Flickr: Who is Looking?

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

Abstract

This article presents a characterization of user behavior on Flickr, a popular on-line photo sharing service that allows users to store, search, sort and share their photos. Based on a sub-set of photos being uploaded during a 10 day window, we track the interest of users in those photos over a period of 50 days. In particular we investigate the user behavior on temporal, social, and spatial dimensions. Results show that the users are able to discover new photos within hours after being uploaded and that 50% of the photo views are generated within the first two days. The social networking behavior of users, and photo pooling are identified as the two major indicators related to a photo's popularity. Finally we show that the geographic distribution is more focussed around a geographic location for the infrequently viewed photos, than for the photos that attract a large number of views.

1. Introduction

Every day, millions of people visit social media sites such as Flickr [2], YouTube [4], and Jumpcut [3] to share their photos and videos, while others enjoy themselves by searching, watching, commenting, and rating the photos and videos. In this article we study the user behavior on Flickr, a popular on-line photo sharing service that allows users to store, search, sort and share their photos. In particular we are interested in the question: "*Who is looking?*". To answer this question, we investigate the temporal, social, and spatial characteristics of the users viewing photos on Flickr, in other words: "*When, why, and from where?*"

There are many different methodologies at hand to perform these kind of studies, such as commonly used in the area of Web Usage Mining [15] and Human-Computer Interaction [12]. In this article we mine the usage of Flickr through the analysis of server access logs to find patterns of user viewing behavior. In particular we focus on the explicit interest of users when viewing one particular photo at a time. For example most of the page views show a number of photos in a series of thumbnails. This allows the user to quickly scan the available photos. The explicit interest of a user in a single photo can be tracked when he/she clicks through on the thumbnail, to see one of its larger versions. It is generally believed that the browsing behavior on Flickr is strongly driven by social motivations [14, 11], which is also reflected in one of the main goals of Flickr: "We want to help people make their photos available to the people who matter to them" [1]. By characterizing the usage of Flickr in different categories, based on the number of views that a photo has generated over a period of 50 days after being uploaded, we aim to get better insights into the user behavior on the temporal, social, and spatial dimensions. The temporal dimension allows to track the user interest in a photo over time, the social dimension focusses on the social incentives of users, and the spatial dimension aims at investigating the geographic distribution of users interest in a photo.

In the remainder of this article we will discuss the related work in Section 2. The data used for this study is described in Section 3. Section 4 describes the general characteristics of the data set, while Section 5, 6, and 7 describe the temporal, social, and spatial characterization, respectively. We briefly discuss potential application domains in Section 8 and present our conclusions in Section 9.

2. Related Work

"Why do we tag?" is a question related to the ones we ask ourselves here. The answer to this question is highly driven by social incentives according to [11, 19, 6]. Also on this subject is a recent article by Marlow [16]. It discusses the Flickr tagging mechanism, and incentives for users to participate in social media services. Although tagging of photos is an important aspect of Flickr, we did not take tag mining into account for the research described in this article. Furthermore, a number of incentives for users are identified there, like for future retrieval, for contribution and sharing, or to attract attention. Flickr also allows tagging of or commenting on other people's photos, which is widely used for expressing opinions about a photo. Von Ahn et al [19] created a competitive edge for the user to photo labeling, when they launched the popular ESP game.

In this article we focus on the receiving end, rather than the supplying end. We are more concerned with the question: "What are the incentives for users to look at a photo?" In [14] social browsing behavior is investigated through the analysis of Flickr data, obtained through the Flickr API. It is concluded there that the browsing behavior is also highly driven by social motivations. In particular the social networking functionality, i.e., the ability to subscribe to photostreams by friends, is considered an important factor to generate photo views. They claim that the contacts (or friends) lists form the social network backbone of social media sites. Social media Web-services, such as Flickr, provide social networks that facilitate new ways of interaction.

The methodology followed in this article is based on web usage mining [15]. The alternative would be to follow a more qualitative approach as is generally adopted in human-computer interaction. In that case we might consider using think-aloud protocols [18] or to use eye-tracking experiments [13]. For future research, these are interesting directions to analyze the "Why?" question in more detail.

3. Data Collection

Flickr is an online photo-sharing service that contains hundreds of millions of photos that are uploaded, tagged and organized 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 [7].

For the research described here, we have collected the server access logs from Flickr for the duration of 60 days. We follow a subset of 1.83 million publicly available photos that were uploaded during the first 10 days over a period of 50 days. To be precise, if a photo is uploaded on day 3, we take the views for that photo into account up to day 53. Furthermore, we only take those page views into account that are explicitly oriented on viewing a single photo. These pages can simply be filtered by using the following regular expression on the access logs: ".*?flickr.com/photos/<owner_id>/<photo_id>/?.*".

Additional information about the selected photos was obtained through the Flickr API. We used the calls flickr.photos.getInfo, flickr.photos.getAllContexts, and flickr.contacts.getPublicList on a 10% subset of the 1.83 million photos. This allowed us to collect additional information that is used to determine the social context of the photo owner, and to track the various ways the photo has been published and used. Finally, we used Yahoo!'s IP address to geographic location conversion service, to determine the origination of the photo view requests for the same 10% subset.

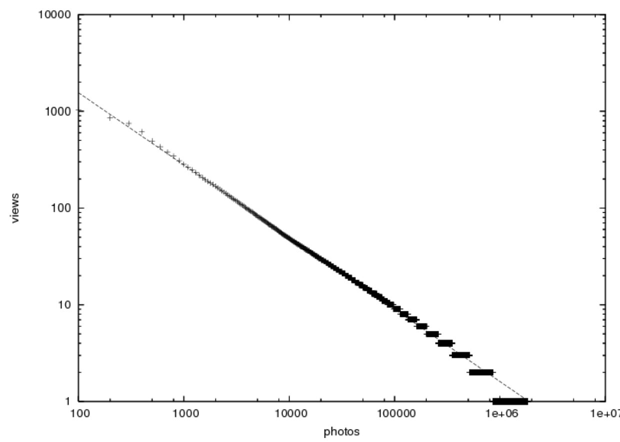


Figure 1. Total number of views per photo.

4. Characterization of Photo Views

After constructing the data collection as described above, we track 1.83 million photos that have generated a total of 6.72 million explicit photo views. Figure 1 shows the distribution of photo views on a log-log scale. The x-axis represents the 1.83 million photos, ordered by the number of views. The y-axis refers to the total number of views per photo. The distribution can be modeled quite accurately by a power law [17, 5], and the probability of having x visits is proportional to $x^{-0.7}$. It shows that the distribution of the photo views is very skewed, i.e. a small fraction of the photos generates most of the views.

Table 1 shows the distribution of photos per bucket, where a bucket groups the photos based on the number of views. For instance, the top most frequent bucket contains 105 photos, which have had at least 1,024 views. The third column shows total number of views, in this case 265,873 views. The most frequently viewed photo in our collection had 9,709 views. In fact, the top 7% of the most frequent viewed photos is responsible for almost 50% of the views, and less than 10% of the photos have more than 8 views.

Table 2 shows a different partitioning of the photos based on their frequency. Here the slices all contain 10% of the photos in the collection. The second column refers to the total number of photo views for a slice. Rather than continuing with the buckets of Table 1, we will use this partitioning for the remainder of this article, because it is more balanced and less biased towards the popular photos. The first slice (0-10%) contains the topmost viewed photos, whereas the the slices from 50% and up contain the photos that are only viewed once.

bucket	size	views
>=1024	105	265,873
512-1024	377	281,695
256-512	664	231,946
128-256	1,694	296,010
64-128	4,219	367,689
32-64	11,268	482,855
16-32	31,803	678,267
8-16	84,382	880,648
4-8	217,150	1,087,870
2-4	509,177	1,182,223
1-2	967,726	967,726
total	1,828,565	6,722,802

Table 1. Number of photos and photo views per bucket.

slice	views
0-10%	3,802,875
10-20%	812,131
20-30%	515,532
30-40%	365,712
40-50%	312,270
50-60%	182,856
60-70%	182,857
70-80%	182,856
80-90%	182,857
90-100%	182,857

Table 2. Number of photo views per slice.

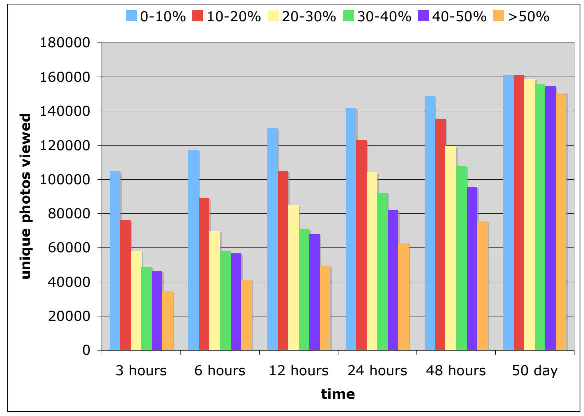


Figure 3. Unique photos discovered over time.

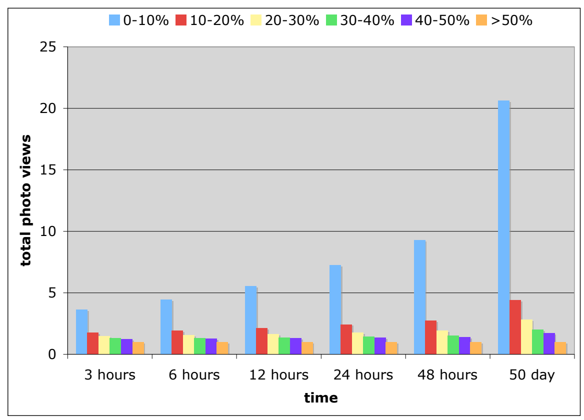


Figure 4. Photos views over time.

5. Temporal Dimension

In this section we investigate the temporal aspects of the photo views. First we present how the views are distributed over time for each of the different slices. We then focus on the first 48 hours after a photo is uploaded, to learn more about the photo discovery by the users and to find trends for the number of views per photo in the first couple of hours.

Figure 2 shows the average number of photo views for the same 10 slices over a 50 day period, as defined in Table 2[1]. Each slice follows the same trend, where it receives a high number of views per photo on the first day, which grows even further on the second day and then steadily declines. The declining trend followed by each of the slices can be modeled by an exponential decay, in which the number of views on day x after being uploaded is proportional to $e^{-1.1x}$. The trends for the slices 0-10%, 10-20% and >50 are included in the figure. The gap between the first two slices is significant and suggests that we can use this property for proxy caching strategies of photos, as discussed in Section 8.

Figures 3 and 4 focus on the first 48 hours after a photo is uploaded, to explore in more detail how quickly users are able to discover new (interesting) photos and how the number of photo views develop during these first two days until they reach their peak in number of views per day.

Figure 3 plots the number of unique photos viewed within x hours after being uploaded for the top 6 slices. It shows that within the first 3 hours already 65% of the photos in the most popular slice (0-10%) have been discovered by the users. This grows up to 92.5% for the first two days. This basically means that only a small fraction of the frequently viewed photos in the 0-10% slice is discovered after the peak. We think that these trends are highly related to the social dimension, as is discussed in the next section.

Figure 4 shows the total number of photo views for the

[1]The average number of views on y-axis are plotted on a logarithmic scale. The figure is depicted on the next page.

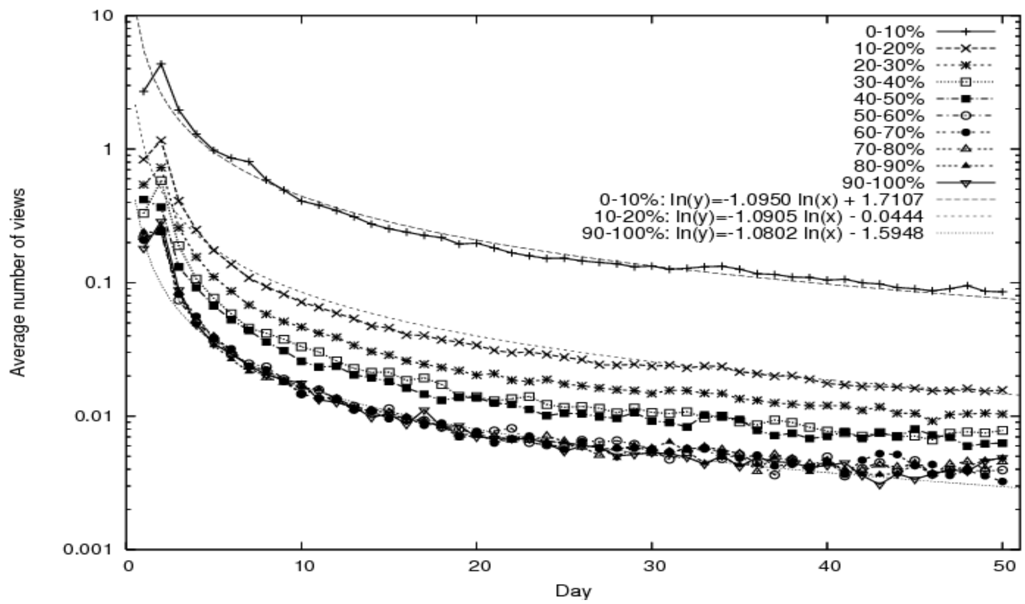


Figure 2. Average number of views per day.

	3 hours		6 hours		12 hours	
slice	avg	std	avg	std	avg	std
0-10%	3.63	8.25	4.44	12.51	5.55	18.66
10-20%	1.77	0.97	1.92	1.05	2.11	1.12
20-30%	1.47	0.67	1.54	0.7	1.62	0.73
30-40%	1.3	0.46	1.33	0.47	1.36	0.48
40-50%	1.25	0.43	1.27	0.44	1.3	0.46
>50%	1	0	1	0	1	0

	24 hours		48 hours		50 day	
slice	avg	std	avg	std	avg	std
0-10%	7.24	26.5	9.28	37.6	20.6	87.7
10-20%	2.43	1.22	2.75	1.28	4.4	0.7
20-30%	1.77	0.77	1.92	0.79	2.8	0.4
30-40%	1.44	0.5	1.52	0.5	2	0
40-50%	1.35	0.48	1.39	0.49	1.7	0.45
>50%	1	0	1	0	1	0

Table 3. Avg. number of views over first 48 hours.

photos after being discovered for the first 48 hours. It shows that the number of views after 3 hours for the first slice already approximates the number of views for the second slice after 50 days. More over, the popular photos will have 45% of their views within the first 48 hours. For the infrequently viewed photos (40-50% slice) this increases up to 82% of the views.

More detailed information can be found in Table 3, which provides the exact information that is used for Figure 4, together with their standard deviations. It shows that the standard deviation for the first slice grows significantly, due to the differences in number of views per photo in the first slice, while for the other slices this remains more or less constant over the first 48 hours[2].

6. Social Dimension

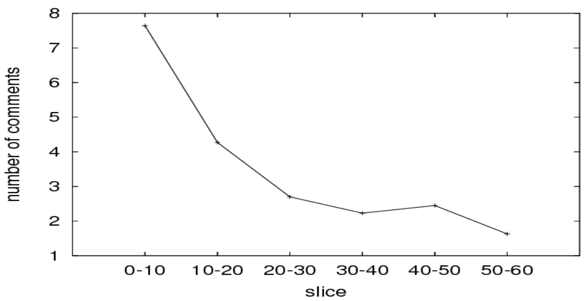
The Flickr Web-site contains many features that are related to the social dimension of the usage characterization, such as the number of groups a photo can appear in, the number of contacts (or the number of reverse contacts as used by [14]), the number of comments a photo will receive over time [16], Flickr opinionated tags (*BRAVO, BIGFAVE*) [16], the interestingness [9], the number of people that have called a photo their favorite, and more. In this section we will focus on the number of comments as a validation step, the number of contacts of the photo owner, and the number of group pools that a photo belongs to.

For this purpose, we have taken a sample from each slice consisting of the 10% most frequent viewed images for that slice. Figure 5 show the results over the different slices (x-axis). Figure 5.a depicts the trend for the average number of comments that each photo receives. It shows that the number of comments for the first slice is almost twice the amount of comments for the second slice (10-20%). A more detailed analysis is needed that also takes the time a com-

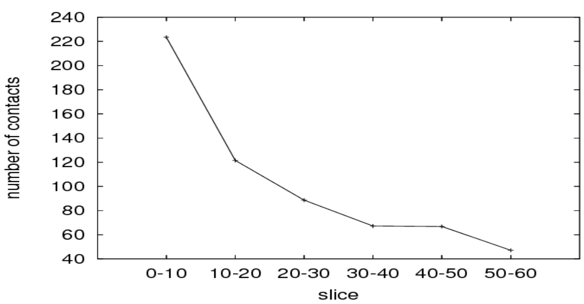
[2] The standard deviation for the photo views in the first slice (0-10%) is very skewed, as the frequencies range from 9706 to 432 views

ment is posted into account to determine if this is a useful *predictor* [8] for the popularity of a photo.

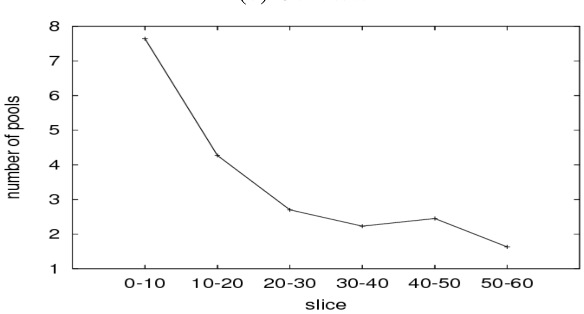
Figure 5.b plots the variation in the average number of contacts of the photo owner for the different slices. It shows a sharp decline in the number of contacts in direct relation to the number of views per photo. It confirms the outcome of related work on social browsing behavior, however it is not necessary to determine the incoming contacts of a photo owner as suggested in [14]. The number of (outgoing) contacts is already a distinctive predictor for the expected popularity of a photo. Moreover, we think that the subscriptions to a user's photo-stream is the most plausible explanation for the behavior found in Section 5, when inspecting the discovery ratio and the number of views a photo receives within the first hours after being uploaded.



(a) Comments



(b) Contacts



(c) Pools

Figure 5. Social context and Flickr participation

Figure 5.c shows the average number of pools that a photo belongs to, for the different slices. A similar trend can be observed as found in Figures 5.a-b. It therefore can be concluded that the number of group pools a photo participates in is also of influence on the number of views a photo receives. However, the photo is often not directly posted in various pools. This might explain the popularity of the remaining 7.5% of the photos that are not discovered within the first 48 hours, but still generate a large number of views (see Section 5). Apart from explicit interest in another user's photos, people are browsing through, or are subscribed to, groups where photos are categorized around a single theme. Given the user interest, he or she will visit these groups and will then be able to discover photos of users that are not in the user's contact list. A more detailed analysis and user studies are needed to validate this hypothesis.

7. Spatial Dimension

The motivation for this final part of the characterization is driven by the following hypothesis:

> *If a photo has many views, the geographic interest in the photo is worldwide, whereas if the photo only attracts a few views, e.g. less than 50 views over the 50 day time window, the geographic distribution is much more clustered around a particular area, which is related to the social contacts of the photo owner.*

In this section we only analyze the geographic distribution of the photo views, without taking the social contacts of the photo owner into account. We use a tool for determining the approximate geographic location in terms of the (longitude,latitude) coordinates of a view, using the IP-address from where the view request is originated.

Measuring geographic distribution The most logical way to derive a measure for the geographic distribution of views for a particular photo would be to take the average distance between all geographic locations in the set of photo views. However this is computationally expensive when it needs to be determined for many photos, with a large number of views per photo. Therefore we use an alternative method to approximate the geographic spreading of photo views.

Given a point set that represents the views for a single photo, such as depicted in Figure 6, we calculate two standard deviations, one for the longitude and one for the latitude. As shown in the figure, these two standard deviations can be used to form a bounding box around the center of gravity, covering most of the points in the point set. A single value representing the geographic spreading of the photo views is then determined by taking the euclidean distance

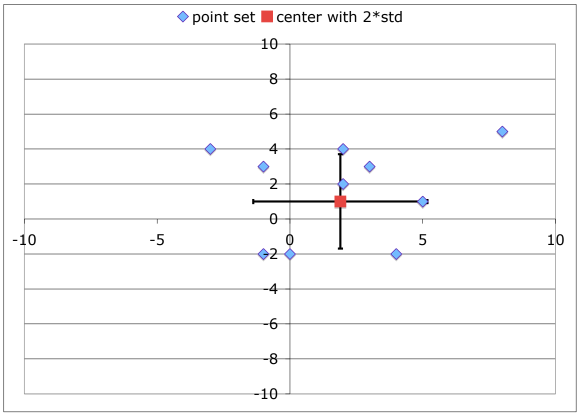


Figure 6. Example point set

over the two standard deviations. Please note that when a point is close to one of the boundaries of the coordinate system, for instance near the polar, a small error is introduced when the center of gravity lays on the opposite quadrant and the actual distance is shorter than the calculated distance. We believe that this does not influence the accuracy of the results, as the standard deviation also levels this error for the number of points in the set.

Geographic characterization In Table 4 the results are shown for the top most frequently viewed slices. Note that we have taken the same 10% sample set of photos as used before for the characterization on the social dimension. The second column of the table contains the average geographic spreading of the photo views over all photos in a slice. It indicates that when the popularity of a photo decreases, the geographic spreading is significantly more focused on a particular geographic location. The third column (std) refers to the standard deviation of the geographic spreading measure and is for the first slice surprisingly small, while for the consecutive slices this is larger, but relatively stable. One explanation for this behavior is that in the slices >20%, we observed that a large group of the photo views where originated from (approximately) the same geographic location. The fourth column indicates how many views on average each photo has within the 10% slice sample.

slice	avg	std	loc
0-10%	28.61	10.44	740.2
10-20%	18.84	19.29	26.4
20-30%	13.23	19.83	5.28
30-40%	9.54	19.57	2
40-50%	9.03	15.99	1.69

Table 4. Spatial interest.

8. Research and Application Areas

In this section we briefly want to address some of the application areas that can benefit from the outcome of the characterization presented here.

[8] studies the problem of predicting the popularity of items in a dynamic environment in which authors continuously post new items and provide feedback on existing items. This problem can be applied to predict popularity of blog posts, or to rank photographs in a photo-sharing system. Both the temporal and social dimensions of the characterization provide useful features for predicting the number of page views an image will receive, and we consider this as part of our future research.

Another area of research would be caching strategies for frequently viewed media objects[5]. Typically photos or streaming video consumes a large part of the bandwidth. Development of intelligent proxy caching strategies will help improving the user experience. In particular, knowledge about the photo discovery time and viewing peaks of popular photos is useful for the design decisions of such strategies.

Related to caching strategies, would be the development of a hybrid model for serving multimedia content that implements a P2P storage strategy for the infrequently viewed photos with little to no replication over the network and content stored close to the user (or its viewers), in combination with a content distribution network for serving the popular photos [10]

9. Conclusions

We have presented the characterization of user behavior on Flickr, an on-line photo-sharing service, to investigate who is looking at a photo. After a general characterization based on photo views we have focused the characterization on the temporal, social, and spatial dimensions to answer when, why, and from where a user is looking.

The analysis is based on a random sub-set of 1.8 million publicly available photos that are uploaded during a ten day window, which we tracked for a period of 50 days. We have shown that the collective user interest follows a perfect power law, where 7% of images generate almost 50% of the views.

When bucketing the images in 10 equal slices based on their view frequencies, we find that they all follow the same exponential decay and have their maximum number of views per day after 48 hours. Inspecting this in more detail, we see that 65% of the most popular images get discovered within 3 hours after being uploaded and that within 48 hours these images get almost 45% of the views they will generate during the 50 day window. For the infrequently viewed photos (40-50% slice) this ratio grows up to 82% of the views.

As an explanation for this behavior we find that the number of contacts, and the number of pools that an image is part of are good indicators for predicting the popularity of a photo. In other words, the social affiliation of users within Flickr is important: people that are highly interconnected will have their photos viewed many times.

Finally, we investigated the geographic distribution of photo views, based on the hypothesis that if a photo has many views, the geographic interest in the photo is worldwide, while if the photo only attracts a limited number of views, e.g. less than 50 views over the 50 day time window, the geographic distribution is more focused on a single geographic area. The results presented support this hypothesis.

The characterization of Flickr usage that is presented in the article, forms a motivation for ongoing research on predicting the popularity of new photos, proxy caching strategies for photos, and the design of fast content delivery mechanisms for rich media services that use partial replication of the photo collection.

References

[1] About flickr. http://www.flickr.com/about.

[2] Flickr - photo sharing. http://www.flickr.com/.

[3] Jumpcut - be good to your video. htpp://www.jumpcut.com/, 2007.

[4] Youtube - broadcast yourself. http://www.youtube.com/, 2007.

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

[6] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980, New York, NY, USA, 2007. ACM Press.

[7] H. Camp. Flickr blog. http://blog.flickr.com/ - May 29, 2007.

[8] C. Castillo, D. Donato, and A. Gionis. Estimating the number of citations using author reputation. Technical report, Yahoo! Research, 2007.

[9] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proceedings of the International World Wide Web Conference*. ACM Press, 2006.

[10] M. J. Freedman, E. Freudenthal, and D. Mazieres. Democratizing content publication with coral. In *Proc. 1st USENIX/ACM Symposium on Networked Systems Design and Implementation*, San Francisco, CA, USA, 2004.

[11] G. W. Furnas, C. Fake, L. von Ahn, J. Schachter, S. Golder, K. Fox, M. Davis, C. Marlow, and M. Naaman. Why do tagging systems work? In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, pages 36–39, New York, NY, USA, 2006. ACM Press.

[12] J. A. Jacko and A. Sears, editors. *Handbook for Human Computer Interaction*. Mahwah: Lawrence Erlbaum & Associates, 2003.

[13] R. J. K. Jacob and K. S. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye: cognitive and applied aspects of eye movement research*, pages 573–605. Boston: North-Holland/Elsevier, 2003.

[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] B. Liu. *Web Data Mining, Exploring Hyperlinks, Contents and Usage Data*. Springer, 2006.

[16] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, December 2006. ACM Press.

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

[18] H. van Oostendorp and S. de Mul. Learning by exploration: Thinking aloud while exploring an information system. *Instructional Science*, 27:269–284, 1999.

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