Flickr Hypergroups

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

Brett Adams
Curtin University of Technology
Australia
b.adams@curtin.edu.au

Dinh Phung
Curtin University of Technology
Australia
d.phung@curtin.edu.au

Svetha Venkatesh
Curtin University of Technology
Australia
s.venkatesh@curtin.edu.au

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

ABSTRACT

The amount of multimedia content available online constantly increases, and this leads to problems for users who search for content or similar communities. Users in Flickr often self-organize in user communities through Flickr Groups. These groups are particularly interesting as they are a natural instantiation of the content + relations social media paradigm. We propose a novel approach to group searching through hypergroup discovery. Starting from roughly 11,000 Flickr groups' content and membership information, we create three different bag-of-word representations for groups, on which we learn probabilistic topic models. Finally, we cast the hypergroup discovery as a clustering problem that is solved via probabilistic affinity propagation. We show that hypergroups so found are generally consistent and can be described through topic-based and similarity-based measures. Our proposed solution could be relatively easily implemented as an application to enrich Flickr's traditional group search.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Human Factors

1. INTRODUCTION

Flickr – a hugely successful social photo management site – had in March 2009 more than 30 million user accounts, who had uploaded and tagged more than 3 billion photos. One of the flagship features in Flickr are *Groups*, that are self-organized user communities. One single example, *Flickr Central*, has at the time of this writing almost 93,000 users and over 2 million photos in its photo pool. This is a somewhat extreme example that nevertheless shows a system-wide reality: users not only upload photos to Flickr, but they also participate in a number of social scenes, usually by tagging their photos and sharing them with groups based on their interests and various social motivations [8, 1]. The way Flickr members make use of the Groups feature is particularly attractive, as it naturally brings together two key aspects of social media: content and relations. Flickr groups have just begun to be studied [7, 8, 5], and overall group dynamics on Flickr are not yet completely understood. Moreover, with more than 200,000 groups, exploring them is not very easy either.

This paper proposes a novel method to discover *hypergroups* in Flickr, i.e., communities consisting of groups of Flickr groups. Our hypothesis is that groups that are similar probably host the same kind of content (in terms of images and associated tags), and depending on their popularity, they may also share an important number of members. Based on this observation, our work has three contributions. First, starting from almost 11,000 groups, we propose to use these two sources of information, content (through photo tags) and relations (through group memberships) in a bag-of-words model to represent groups in Flickr. In particular, we propose a novel angle to modeling relations. While traditional approaches to social networks have mainly examined a user's explicit contacts, participation in the same groups can also be viewed as an implicit social link; this is how we will approach relations in this paper. Second, using a probabilistic topic model, we build three comparable topic-based representations, one based on content, one based on binary membership links, and a hybrid, based on membership links weighted by the content-wise contributions of the user to the group. Third, we employ a state-of-the-art clustering algorithm that discovers cohesive hypergroups, and analyze and compare our methodology from several viewpoints. Overall, our approach provides a prototype solution to the problem of how users can find interesting groups, as it allows users to find potentially obscure groups, that are still relevant to their search, based on how similar the target groups are to an example group a user would provide.

MM'09, October 19–24, 2009, Beijing, China.

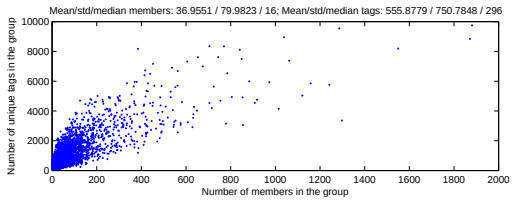

Figure 1: The number of members vs. the number of tags for each group. The bigger the group in terms of members, the larger the group tag vocabulary.

We briefly describe Flickr groups in Section 2. Our approach is described in detail in Section 3, and the analysis of our method's results is presented in Section 4. Finally, we conclude in Section 5.

2. FLICKR GROUPS

Flickr groups are self-managed, user-created communities revolving around a common interest or goal. They can be **geographical** in nature, bringing together users photographing the same area, such as *New York Photography* or *Alaska*, **thematic**, centered on a specific photographic technique or subject, such as *Insect Macro Photography* or *Concert Photography*, or simply **social**, with no other purpose than to bring together people from all over the world, such as *FlickrCentral*. A fourth, rather distinct type of groups, is what we may call **exposure and awards** groups, such as *Views 2000* or *Better than Good (Invited Images Only–Give 2 Awards)*, focused on the number of views a photo is exposed to, and/or on the perceived quality of a photo. Of course these categories are not a full taxonomy, and within a group they often overlap, but they serve as a general indication of the purpose of a group.

In our recent work [7, 8] we have shown that users who like sharing photos with groups do so significantly, and that users will often share the same photo with a number of groups. This is one of the reasons for our hypothesis: similar groups will have the same kind of content, and sometimes even the same users. By clustering together groups based on their similarity, we could thus find relevant hypergroups.

The dataset used in this paper was obtained from [8], and it consists of 10,800 groups and a sample of their members, for a total of 8,000 users. These users contribute more than 1 million photos to the groups. The total number of tags in the group photo pools is around 38.6 million. Similar to [8], we have only kept tags that appeared in a list of 10,236 tags.

A quick look at the groups' structure reveals a large correlation (with correlation coefficient $r = 0.8195$) between the number of members in a group and the number of unique tags in the group's vocabulary (see Fig. 1), that is, bigger groups in terms of members tend to also have larger tag vocabularies. This is not surprising, as more users may tag differently the same kind of content than fewer users. It is however an early indication that using membership information may be useful in modeling groups.

3. OUR APPROACH

Finding groups in Flickr is relatively easy for popular groups whose names and/or descriptions include the keywords used for searching. However, when these keywords are not present, or when the group is not very popular, finding groups can be problematic. We propose as solution to this problem the automatic discovery of hypergroups, or groups of groups, a process that allows a user to find interesting groups starting from one group he or she considers relevant.

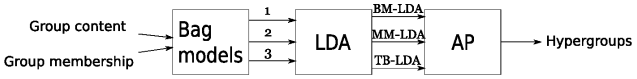

Figure 2: Hypergroup discovery: from group content and membership we create bags of words, then we learn LDA models for each bag model, and finally we obtain hypergroups through AP.

In order to test our hypothesis (similar groups have similar content and/or members), we develop three topic models: one based on a bag-of-tags representation for each group, and the other two on two different bag-of-members representations. We will describe these into more detail in the following subsections. The conceptual workflow for hypergroup discovery is illustrated in Fig. 2. We start by creating bag representations for the groups, which are then used to learn probabilistic topic models. Finding hypergroups is then cast as a clustering problem.

3.1 Bag Models

We construct three bag representations for the documents in our corpus, namely the groups:

1. a bag-of-users representation, by counting once each member of a given group; this is a binary-membership bag;
2. a bag-of-weighted-users representation, by counting for each user in a group all the unique tags they contributed to the group; thus this represents a membership bag too, but weighted by content, with multiple occurrences for the same user based on his or her contribution to the group vocabulary;
3. a bag-of-tags representation, by counting all the occurrences of a given tag in a given group's photo pool.

These three representations are then used for three different topic models, using Latent Dirichlet Allocation (LDA) [2].

We shall name these models BM-LDA for binary-membership, MM-LDA for multiple-occurrence membership, and TB-LDA for the tag-based representation respectively, and we will describe them next.

3.2 Content and Membership LDA

Latent Dirichlet Allocation (LDA) is a fully generative probabilistic model [2] that works under the assumption 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 a document corpus. The latter can be used to represent a document based on its topic distribution. Because exact inference in LDA is known to be intractable, we use Gibbs sampling as proposed in [4], with 5000 iterations for all three models. The last sample is used to compute the word-topic and topic-document distributions. In our problem, documents are Flickr groups.

For the membership-based representations we learned the LDA models starting from the two bags described in Section 3.1, i.e., binary-membership (BM-LDA), and multiple-occurrence membership (MM-LDA). The words in these two topic models are therefore users. The same number of topics is set as $N = 100$ hidden topics for both models. Each topic

is characterized by a probability distribution over users, so their meaning is linked to what those users have in common, in terms of co-occurrences. Each group is now characterized by a probability distribution over topics, given by $P(z_u \mid G)$, where z_u is the notation for the user-based topics.

For the content-based representation, we learned an LDA model with $N = 100$ hidden topics starting from the bag-of-tags defined previously. Each document is then characterized by a distribution over topics, given by $P(z_t \mid G)$, where z_t is the notation for the tag-based topics. In the case of this model, the learned topics are mostly topics of interest, described by semantically similar tags. As observed in previous work relying on similar models (PLSA) [6, 8], tag-based topics are quite consistent, and this is the case for the LDA model as well (results omitted for space reasons).

3.3 Clustering

We rely on a pairwise measure of similarity S between any two given groups starting from their topic-based representations. A few distribution measures were explored, including Kullback-Leibler divergence, and a parameterized (and thus generally asymmetrical) Jensen-Shannon divergence.

The similarity measure was calculated for every pair of groups, yielding a N_GxN_G similarity matrix, where N_G is the total number of groups. Hypergroup discovery is now cast as a clustering problem on this similarity matrix. Any number of clustering algorithms could be used, we chose the recently proposed Affinity Propagation method (AP) [3], as it has good properties: it is non-parametric, the number of clusters is automatically determined, and it does not assume the similarity function to be a metric, or symmetric. An additional benefit of AP is the discovery of *exemplars* as a by-product of the clustering process. Exemplars are the "most representative" members of a cluster, and hence provide a ready-made description of a hypergroup. For a detailed description of AP, we refer the reader to [3].

4. RESULTS

Our goal is to find through clustering hypergroups that bring together semantically similar groups that do not necessarily have the same keywords in their names or descriptions. But what is a good clustering outcome? Which model performs best? Since no ground truth exists for Flickr groups' similarity, results are inherently subjective. In this section we present some of the hypergroups discovered for each of the three models, then we analyze size and topic-driven statistics for each clustering outcome, and finally we propose and analyze a measure of homogeneity for hypergroups. For all results presented hereafter, the negative Jensen-Shannon (JS) divergence was used as the similarity measure: the smaller the value, the more similar two groups are.

First we show in Table 1 a couple of examples of hypergroups whose size is around the mean and median of each model's clustering outcome. Listed on the first line of each cell and in bold-face is the hypergroup exemplar (the group that defines the hypergroup), and listed under it are the other groups belonging to that hypergroup. We also show the number of members and the size of the vocabulary for each group. We observe that all models seem to produce relatively homogeneous hypergroups, with interesting results like the grouping of *RUSTY and CRUSTY* and *Things that Moved*, a group about "Past Tense. Things that moved but don't anymore. Broken down and retired vehicles. Planes, trains, automobiles, riding mowers, dead weasles, etc, etc, etc.", or *Toysaholic Anonymous* and *Urban Vinyl Fiend*, a group dedicated to "photographs of toys from the designer urban vinyl scene or toys with a flair."

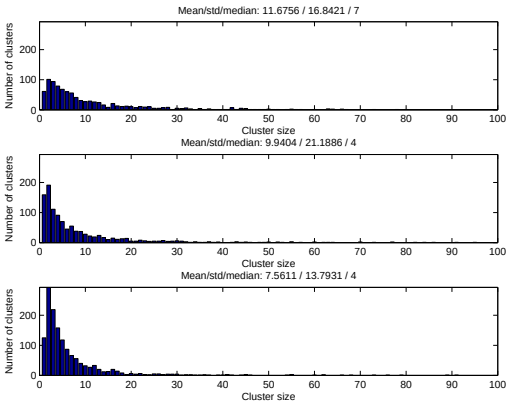

Figure 3: Histograms of hypergroup sizes for each of the models. Top: BM-LDA; middle: MM-LDA; bottom: TB-LDA. The latter two models generate more hypergroups of smaller sizes than the binary-membership model.

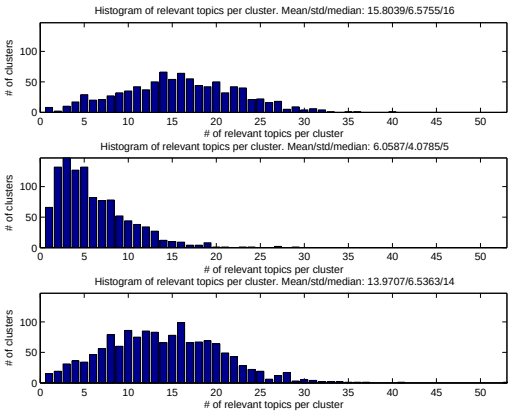

Figure 4: Histograms of the number of relevant topics per hypergroup for each model. Top: BM-LDA; middle: MM-LDA; bottom: TB-LDA.

Second, we look at some statistics of the discovered hypergroups. The total numbers of hypergroups for each model are 928 for BM-LDA, 1090 for MM-LDA, and 1433 for TB-LDA. In Fig. 3 we show the histogram of hypergroup sizes for the three models. We observe that MM-LDA and TB-LDA tend to generate more hypergroups of rather smaller sizes (medians of 4 as opposed to 7 for BM-LDA). A double tail t-test at $\alpha = 0.01$ for all three models shows that, despite these apparent differences, hypergroup sizes for the two membership-based models are likely to have been drawn from the same distribution, while the sizes for the tag-based model TB-LDA are significantly different.

Starting from the LDA representations, we define *relevant topics* for a group to be those topics that account together for over 80% of the probability mass in its topic-based representation. The number of relevant topics for a hypergroup is further defined as the total number of distinct relevant topics found in its component groups, and it can be seen as a measure of the diversity of the hypergroup topics. At the group level, MM-LDA appears to generate much more fo-

BM-LDA: median 7, mean 11

Hypergroup #3	Mem.	Voc.
Window seat please	105	656
Aerials	59	621
Cambodia Images	21	329
Central Park	43	321
Bangkok	21	310
Thailand Travel	6	248
Monkeys	37	192

Hypergroup #431	Mem.	Voc.
HDR	275	2750
28mm or wider	113	2383
Photojournalism	101	1953
Photomatix	93	1498
Quality HDR	56	858
TTHDR (True Tone High Dynamic Range)	47	761
HDR Skies (please read the rules!!!!)	47	642
The Moon [*current* photos only]	113	485
Moon/Lua	65	321
HDaRt	10	238
HDR Rides	27	209

MM-LDA: median 4, mean 10

Hypergroup #25	Mem.	Voc.
NYC Photobloggers	56	2095
Hello Brooklyn	17	507
Uneasy	9	381
Lonely Moment	8	258

Hypergroup #37	Mem.	Voc.
Toysaholic Anonymous	29	1022
Unbearable Cuteness	20	500
Traveling Toys	16	477
Urban Vinyl Fiend	13	417
Via Alley	5	342
My new Toys and my growing collection	3	246
Little Friends Around the World	4	227
Winnie the Pooh and Friends	3	150
Space-Invaders	25	126

TB-LDA: median 4, mean 7

Hypergroup #19	Mem.	Voc.
Patterns and Designs	128	1624
Symmetry	34	1141
Curves vs. Straight Lines	63	1100
A symmetry A	14	362

Hypergroup #567	Mem.	Voc.
RUSTY and CRUSTY	443	2725
Wonders of Oxidation	159	1290
all things rusty	87	904
The Rust Bucket	84	885
Things that Moved	73	689
Rusted	37	516
RUSTY	21	257

Table 1: Two examples of hypergroups for each of the three models (with sizes around the mean and median). Hypergroups are in general quite homogeneous across all three models. The top group in each hypergroup (in bold) is the found exemplar. We also show the number of members and the size of the vocabulary for each group in the corresponding hypergroup.

	BM-LDA	MM-LDA	TB-LDA
JS-BM	0.557 / 0.622	0.491 / 0.547	0.484 / 0.521
JS-MM	0.549 / 0.602	0.372 / 0.420	0.388 / 0.411
JS-TB	0.512 / 0.555	0.431 / 0.494	0.408 / 0.436

Table 2: Mean/median hypergroup homogeneities for the three topic models using cross-model similarity measures.

cused topic-based representations, with a mean of around 3 topics per group, as opposed to the BM-LDA and TB-LDA models, which both have means of around 9 topics (details not shown for space considerations). This is also observed at the hypergroup level (see Fig. 4), where aggregating all distinct relevant topics in the hypergroup yields a mean of 6 for the MM-LDA model, while the BM-LDA and TB-LDA have means around 16 and 14 topics respectively. The MM-LDA representation is overall more spartan.

Finally, we define a measure of homogeneity for a hypergroup based on the intra-cluster similarity, by averaging the pair-wise similarities for all groups in a hypergroup. For each of the three LDA models we use a Jensen-Shannon similarity measure, dubbed JS-BM, JS-MM, and JS-TB for the similarity derived from each of the three LDA models. These are the same similarities used for the AP clustering algorithm. We then analyzed the effect of each similarity measure on the homogeneity of hypergroups discovered by a given model. We present these measurements in Table 2. In this table, lower JS distances mean higher homogeneity of the hypergroups. We note that hypergroups based on the BM-LDA model tend to be less homogeneous than hypergroups discovered by the other two models, regardless of the similarity measure used. These differences are statistically significant at $\alpha = 0.01$. This suggests that hypergroups defined based solely on binary-membership links may generally be less consistent. These results are likely explained (at least partially) by the fact that BM-LDA produces less hypergroups, which in turn leads to less homogeneity due to the larger hypergroup size.

Overall, we observe that hypergroups obtained from the multiple-occurrence membership and tag-based models are most homogeneous when the distance JS-MM is used, which suggests that capturing the relations (through membership) and content (through the size of the contributed vocabulary) might indeed be beneficial for hypergroup modeling.

5. CONCLUSIONS

We have proposed a method to discover groups in Flickr. By finding groups of similar groups we enable users to find somewhat obscure groups that do not show up at the top of traditional search results. We have shown that the affinity propagation clustering algorithm yields rather homogeneous hypergroups, regardless of the underlying model used. Hypergroups found this way tend to be of relatively small sizes. Manual inspection shows that the discovered hypergroups are indeed meaningful, and confirms our hypothesis that similar groups share content and/or members. We have also shown that using information derived from topic models, such as number of relevant topics, can give insights into the structure and quality of the hypergroups. We have also proposed a method to assess the homogeneity of discovered hypergroups based on similarity measures employed by the clustering process. Our results seem to encourage the use of fused information coming from content and relations, such as is the case for the MM-LDA model. A prototype of group search-through-hypergroups which contains a number of challenges for effective visualization and discovery is subject of future work.

Acknowledgements

This research has been supported by the Swiss National Science Foundation through the MULTI project.

6. REFERENCES

[1] 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.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 2003.

[3] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[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, Dec 2006.

[6] K. Lerman, A. Plangprasopchok, and C. Wong. Personalizing Image Search Results on Flickr, Apr 2007.

[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] R. A. Negoescu and D. Gatica-Perez. Topickr: Flickr Groups and Users Reloaded. In *MM '08: Proc. of the 16th ACM Intl. Conf. on Multimedia*, Vancouver, Canada, Oct. 2008.