IDIAP RESEARCH REPORT

TOPICKR: FLICKR GROUPS AND USERS RELOADED

Radu-Andrei Negoescu [1]
Daniel Gatica-Perez [1]

IDIAP–RR 08-61

JULY 2008

[1] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, {negora,gatica}@idiap.ch

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-61

Topickr: Flickr Groups and Users Reloaded

Radu-Andrei Negoescu   Daniel Gatica-Perez

July 2008

Abstract. With the increased presence of digital imaging devices there also came an explosion in the amount of multimedia content available online. Users have transformed from passive consumers of media into content creators. Flickr.com is such an example of an online community, with over 2 billion photos (and more recently, videos as well), most of which are publicly available. The user interaction with the system also provides a plethora of metadata associated with this content, and in particular tags. One very important aspect in Flickr is the ability of users to organize in self-managed communities called groups. Although users and groups are conceptually different, in practice they can be represented in the same way: a bag-of-tags, which is amenable for probabilistic topic modeling. We present a topic-based approach to represent Flickr users and groups and demonstrate it with a web application, Topickr, that allows similarity based exploration of Flickr entities using their topic-based representation, learned in an unsupervised manner.

1 Introduction

Information management systems face a tough challenge in the wake of social media repositories involving images, video, and text. The amount of data is enormous and users create, by interacting with media, additional metadata such as tags, comments, and ratings. Users of such systems are now producing, viewing, sharing and repurposing content in a number of different social scenarios. This presents the multimedia community with the unique opportunity of leveraging on the data that collective behaviors of people interacting with content bring to light [5, 6]. One example of a social media system is Flickr.com, which hosts over 2 billion images [1]. Users of Flickr organize themselves in self-managed communities with common standards and interests, called groups. Users participate in groups by sharing and commenting on photos, most often on specific topics or themes, like a popular event, location, or photographic style. Such collective behaviors offer viable new alternatives to understand, represent, and manage visual content [7].

There has been previous work analyzing Flickr data, and in particular tags. Marlow et al. [6] have analyzed tagging systems in general, and a taxonomy of users' motivations to tag has been proposed by Ames and Naaman in [3]. Some other studies have analyzed the sharing practices, motivations, and privacy concerns of users [2, 10]. In [7], we proposed a topic-based representation for groups and showed it allowed a novel way of characterizing and searching for groups. Users however were not included in our analysis and the model was limited to groups.

Although users and groups are fundamental components of Flickr, their interrelations are, to our knowledge, not completely understood or fully exploited. Ideally, we would like to be able to discover similar users or groups beyond direct tag-based strategies. The use of higher-level information (e.g. topics) could be an attractive alternative.

In this paper, we propose that groups and users be treated as equal *entities*, and a common representation be attempted, allowing direct comparison between groups and users. The key concept is the (bold, and clearly simplifying) assumption that groups and users in Flickr can be reasonably modeled as if they were equivalent entities, and that their direct joint modeling is beneficial despite the complex ways in which Flickr groups are created through users' contributions. First, we jointly analyze Flickr groups and users from the perspective of their tagging patterns. Based on a snapshot of the Flickr collection, our analysis reveals a number of fundamental similarities, as well as differences, with respect to vocabulary size and vocabulary overlap between the two types of entities. Second, we propose a joint user-group topic-based representation, which is learned in a probabilistic, unsupervised manner, from the groups' and users' tags. We demonstrate, using a simple web application, that our topic-based representation facilitates the discovery of expert entities for specific topics (be they users or groups), it allows the creation of new methods of group and user discovery, and it is also useful for further structural analysis of the groups and users at a higher semantic level. We therefore contribute to a better understanding of the behavior of communities, as well as to the development of useful applications.

The paper is organized as follows. Section 2 presents our analysis of the users' and groups' tag usage, which motivates a joint model over these two types of entities. Section 3 introduces our proposed topic-based user and group representation. Section 4 presents a topic-based analysis of Flickr groups and users, highlighting some of its further uses, and in particular the Topickr application. Finally we present our conclusions in Section 5.

2 The Users-Groups Ecosystem

We have collected the data used in this study using Flickr's API. All the information extracted about a particular user is publicly available, but statistics linked to the number of photos and tags may vary if users employ restrictive privacy settings for their photos. We had no access to this private data.

Our dataset consists of approximately 22,000 registered Flickr users, roughly 7 million photos belonging to these users (the most recent 500 photos per user), and about 23 million tags belonging to these photos.

For this study we filtered the original dataset in a number of ways. We concentrated on a vocabulary of the most common 10,236 tags by removing tags that contained among others numeric characters (e.g. dates or years), or that were used by only one user. Further constraints were imposed on the groups and users, more specifically, a vocabulary overlap of at least 150 tags (i.e. the group or user bag-of-tags should contain at least 150 unique tags from the vocabulary, a mere 1.5% vocabulary overlap). We can summarize this dataset D_R as follows:

unique tags	$T = \{T_i\}$ with $N_t = 10,236$
users	$U = \{U_i\}$ with $N_u = 6,144$
groups	$G = \{G_i\}$ with $N_g = 8,786$
photos	$P = \{P_i\}$ with $N_p = 766,056$

The total number of tag occurrences for users is roughly 46 million, and for groups about 30 million.

We previously analyzed [7] the photo sharing practices of users and devised a novel representation for Flickr groups using a Probabilistic Latent Semantic Analysis (PLSA) model on group tags in order to obtain a topic decomposition for each group. Considering that the group tags are contributed by users who are members of the group by adding photos to the group pool, a joint topic representation for groups and users appears to be not only justified but also useful for direct comparisons between these two types of entities.

We examine both tag occurrences and unique tags in the dataset D_R. Figure 1 shows the cumulative sums over unique tags and tag occurrences for both types of entities. Both groups and users follow similar curves for the distribution of unique tags, but that is not the case for the total number of tags. Thus, for a vocabulary size of at most 2000 tags we obtain 95% of the groups and 98% of the users. On the other hand, 41% of the users have more than 5000 tag occurrences, while for groups the percentage is much smaller, about 14.7%. These numbers confirm our earlier observations that users contribute only a relatively small part of their collections to groups, however it is very interesting to observe that these contributions create comparable tag vocabularies for groups. Based on these observations a joint model of the two types of entities seems justified. We will describe this model next.

3 Joint PLSA Model for Users and Groups

We can think of Flickr *entities* (groups and users) as being a collection of text documents. The content of these documents are the tags associated with the entity photos. An intuitive way to describe a text document is by considering the different topics it is about. These topics are not always explicit but can be derived from the document and represent an accurate and compact summary of the original content.

PLSA [4] assumes the existence of a latent topic z_k ($k \in 1, ..., N_z$) in the generative process of each tag t_j ($j \in 1, ..., N_T$) in an entity E_i ($i \in 1, ..., N_E$). Each occurrence t_j is independent from the document it belongs to given the latent variable z_k. This corresponds to the joint probability:

$$P(t_j, z_k, E_i) = P(E_i)P(z_k \mid E_i)P(t_j \mid z_k). \tag{1}$$

The joint probability of the observed variables is the marginalization over the N_z latent topics z_k:

$$P(t_j, E_i) = P(E_i) \sum_{k}^{N_z} P(z_k \mid E_i)P(t_j \mid z_k). \tag{2}$$

In our model this is equivalent to the following generative process: an entity E is selected, then a hidden topic z_k is sampled from $P(z \mid E)$. Given topic z_k, a tag t_j is selected based on $P(t \mid z_k)$.

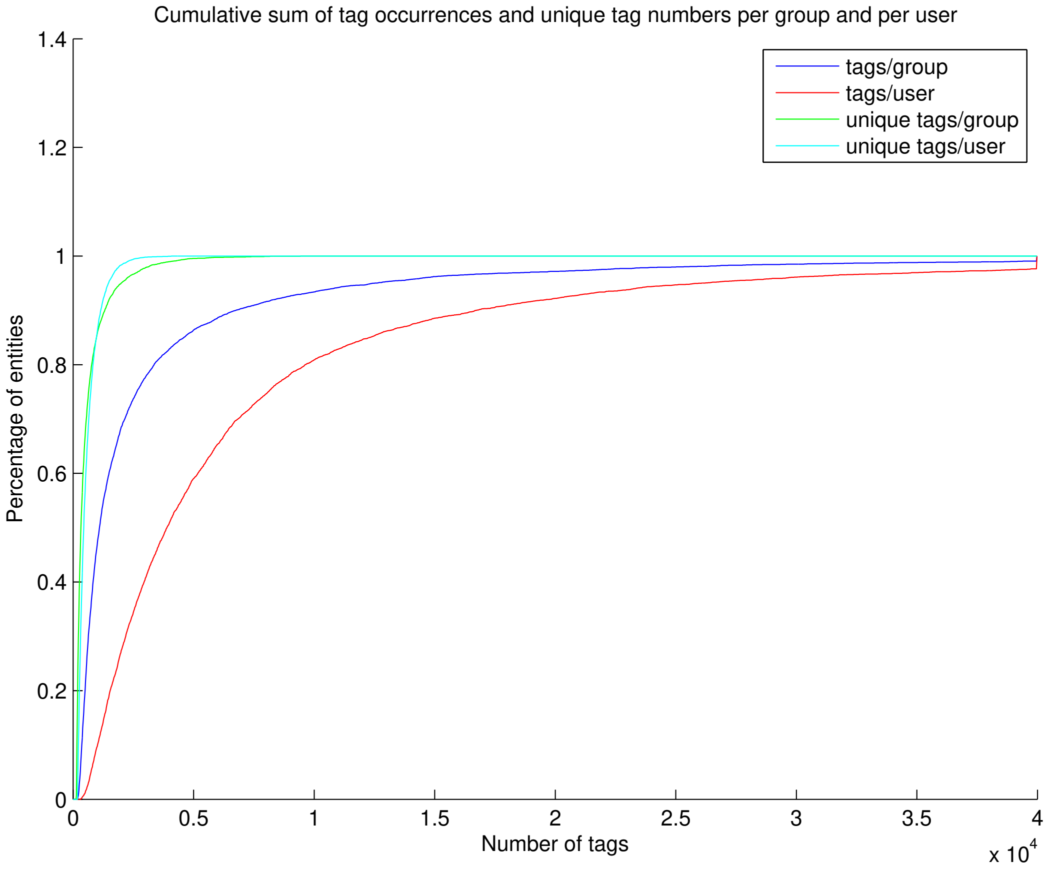

Figure 1: Comparison of the unique and total number of tags for groups and users.

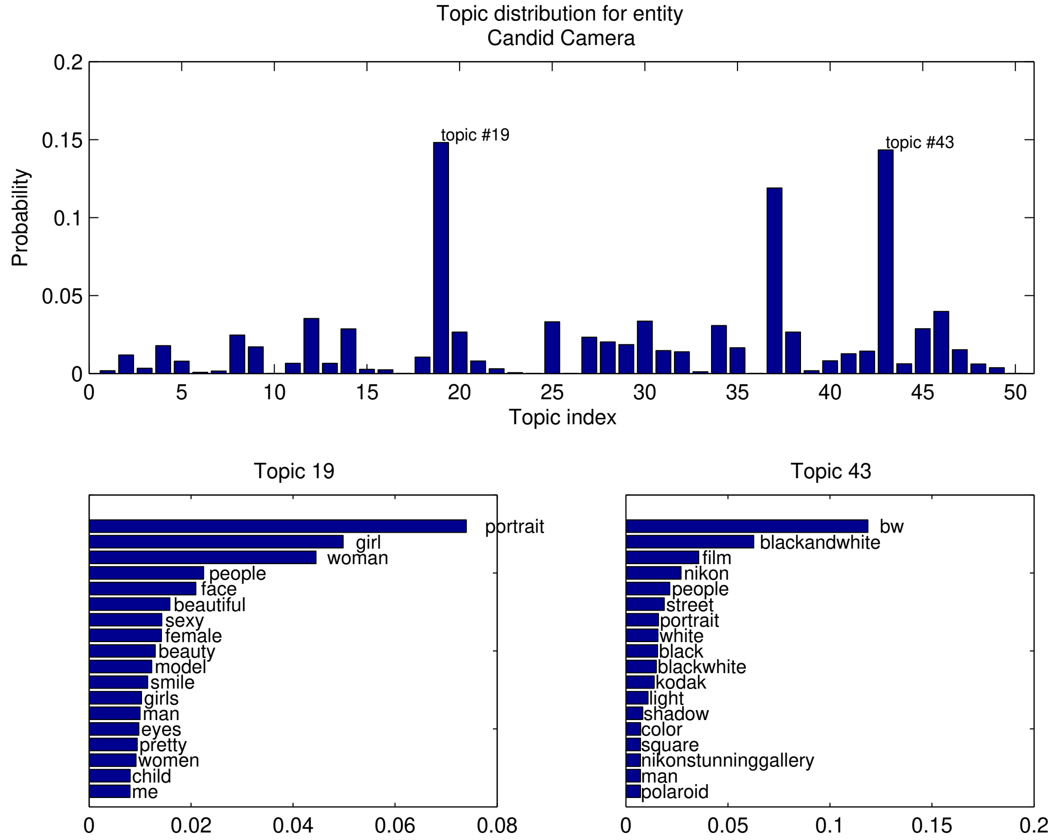

Figure 2: Topic distribution for the entity *Candid Camera* (a Flickr group).

The conditional probability distributions $P(t \mid z_k)$ and $P(z \mid E_i)$ are multinomial given that both z and t are discrete random variables. The parameters of these distributions are estimated by the Expectation-Maximization algorithm [4].

The two steps of the algorithm are the following:

E-step: the conditional probability distribution of the latent topic z_k given the observation pair (E_i, t_j) is computed from the previous estimate of the model parameters.

M-step: The parameters of the multinomial distribution $P(t \mid z)$ and $P(z \mid E)$ are updated with the new expected values $P(z \mid E, t)$.

We represent each group G_i and each user U_i (from now on referred to as entity E_i) as bags-of-tags, i.e. vectors $t_i = (t_{i1}, ..., t_{ij}, ..., t_{iN_t})$ of size N_t (the number of distinct tags). Here t_{ij} represents the number of times tag j occurs in entity's E_i bag-of-tags. The PLSA model described above is trained on the bag-of-tags representation of groups and users regardless of their type. We show in Figure 2 the topic distribution for a group called *Candid Camera* and an example of some of the learned latent topics is shown in the lower half of Figure 2.

Other topic-based formulations that involve (implicitly or explicitly) the existence of users and groups characterized by their content have been proposed in the text modeling literature [9, 11]. To our knowledge, none of these options have been investigated to model Flickr groups and users and their content. While these options are potentially interesting, the complexity of most of these models is higher and their applicability (e.g. in the case of the group-topic model [11]) might not be straightforward given the type of user-to-group membership evidence that is assumed.

In contrast, we advocate for a simpler computational modeling option that is nevertheless powerful. The key idea is the assumption that groups and users in Flickr can be reasonably modeled as if they were comparable entities and that their direct joint modeling is beneficial despite the complex ways

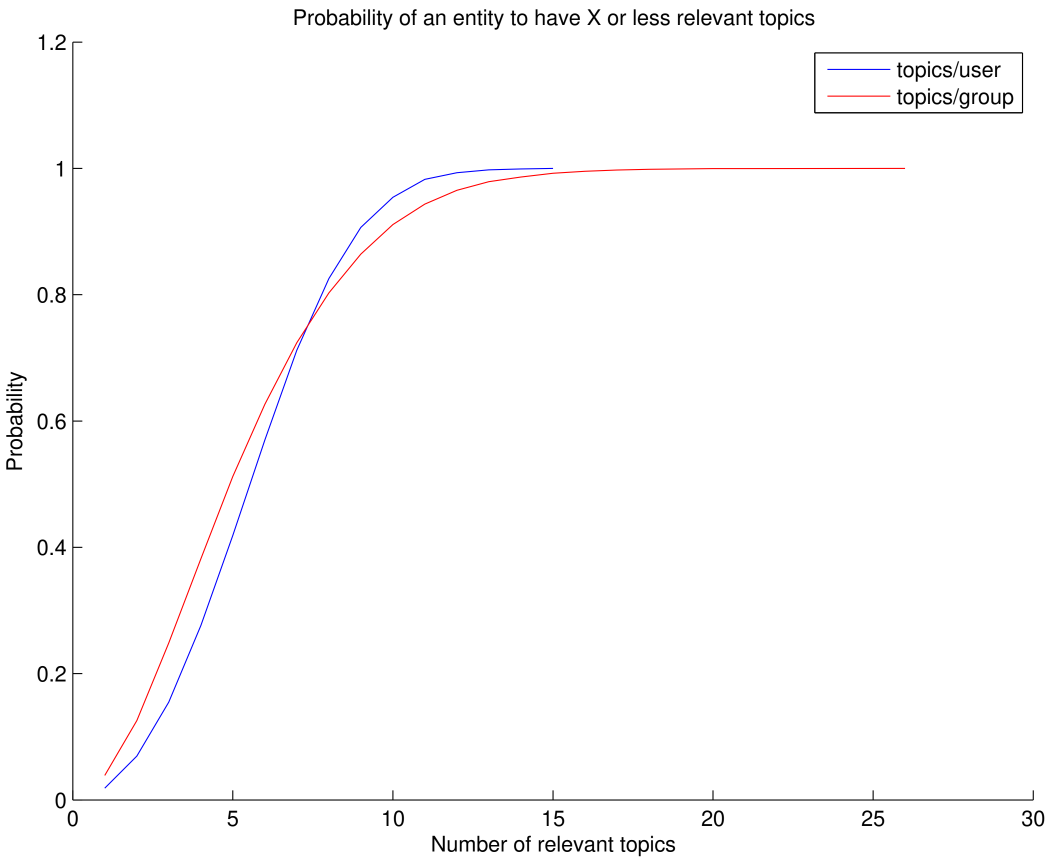

Figure 3: Probability of an entity to be defined by x or less relevant topics.

in which Flickr groups are created.

4 Applying User-Group Topic Representations

Apart from the obvious advantage of being able to directly compare entities, another benefit of the topic representation is the fact that entities can now be characterized by their topic distributions. This means that the fewer relevant topics in the entity distribution, the higher that entity's concentration on specific (possibly photographic) themes is, and vice versa. We show in Figure 3 the probabilities of the number of relevant topics for users and groups in the jointly learned topic model for $N_z = 50$. In this context, relevant is a term that describes the topics (ranked by their probability mass) that account for an accumulated amount $\epsilon = 0.8$ of the total probability mass. It can be seen that users have a maximum of 15 topics, while some groups may even have more than 20 topics. The 9 groups with more than 20 topics are what we call *catch-all* groups, like *FlickrCentral*, *Utata*, *10 Million Photos*, and *The Biggest Group! - Playground for Psychotics!*, which are very large in terms of members and tag vocabularies. About 1% of groups have vocabularies larger than those of any of the users. Not surprisingly, the number of relevant topics in their topic decompositions is also relatively high. Figure 4 shows the histogram of the relevant topics in the decompositions of this 1% subpopulation (85 groups), with most groups having more than 10 topics.

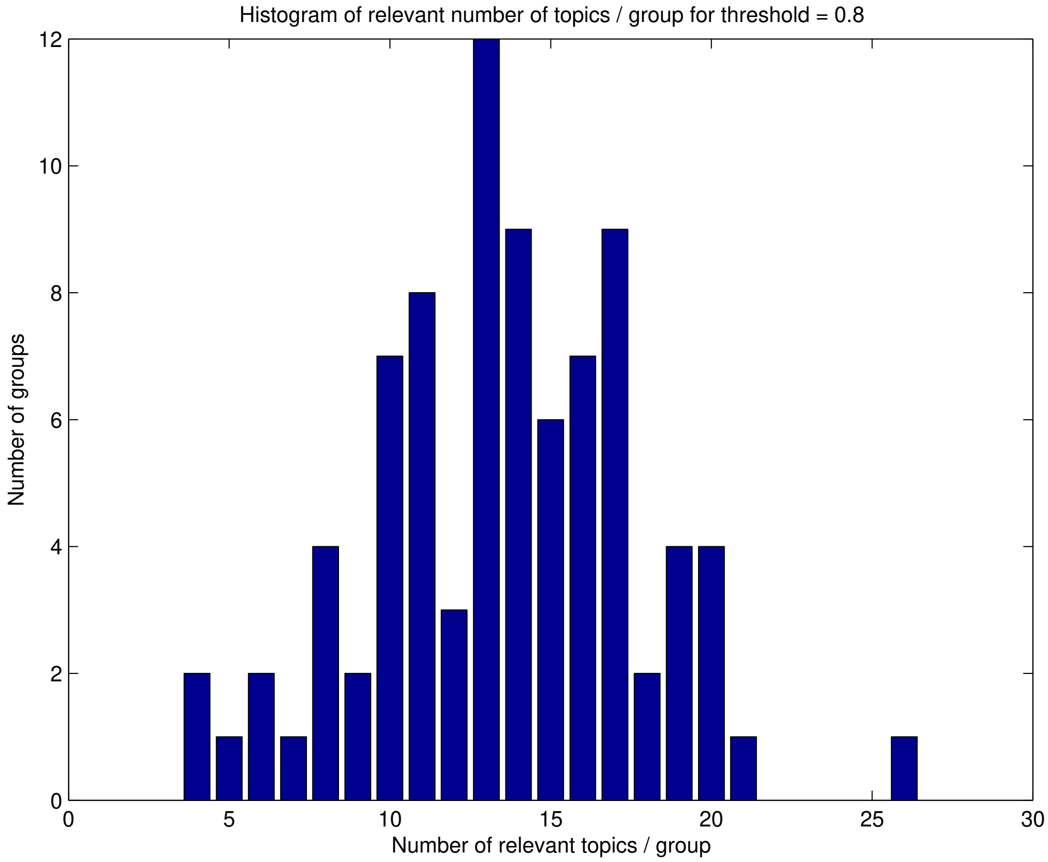

Figure 4: Histogram of relevant number of topics in the topic distributions of the 85 “big vocabulary” groups.

Figure 5: A latent topic learned by the PLSA model, and expert entities (mixed users and groups), groups, and users (not already selected in the top entities) for this specific topic.

Topickr: Similarity-based Exploration

One of Flickr's most addictive features is the opportunity to explore quasi-random photographs. Using a proprietary algorithm Flickr provides a ranking measure called "interestingness", which is then used to display photos from people the user doesn't necessarily know.

This exploration mechanism can be very well used with our topic-based representation model. Instead of ranking photos, we rank users and groups with respect to each other based on a similarity measure. Because users and groups share the same representation, i.e. the topic decomposition, a simple similarity measure can be obtained from the distance between two distributions. Having computed the topic distribution for every entity in our dataset, we pre-compute the Bhattacharyya pair-wise distances for the approximately 15,000 entities and store them in a MySQL database. Our Topickr application[1] allows us to explore the topic model visually: starting from any given topic in the model, described by the most characteristic tags, we present the most probable expert entities for that topic. We also present the most probable group experts and the most probable user experts in order to provide balanced exploration options, selecting groups and users that hadn't already been selected in the top 10 entities. Some entities have a very spiky topic distribution, showing a strong interest in a certain photographic technique or subject, others have a few more topics, while others have a more uniform distribution over topics, probably indicating a very wide range of interests or no particularly strong interests. Figure 5 shows topic 6, which seems mostly related to food, and a few of its top entities (mixed users and groups), top groups, and top users, which hadn't already been selected in the top entities. For this particular topic, there is an overwhelming domination of groups in the top 10 entities. A possible explanation is that probably such groups are much more focused on the content than the individual users, who may have photographic side interests as well. For any given entity (be it a group or a user) we can then see the most similar users and most similar groups. We use groups' and users' Flickr icons for display and allow the user to either select an entity for topic-based exploration or to go directly to that entity's Flickr page.

5 Conclusions

We have shown in this paper that a joint topic-based representation for Flickr groups and users is beneficial despite inherent differences between these two types of entities. We have proposed an application that allows similarity-based exploration and recommendation. Topickr shows some of the advantages of the topic-based approach model, particularly the ability to provide suggestions to users about groups or users they might be interested in exploring. While the most straightforward way is to recommend groups or users with very similar topic distributions (particularly useful when the user is looking for groups or users similar to a group or user they already know) the system can also recommend the topic experts for the relevant topics of the current user.

In our future work we intend to investigate extensions of topic models that will allows us to scale up our application and handle large-scale data, as studied by Newman et al. [8].

[1] see demo at http://www.idiap.ch/~negora/acmmm08

Acknowledgements

This research has been supported by the Swiss National Science Foundation through the MULTI project. The authors would also like to thank Dr. Christopher McCool (Idiap) for his insightful comments.

References

[1] Flickr Blog, 13 Nov. 2007. http://flickr.com/blog.

[2] S. Ahern, D. Eckles, N. S. Good, S. King, M. Naaman, and R. Nair. Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In *CHI '07: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, San Jose, CA, USA, 2007.

[3] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, San Jose, CA, USA, 2007.

[4] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001.

[5] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections. In *MULTIMEDIA '07: Proc. of the 15th ACM Intl. Conf. on Multimedia*, Augsburg, Germany, 2007.

[6] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proc. of the 17th Conf. on Hypertext and Hypermedia*, 2006.

[7] R. A. Negoescu and D. Gatica-Perez. Analyzing Flickr Groups. In *CIVR '08: Proc. of the Intl. Conf. on Image and Video Retrieval*, Niagara Falls, Canada, July 2008.

[8] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed inference for latent dirichlet allocation. In *NIPS'07: Proc. of the 21st Conf. on Advances in Neural Information Processing Systems*, Cambridge, USA, 2007.

[9] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Auai '04: Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence*, Arlington, Virginia, United States, 2004.

[10] N. A. Van House. Flickr and public image-sharing: distant closeness and photo exhibition. In *Chi '07: Extended Abstracts on Human Factors in Computing Systems*, San Jose, CA, USA, 2007.

[11] X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and their attributes. In *Nips'05: proc. of the 19th conf. on Advances in Neural Information Processing Systems*, Vancouver, Canada, 12 2005.