IDIAP RESEARCH REPORT

KODAK MOMENTS AND FLICKR DIAMONDS: HOW USERS SHAPE LARGE-SCALE MEDIA

Radu-Andrei Negoescu Alexander Loui
Daniel Gatica-Perez

Idiap-RR-20-2010

JULY 2010

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Kodak Moments and Flickr Diamonds: How Users Shape Large-scale Media

Radu-Andrei Negoescu
Idiap Research Institute
EPF Lausanne
Switzerland
radu.negoescu@idiap.ch

Alexander C. Loui
Kodak Research Labs
Rochester, NY
alexander.loui@kodak.com

Daniel Gatica-Perez
Idiap Research Institute
EPF Lausanne
Switzerland
gatica@idiap.ch

ABSTRACT

In today's age of digital multimedia deluge, a clear understanding of the dynamics of online communities is capital. Users have abandoned their role of passive consumers and are now the driving force behind large-scale media repositories, whose dynamics and shaping factors are not yet fully understood. In this paper we present a novel human-centered analysis of two major photo sharing websites, Flickr and Kodak Gallery. On a combined dataset of over 5 million tagged photos, we investigate fundamental differences and similarities at the level of tag usage and propose a joint probabilistic topic model to provide further insight into semantic differences between the two communities. Our results show that the effects of the users' motivations and needs can be strongly observed in this large-scale data, in the form of what we call Kodak Moments and Flickr Diamonds. They are an indication that system designers should carefully take into account the target audience and its needs.

1. INTRODUCTION

It is no longer a secret: multimedia content generated by users of online (social or not) media outlets is growing by the day to staggering amounts. A real need to understand the organization and growth of such systems is evident, as this will ultimately help users in getting to the right content, a less than trivial task. There is an increasing interest in the human centered computing community to understand social media phenomena *across* media sites. For instance, Mislove et al. [8] analyzed the connectivity network properties of four major websites, Flickr, YouTube, LiveJournal, and Orkut, while Leskovec et al. [6] investigated those of Flickr, del.icio.us, Yahoo! Answers, and LinkedIn. In a work more related to ours, Schifanella et al. [10] analyzed Last.fm and Flickr from a social and semantic interplay perspective, showing that a substantial level of local lexical and topical alignment can be observed among users in proximity in the social network. However, despite these initial works, the large-scale differences across photo repositories (or social media websites in general) in terms of tagging behavior and tagging content are not yet fully understood.

Among the existing online photo sites, Kodak Gallery (www.kodakgallery.com), formerly known as "Ofoto" and owned since 2001 by Eastman Kodak Company, is one of the leading online digital photo-developing services, operating a number of international sites including Canada, the main US site, and Europe. Apart from Kodak prints of digital pictures, Kodak Gallery offers several additional services around the digital images, such as online photo storage and sharing options, personalized photo gifts, photo books, and mobile phone access to stored photos. It also allows a user to share individual photos or entire albums directly from their Gallery account through social networking sites, such as Facebook. Users are able to provide free form captions of their assets, both at the image level or album level. Kodak Gallery has over 60 million users and storages of billions of images. It is estimated that in 2009 they were averaging in excess of 2 million image uploads a day.

Flickr (www.flickr.com) is one of the most popular online photo sharing websites. It was founded in 2004 and was acquired in 2005 by Yahoo!, quickly becoming one of the largest photo repositories online, with a thriving user community. Flickr offers access to their data through an Application Programming Interface (API), and third party developers and researchers have taken full advantage of it. The first ones created web applications and tools that bring added value to Flickr, the second ones analyzed and described the Flickr ecosystem and its numerous facets. Flickr has over 30 million usernames and received on its servers the 4 billionth photo upload in November 2009. Taking into account the time it took to upload the last billion photos, Flickr averaged during 2009 roughly 2.8 million images per day.

In this paper we present what to our knowledge is the first large-scale comparative analysis of these two online photo services, which differ in their design and affordances, and as a result may cater to different needs of their (maybe overlapping) audiences. In Flickr the accent is placed on sharing images with the world, and a real tagging system is employed by users in order to annotate their images and make them searchable in the system. Previous research has shown that the main motivations for tagging on Flickr [1, 9, 11] come from the social involvement within the online community. Many users involve in showcasing high quality photographs, often by joining online Flickr groups, such as *Diamond Stars, Flickr Diamond, The Best of Flickr, Shield of Excellence*, etc. In Kodak Gallery, on the other hand, the

MM'10, October 25–29, 2010, Firenze, Italy.

focus is placed on getting physical copies of digital photos, and then on sharing photos mostly with family and friends. The motivations in this case are most likely different, as is suggested by results of an ethnographic study [7] with 10 Flickr users. Miller and Edwards suggested that two distinct categories of users could be found, based on their sharing behavior: *Snaprs* and *Kodak Culture* sharers. The first group takes photos with the primary objective of sharing them with the world, while the second group takes photos to share with a small existing social group, and to archive. In contrast to ethnographic works which use traditionally a small number of users, we approach human centered computing from the other end of the scale, in a study using several orders of magnitude more data.

In this first comparative study, using over 5 million tagged photos from both these sites, we analyze the differences and similarities of Kodak and Flickr users and their tag usage, bearing in mind that two major components are in constant interplay: on one side, from an HCI perspective, the impact of the system design on the actual behavior of the users in terms of media usage, and on the other, from a sociological perspective, the impact of the users' motivations and needs on how the systems are actually used. Our results suggest that "Kodak Moments" and "Flickr Diamonds" are indeed two phenomena that impact the large-scale content generation by users.

We present the two datasets in the following section, then we look at some of the basic features, similarities, and differences in Section 3. We then jointly model Flickr and Kodak users using a probabilistic topic model and describe this analysis briefly in Section 4. We end with a discussion of the analysis in the final section.

2. DATASETS

The Kodak dataset is made up of 3,941,463 photos with free-text descriptions. In total, these photos come from 21,238 different users. A total of 2,681,901 empty captions appear in the dataset, which means more than 65% of the photos have no description at all. Furthermore, almost 697,000 captions contain the camera standard filename, and 19,337 of them are filenames entered by the user, such as *family.jpg, All the grandchildren.jpg,* or *Riding party.jpg*, etc.Since in Kodak Gallery the concept of tags per se does not exist, we preprocessed photo captions, extracting words and using them as tags. As already mentioned, an important number of photos have as caption their filename, and this leads to artifact tags which are quite popular, such as *img, jpg, copy*. In order to get a clearer idea of the actual words used by users, we have filtered them based on this observation. Additionally, as these tags are extracted from free text, stopwords are quite popular. We have therefore also removed stopwords from the list of tags, using the MySQL list of stopwords[1]. After preprocessing, the Kodak dataset counts 50,000 distinct words.

The Flickr dataset is made of 4,794,868 million photos from 32,751 users. The users were picked at random, by sampling a moment in time between December 2004 and April 2007 and getting the users who uploaded the first 4,000 photos from that point in time. These photos are tagged with roughly 23.9 million tags. For our study, we decided to keep an equivalent number of tags for each dataset. We ordered each dataset's tags by popularity (that is, the number of distinct users who employed them), and then kept the most popular 10,000 of them. We present in Table 1 some statistics of the two filtered datasets. For Kodak, we have a total of 900,000 tag occurrences from 5,400 users and 413,000 photos. For Flickr, we have a total of 13 million tag occurrences from 25,800 users and 4.6 million photos. We can see that the average number of photos per user is 76 for Kodak and 157 for Flickr, while the average number of unique tags per user is 34 for Kodak and 81 for Flickr.

[1]http://dev.mysql.com/tech-resources/articles/full-text-revealed.html#stopwords

Statistic	Flickr	Kodak
Total photos	4.6M	413,000
Total tag occurrences	13M	900,000
Total users	25,800	5400
Photos / user	157	76
Unique tags / user	81	34

Table 1: Statistics of Flickr and Kodak vocabularies.

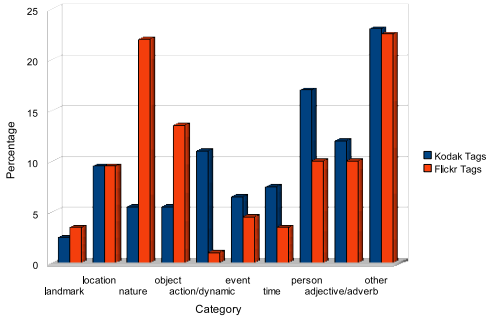

Figure 1: Ten-category tag taxonomy and the percentage of the top 200 most popular tags that belong to each category.

3. SPEAKING THE SAME LANGUAGE?

In order to be able to understand the two vocabularies better, we manually annotated the most popular 200 tags of each vocabulary. We designed a simple taxonomy of 9 categories (*landmark, location, nature, object, action/dynamic, event, time, person, adjective/adverb*), and a 10^{th} catch-all one, labeled *other*. The distributions of tags over categories for both datasets are shown in Figure 1. First off we notice that roughly 23% of each vocabulary falls in the *other* category, which is a reflection of the wide variety of subjects. The two vocabularies also show comparable tag frequencies for three other categories, namely *landmark* (with tags *church, bridge, house, building*, etc.), *location* (tags *home, street, museum, city,* etc.), and *adjective/adverb* (*cute, black, green, happy,* etc.). In contrast, the remaining categories display quite important differences between the two vocabularies: *nature* is represented 4 times more often in the Flickr vocabulary than in the Kodak one, while tags belonging to the *action/dynamic* category appear 5 times more often in the Kodak vocabulary than in the Flickr one. Flickr also shows a higher percentage of *objects*, at around 13%, as opposed to just about 5% in Kodak. Tags belonging to the *time*, *event*, and *person* categories appear much more frequently in the Kodak vocabulary.

While these statistics are computed on only the top 200 tags of each vocabulary, they are likely a good indicator of the inherent differences between the two sets. At the larger scale of our dataset, Kodak photos are more about *events*

Flickr		Kodak	
Tag	**% users**	**Tag**	**% users**
Category: **landmark**			
bridge	21.5	house	8.6
house	20.7	bridge	3.9
church	19.7	church	3.4
Category: **nature**			
sunset	30.9	beach	7.8
beach	29.8	water	5.4
tree	28.7	tree	5.0
Category: **action/dynamic**			
work	13.0	ride	4.8
dance	10.8	playing	4.6
		waiting	4.4
Category: **event**			
christmas	24.2	birthday	7.8
birthday	21.2	party	7.4
party	21.1	christmas	7.1
Category: **person**			
family	22.4	i	14.8
me	22.1	family	10.8
portrait	19.8	mom	10.4
Category: **adjective/adverb**			
red	27.4	big	8.1
blue	26.8	happy	7.2
green	25.5	good	6.7

Table 2: Top 3 words per category, and percentage of users using them for the two vocabularies.

and *persons* taking part in them, while many of the Flickr photos seem to be about *nature*. Also, because of the fact that in Flickr tags are used as search keywords, there is a higher number of content descriptive tags, most of which belong to the *object* category. In other words, the "Kodak Moment" concept (family events) and the "Flickr Diamond" one (artistic photos) do show up in the data when taken at large scale. This result therefore backs up the results from [7] but with several orders of magnitude more data and users.

As an illustration, we show in Table 2 the most popular three tags for some of the categories. For some categories, the most popular tags are common to Flickr and Kodak users. This is the case for *locations*, *events*, and *nature*. Some differences can be observed for *action/dynamic*, where Flickr has only two tags in the top 200, as well as for *adjective/adverb* where, in contrast to Kodak tags which mainly relate to persons, Flickr tags are dominated by color names.

Going back to the full 10K vocabularies, we are interested in understanding how they compare at the word level. Interestingly, we observe a linear relationship between the size of the vocabulary and the amount of overlap (not shown for space reasons), with overlaps between 50-60%, and the overlap for the full vocabularies at 56%. This shows that although most words are common to the two vocabularies, a significant amount of words is different across the datasets. A look at the most popular "missing" tags from each vocabulary shows tags like *macro, selfportrait, blackandwhite, photoshop, insect, flickr, abigfave, impressedbeauty, geotagged* missing from the Kodak vocabulary, while the Flickr one misses tags like *enjoying, lots, put, showing, giving, checking, heading, loved, weeks, visiting, dressed, wearing*. The first set of tags represents, more or less, Flickr jargon: photography techniques (*macro* and *blackandwhite*), Flickr groups' tags (*abigfave, impressedbeauty*), and "modern photographer"

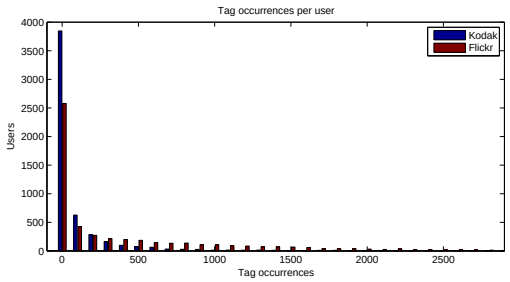

Figure 2: Number of tag occurrences by user split by their belonging to either the Kodak or Flickr dataset.

related activities, such as *geotagging*. The second set of tags is clearly dominated by *action/dynamic* tags, a by-product of the free-text descriptions of the Kodak dataset and the general orientation of the Kodak users towards events and people.

4. TOPIC-BASED USER ANALYSIS

Flickr, through its users and its groups, has been previously analyzed in the context of topic models [5, 3, 9]. In particular, in our own previous work [9] we have modeled a subset of Flickr of similar scale using a probabilistic topic model and compared two types of entities, namely Flickr users and Flickr Groups. In contrast to that work, which used Probabilistic Latent Semantic Analysis (PLSA), we use in this study Latent Dirichlet Allocation (LDA), a generative probabilistic topic model proposed by Blei et al. [2]. Additionally, we jointly model Kodak and Flickr users, merging the two datasets at the user level.

LDA is a fully generative model that assumes that documents in a corpus are a low-dimensional mixture of hidden topics of interest. LDA learns, in an unsupervised way, a word-topic and a topic-document distribution from the corpus. Because exact inference in LDA is known to be intractable, we used Gibbs sampling with 5000 iterations, as proposed in [4]. The last sample is used to compute the word-topic and topic-document distributions. In our case documents are Flickr and Kodak users, represented as bags-of-tags. As we have seen previously, the two vocabularies only share 56% of the tags, therefore the joint vocabulary is composed of almost 15,000 words. In order to avoid an unbalanced dataset due to the much larger number of Flickr users, we randomly sample 5,400 users from Flickr to match the 5,400 users from Kodak, and we thus have a total of 10,800 users.

In Figure 2 we show the histograms of tag occurrences per user, split by their original dataset. Kodak users have a median of 34 tags, and standard deviation 525, while Flickr users have a median of 124 tags, with standard deviation 867. These statistics show there is a high variability even within each of the two datasets. In terms of vocabulary size per user (number of unique tags), Kodak and Flickr users have a median of 8 and 15 tags respectively, with standard deviations of 94 and 148.

We trained the LDA model using a bags-of-tags representation for all users, counting for each of them the number of times any given tag was used. The model parameters are the number of topics $T = 150$, the parameters of the Dirichlet priors on the per-user topic distribution $\alpha = 50/T$, and the per-topic word distribution $\beta = 0.01$. Although we trained

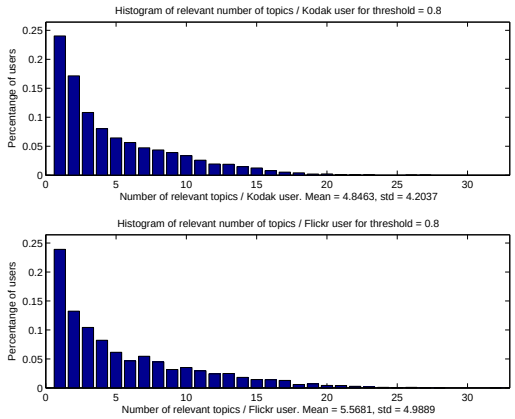

Figure 3: Number of relevant topics per user, for the Kodak (top) and Flickr (bottom) datasets.

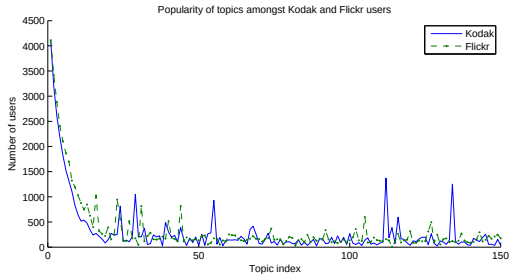

Figure 4: Topic popularity among users of the two communities. Popularity is computed from the number of times a topic is deemed as relevant for a user.

a joint model, we hypothesize that the inherent differences of the two populations should show up at the topic level.

As the output of the LDA model we have the distributions over topics for each user, or $P(Z \mid U)$, where Z and U represent the hidden topics and the users respectively, as well as the distributions over words for each topic, denoted by $P(W \mid Z)$, where W represents the words, and Z the hidden topics. For each user it is then possible to compute which are the most relevant topics, by setting an arbitrary threshold $\tau = 0.8$ on the cumulative sum of the most probable topics. We show in Figure 3 the histograms of the number of relevant topics per user for each of the two datasets. We observe that Kodak users are more likely to have fewer topics than their counterparts from Flickr. On average Kodak users are about 4.8 topics, while Flickr users are about 5.6. This difference is statistically significant at $\alpha = 0.001\%$ in a two-tailed test.

In Figure 4 we show a plot of the topic popularity among Flickr and Kodak users. A given topic's popularity is represented by the number of times this topic is kept as relevant for any given user. Here we see that the model learned a few very popular topics, which score very high for both datasets. However, it is also interesting to observe that around 12 topics are much more popular with Flickr users, and also 5 other topics are much more popular with Kodak users. From the distributions $P(W \mid Z)$ we can extract the most probable 10 words for some of these "special" topics. Two of the mainly-Flickr topics are topic #23, which is characterized by the words *red, blue, green, sky, white, light, yellow, black, orange, clouds*, and topic #16, defined by the words *sky, night, clouds, sunset, water, light, trees, sun, tree, lights*. From the Kodak camp, topic #134 is about *mom, dad, edited, grandma, aunt, grandpa, uncle, jim, taylor, cake*, and topic #55 about *john, chris, irene, david, mary, fiesta, scott, kim, eric, dan*. We obviously have in this case the nature oriented users of Flickr and the family oriented users of Kodak.

5. DISCUSSION

We presented in this study a novel large-scale analysis of Flickr and Kodak Gallery, two large online photo-sharing communities which so far have not been jointly analyzed. We have observed, despite inherent differences induced by the underlying users, their motivations and their needs, as well as system design and affordances, certain similarities at the vocabulary level, as well as at a more abstract, semantic level, through topic modeling. At the same time, as Miller and Edwards [7] found in their 10 user study, we have verified, through the joint topic modeling of our two datasets with 5 million images and more than 10,000 users, that we can talk about two types of emerging phenomena: *Flickr Diamonds*, the product of *Snaprs* who tend to take photos and share them with the world, and *Kodak Moments*, a product of the *Kodak Culture* users, who take photos mostly at family events, and mainly share them within their existing social circle.

We believe that this study also points out one of the great potentials that social computing holds for future research. Large-scale studies are nowadays much easier to perform, unlike ethnographic studies which are usually small-scale, given their time and subject-effort intensive nature. Social computing may therefore become the first step in the research process, with in-depth ethnographic studies as a second step once a preliminary hypothesis has been chosen for verification.

Acknowledgements

This research has been supported by the Swiss National Science Foundation through the MULTI project. The authors would like to thank David Anantharajan of Kodak Gallery for his help with the dataset.

6. REFERENCES

[1] M. Ames and M. Naaman. Why We Tag: Motivations for Annotation in Mobile and Online Media. In *CHI '07*, San Jose, CA, USA, 2007.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal Machine Learning Research*, 3, 2003.

[3] M. De Choudhury, H. Sundaram, Y.-R. Lin, A. John, and D. Duncan Seligmann. Connecting Content to Community in Social Media via Image Content, User Tags and User Communication. In *ICME '09*, New York, NY, USA, 2009.

[4] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, April 2004.

[5] K. Lerman and L. Jones. Social Browsing on Flickr. In *ICWSM '07*, Boulder, CO, U.S.A., March 2007.

[6] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic Evolution of Social Networks. In *KDD '08*, Las Vegas, NV, USA, 2008.

[7] A. D. Miller and W. K. Edwards. Give and Take: a Study of Consumer Photo-sharing Culture and Practice. In *CHI '07*, San Jose, CA, USA, 2007.

[8] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *IMC '07*, New York, NY, USA, 2007. ACM.

[9] R. A. Negoescu and D. Gatica-Perez. Topickr: Flickr Groups and Users Reloaded. In *MM '08*, Vancouver, Canada, Oct. 2008.

[10] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in Folksonomies: Social Link Prediction from Shared Metadata. In *WSDM '10*, New York, NY, USA, 2010.

[11] N. A. Van House. Flickr and Public Image-sharing: Distant Closeness and Photo Exhibition. In *CHI '07*, San Jose, CA, USA, 2007.