Flickr Tag Recommendation based on Collective Knowledge

Börkur Sigurbjörnsson
Yahoo! Research
C/Ocata 1
08003 Barcelona
Spain
borkur@yahoo-inc.com

Roelof van Zwol
Yahoo! Research
C/Ocata 1
08003 Barcelona
Spain
roelof@yahoo-inc.com

ABSTRACT

Online photo services such as Flickr and Zooomr allow users to share their photos with family, friends, and the online community at large. An important facet of these services is that users manually annotate their photos using so called tags, which describe the contents of the photo or provide additional contextual and semantical information. In this paper we investigate how we can assist users in the tagging phase. The contribution of our research is twofold. We analyse a representative snapshot of Flickr and present the results by means of a tag characterisation focussing on how users tags photos and what information is contained in the tagging. Based on this analysis, we present and evaluate tag recommendation strategies to support the user in the photo annotation task by recommending a set of tags that can be added to the photo. The results of the empirical evaluation show that we can effectively recommend relevant tags for a variety of photos with different levels of exhaustiveness of original tagging.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Algorithms, Experimentation, Performance

Keywords

Flickr, tag characterisation, tag recommendation, photo annotations, collective knowledge, tag co-occurence, aggregated tag suggestion.

1. INTRODUCTION

In recent years, tagging – the act of adding keywords (tags) to objects – has become a popular means to annotate various web resources, such as web page bookmarks [8], academic publications [6], and multimedia objects [11, 25]. The tags provide meaningful descriptors of the objects, and allow the user to organise and index her content. This becomes even more important, when dealing with multimedia objects that provide little or no textual context, such as bookmarks, photos and videos.

The availability of rich media annotations is essential for large-scale retrieval systems to work in practice. The current state-of-the-art in content-based image retrieval is progressing, but has not yet succeeded in bridging the semantic gap between human concepts, e.g., keyword-based queries, and low-level visual features that are extracted from the images [22]. However, the success of Flickr proves that users are willing to provide this semantic context through manual annotations. Recent user studies on this topic reveal that users do annotate their photos with the motivation to make them better accessible to the general public [4].

Photo annotations provided by the user reflect the personal perspective and context that is important to the photo owner and her audience. This implies that if the same photo would be annotated by another user it is possible that a different description is produced. In Flickr, one can find many photos on the same subject from many different users, which are consequentially described by a wide variety of tags.

For example, a Flickr photo of *La Sagrada Familia* – a massive Roman Catholic basilica under construction in Barcelona – is described by its owner using the tags *Sagrada Familia*, and *Barcelona*. Using the collective knowledge that resides in Flickr community on this particular topic one can extend the description of the photo with the tags: *Gaudi*, *Spain*, *Catalunya*, *architecture*, and *church*. This extension provides a richer semantical description of the photo and can be used to retrieve the photo for a larger range of keyword queries.

The contribution of this paper is twofold. First we analyse "how users tag photos" and "what kind of tags they provide", based on a representative snapshot of Flickr consisting of 52 million publicly available photos. Second, we present four different tag recommendation strategies to support to the user when annotating photos by tapping into the collective knowledge of the Flickr community as a whole. With the incredible amount of photos being tagged by users, we can derive relationships between tags, using global co-occurrence metrics. Given a user-defined tag and a photo, tags co-occurring with the user-defined tag are usually good candidates for recommendation, but their relevance of course depends on the photo. Likewise, for a given *set* of user-defined tags and a photo, tags co-occurring with tags in the set are good candidates. However, in this case a tag aggregation step is needed to produce the short list of tags that will be recommended.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

We evaluate the four tag recommendation strategies in an experimental evaluation, by implementing a blind pooling method to collect candidate tags for a given photo with user-defined tags. We repeatedly measure the performance on 200 randomly selected photos with a varying number of user-defined tags per photo. The number of user-defined tags per photo range from a single tag for sparsely annotated photos to more than six tags for exhaustively annotated photos. We measure the effectiveness of the recommendation strategies using four different metrics to gain detailed insight in the performance.

We envision two potential applications for the recommendation strategies. In one application, the recommendations are presented to the user, who can select the relevant tags and add them to the photo. Alternatively, the recommended tags are directly used to enrich the index of an image retrieval system.

The remainder of the paper is structured as follows. We start with discussing the related work in Section 2, followed by the analysis of tag behaviour in Flickr in Section 3, where we focus on tag frequencies and tag semantics. In Section 4 we present the four tag recommendation strategies for extending photo annotations in Flickr. The setup of the experimental evaluation is described in Section 5, while the results of the experiment are presented in Section 6. Finally, in Section 7 we come to the conclusions and explore future directions.

2. RELATED WORK

Tagging is a popular means of annotating objects on the web. A detailed account of different types of tagging systems can be found in [13] and [16]. The tags have been shown to be useful to give improved access to photo collections both using temporal information [9] and geographic information [3]. The methods we present in this paper extend the tagging of individual photos making them even more useful for the visualisation applications above.

The usefulness of tagging information depends on the motivation of users. Ames and Naaman explore the motivations for tagging photographs in mobile and online media [4]. Their investigation focuses on the use of the ZoneTag [2, 24] application in combination with Flickr, where users can upload and annotate photos to Flickr using their mobile phones. They find that most users are motivated to tag photos for organisation for the general public. They conclude that the tag-suggestion option included in ZoneTag encourages users to tag their photos. However, suggesting non-obvious tags may be confusing for users. Furthermore, users may be inclined to add suggested tags, even if they are not immediately relevant.

Various methods exist to (semi-)automatically annotate photographs. In the image processing and machine learning communities there is work on learning mappings from visual features to semantic labels [5, 15]. The methods take as input a set of labelled images and try to learn which low level visual features correspond to higher level semantic labels. The mapping can then be applied to suggest labels for unlabelled images based on visual features alone. For a more detailed account of content-based analysis for image annotation we refer to a recent overview paper by Datta et al. [7]. The ESP game is a tool for adding meaningful labels to images using a computer game [23]. Users play the game by suggesting tags for photos that appear on their screen and earn points suggesting same tags as another player. The mobile photo upload tool ZoneTag provides tag suggestion based on personal history, geographic location and time [24]. The different approaches work on different input data and thus complement each other. The methods we present here complement the ones above since we use yet a different input data, namely, the tags assigned originally by the photo owner. Our method can be applied on top of any of the tagging methods described above.

Our co-occurrence analysis is related to the construction of term hierarchies and ontologies that have been studied in the information retrieval and semantic web communities [20, 17, 21]. However, in the case of Flickr, the vocabulary is unlimited, and relations between nodes in the graph have an uncontrolled nature. Despite these two aspects, we use similar concepts to analyse the tag relations.

There has been some previous work on adding semantic labels to Flickr tags. Rattenbury et al. describe an approach for extracting event and place semantics of tags [18]. The intuition behind their methods is that event- and place-tags "burst" in a specific segments of time or regions in space, respectively. Their evaluation is based on a set of geotagged Flickr photographs. Using the method described above they were able to achieve fairly high precision of classifying tags as either a place or event. The semantic tag analysis presented in this paper we complement this method using WordNet to add a richer set of semantic tags.

3. TAG BEHAVIOUR IN FLICKR

In this section we describe the Flickr photo collection that is used for the evaluation, and we provide insights in the photo tagging behaviour of users. In particular we are interested in discovering "How do users tag?" and "What are they tagging?". Besides these two aspects, a third aspect is of importance, when studying tag behaviour in Flickr: "Why do people tag?". This aspect is studied thoroughly in [23, 16, 14, 4]. There it is concluded that users are highly driven by social incentives.

3.1 Flickr Photo Collection

Flickr is an online photo-sharing service that contains hundreds of millions of photos that are uploaded, tagged and organised by more then 8.5 million registered Web-users. To get some feeling for the size of the operation, during peak times up to 12,000 photos are being served per second, and the record for number of photos uploaded per day exceeds 2 million photos [12]. For the research described in this paper we have used a random snapshot from Flickr of 52 million publicly available photos with annotations. The photos were uploaded between February 2004 and June 2007 and each photo has at least one user-defined tag.

3.2 General Tag Characteristics

When developing tag recommendation strategies, it is important to analyse why, how, and what users are tagging. The focus in this section is on *how* users tag their photos.

The collection we use in this paper consists of over 52 million photos that contain about 188 million tags in total, and about 3.7 million unique tags. Figure 1 shows the distribution of the tag frequency on a log-log scale. The x-axis represents the 3.7 million unique tags, ordered by descending tag frequency. The y-axis refers to the tag frequency. The distribution can be modeled quite accurately

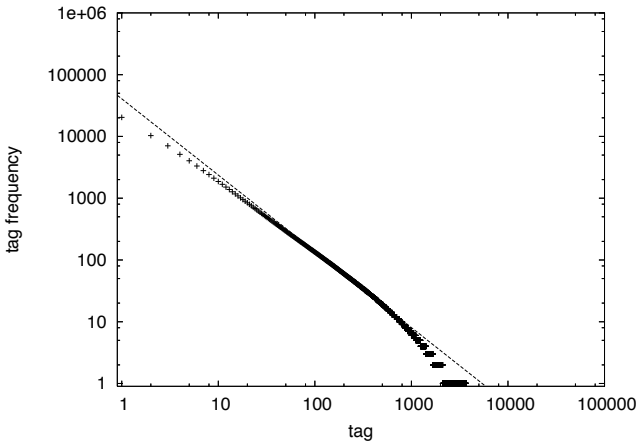

Figure 1: Distribution of the Tag Frequency in Flickr.

by a power law [19, 1], and the probability of a tag having tag frequency x is proportional to $x^{-1.15}$. With respect to the tag recommendation task, the head of the power law contains tags that would be too generic to be useful as a tag suggestion. For example the top 5 most frequent occurring tags are: *2006*, *2005*, *wedding*, *party*, and *2004*. The very tail of the power law contains the infrequent tags that typically can be categorised as incidentally occurring words, such as mis-spellings, and complex phrases. For example: *ambrose tompkins*, *ambient vector*, and more than 15.7 million other tags that occur only once in this Flickr snapshot. Due to their infrequent nature, we expect that these highly specific tags will only be useful recommendations in exceptional cases.

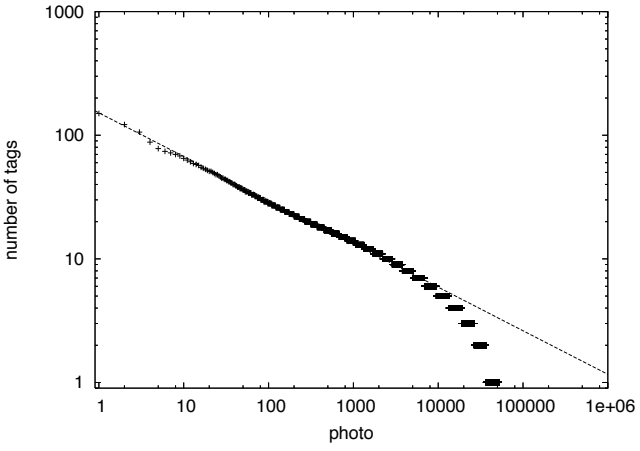

Figure 2: Distribution of the number of tags per photo in Flickr.

Figure 2 shows the distribution of the number of tags per photo also follows a power law distribution. The x-axis represents the 52 million photos, ordered by the number of tags per photo (descending). The y-axis refers to the number of tags assigned to the corresponding photo. The probability of having x tags per photo is proportional to $x^{-0.33}$. Again, in context of the tag recommendation task, the head of the power law contains photos that are already exceptionally exhaustively annotated, as there are photos that have more than 50 tags defined. Obviously, it will be hard to provide useful recommendations in such a case. The tail of the power law consists of more than 15 million photos with only a single tag annotated and 17 million photos having only 2 or 3 tags. Together this already covers 64% of the photos. Typically, these are the cases where we expect tag recommendation to be useful to extend the annotation of the photo.

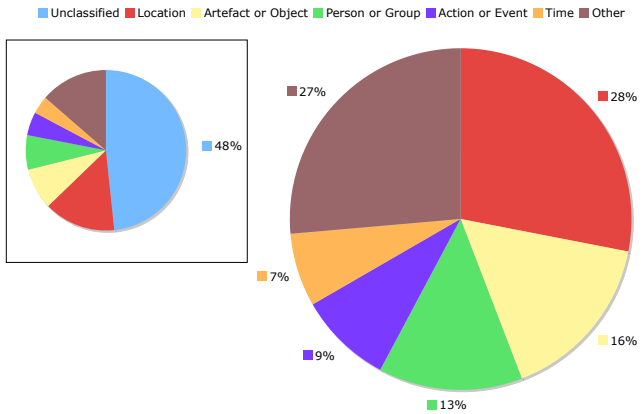

Figure 3: Most frequent WordNet categories for Flickr tags.

To analyse the behaviour of the tag recommendation systems for photos with different levels of exhaustiveness of the original annotation, we have defined four classes, as shown in Table 1. The classes differentiate from sparsely annotated to exhaustively annotated photos, and take the distribution of the number of tags per photo into account as is shown in the last column of the table. In Section 6, we will use this categorisation to analyse the performance for the different annotation classes.

	Tags per photo	Photos
Class I	1	$\approx$ 15,500,000
Class II	2 – 3	$\approx$ 17,500,000
Class III	4 – 6	$\approx$ 12,000,000
Class IV	> 6	$\approx$ 7,000,000

Table 1: The definition of photo-tag classes and the number of photos in each class.

3.3 Tag Categorisation

To answer the question "What are users tagging?", we have mapped Flickr tags onto the WordNet broad categories [10]. In a number of cases, multiple WordNet category entries are defined for a term. In that case, the tag is bound to the category with the highest ranking. Consider for example the tag *London*. According to WordNet, *London* belongs to two categories: `noun.location`, which refers to the city London, and `noun.person`, referring to the novelist Jack London. In this case the location category is ranked higher than the person. Hence, we consider the tag London to refer to the location.

Figure 3 shows the distribution of Flickr tags over the most common WordNet categories. Following this approach, we can classify 52% of the tags in the collection, leaving 48%

of the tags unclassified, as depicted in the in-set of Figure 3. When focussing on the set of classified tags, we find that *locations* are tagged most frequent (28%); followed by *artifacts or objects* (16%), *people or groups* (13%), *actions or events* (9%), and *time* (7%). The category *other* (27%) contains the set of tags that is classified by the WordNet broad categories, but does not belong any of the before mentioned categories. From this information, we can conclude that users do not only tag the visual contents of the photo, but to a large extent provide a broader context in which the photo was taken, such as, location, time, and actions.

4. TAG RECOMMENDATION STRATEGIES

In this section we provide a detailed description of the tag recommendation system. We start with a general overview of the system architecture, followed by an introduction of the tag co-occurrence metrics used. Finally, we explain the tag aggregation and promotion strategies that are used by the system and evaluated in the experiment.

4.1 Tag Recommendation System

Figure 4 provides an overview of the tag recommendation process. Given a photo with user-defined tags, an ordered list of m candidate tags is derived for each of the user-defined tags, based on tag co-occurrence. The lists of candidate tags are then used as input for tag aggregation and ranking, which ultimately produces the ranked list of n recommended tags. Consider the example given in Figure 4, there are two tags defined by the user: *Sagrada Familia* and *Barcelona*. For both tags, a list of 6 co-occurring tags is derived. They have some tags in common, such as *Spain*, *Gaudi*, and *Catalunya*, while the other candidate tags only appear in one. After aggregation and ranking 5 tags are being recommended: *Gaudi*, *Spain*, *Catalunya*, *architecture*, and *church*. The actual number of tags being recommended should of course depend on the relevancy of the tags, and varies for each different application.

4.2 Tag Co-occurrence

Tag co-occurrence is the key to our tag recommendation approach, and only works reliable when a large quantity of supporting data is available. Obviously, the amount of user-generated content that is created by Flickr users, satisfies this demand and provides the collective knowledge base that is needed to make tag recommendation systems work in practise. In this sub-section we look at various methods to calculate co-occurrence coefficients between of two tags. We define the co-occurrence between two tags to be the number of photos [in our collection] where both tags are used in the same annotation.

Using the raw tag co-occurrence for computing the quality of the relationship between two tags is not very meaningful, as these values do not take the frequency of the individual tags into account. Therefore it is common to normalise the co-occurrence count with the overall frequency of the tags. There are essentially two different normalisation methods: symmetric and asymmetric.

Symmetric measures. According to the Jaccard coefficient we can normalise the co-occurrence of two tags t_i and t_j by calculating:

$$J(t_i, t_j) := \frac{|t_i \cap t_j|}{|t_i \cup t_j|} \quad (1)$$

The coefficient takes the number of intersections between the two tags, divided by the union of the two tags. The Jaccard coefficient is know to be useful to measure the similarity between two objects or sets. In general, we can use symmetric measures, like Jaccard, to induce whether two tags have a similar meaning.

Asymmetric measures. Alternatively, tag co-occurrence can be normalised using the frequency of one of the tags. For instance, using the equation:

$$P(t_j|t_i) := \frac{|t_i \cap t_j|}{|t_i|} \quad (2)$$

It captures how often the tag t_i co-occurs with tag t_j normalised by the total frequency of tag t_i. We can interpret this as the probability of a photo being annotated with tag t_j given the it was annotated with tag t_i. Several variations of asymmetric co-occurrence measure have been proposed in literature before to build tag (or term) hierarchies [20, 17, 21].

To illustrate the difference between symmetric and asymmetric co-occurrence measures consider the tag *Eiffel Tower*. For the symmetric measure we find that the most co-occurring tags are (in order): *Tour Eiffel*, *Eiffel*, *Seine*, *La Tour Eiffel* and *Paris*. When using the asymmetric measure the most co-occurring tags are (in order): *Paris*, *France*, *Tour Eiffel*, *Eiffel* and *Europe*. It shows that the Jaccard symmetric coefficient is good at identifying equivalent tags, like *Tour Eiffel*, *Eiffel*, and *La Tour Eiffel*, or picking up a close by landmark such as the *Seine*. Based on this observation, it is more likely that asymmetric tag co-occurrence will provide a more suitable diversity of candidate tags than its symmetric opponent.

4.3 Tag Aggregation and Promotion

When the lists of candidate tags for each of the user-defined tags are known, a tag aggregation step is needed to merge the lists into a single ranking. In this section, we define two aggregation methods, based on voting and summing that serve this purpose. Furthermore, we implemented a re-ranking procedure that promotes candidate tags having certain properties.

In the this section we refer to three different types of tags:

- **User-defined tags** U refers to the set of tags that the user assigned to a photo.
- **Candidate tags** C_u is the ranked list with the top m most co-occurring tags, for a user-defined tag $u \in U$. We denote C to refer to the union of all candidate tags for each user-defined tag $u \in U$.
- **Recommended tags** R is the ranked list of n most relevant tags produced by the tag recommendation system.

For a given set of candidate tags (C) a tag aggregation step is needed to produce the final list of recommended tags (R), whenever there is more than one user-defined tag. In this section, we define two aggregation strategies. One strategy is based on *voting*, and does not take the co-occurrence values of the candidate tags into account, while the *summing* strategy uses the co-occurrence values to produce the final ranking. In both cases, we apply the strategy to the top m co-occurring tags in the list.

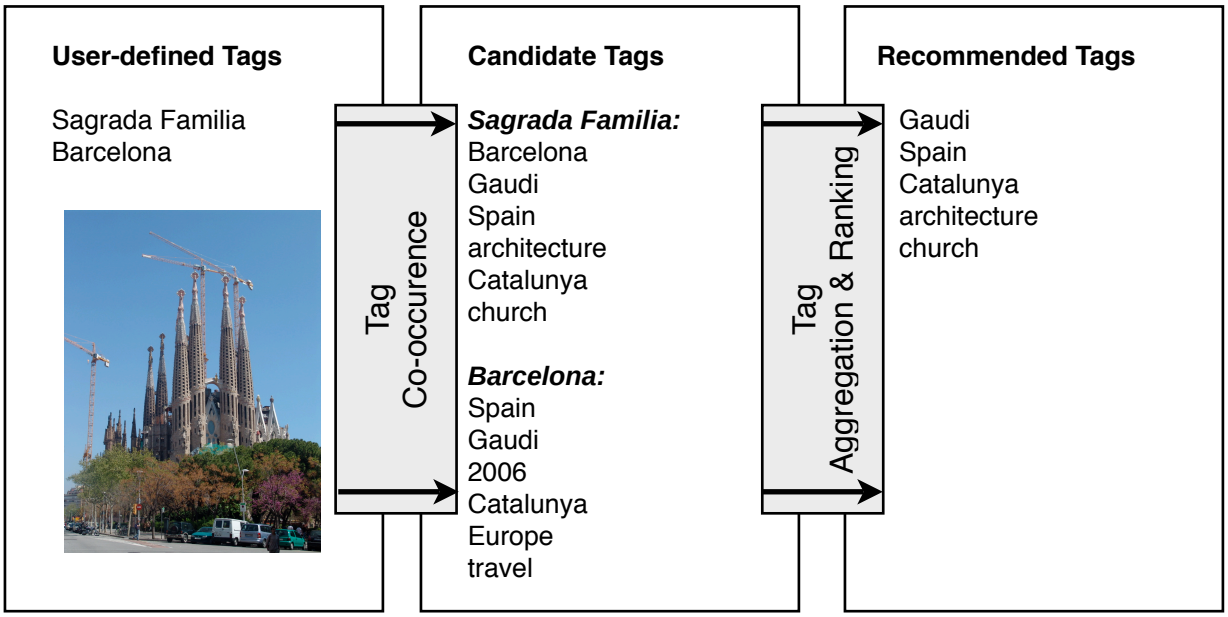

Figure 4: System overview of the tag recommendation process.

Vote. The voting strategy computes a score for each candidate tag $c \in C$, where a vote for c is cast, whenever $c \in C_u$.

$$vote(u,c) = \begin{cases} 1 & \text{if } c \in C_u \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

A list of recommended tags R is obtained by sorting the candidate tags on the number of votes. A score is therefore computed as:

$$score(c) := \sum_{u \in U} vote(u,c), \quad (4)$$

Sum. The summing strategy also takes the union of all candidate tag lists (C), and sums over the co-occurrence values of the tags, thus the score of a candidate tag $c \in C$ as calculated as:

$$score(c) := \sum_{u \in U} (P(c|u) \quad \text{, if } c \in C_u) \quad (5)$$

The function $P(c|u)$ calculates the asymmetric co-occurrence value, as defined in Equation 2. Note that the score of candidate tag c is obtained by only summing over the tags $c \in C_u$.

We will use these two aggregation strategies as the baseline for our evaluation as is presented in Section 6.

Promotion. In Section 3 we have made a number of observations with respect to tagging behaviour. In this section, we translate these observations into a "promotion function" to promote more descriptive tags for recommendation.

From the tag frequency distribution presented in Figure 1, we learnt that both the head and the tail of the power law would probably not contain good tags for recommendation. Tags in the tail were judged to be unstable descriptors, due to their infrequent nature. The head on the other hand contained tags that would be too generic to be useful (*2006*, *2005*, *wedding*, etc.).

- **Stability-promotion**. Considered that user-defined tags with very low collection frequency are less reliable than tags with higher collection frequency, we want to promote those tags for which the statistics are more stable. This is achieved with the following function:

$$stability(u) := \frac{k_s}{k_s + abs(k_s - log(|u|))} \quad (6)$$

In principle this is a weighting function that weights the impact of the candidate tags for a given user-defined tag. $|u|$ is the collection frequency of the tag u and k_s is a parameter in this function, which is determined by training. The function $abs(x)$ returns the absolute value of x.

- **Descriptiveness-promotion**. Tags with very high frequency are likely to be too general for individual photos. We want to promote the descriptiveness by damping the contribution of candidate tags with a very high-frequency:

$$descriptive(c) := \frac{k_d}{k_d + abs(k_d - log(|c|))} \quad (7)$$

This is another weighting function, now only applied to re-value the weight of a candidate tag. k_d is parameter in this function, and is configured by training.

- **Rank-promotion**. The co-occurrence values of tags provide good estimates of the relevance of a candidate tag for a user-defined tag. In principle, this is already used by the aggregation strategy for summing, but we observed that the co-occurrence values decline very fast. The rank promotion does not look at the co-occurrence value, but at the position r of the candidate tag $c \in C_u$ for a given user-defined tag u:

$$rank(u,c) = \frac{k_r}{k_r + (r-1)} \quad (8)$$

In the equation above, k_r is a damping parameter.

The combined promotion function we apply on a tag pair (u, c) is the following:

$$promotion(u, c) := rank(u, c) \cdot stability(u) \cdot descriptive(c) \quad (9)$$

When applying the promotion function in combination with either the voting or summing aggregation function, the score function is update as presented below for the voting case:

$$score(c) := \sum_{u \in U} vote(u, c) \cdot promotion(u, c) \quad (10)$$

The tag recommendation system now contains a set of parameters (m, k_r, k_s, k_d) which have to be configured. We use a training set, as described in the next section to derive the proper configuration of these parameters. Furthermore, we will evaluate the performance of the promotion function, with respect to the two aggregation strategies in Section 6. I.e., we evaluate the four different strategies as presented in Table 2.

	vote	**sum**
no-promotion	vote	sum
promotion	vote$^+$	sum$^+$

Table 2: The four tag recommendation strategies explored in this paper.

5. EXPERIMENTAL SETUP

In the following experiment we compare the four different tag recommendation strategies through an empirical evaluation. In this section we define the experimental setup and shortly present the system optimisation results, while the evaluation results are presented in Section 6.

5.1 Task

We have defined the following task: Given a Flickr photo and a set of user-defined tags the system has to recommend tags that are good descriptors of the photo. In our evaluation we set this up as a ranking problem, i.e., the system retrieves a list of tags where the tags are ranked by decreasing likelihood of being a good descriptor for the photo. In an operational setting, such a system is expected to present the recommended tags to the user, such that she can extend the annotation by selecting the relevant tags from the list.

5.2 Photo Collection

For the evaluation we have selected 331 photos through the Flickr API. The selected photos are based on a series of high level topics, for example "basketball", "Iceland", and "sailing", that were chosen by the assessors to ensure that they possessed the necessary expertise to judge the relevancy of the recommended tags in context of the photo.

In addition, we ensured that the photos were evenly distributed over the different tag classes as defined in Table 1 of Section 3, to have variation in the exhaustiveness of the annotations. Despite these two manipulations, the photo selection process was randomised.

Finally, we have divided the photo pool in a training set and a test set. For training we used 131 photos and the test set consists of 200 photos.

	m	**k$_s$**	**k$_d$**	**k$_r$**	**MRR**	**P@5**
sum	10	-	-	-	.7779	.5252
vote	10	-	-	-	.6824	.4626
sum$^+$	25	0	12	3	.7920	.5405
vote$^+$	25	9	11	4	.7995	.5527

Table 3: Optimal parameter settings and system performance for our tag recommendation strategies.

5.3 Assessments

The ground truth is manually created through a blind review pooling method, where for each of the 331 photos, the top 10 recommendations from each of the four strategies was taken to construct the pool. The assessors were then asked to assess the descriptiveness of each of the recommended tags in context of the photo. To help them in their task, the assessors were presented the photo, title, tags, owner name, and the description. They could access and view the photo directly on Flickr, to find additional context when needed.

The assessors were asked to judge the descriptiveness on a four-point scale: *very good*, *good*, *not good*, and *don't know*. The distinction between *very good* and *good* is defined, to make the assessment task conceptually easier for the user. For the evaluation of the results, we will however use a binary judgement, and map both scales to good. In some cases, we expected that the assessor would not be able to make a good judgement, simply because there is not enough contextual information, or when the expertise of the assessor is not sufficient to make a motivated choice. For this purpose, we added the option *don't know*.

The assessment pool contains 972 *very good* judgements, and 984 *good* judgements. In 2811 cases the judgement was *not good*, and in 289 cases it was undecided (*don't know*).

5.4 Evaluation Metrics

For the evaluation of the task, we adopted three metrics, that capture the performance at different aspects:

Mean Reciprocal Rank (MRR) MRR measures where in the ranking the first relevant – i.e., descriptive – tag is returned by the system, averaged over all the photos. This measure provides insight in the ability of the system to return a relevant tag at the top of the ranking.

Success at rank k (S@k) We report the success at rank k for two values of k: S@1 and S@5. The success at rank k is defined as the probability of finding a good descriptive tag among the top k recommended tags.

Precision at rank k (P@k) We report the precision at rank 5 (P@5). Precision at rank k is defined as the proportion of retrieved tags that is relevant, averaged over all photos.

5.5 System Tuning

We used the training set of 131 photos to tune the parameters of our system. Recall from the previous section that our baseline strategies have one parameter m and our promotion strategies have additional three parameters k_s, k_d, and k_r. We tuned our four strategies by performing a parameter-sweep and maximising system performance both in terms of MRR and P@5. Table 3 shows the optimal parameter settings and system performance for the four tag

	MRR	S@1	S@5	P@5
Baseline strategies				
sum	.7628	.6550	.9200	.4930
vote	.6755	.4550	.8750	.4730
Promotion strategies				
sum$^+$	.7718	.6600	.9450	.5080
vote$^+$	.7883	.6750	.9400	.5420
Improvement of promotion				
vote$^+$ vs sum	3.3%	3.1%	2.2%	9.9%

Table 4: Evaluation results for our four tag recommendation strategies using the test collection. The improvement of promotion is calculated using our better performing baseline run (sum) and better performing promotion run (vote$^+$).

recommendation strategies. In the next section we use the same parameter settings when we evaluate the system using the test collection.

6. EVALUATION RESULTS

The presentation of the evaluation results is organised in four sections. First we report the results for the two aggregation strategies, and in Section 6.2 we examine the performance of the promotion function. Section 6.3 discusses the results for the different tag classes. Finally, in Section 6.4, we analyse the type of tags that are recommended and accepted, in comparison to the user-defined tags based on the WordNet classification.

6.1 Aggregation Strategies

In this section we evaluate the performance of the aggregation strategies sum and vote. The top section of Table 4 shows the results for the two aggregation methods on the test collection.

First, we inspect the absolute performance of the two strategies. Based on the metric success at rank 1 (S@1), we observe that for more than 65% of the cases our best performing aggregation strategy – i.e., sum – returns a good descriptive tag at rank 1. For the success at rank 5 (S@5), we see that this percentage goes up to 92%.

For the precision at rank 5 (P@5), we measure a precision of 0.49 for the sum aggregation strategy, which indicates that on average for this strategy 50% of the tags recommended are considered useful. We can thus safely argue that the sum aggregation strategy performs very well and would be a useful asset for users who want support when annotating their photos.

When looking at the relative difference in performance between the two aggregation strategies, vote and sum, we observe that for all metrics the sum strategy outperforms the voting strategy. This is particularly evident for the very early precision (MRR and S@1) where the voting strategy is clearly inferior. The intuition behind this behaviour is that the voting strategy does not distinguish between tags that occur at different positions in the ranking of the candidate lists. I.e., it considers the top co-occurring tag just as a good candidate as the tenth. To the contrary, the sum strategy takes the co-occurrence values into account and thus treats a first co-occurring tag as a better candidate than the tenth co-occurring tag.

6.2 Promotion

We will now turn our attention to the performance of our promotion function. The mid-section of Table 4 shows the results of the promotion function in combination with the sum or vote aggregation strategies.

First, we inspect at the absolute performance of our promotion method. In terms of success at rank 1 (S@1) we see that for more than 67% of the photos the vote$^+$ strategy returns a relevant tag at rank 1. Expanding to the top 5 recommending tags (S@5) we see the performance goes up to 94%. In terms of precision at rank 5, P@5, we also observe that the vote$^+$ strategy achieves a precision of 0.54, which says that on average 2.7 of the top 5 recommended tags were accepted as being good descriptors for the photo.

If we compare the relative performance between the two aggregation strategies, sum$^+$ and vote$^+$, we observe that the two strategies behave rather similar, except in terms of precision at 5, where the vote$^+$ strategy outperforms the sum$^+$ method. This indicates that there is an interaction effect between the sum strategy and the vote$^+$ strategy, showing that the promotion function has a significant positive effect on the effectiveness of the recommendation. As a matter of fact, statistical significance tests, based on Manova repeated measurements with a general linear model show that the sum, sum$^+$, and vote$^+$ strategies all perform significantly better than the vote strategy ($p < 0.05$). And likewise for the vote$^+$ strategy, which is significantly performing better than sum, and sum$^+$.

In addition, when comparing the relative improvement, as shown in the bottom section of Table 4 for the best promotion strategy (vote$^+$) compared to the sum strategy. We find that for all metrics there is improvement. The improvement is marginal for MRR, S@1, and S@5, and as reported before, for the precision at 5, P@5, the improvement is significant (9.9%). We can thus argue that our promotion strategy is good at retrieving useful recommendations in the top 5 of the ranking without negatively affecting the performance very early in the ranking. This effect continues if we look beyond rank 5. For P@10 we measure that the vote$^+$ strategy continues to improve, showing a 10.1% improvement compared to the sum strategy, although the absolute precision goes down to 0.46.

6.3 Tag Classes

In this subsection we look at the performance of our system over different classes of photos, where we classify the photos, based on the criteria as defined in Section 3.2 (Table 1). I.e., we look at classes of photos with 1 tag, photos with 2–3 tags, 4–6 tags, and more than 6 tags, respectively. Table 5 shows the evaluation results of the sum strategy in comparison to the vote$^+$ strategy. On the sum strategy the performance is not evenly distributed over the different classes. The performance is better when the photo annotation is sparse (classes I and II) than for the photos with a richer annotation (classes III and IV). For the vote$^+$ strategy, we find that the the performance is more evenly distributed over the different classes. Which is reflected in the bottom section of the table where the relative comparison of the sum and vote$^+$ strategies shows a larger improvement for the classes III and IV. We observe that promotion has a marginal effect on the photos with only a few user-defined tags. However, for the photos with richer annotations the improvement is significant. Hence we conclude that the pro-

	MRR	S@1	S@5	P@5
Baseline (using sum)				
Class I	.7937	.7000	.9400	.5160
Class II	.7762	.6800	.9000	.5400
Class III	.7542	.6400	.9600	.5160
Class IV	.7272	.6000	.8800	.4000
Promotion (using vote$^+$)				
Class I	.7932	.7000	.9600	.5120
Class II	.8040	.7200	.9000	.5640
Class III	.7887	.6800	.9200	.5280
Class IV	.7673	.6000	.9800	.5640
Improvement				
Class I	-0.1%	0.0%	2.1%	-0.8%
Class II	3.6%	5.9%	0.0%	4.4%
Class III	4.6%	6.3%	-4.2%	2.3%
Class IV	5.5%	0.0%	11.4%	41.0%

Table 5: Performance of our system over different classes of topics.

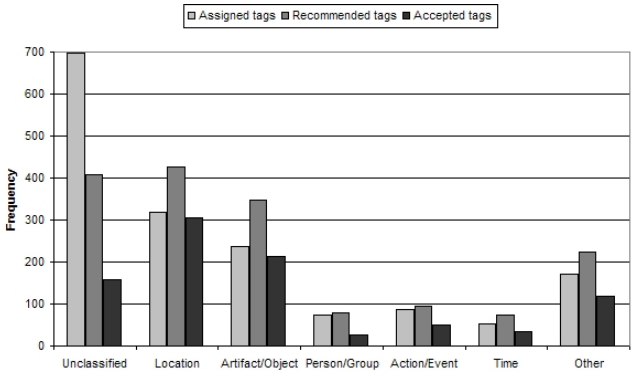

Figure 5: WordNet categories of initially assigned tags, recommended tags, and accepted recommendations.

motion has an overall positive effect, but mainly increases the performance of our system on photos that have more user-defined tags.

6.4 Semantic Analysis

We finish the evaluation of the tag recommendation system by analysing what type of tags are being recommended and accepted, to follow up on the tag characterisation presented in Section 3. We will perform this analysis using our best performing strategy, based on vote aggregation and promotion (vote$^+$). We turn our attention to the WordNet categories of the tags that are visible to the user in the recommendation application: the user-defined tags, recommended tags, and accepted tags.

Figure 5 shows the WordNet categories of all the tags that took part in the tag recommendation process. The figure shows results for the combination of training and testing sets. The first column in each group shows the categories of the tags initially assigned by the Flickr photo owners, the next column shows the categories of the top 5 recommended tags, and the third column shows the categories of the accepted tags (i.e., the tags judged as good or very good). It can be seen that there exists a gap between user-defined and accepted tags for those tags which can not be classified using WordNet, but that these two types of tags are well balanced for the other categories.

WordNet	Acceptance ratio
Unclassified	39%
Location	71%
Artifact or Object	61%
Person or Group	33%
Action or Event	51%
Time	46%
Other	53%

Table 6: Acceptance ratio of tags of different WordNet categories.

Table 6 shows the acceptance ratio for different WordNet categories. From the figure and the table we see that locations, artifacts, and objects have a relatively high acceptance ratio. However, people, groups and unclassified tags (tags that do not appear in WordNet) have relatively low acceptance ratio. We conclude that our system is particularly good at recommending additional location-, artifact-, and object-tags.

6.5 Summary

We conclude this section by recapitulating the main results of the evaluation results presented in this section. *First*, we have shown that the proposed strategies are effective, i.e., the recommended tags contain useful additions to the user-defined tags. For almost 70% of photos we give a good recommendation at the first position in the ranking (S@1) and for 94% of the photos we provide a good recommendation among the top 5 ranks. If 5 tags are recommended for each photo, than on average more than half of our recommendations are good. *Second*, we proved that our promotion function has a positive effect on the performance in general, and in particular on the precision at rank 5. We found a significant increase in the number relevant tags in the top five recommended tags. *Third*, we have shown that our best strategy (vote$^+$) has a stable performance over different classes of photos. *Fourth*, we reported that our system is particularly good at recommending locations, artifacts and objects, both in terms of volume and acceptance ratio.

7. CONCLUSIONS

Annotating photos through tagging is a popular way to index and organise photos. In this paper we first presented a characterisation of tag behaviour in Flickr, which forms the foundation for the tag recommendation system and evaluation presented in the second part of the paper.

Tag behaviour in Flickr. We have taken a random snapshot of Flickr consisting of 52 million photos to analyse *how users tag their photos* and *what type of tags they are providing.*

We found that the tag frequency distribution follows a perfect power law, and we indicated that the mid section of this power law contained the most interesting candidates for tag recommendation. Looking at the photo-tag distribution, we observed that the majority of the photos is being annotated with only a few tags. Yet, based on a mapping of tags on the WordNet classification scheme, we discovered that the

Flickr community as a whole annotates their photos using tags that span a broad spectrum of the semantic space, i.e., they annotate *where* their photos are taken, *who or what* is on the photo, and *when* the photo was taken. This motivated us to investigate whether the collective knowledge of the community as a whole could be used to help user extend their annotations of individual photos.

Extending Flickr photo annotations. Based on our observations, we introduced a novel and generic method for recommending tags, i.e., our approach deploys the collective knowledge that resides in Flickr without introducing tag-class specific heuristics. Based on a representative sample of Flickr, we have extracted tag co-occurrence statistics, which in combination with the two tag aggregation strategies, and the promotion function allowed us to build a highly effective system for tag recommendation.

We have evaluated the four tag recommendation strategies in an empirical experiment using 200 photos which are also available on Flickr. The evaluation results showed that both tag aggregation strategies are effective, but that it is essential to take the co-occurrence values of the candidate tags into account when aggregating the intermediate results in a ranked list of recommended tags.

We showed that the promotion function is an effective way to incorporate the ranking of tags and allows us to focus on the candidate tag set, where we expect to find good descriptive tags. Furthermore, the promotion function further improves the results, and has a highly positive effect of the precision at rank 5. The best combination, the vote$^+$ strategy, gives a relevant tag on the first position in the ranking in 67% of the cases, and we find a relevant tag in 94% of the cases when looking at the top 5. On average, more than 54% of the recommended tags in the top 5 is accepted as a useful tag in context of the photo. The vote$^+$ strategy also shows to be a very stable approach for different types of tag-classes. Finally, we have shown that our system is particularly good at recommending locations, artifacts and objects.

Open tagging systems like Flickr have continuously evolving vocabularies. Our method is based on the statistics of Flickr annotation patterns and our co-occurrence model can be incrementally updated when new annotations become available. Hence our method can gracefully handle the evolution of the vocabulary.

Future Work. Our future work includes implementing an online system where users can be aided in extending the annotations of their own photos. Having such a system allows us to evaluate the tag recommendation task more extensively in an on-line usability experiment.

Our method is complementary to previously explored approaches using either content-based methods [5, 15] or the spatial, temporal and social context of the user [2, 24]. A combination of different complimentary methods is likely to give a more robust performance. Further research into this is left as future work.

Acknowledgments

This research is partially supported by the European Union under contract FP6-045032, "Search Environments for Media – SEMEDIA" (`http://www.semedia.org`).

8. REFERENCES

[1] L. A. Adamic. Zipf, power-laws, and pareto - a ranking tutorial. `http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html`, 2002.

[2] S. Ahern, S. King, M. Naaman, R. Nair, and J. H.-I. Yang. ZoneTag: Rich, community-supported context-aware media capture and annotation. In *Mobile Spatial Interaction workshop (MSI) at the SIGCHI conference on Human Factors in computing systems (CHI 2007)*, 2007.

[3] S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the Seventh ACM/IEEE-CS Joint Conference on Digital Libraries, (JCDL 07)*, 2007.

[4] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI 2007)*, San Jose, CA, USA, 2007.

[5] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[6] CiteULike. http://www.citeulike.org.

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40, 2008. to appear.

[8] del.icio.us. http://www.del.icio.us.

[9] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 193–202. ACM Press, 2006.

[10] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[11] Flickr. http://www.flickr.com.

[12] Flickr blog: We're going down. http://blog.flickr.com/en/2007/05/29/were-going-down/.

[13] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. http://www.hpl.hp.com/research/idl/papers/tags/, 2006.

[14] K. Lerman and L. Jones. Social browsing on Flickr. In *Proceedings of International Conference on Weblogs and Social Media*, Boulder, Co, USA, 2007.

[15] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. In *Proceedings of the ACM Multimedia Conference*, pages 911–920, 2006.

[16] C. Marlow, M. Naaman, M. Davis, and D. Boyd. HT06, tagging paper, taxonomy, Flickr, academic article, toread. In *HT'06: Proceedings of the seventeenth ACM conference on Hypertext and hypermedia*, 2006.

[17] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference (ISWC 2005)*, volume 3729 of *LNCS*. Springer-Verlag, 2005.

[18] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *Proceedings of the Thirtieth International ACM SIGIR Conference, (SIGIR 07)*, 2007.

[19] W. J. Reed. The Pareto, Zipf and other power laws. *Economics Letters*, 74(1):15–19, December 2001.

[20] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213. ACM Press, 1999.

[21] P. Schmitz. Inducing ontology from Flickr tags. In *Proceedings of the Collaborative Web Tagging Workshop (WWW'06)*, 2006.

[22] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions Pattern Analysis Machine Intelligence*, 22(12):1349–1380, 2000.

[23] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM Press.

[24] Zonetag. http://zonetag.research.yahoo.com/.

[25] Zooomr. http://www.zooomr.com/.