Analyzing Favorite Behavior in Flickr

Marek Lipczak[1]**, Michele Trevisiol[2], and Alejandro Jaimes[2]

[1] Dalhousie University, Halifax, Canada, B3H 1W5,
lipczak@cs.dal.ca
[2] Yahoo! Research, Barcelona, Spain,
trevi,ajaimes@yahoo-inc.com

Abstract. Liking or marking an object, event, or resource as a favorite is one of the most pervasive actions in social media. This particular action plays an important role in platforms in which a lot of content is shared. In this paper we take a large sample of users in Flickr and analyze logs of their favorite actions considering factors such as time period, type of connection with the owner of the photo, and other aspects. The objective of our work is, on one hand to gain insights into the "liking" behavior in social media, and on the other hand, to inform strategies for recommending items users may like. We place particular focus on analyzing the relationship between recent photos uploaded by user's connections and the favorite action, noting that a direct application of our work would lead to algorithms for recommending users a subset of these "recently uploaded" photos that they might favorite. We compare several features derived from our analysis, in terms of how effective they might be in retrieving favorite photographs.

1 Introduction

Sharing and marking objects as favorites are fairly new phenomena in social media and many questions remain open on the behavior of users in relation to liking or marking an object as a favorite. Questions include *what* they favorite, *when* they favorite, and *how* they are related to the owner[3] of such object (e.g., in Flickr users can be "contacts", "friends", or "family").

The problem of understanding the dynamics of such actions is of extreme importance given the pervasiveness of sharing and like/favorite actions in many social media platforms. It is important because when users express such preferences explicitly, they are implicitly contributing to the building of more accurate user models of themselves. Such models have applications in a wide range of areas: they can be used to recommend content, to improve user experience in terms of interaction design, for advertising purposes, and for recommending other users. In addition, one of the basic functionalities of social media platforms is providing easy access to content added by friends and other types of connections. It is common in some social media platforms (e.g., Facebook, Twitter) to

** Research conducted during an internship at Yahoo! Research Barcelona.

[3] We use the term "owner" to refer to the user who uploads the photo.

rank photos and updates based on their recency and other features, but due to the increasing amount of shared content, and the size of personal networks, a simple recency based ranking is insufficient. Gaining insights into the favorite actions can contribute to designing novel ranking and recommendation algorithms, and to developing new functionalities around surfacing content users may like or favorite.

In this paper we present an analysis of favorite behavior on a large Flickr dataset. We analyze over 110 million favorite actions, focusing most of our study on a set of 24,000 users[4]. In particular, we examine temporal factors, user profiles derived from tags, and photo and photo-owner features, as well as the relationship between favorite actions and different link types between the users performing the actions and the owners of the photos. Finally, we perform experiments using several features to gain insights into their suitability for building algorithms for recommending photos to "favorite."

We examine the following: (1) whether users tend to favorite photos of people connected to them more than of people who are not connected, (2) whether users tend to favorite recent photos more than non-recent photos, and (3) whether the favorite activity happens in bursts. In addition, we evaluate several features for predicting favorite actions.

2 Related Work

Valafar *et al.*[13] performed a study of favorites in Flickr and found that 10% of users are responsible for $80 - 90\%$ of all favorites, and that the favorite action exhibits 50% overlap and 15% reciprocity between users. These statistics are confirmed by many other studies (*e.g.* [2, 9, 11]). Cha *et al.*[2] investigated how an image spreads through the social network and highlighted how propagation varies considerably with the duration of exposure to new photos. In some cases, it takes a long time for photos to propagate from one user to another (*i.e.* [1, 3]) as there is an initial phase of exponential growth in the number of users that favorite a photo, followed by a phase of slow and linear growth over the years.

Lee *et al.*[8] studied reciprocity in Twitter, and also in Flickr around favorites by dividing users into three groups: those that only browse, those that also upload photos, but do not participate in social activities, and those that participate in social activities. van Zwol *et al.*[15] presented a multi-modal, machine learned approach that combines social, visual and textual signals to predict favorite photos, while Lu and Li [10] exploited the photos previously marked as favorites by friends, in order to build a personalized search model to assist users in getting access to photos of interest.

Wonyong *et al.*[6] recommend tags for newly uploaded images, taking advantage of the tags assigned to favorite images of the user who uploaded the image, and combining tags with visual similarity. A similar work presented by Chen *et al.*[16] used favorite photos in order to extract representative tags, under the assumption that favorite images are better annotated.

[4] All analysis was aggregate, anonymous, and only on public photos.

Gursel and Sen [7] proposed an online photo recommendation system based on metadata and comments, assuming these two sources are highly related to the user's interests. De Choudhury *et al.*[5] developed a recommendation framework to connect image content with communities in online social media. They used visual features, user generated tags, and social interaction (*i.e.* comment actions) in order to recommend the most suitable group for a given image. Finally, Trevisiol *et al.*[12] presented an image ranking technique based on aggregating browsing patterns of about one million Flickr users.

A significant number of papers have been published using Flickr data, so we focused only on citing those that specifically deal with favorites and that are more relevant to our work. We are not aware, however, of any large-scale favorite action analysis such as the one we present in this paper.

3 The Dataset

Our dataset consists of a snapshot of Flickr until May 2008, which includes the explicit social network at the time and all interactions on public Flickr photos: over 110 million favorite actions made by over 1 million users.

Most users favorited photos only a few times (expected long tail of the distribution), more than $140K$ users favorited at least 100 photos, and the most active users favorited almost 100 thousand photos (heavy tail of the distribution). Sine long tail and heavy tail users are not representative for most of the favorite actions, we discarded them from most of the experiments. As we show later, the origin and recency of photos are very important factors. Therefore, with some exceptions, for the rest of the paper we consider only favorite actions made on photos uploaded to the system by the user's *connections* within 10 hours of the favorite action recorded. In order to limit the impact of the long and heavy tail we constrain the set of users for which we run the experiments to users with more than 100 and less than 2,500 favorites. We refer to this set as the "sample." We end up with 24,000 users that chose 8.6 million favorites among 1.194 billion photos.

4 Data Analysis

In Flickr, each photo has an *owner*, and users can be linked by more than one relationship type (*contacts*, *friends*, *family*, or any combination of those three). In the rest of the paper we will use the word *connection* to refer to any of the relationship types, but when we use the word *contact* we refer only to the contact relationship.

4.1 Photo Origin

We calculated the number of favorites with respect to link types (Table 1). The largest number of favorite photos come from *contacts*, both in terms of absolute number and average number of favorites per link, while *family* links have the fewest number of favorites. At the same time, nearly half of all favorites come from linked users: users tend to favorite photos of users that they are linked to, especially of their *contacts*.

Table 1. Social links statistics. *All favorites* includes favorites from users that are not linked by any of the relationship types to the user performing the favorite action.

links type	nr of favorites	avg favorites per link
contacts	29,642,943	0.90
friends	28,125,595	0.79
family	4,577,669	0.63
any link type	59,206,180	0.83
all favorites	112,177,317	10^{-5} (estimated)

4.2 Recency

Recency of a photo has been found to be important in image retrieval (van Zwol [14]). An analysis of our entire dataset of 110 million favorite actions shows that 20% of favorites happen within 10 hours, 30% within 24 hours, and 50% within a week from upload time.

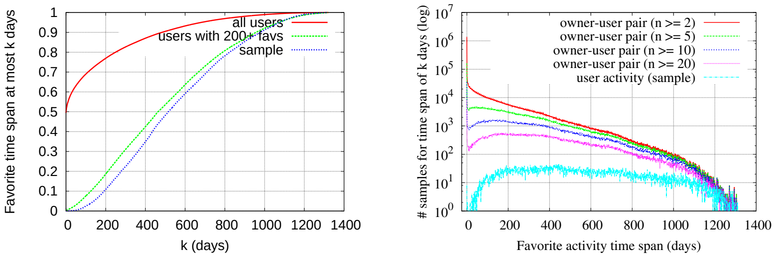
4.3 Time span of favorite actions

With *time span* we denote the number of days between the first and last time a user favorites photos. We examined the following sets of users:

i. **all users**: the entire set of users in the initial dataset (over 1 million users).
ii. **users with over 200 favorites**: 80% of all favorite actions are performed by the users in this set, which is obtained by filtering the 1 million users by selecting only those that have more than 200 favorites.
iii. **sample**: the set of $24K$ users described in Section 3, where we considered only favorite actions performed within 10 hours of uploading of a favorited photo.

Fig. 1(a) shows the cumulative distribution of users who performed favorite actions in a time span of k days. The distribution for *all users* is strongly biased by the long tail of users with a small number of favorites. In group (ii), the distribution resembles a normal distribution with high variance, where 80% of users have a time span of over 200 days. In group (iii) the ratio is even higher, around 90%.

Time span of owner-user interactions. We created a histogram of favorite actions for all user-owner pairs in our data set. Note that with our notation,



(a) Cumulative distribution of the ratio of time span of favorite action

(b) Time span of interaction between the owner of the photo and the user performing favorite actions

Fig. 1. Time span of user's activity.

a user selects a photo as a favorite and an owner is the one who uploads that photo.

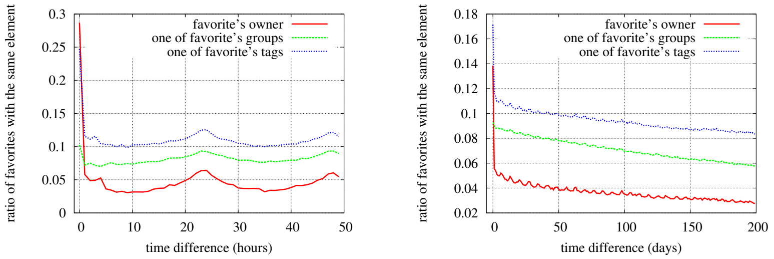
We analyzed only user-owner pairs with at least n favorites in total. As Fig. 1(b) shows, there are high peaks in very short time periods (less than a day). Note that we are considering only users that are connected by any of the relationship types, therefore the analysis shows that the favorite action happens in bursts and in many cases users do not return to favorite more photos of those owners: users tend to favorite photos of connections in short bursts.

.

Temporal locality of user's interests We analyzed favorite actions that were close to each other in time and observed how many of the favorites shared a particular feature (*e.g.*, were uploaded by the same owner). In Fig. 2(a) we see that a large number of photos favorited in less than an hour are likely to be from the same owner, or group. Unexpectedly, in Fig. 2(a) we can observe subsequent daily peaks for owners and tags. In Fig. 2(b) similar peaks are observed for weeks. One interpretation of this is that when a user is interested in a picture with a certain feature, pictures that share this feature are more likely to be favorited.

4.4 Favorite sessions

Another interesting aspect of favorite actions is their *burstiness*, in other words, measuring whether favorite actions occur uniformly over time or in bursts. We analyzed favorite actions within sessions, assuming that favorite actions are performed in the same session if the time difference between each pair of consecutive actions was lower than 30 minutes. Given this constraint, we measured the size of each session, comparing the two last groups described in Section 4.3: users with more than 200 favorites in total, and the 24k sampled users. The size of



(a) Short time period

(b) Long time period

Fig. 2. Likelihood of favoriting a photo with the same user, group or tags.

the sessions is much smaller for the sampled users, for which we considered only favorites performed within 10 hours of adding a photo.

We found that approximately 70% of the sessions have favorited photos of no more than 3 different owners. In almost 21% of the sessions the photos selected as favorites are from a single owner, and in about 35% of the sessions from two owners.

5 Computational Features

Below we describe a number of features derived from the outcomes of the data analysis and perform a simple evaluation of their suitability in terms of how well they are able to recall favorited photos. We also build a baseline favorite recommender as a proof of concept.

Since we have information on which photos have been favorited, we perform the evaluation by assuming that we want to predict a particular favorite action. In other words, let's say that at time t a user favorites a photo p. When the user favorites that photo, he choses it from a set of photos S. In our analysis we simply consider all photos in set S and examine which features might be more useful in predicting photo p, the one that was selected as a favorite.

We will use the following notation:

- **Recipient** – a user who is receiving photo recommendations.
- **Owner** – the owner of the photo that is recommended to the recipient.
- **Recommendation event** – the moment in time (t) in which the system presents the recommendation to the recipient. We assume that the favorite action takes place when a photo is recommended (i.e., not much later, for instance, not a week later).
- **Search space** – the set of photos S that are considered for recommendation in a single recommendation event. In this analysis we focus on photos that were uploaded by *connections* of the recipient, at most 10 hours before the recommendation event.

5.1 Photo based features

Photo based features are extracted from each photo that is considered for recommendation (i.e., each photo in the search space S defined above).

- **Photo recency** – time stamp of the upload of the photo (users are more likely to favorite recently uploaded photos (see Section 4)).
- **Number of favorites** – the number of times a photo was favorited by other users prior to the recommendation event.
- **Number of comments** – could be indicative of interest in the photo.

5.2 Photo owner features

Owner based features are related to the owner of the photo that is considered for recommendation, so all photos uploaded by a single user (owner) have the same owner features.

- **Likelihood of favoriting owner's photo** – the number of times a photo from the user was favorited divided by the total number of user's uploads.
- **Inverted batch size** – the inverted number of photos by the same user in the search space.
- **Recency of connection link** – time stamp of establishing a connection between two users: users might be more curious about photos of recent connections.

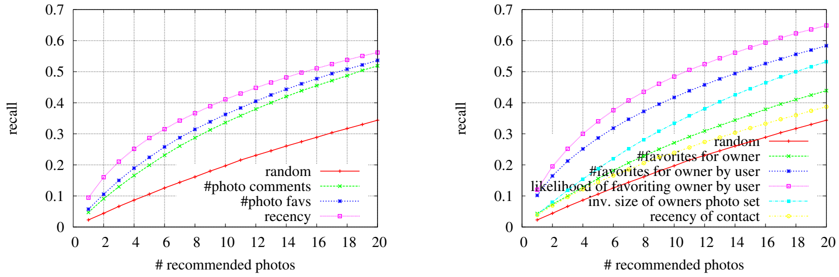
5.3 Feature evaluation

In the following experiments we used the favorite actions of 100 users randomly chosen from the sample set of 24k users described in Section 4.3. Therefore, we considered only favorites done on photos from *connections* uploaded at most 10 hours before the favorite action.

The objective is to rank all the photos that were uploaded within that time frame by the user's connections, so the favorited photo is in the top of the ranking. Each favorite action was considered a separate recommendation event. In this setting, the dataset contained $38,211$ recommendation events in which the total number of photos was $4,632,013$ (on average 121 photos per recommendation event). We split the data into training and test sets, where the first 80 favorites of each user correspond to the training set (8000 recommendation events).

Since photos uploaded within 10 hours are considered, it is possible that some of them may have already been marked as favorites by the user. Given that photos cannot be favorited more than once by the same user, these were omitted (on average 1.8 photo per recommendation event). All feature values in training and test sets were calculated on the information that was available prior to the recommendation event (following the timestamps of user actions).

We used the average recall@k metric to determine the accuracy of features. Recall@k is the number of true positive instances among the first k results of the ranking, divided by the total number of positive instances. In each test case there is always one true positive instance (the favorited photo), and averaged recall@k represents the ratio of recommendation events for which the ranking was able to place the favorited photo among the first k photos from the search space.



(a) Photo based features.

(b) User based features.

Fig. 3. Accuracy of features in photo recommendation task.

Accuracy of photo based features Among the three photo based features (i.e., *photo recency, number of favorites, number of comments*), the recency of the photo turns out to be the most accurate (Fig. 3(a)). This feature is also the third most accurate among all tested features. Good performance of this feature could be predicted observing the relation between the recency and the ratio of favorited photos. However, it is also possible that the performance of the recency feature is biased by the fact that the most recent photos of a user are shown first in the Flickr interface.

The number of favorites and number of comments prior to the recommendation event represent actions by other users (i.e., those that do not own the photo and those that are not receiving the recommendation). Such actions are commonly used in standard recommendation techniques based on collaborative filtering. Both features have lower accuracy than the recency of a photo. 90% of the photos in the search space that are already marked as favorites were favorited less than 10 times. This is expected given that we consider only photos that were uploaded at most ten hours before the recommendation event.

Accuracy of user based features The main feature describing an owner of a photo is the number of favorites of his/her photos prior to the recommendation event. We tested three features: total number of favorites for the owner's photos (#favorites for owner), total number of favorites by the recipient for the owner (#favorites for owner by recipient), and likelihood of owner's photo being favorited by a specific user given the photos in the search spaces of all recommendation events prior to the current one (likelihood of favoriting owner by user).

The first metric has the highest coverage, the last is most likely to have the best precision. Surprisingly, we can observe a very large difference between the total count of favorites and the personal count of favorites. The former has very low accuracy, which is unexpected.

The third feature is clearly superior. It measures the likelihood of an owner's photo being favorited by a recipient. The feature is personalized, which means that a separate likelihood value is calculated for each recipient. The high accuracy of this feature suggests that users tend to have a set of owners whose photos they frequently favorite. Indeed, in the sample, 40% of favorites are from owners who were favorited five times already (the total ratio of photos in the search space from these users is 13%). On the other hand, owners with no prior favorites contribute with 41% photos in the search space, but only 20% of favorites come from them.

The inverted size of owners' photos has reasonably good accuracy, suggesting that users who submit large sets of photos are in general, less likely to submit interesting photos. The last user based feature – the recency of user connections has very low accuracy.

Accuracy of similarity based features In addition to photo and user based features we tested a range of features calculated based on the similarity of users and photos. We found that comparing to other features similarity between recipients and owners/photos has low favorite photo prediction accuracy. It appears that that tags and groups are not as important in choosing favorited photos as who the photo owners are. Two additional reasons for low accuracy of similarity based metrics is the sparsity of tags and groups and the fact that users often assign the same set of tags to a large group of photos.

6 Discussion

7 Conclusions and Future Work

We presented an analysis of the favorite action in Flickr. The results show that users tend to favorite recent pictures of their connections, and in particular of their contacts; favorite actions tend to happen in bursts, particularly when considering individual user-owner pairs (i.e., it is common for a user who favorites a photo of a connection to favorite several photos of that connection in a very short period of time). We also examined different features (for owners and photos) to determine how useful they might be in a recommendation task. In particular, we used the results of the analysis to build a number of computational features and tested their suitability in determining which photographs may be marked as favorites. Our work contributes to gaining insights into the "liking" behavior in social media (at least in the specific case of Flickr), and to inform strategies for recommending items users may like.

Future work includes performing user studies and surveys to gain deeper insights into the reasons people favorite photos (e.g., as a form of appreciation with their connections, in order to collect photos, etc.), as well as building and evaluating a recommender system based on the findings of this study. The analysis itself could also be expanded and, for example, we could maybe find that different types of Flickr users exhibit different types of behavior.

Acknowledgements. This research is partially supported by the European Community's Seventh Framework Programme FP7/2007-2013 under the ARCOMEM and SOCIALSENSOR projects, and by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037 (www.cenitsocialmedia.es), "Social Media."

References

1. M. Cha, F. Benevenuto, Y.-Y. Ahn, and K. P. Gummadi. Delayed information cascades in flickr: Measurement, analysis, and modeling. *Computer Networks*, 56(3):1066 – 1076, 2012.
2. M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. *Proceedings of the first workshop on Online social networks - WOSP '08*, page 13, 2008.
3. M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. *Proceedings of the 18th international conference on World wide web WWW 09*, page 721, 2009.
4. J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1185–1194, New York, NY, USA, 2010. ACM.
5. M. De Choudhury, H. Sundaram, Y.-R. Lin, A. John, and D. D. Seligmann. Connecting content to community in social media via image content, user tags and user communication. *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1238–1241, June 2009.
6. W. Eom, S. Lee, W. De Neve, and Y. M. Ro. Improving image tag recommendation using favorite image context. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2445 –2448, sept. 2011.
7. A. Gürsel and S. Sen. Producing timely recommendations from social networks through targeted search. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 805–812. International Foundation for Autonomous Agents and Multiagent Systems, 2009.
8. J. G. Lee, P. Antoniadis, and K. Salamatian. *Faving Reciprocity in Content Sharing Communities: A Comparative Analysis of Flickr and Twitter*, pages 136–143. IEEE, 2010.
9. K. Lerman and L. Jones. Social browsing on flickr. *International Conference on Weblogs and Social Media*, pages 1–4, 2007.
10. D. Lu and Q. Li. Personalized search on Flickr based on searcher's preference prediction. In *WWW*, pages 81–82, 2011.

11. C. Prieur, D. Cardon, J.-S. Beuscart, N. Pissard, and P. Pons. The Stength of Weak cooperation : A Case Study on Flickr. *arxiv.org*, 65(8):610–613, 2008.
12. M. Trevisiol, L. Chiarandini, L. M. Aiello, and A. Jaimes. Image ranking based on user browsing behavior. In *SIGIR*, pages 445–454, 2012.
13. M. Valafar, R. Rejaie, and W. Willinger. Beyond friendship graphs: a study of user interactions in Flickr. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 25–30. ACM, 2009.
14. R. van Zwol. Flickr: Who is looking? In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 184–190, Washington, DC, USA, 2007. IEEE Computer Society.
15. R. van Zwol, A. Rae, and L. G. Pueyo. Prediction of favourite photos using social, visual, and textual signals. In *ACM Multimedia'10*, pages 1015–1018, 2010.
16. C. Xian and S. Hyoseop. Extracting Representative Tags for Flickr Users. In *IEEE International Conference on Data Mining*, pages 312–317, 2010.