conditional probability of two tags co-occurring are formulated as:

$$p(t_i) = \frac{o(t_i)}{\sum_{t \in T} o(t)} \tag{1}$$

$$p(t_i|t_j) = \frac{c(t_i, t_j)}{o(t_j)} \tag{2}$$

To produce a set of recommendations for a given set of input query tags, each query tag is first used to generate a intermediate set of recommendations and these sets are then combined. The set of recommendations $s \in S$ for a given query tag for a given context is the complete set of tags that co-occur with that tag in that context's network. We emphasise final recommendations in terms of their rank position by penalising those tags that are not recommended by all query tags.

So, to calculate a set of recommendations given a set Q of input query tags, the probability of an intermediate suggestion given Q in context x is first calculated for each s:

$$p_x(s|Q) = p_x(s) \prod_{q \in Q} \begin{cases} p_x(s|q), & \text{if } p_x(s|q) > 0 \\ \epsilon, & \text{otherwise} \end{cases} \tag{3}$$

where ϵ is a non-zero value significantly smaller than the lowest conditional probability in the complete set of all conditional probabilities that allows us to avoid the zero-probability problem. Without the use of ϵ, in cases where recommended tags doesn't co-occur with all input query tags, any instance of non co-occurrence would reduce the overall probability of the recommended tag given the query tags to zero. This would completely ignore the contribution of the tags that did co-occurr.

Each resultant probability $p_x(s \in S|Q)$ is then used to produce an ordered list of tags in descending order of probability. The top N tags are then the final recommendations as given by that context's network of tags for a given query tag set. This method can be used in an identical manner for any similarly structured network of tags. In the experiments in this paper, we describe applying this method to four different networks.

4.2 Personal Context (PC)

The personal network of tags for a given user is made up of all instances of tags used on all the images that the user has uploaded. These sets vary between users, but consist solely of information relevant to that particular user. These sets tend to be far smaller and less comprehensive than that of the general tag cloud discussed in Section 4.5, but better reflect a user's personal ontology of keywords, or *personomy.* It is this user-specific nature of the Personal Context that should allow it to make more relevant annotation recommendations to particular users.

4.3 Social Contact Context (SCC)

A user in Flickr can explicitly connect themselves to other users by giving them the label 'Contact'. These inter-personal connections form a social graph between many of the users in the system. We produce a tag network from this data by taking all the photos from the contacts of the user for

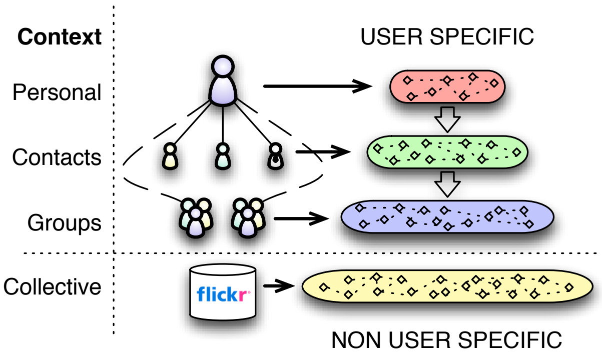


Figure 4: Hierarchical ordering of contexts going from most personalised to most general.

whom recommendations are being generated and aggregating them, excluding the tags from the photos of the user themselves. These tag networks capture the vocabulary not of the user but of their social group, possibly sharing attributes like language, geographical proximity and to some degree photographic interests, which are considered to be helpful in providing a more focused set of recommendations.

4.4 Social Group Context (SGC)

Users on Flickr can interact with each other through becoming members of shared interest groups and share photos with others who have done the same. There are therefore images associated with such groups and the tags annotating these images may have a common theme—the topic of the group. These group topics vary immensely from visual themes (e.g. black and white, High Dynamic Range) to subject themes (e.g. landscape, portraiture) and activities (e.g. A Photo A Day, reportage of real word events). We aggregate the tags of the photos associated with the groups of which a user is a member to form another tag network which can also be used to derive possible tags for recommendation. These recommendations should more closely represent the interests of the user in terms of the photos they interact with as opposed to their attributes, better described by the Social Contact Context.

4.5 Collective Context (CC)

Whereas the previously defined tag networks have been selected subsets of the entire collection of photos available in Flickr to better reflect certain aspects of the user requiring recommendations, the Collective Context aggregates the tags from all photos. This forms a very large network of tags that encapsulates the tag usage of all users. While it is not user specific, it does provide an extensive (yet potentially noisy) data set from which to make recommendations. It also has the advantage of being able to provide recommendations when the user is not very socially active (i.e. has few contacts or is not a member of many groups etc.) which would restrict the capacity of the personalised contexts to provide relevant results.

4.6 Aggregation methods

We wished to combine the four individual ranks produced from the tag networks described previously to maximise performance. We initially evaluated a number of methods, including simple rank concatenation using an ordered hierarchy, linear combination based on probability values and a Multi-Layered Perceptron. We found that a rank aggregation method based on the Borda Count provided the best performance in most of our tests. We therefore chose this for the experiment.

The Borda Count is a group consensus function that combines voting ranks by assigning descending consecutive integer scores to each element of the individual ranks and summing (or averaging) values to produce a new ordered rank, as described in the work of van Erp and Schomaker [12]. The basic Borda Count method treats each input rank equally and uses linear scoring. There are issues when dealing with ranks of differing lengths as this method is based on the assumption of additive independence, which is not fully justified in this case. For example, the top score recommendation from one rank may be considerably worse than the top score recommendation from another, but they would be treated as equivalently good recommendations by this implementation of the Borda Count method.

This method emphasises those suggestions that are common to more than one constituent rank and can therefore also penalise relevant recommendations that were only produced by a single input rank.

In our implementation, the scores assigned to the ordered ranks start with the first element of each rank being given the same value equal to the length of the longest of all the input ranks.

5. EVALUATION

5.1 Evaluation Setup

The experimental task we define here tests how well the different types of tag contexts—personalized, social and collective, as well as their combinations—can be used to predict tags that a particular user would apply to an untagged photo, given two tags that the user has used already. As described in Section 3.2 we simulate this task by taking a tagged photo, randomly choose two tags as input and measure how well we can predict the remaining tags (See Figure 3). We refer to Section 3.2 for more details on the evaluation task and collection.

We use the *trec_eval* tool[5] to calculate the performance of our algorithm in recreating the prediction set. We measure the performance using standard information retrieval metrics: Precision of the top 5 recommended tags (P@5), Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP). It should be noted that while the system might recommend tags relevant to the photo, the metrics will only take into account the exact matches with the target tags, as specified by the user. They must therefore be interpreted within the particular context of the experiment and should only be seen as relative performance indicators.

The results were tested for statistical significance using the Student's T-test as this has been found to be reliable for this type of information retrieval experiment [9]. Those results that are statistically significant are marked in Tables 2 and 3. All significance tests are performed relative to our baseline.

Furthermore, in order to observe the influence of the size of a user's available social context on performance, we divide the users into buckets based on their attributes and

[5] http://trec.nist.gov/trec_eval/

compare between them. For the personal context we group users based on how many photos they have; for the social contact context we divide into buckets based on the number of contacts the users have; and for the social group context we use the number of groups to which they belong.

5.2 Evaluation of Results

We start with evaluating the performance of our framework using different contexts in isolation and then evaluate the contexts in combination. We use different baselines for these two stages of the evaluation. For the first stage, the Collective Context (CC) is used since it is comparable to the system presented in [10] and is non-personalised. This allows us to examine the effect of personalisation more distinctly. For the combination runs we adopt the combination of Personal and Collective Contexts as a baseline as this allows easier inspection of the effect of our more novel use of the Social Contacts and Social Groups Contexts. The baseline for the combination runs is comparable with the system presented in [1].

5.2.1 Performance of individual Contexts

The results of evaluating different contexts in isolation are shown in Table 2. It can be seen that, when measured over all users and their queries, the personalised contexts mostly perform significantly worse than the non-personalised Collective Context. The Social Contacts Context is particularly bad when considered on its own. The MAP of our Personal Context run is, however, significantly higher than for the Collective Context.

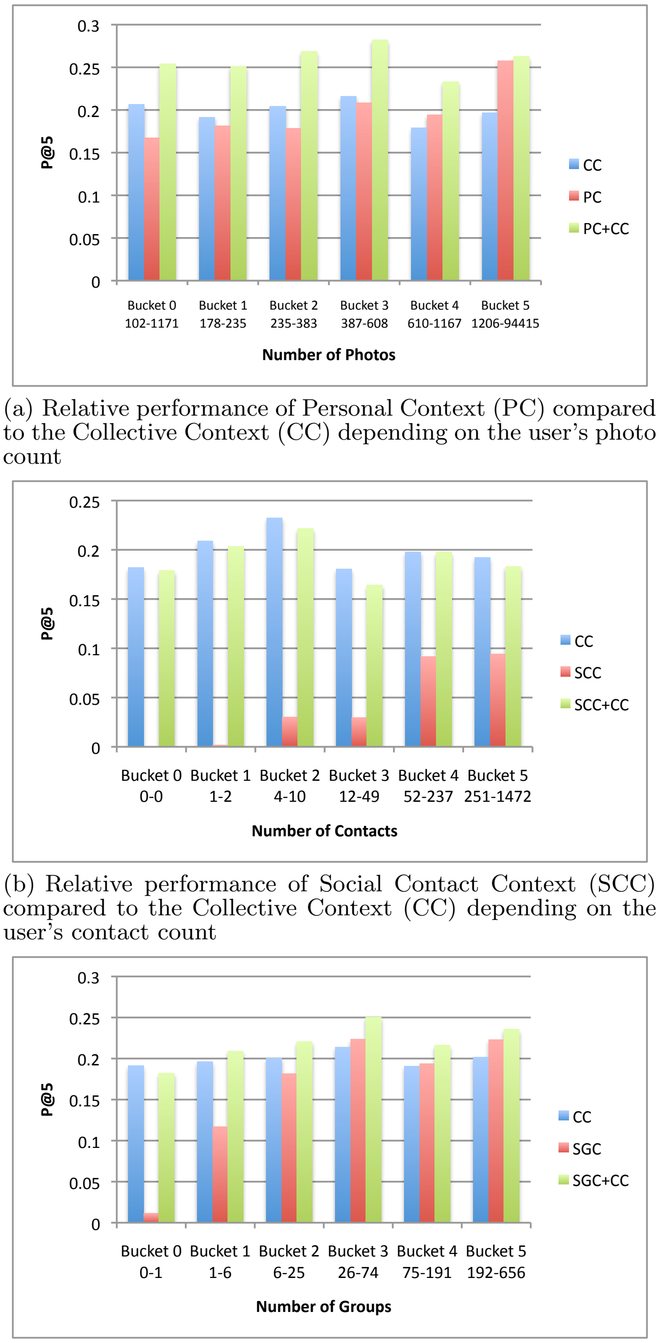
The results in Table 2 don't describe the relative proficiencies of the contexts - are some contexts better than others for certain types of users? We now extend our analysis by looking at sub-sets of users based on social criteria.

We feel that in the scenario of suggesting tags to users, higher priority should be given to early precision than for recall. Therefore in the following analysis we focus on the performance metric of 'Precision at 5'.

Figure 5 shows the relative performance with respect to P@5 of the Personal Context, Social Contact Context, and Social Group Context compared to the Collective Context and their combination with the Collective Context. Using topic sets partitioned on the users based on photo count, contact count and group membership count, we divide our 300 users into 6 equally sized buckets based on increasing "count".

Figure 5 (a) shows the performance of the Personal Context compared to the Collective Context and their combination, for users with increasing number of photos. Bucket 0 contains the 50 users with fewest photos and bucket 5 contains the 50 users with the greatest number of photos. We see that for users with relatively few photos the Collective Context outperforms the Personal Context. However, for users with many photos (buckets 4 and 5) the Personal Context outperforms the Collective Context. This might suggest that a user's personal tag dictionary, whilst personalised, does not become more useful for tag recommendation than collective knowledge until it reaches some critical size. From then on it is sufficiently large as well as tailored to the vocabulary of the given user and is capable of providing better tag recommendations.

Figure 5 (b) shows the performance of the Social Contact Context compared to the Collective Context and their combination for users with increasing number of contacts (i.e., bucket 0 contains the users with the fewest number of contacts and bucket 5 contains users with the greatest number



(a) Relative performance of Personal Context (PC) compared to the Collective Context (CC) depending on the user's photo count

(b) Relative performance of Social Contact Context (SCC) compared to the Collective Context (CC) depending on the user's contact count

(c) Relative performance of Social Group Context (SGC) compared to the Collective Context (CC) depending on the user's group count

Figure 5: Evaluation of performance of different contexts depending on the user characteristics. The performance is measured in terms of P@5. Columns signify equally sized buckets where each bucket contains 50 users. The lowest and highest values of bucket criterion are also shown.

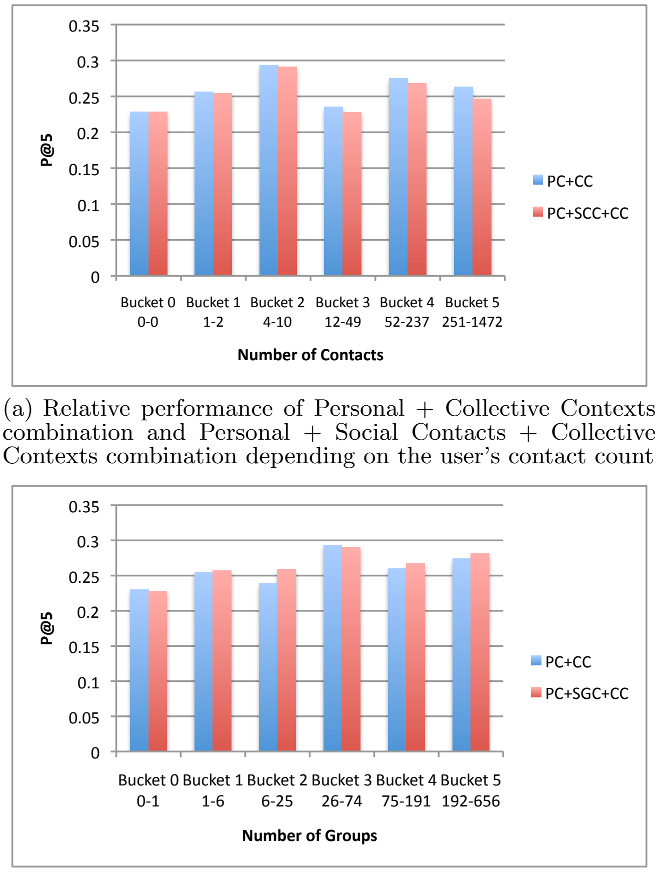
Table 2: Evaluation results for the individual contexts. Improvement is calculated relative to the Collective Context (CC) baseline. Values marked with †are significant with $p < 0.05$ and those with ‡with $p < 0.01$.

Run	MRR		P@5		MAP	
Collective Context (CC)	0.4473	–	0.1991	–	0.0934	–
Personal Context (PC)	0.3459	-22.7% ‡	0.1979	-0.6%	0.1034	10.7% ‡
Social Contacts Context (SCC)	0.0997	-77.7% ‡	0.0413	-79.3% ‡	0.0171	-81.7% ‡
Social Groups Context (SGC)	0.3395	-24.1% ‡	0.1585	-20.4% ‡	0.0777	-16.8% †

of contacts). We see that the Social Context is poor for all groups and always detrimentally affects the combination run (discussed further in Section 5.2.2). This seems to suggest that the tagging behaviour of a user's contacts poorly reflects that of the user, and so is unhelpful when making tag recommendations.

Figure 5 (c) shows the performance of the Social Group Context compared to the Collective Context and their combination for users with increasing number of group memberships (i.e., bucket 0 contains the users who are members of the fewest groups and bucket 5 contains the users who are members of the largest number of groups). We see that for users who are members of few groups the Social Group Context is clearly inferior to the Collective Context. However, as group membership increases, performance tends to increase. For users who are members of many groups (buckets 3 – 5) the Social Group Context does improve over the Collective Context. This suggests that for a sufficiently large collection of groups from which to mine tags, useful recommendations can be made. It also seems to lend support to the intuition that groups are likely to reflect the interests of a user, that ultimately affect or reflect their tagging behaviour.

Similar trends as described above are reported by [5], where in the context of music recommendation, the music taste of one's friends is less likely to positively correlate with their music taste. Conversely it is possible to make good recommendations based on other users that share the same taste.



(a) Relative performance of Personal + Collective Contexts combination and Personal + Social Contacts + Collective Contexts combination depending on the user's contact count

(b) Relative performance of Personal + Collective Contexts combination and Personal + Social Group + Collective Contexts cobination depending on the user's group count

Figure 6: Evaluation of performance of different contexts depending on the user characteristics. The performance is measured in terms of P@5. Columns signify equally sized buckets where each bucket contains 50 users. The lowest and highest values of bucket criterion are also shown.

5.2.2 *Combination of Contexts*

Combining different contexts has been shown to be useful for tag recommendation [1]. Table 3 shows the results of combining various contexts using the Borda Count method outlined in Section 4.6. We show their individual results as well as comparing them to a baseline of the combination of the Personal and Collective context. The aim of this part of our evaluation is to investigate whether the Social Group Context and Social Contact Context can add to the performance of the system when used in combination with the more conventional Personal and Collective contexts. We see that combining the Personal Context and the Collective Context gives a highly performing baseline with which to compare our other runs. Referring back to our example scenario illustrated in Figure 3, a P@5 of around 25% as we have here implies being able to exactly match 2.25 target tags out of 9. The other tags suggested are highly likely to also be relevant, but due to our experimental evaluation method we only count exact matches.

If the Social Contact Context is combined with the Collective and Personal Contexts, we see a statistically significant degradation in performance for the combined run with p-value < 0.01 for all our metrics. The Social Contact Context appears to perform so badly that it is in fact harmful to overall performance when used in combination with other contexts. This further supports the findings in the previous section that the tagging behaviour of contacts is unhelpful for making tag suggestions.

When the Social Group Context is combined with the Personal and Collective Contexts a marginal improvement can be observed, but only statistically significant for MRR and P@5. They suggest that there is value in using the Social Group Context.

By combining all contexts together we see a statistically insignificant changes in performance over the combined base-

Table 3: Evaluation results for the combined contexts. Improvement is calculated relative to the PC + CC baseline. Values marked with †are significant with $p < 0.05$ and those with ‡with $p < 0.01$.

Run	MRR		P@5		MAP	
PC+CC	0.5307	–	0.2587	–	0.1347	–
PC+SCC+CC	0.5189	-2.2%‡	0.2527	-2.4%‡	0.1300	-3.5%‡
PC+SGC+CC	0.5406	1.9%‡	0.2638	2.0%†	0.1351	0.3%
PC+SCC+SGC+CC	0.5260	0.9%	0.2591	0.2%	0.1319	-2.1%†

line for MRR and P@5, and a significant decrease in MAP. The inclusion of the harmful Social Contacts Contexts would explain the decrease in performance when compared to the Personal, Social Groups and Collective combination.

6. CONCLUSIONS

We have demonstrated how personal tag co-occurrence data can be used to provide more relevant recommendations of tags to a user when annotating photos than our baseline system. We have further shown that by combining this personalised data with data from all users of Flickr, we can significantly improve our performance. In addition, and most interestingly, we have demonstrated the considerable usefulness of additional social contextual data, in this case our Social Group context.

We have presented a framework for extracting tag networks from different 'strata' of a user's social graph from Flickr and shown how this can be evaluated with respect to established information retrieval performance measures. The framework can be extended with additional contexts–activity we hope to undertake in the future—to gain a better understanding of the relative usefulness of social graphs defined by different inter-user relationships.

The model we have presented has benefits for users who do not use English while interacting with Flickr. We are able to make relevant recommendations in their own language by virtue of their past interactions that make up their personal tag set and the interactions of their social groups, in addition to the most popular (usually English) tags contributed by the generalised data. We have also shown the difficulty in selecting other types of social data from Flickr that are ultimately useful when trying to boost performance for this particular user task.

We are confident that through further exploration of the rich social data available within online media sharing sites like Flickr, we can improve performance further still. We also think that learning weightings for the combination of our different contexts can be done on a more sophisticated, per user level which could also increase our ability to make good tag recommendations—an area we will investigate in future.

7. ACKNOWLEDGEMENTS

Part of this research was supported by the WeKnowIt project, partially funded by the European Commission, under contract number FP7-215453 and was also supported by a UK Engineering and Physical Sciences Research Council CASE studentship.

8. REFERENCES

[1] N. Garg and I. Weber. Personalized, interactive tag recommendation for flickr. In *ACM Conference on Recommender Systems, 2008*, pages 67–74.

[2] P. Heymann, D. Ramage, and H. G. Molina. Social tag prediction. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538, New York, NY, USA, 2008. ACM.

[3] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *European conference on Principles and Practice of Knowledge Discovery in Databases*, pages 506–514. Springer, 2007.

[4] R. Kern, M. Granitzer, and V. Pammer. Extending folksonomies for image tagging. In *Workshop on Image Analysis for Multimedia Interactive Services, 2008*, pages 126–129, May 2008.

[5] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202, New York, NY, USA, 2009. ACM.

[6] K. Lerman, A. Plangprasopchok, and C. Wong. Personalizing image search results on flickr. *American Association for Artificial Intelligence*, Apr. 2007.

[7] Lindstaedt, Pammer, Mörzinger, Kern, Mülner, and Wagner. Recommending tags for pictures based on text, visual content and user context. In *International Conference on Internet and Web Applications and Services*, page 506–511, 2008.

[8] G. Mishne. Autotag: A collaborative approach to automated tag assignment for weblog posts. In *Proceedings of the 15th international conference on World Wide Web, WWW 2006*, 2006.

[9] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, 2005.

[10] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *International conference on World Wide Web*, pages 327–336, 2008.

[11] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W. C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 515–522, New York, NY, USA, 2008. ACM.

[12] M. van Erp and L. Schomaker. Variants of the Borda Count method for combining ranked classifier hypotheses. *International Workshop on Frontiers in Handwriting Recognition. 2000. Amsterdam Learning Methodology Inspired by Human's Intelligence*, pages 443—452, 2000.