CONNECTING CONTENT TO COMMUNITY IN SOCIAL MEDIA VIA IMAGE CONTENT, USER TAGS AND USER COMMUNICATION

Munmun De Choudhury *Hari Sundaram* *Yu-Ru Lin* *Ajita John* *Doree Duncan Seligmann*
Arts, Media & Engineering
Arizona State University, Tempe, AZ 85281
Collaborative Applications Research
Avaya Labs, Lincroft, NJ 07738
Email: {munmun.dechoudhury,hari.sundaram,yu-ru.lin}@asu.edu, {ajita,doree}@avaya.com

ABSTRACT

In this paper we develop a recommendation framework to connect image content with communities in online social media. The problem is important because users are looking for useful feedback on their uploaded content, but finding the right community for feedback is challenging for the end user. Social media are characterized by both content and community. Hence, in our approach, we characterize images through three types of features: visual features, user generated text tags, and social interaction (user communication history in the form of comments). A recommendation framework based on learning a latent space representation of the groups is developed to recommend the most likely groups for a given image. The model was tested on a large corpus of Flickr images comprising 15,689 images. Our method outperforms the baseline method, with a mean precision 0.62 and mean recall 0.69. Importantly, we show that fusing image content, text tags with social interaction features outperforms the case of only using image content or tags.

1. INTRODUCTION

There has been an unprecedented increase in the number of social media websites (e.g. Flickr [1], YouTube, Slashdot, Digg, del.icio.us) in the past few years which have allowed users to create, share and consume rich media very easily. Social media sites are popular not just for the content, but also due to the accompanying social interaction. In popular image sharing sites such as Flickr, enthusiastic photographers are interested in receiving critical comments on their photos. Note that simply uploading an image onto Flickr does not ensure rich social interaction or *reachability* to other users for critical feedback.

Flickr allows people to connect their images to communities, through the mechanism of image 'groups' (also known as image pools). A Flickr group is a repository of images shared by a set of users and is usually organized under a certain coherent theme (e.g. the group "The Magic of Nature"). However, *finding the right community* that will give useful comments is not easy. Simple text based search for a group will reveal a large number of similar communities (also known as image groups / pools on Flickr) e.g. "Travel / Travel Photography / Travel in Asia" etc. The fundamental challenge addressed in this paper is to connect user content to the correct community – i.e. given an image, recommend the relevant group(s) that would enable social interaction and enhance the reachability to other users.

Related Work: There has been considerable work in recommendation of items (e.g. books, movies) [9] to users as well as on recommending tags to media objects [3,10]. A fundamental distinction between prior work and our own, is that prior work has tended to pay close attention to the content (e.g. automated tag recommendation systems for images), while paying less attention to the social interaction, a key component of social media. Since we are interested in 'connecting' content (image) with community (groups, where significant social interaction occurs), we pay close attention to social interaction (user-user comments), *in addition* to content based visual features and text tags.

In a recent work on group and tag recommendation [2], authors use appearance based image concepts, but they do not incorporate social interaction in their analysis. In another work on analysis of Flickr groups [8], the authors have analyzed user behavior in these groups. While the group representation framework is rich, it does not incorporate social interaction, and has not been used for connecting content with community. To the best of our knowledge, this is one of the first works where image content is connected to social media communities, through the use of appearance based features, text tags and social interaction history.

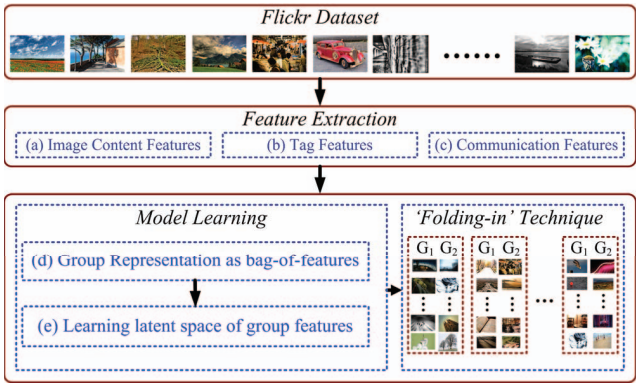

Figure 1: The overall system overview of our group recommendation framework.

Our Approach: A system overview of our group recommendation framework is shown in Figure 1. The framework is based on three ideas: (a) users can associate their images with groups whose themes they consider fit to the image, or (b) are interested in the concepts (tags) or (c) the on-going communication (comments) among the group members. Hence at the first step (feature extraction), three types of features are extracted for all images in the dataset: image content, tag and communication features. Tag features are given by a vector of the frequency counts of the tags the owner of the image has used over the past. Communication features are given by the frequency of comments written by the owner of the image on different groups. In the second step (model learning), bag-of-features based representations of the groups are generated and a model is learnt to represent the groups in a latent space. Finally, we use the learnt model parameters to recommend k groups for each image.

We have performed extensive experimentation on a dataset (15,689 images) crawled from Flickr. Our method yields good results in recommending groups to images compared to a k-

Nearest Neighbor based baseline framework. In our dataset, each image on an average is associated with three groups, and we observe that our results yield high precision and recall of 0.65 and 0.69 respectively for k=3.

The rest of the paper is organized as follows. We describe image content, tag and communication features in sections 2-4. In section 5 we present our group recommendation framework. Section 6 discusses our dataset and experimental results. Finally we conclude in section 7 with our major contributions.

2. IMAGE CONTENT FEATURES

In this section, we briefly discuss different content based features that have been used to characterize the images.

Color: There are two color-based features of interest – color histogram and color moments, which capture the distribution of different colors in the images.

Texture: Flickr images show diversity of texture. We use two texture features: gray level co-occurrence matrix, GLCM, and a texture detector for arbitrary "blobs" in images – called phase symmetry [11]. Phase symmetry is based on determining local symmetry and asymmetry across an image from phase information. Given an image, phase-based symmetry detector (PSD) maps a pixel, p, an orientation, o, and a scale, n, to a phase congruency value $PC_{no}(p)$ and a special phase, $\phi_{no}(p)$:

$$PSD(p,n,o) = \left(PC_{no}(p)\phi_{no}(p)\right). \tag{1}$$

Here,

$$PC_{no}(p) = \frac{\text{sum}E(p)}{\text{sum}A(p)+\varepsilon} = \frac{\sum_{k,q} E_{n_k o_q}(p)}{\sum_{k,q} A_{n_k o_q}(p)+\varepsilon}, \tag{2}$$

where sum$E(p)$ is the total energy when phases are congruent under all scales and orientations and phases are zero; sum$A(p)$ is the total amplitude when phases are congruent under all scales and orientations and phases are zero; ε is a positive constant.

Shape: Images are often characterized by central themes or concepts. To extract such features, we use two shape features – radial symmetry [7] and phase congruency [5]. The radial symmetry feature is based on the idea of detecting points of interest in an image. Phase congruency is an illumination and contrast invariant measure of feature significance. For a given image, phase congruency $PC(x)$ at some location x is expressed as the summation over orientation o and scale n:

$$PC(x) = \frac{\sum_o \sum_n W_o(x)\lfloor A_{no}(x)\Delta\Phi_{no}(x) - T_o \rfloor}{\sum_o \sum_n A_{no}(x)+\varepsilon}, \tag{3}$$

where A_n represents the amplitude of the n^{th} component (of the image) in the Fourier series expansion, $W_o(x)$ is the convolution of the given image with an even / odd filter, ε is added for cases of small Fourier amplitudes, T_o is a compensating measure for the influence of noise and $\Delta\Phi_n(x)$ is a sensitive phase deviation.

SIFT: SIFT or Scale Invariant Feature Transform [6] is a content-based image feature that detects stable keypoint locations in scale space of an image. It computes the following function from the difference of two nearby scales separated by a constant multiplicative factor κ:

$$\begin{aligned} D(x,y,\sigma) &= \left(G(x,y,\kappa\sigma) - G(x,y,\sigma)\right) * I(x,y) \\ &= L(x,y,\kappa\sigma) - L(x,y,\sigma), \end{aligned} \tag{4}$$

where $L(x,y,\sigma)$ is the scale space of image $I(x,y)$ and is produced from the convolution of a variable-scale Gaussian having scale (σ) with $I(x,y)$.

The details of these features may be referred to in [5,6,7,11]. We discuss tag features to characterize images in the next section.

3. TAG FEATURES

We develop a set of tag features for each image based on the tagging activity of the owner of the image. Tag feature extraction is useful because: (1) users develop a set of 'favorite' concept spaces (in the form of tags) over time to describe their images, and (2) an image is likely to be associated by the user to a group whose concepts are similar to the high frequency tags she has used in the past.

Hence, if the tag distribution given by the prior tagging activity of the owner of an image is close to the distribution of tags in a group, this group should be assigned a high probability of recommendation. We construct a vector of the frequency counts of the tags the owner of the image has used in the past (on other images), *prior to* the date of upload of the image. Let for image i the timestamp of its upload be t_i and τ_u be the set of all unique tags that user u (owner of i) has used for her other images from time 0 to t_i. The frequency count $n_{u,j}$ of usage of each tag j in τ_u gives the tag feature vector $\mathbf{T}_i$ for image i:

$$\mathbf{T}_i = \left[n_{u,1} \; n_{u,2} \; \cdots \; n_{u,L} \right] \text{ s.t. } L = |\tau_u|. \tag{5}$$

We now discuss communication features to characterize images in the next section.

4. COMMUNICATION ACTIVITY FEATURES

We develop communication based features for each image based on the frequency of comments written by the owner of the image on different groups. A user participating in the communication in a certain set of groups through comments is likely to be interested in those groups. Hence recommending them to the user is useful. Let Alice be interested in groups related to travel and frequently leave comments on the images in such groups. Further suppose Alice uploads a new image on Grand Canyon for which she intends to find suitable groups. Since Alice has earlier expressed interest in the travel related groups through her communication, our framework should recommend her such groups. Let $n_{u,j}$ be the number of comments written by user u on group j ($1 \le j \le M$) in the time period from 0 to t_i where t_i is the timestamp of posting image i by u. Then the communication activity feature $\mathbf{O}_i$ for image i is given by the vector of $n_{u,j}$ over all groups j:

$$\mathbf{O}_i = \left[n_{u,1} \; n_{u,2} \; \cdots \; n_{u,M} \right]. \tag{6}$$

Based on the three types of features, we now characterize all the N images in the dataset. Suppose D is the dimensionality of the feature vector of each image i, then, $\mathbf{f}_i \in \Re^{1\times D}$, $\mathbf{f}_i = [\alpha_1 \mathbf{C}_i \; \alpha_2 \mathbf{T}_i \; \alpha_3 \mathbf{O}_i]$, where $\mathbf{C}_i$, $\mathbf{T}_i$ and $\mathbf{O}_i$ are the image content, tag and communication feature vectors; and α_1, α_2 and α_3 are the weights determining the impact of each kind of feature. Based on these features, we now discuss the group recommendation framework.

5. GROUP RECOMMENDATION FRAMEWORK

We now present our group recommendation framework. First, we present the main idea. Second, we discuss the learning of the model parameters. Finally, we discuss the folding-in technique where we determine top k recommended groups.

5.1 Main Idea

The goal of the recommendation framework is to determine the following probability over all M groups G_j for a given image i:

$$P(G_j \mid i) \propto P(G_j, i). \tag{7}$$

In order to compute the above joint probability, we develop a mixture model representation of each image over a set of latent states, motivated from pLSA (probabilistic latent semantic analysis) [4]. This latent space could be impacted by different factors – the content of the image, owner's prior tagging activity or her prior commenting activity over different groups. The joint probability $P(G_j,i)$ above can thus be represented as:

$$P(G_j, i) = \sum_z P(z).P(i \mid z).P(G_j \mid z) \propto \sum_z P(z \mid i).P(G_j \mid z), \tag{8}$$

where z are the latent states. Hence determining the group recommendation probabilities given an image can be reduced to computing the two conditional probabilities: $P(G_j|z)$ and $P(z|i)$. Computing $P(G_j|z)$ can be considered as training or model learning (section 5.2), as it is independent of the image, given the latent states. While, computing $P(z|i)$ (section 5.3) directly depends on the image whose recommendations we are seeking, and hence can be considered as testing or folding-in technique.

5.2 Model Learning

We first discuss the construction of the training set based on computing feature vector representations for each group; and second, learning the model. We assume that each group is a 'bag-of-features', comprising its constituent image content, tags and user comments. Let Q images be used for the training set and $N–Q$ images for the test set. Using these Q images, the training set, $\mathbf{Y} \in \Re^{D \times M}$ is defined over M groups where each group G_j is represented by a feature space of its associated images (centroid). The feature vector for the j^{th} group in $\mathbf{Y}$ is thus computed as:

$$\mathbf{Y}_j = \frac{1}{|G_j|} \sum_{i \in G_j} \mathbf{f}_i, \tag{9}$$

where $\mathbf{f}_i$ is the i^{th} image in the group G_j. Using this training set, we now discuss learning the conditional probability $P(G_j|z)$ based on the EM-algorithm. If F_m is the m^{th} feature attribute in $\mathbf{Y}$, the update equations of EM [4] are given as:

$$\begin{aligned} E\text{-step: } & P(z \mid F_m, G_j) = \frac{P(z).P(F_m \mid z).P(G_j \mid z)}{\sum_{z'} P(z').P(F_m \mid z').P(G_j \mid z')}. \\ M\text{-step: } & P(G_j \mid z) \propto \sum_m \mathbf{Y}_{m,j}.P(z \mid F_m, G_j), \\ & P(F_m \mid z) \propto \sum_j \mathbf{Y}_{m,j}.P(z \mid F_m, G_j), \\ & P(z) \propto \sum_m \sum_j \mathbf{Y}_{m,j}.P(z \mid F_m, G_j). \end{aligned} \tag{10}$$

Now we discuss the folding-in technique for determining the groups to be recommended to a given new image (test image).

5.3 Folding-in Technique

We discuss learning the conditional probability $P(z|i)$ for a given new image (test image) in $\mathbf{Z} \in \Re^{D \times (N-Q)}$ and how based on the learnt parameters we can determine the top k recommended groups for the images. The basic idea of computing the probability $P(z|i)$ is based on how we can "fold-in" [4] a new image in our existing data to predict its probability of being recommended to different groups. We again use the following EM update rules to determine the probability:

$$\begin{aligned} & E\text{-step: } P(z \mid i, F_m) \propto P(F_m \mid z).P(z \mid i). \\ & M\text{-step: } P(z \mid i) \propto \sum_m \mathbf{Z}_{m,i}.P(z \mid i, F_m), \end{aligned} \tag{11}$$

where F_m is the m^{th} feature attribute of test image i. Substituting the learnt probabilities $P(G_j|z)$ and $P(z|i)$ from eqns. (10) and (11) in eqn. (8) and finally in eqn. (7), we can determine the probability of recommending a group G_j to test image i. Hence the top k recommendations $\mathbf{g}_i \in \Re^{1 \times k}$ for a test image i is given by those k groups among M for which $P(G_j|i)$ is maximum. Let us now discuss the experimental results.

6. EXPERIMENTAL RESULTS

In this section we discuss the experimental results. First we present a brief overview of the Flickr dataset. Second we discuss the results and finally we present a brief discussion of the limitations of our results.

Flickr Dataset: We have tested our group recommendation framework on a dataset crawled from the popular media-sharing site, Flickr. We downloaded images ranked by Flickr's proprietary "interestingness" criterion. The dataset comprises 15,689 images which belong to 925 groups; each group on an average consisting of 17 images. The mean number of tags per image is six, that of groups per image is three and comments per image is 14. The upload time period of these images ranged from March 21, 2008 to August 20, 2008. About 80% of the images (randomly selected) were used to construct the training set (12,551 images) and the rest of the 3,138 images constituted the test set.

Table 1: Evaluation of our method (M_1) against k-Nearest Neighbor (M_2) using Precision, Recall and F1-measure. Metrics are computed at k=1, 3, 10 and for three cases: precision, recall and F1-measure over all images, over images owned by each user and over images belonging to each group. Our method outperforms k-NN.

		Precision		Recall		F1-measure	
		M_1	M_2	M_1	M_2	M_1	M_2
Image-based	k=1	**0.68**	0.50	**0.63**	0.47	**0.66**	0.49
	k=3	**0.63**	0.49	**0.69**	0.52	**0.64**	0.50
	k=10	**0.52**	0.47	**0.71**	0.55	**0.59**	0.51
User-based	k=1	**0.64**	0.48	**0.59**	0.46	**0.61**	0.48
	k=3	**0.59**	0.46	**0.62**	0.48	**0.60**	0.47
	k=10	**0.51**	0.43	**0.68**	0.52	**0.58**	0.47
Group-based	k=1	**0.74**	0.52	**0.72**	0.52	**0.72**	0.53
	k=3	**0.69**	0.50	**0.71**	0.54	**0.70**	0.52
	k=10	**0.61**	0.48	**0.81**	0.57	**0.69**	0.55

Evaluation against k-NN: The results of validation of our group recommendation framework (M_1) against a baseline method k-Nearest Neighbor (M_2) is shown in Table 1. The performance of our method is evaluated using three metrics: precision, recall and F1-measure. We compute these metrics for different values of the number of recommended groups (k=1, 3 and 10). Further, precision, recall and F1-measure are computed for three cases: (a)

over all images, (b) over images from each user and (c) over images in each group. The motivation for (b) lies in the fact that since our method uses communication features, the recommendation performance is going to be different for different users. Similarly for (c), we want to be able to analyze our method at the group level, since for some groups we might be able to yield recommendations better than others. Note, all features are used in both the methods.

The results yield interesting insights. First, we observe that precision gradually decreases, while recall increases with increase in *k*. This is because with increase in the number of recommended groups, more false positives are likely to be returned (because recall, the average number of groups per image is three), resulting in low precision. While, more ground truth groups are likely to be returned for larger values of *k*, yielding high recall. Second, maximum precision and recall occur in the group-based case, while minimum for the user-based case. This is explained by the fact that groups are likely to comprise images which are consistent content-wise or tag feature-wise. Whereas, users often associate their images to different groups due to personal preferences, apart from image content, tags or commenting behavior, resulting in diversity of image-group association. Third, comparing with the baseline method *k*-NN our method seems to yield higher precision and recall; mean precision is 0.62 against 0.49 for *k*-NN; and mean recall is 0.69 against 0.59 for *k*-NN. Also, overall, we observe that the mean F1-measure (that combines precision and recall together) for our method is approx. 0.65, while for *k*-NN is 0.51. Thus our method outperforms the baseline method.

Table 2: Evaluation of the three types of features (image content, tags and communication features) used for image characterization against our optimal method (all features).

	Precision	Recall	F1-measure
Image Content features	0.43	0.48	0.46
Tag features	0.61	0.66	0.63
Communication features	0.57	0.64	0.61
Optimal method (all features)	**0.62**	**0.69**	**0.65**

Evaluation of feature types: The results of evaluation of the different types of features extracted in this paper have been shown in Table 2. We observe that the combination of all features yields the highest values of precision, recall and F1-measure. Among the three types of features, we observe that content features perform the worst while the tag features perform the best. This is explained by the nature of groups on Flickr. Several groups are organized along certain concept spaces / themes (e.g. "The South-west of United States"); as a result they consist of images which are visually quite diverse; however are likely to contain consistent tags like "Grand Canyon" or "Arizona" which might be reflected in the corresponding user activity over the past. Interestingly, communication features perform well; implying that comments do indeed impact users' intent to associate groups to images.

Discussion: Online social media are not mere repositories of diverse content. Hence group recommendation to images on social media is an extremely challenging problem unlike traditional classification; because it needs to account for the inter-user interactions in the groups. However such interactions might always not be directly observable from the image content, tagging or communication activity of the users. Users could have intrinsic motivations affecting these interactions, which in turn might be responsible for associating an image to a group. Despite this, our paper gives a novel approach to characterize images along several feature types and yields promising results against methods incorporating image content or tags alone.

7. CONCLUSIONS

In this paper, we developed three kinds of features to characterize images in online social media: image content, user tagging activity and user communication activity. A group recommendation framework based on learning a latent space for the groups was developed which recommended *k* most likely groups to a given image. Experiments on the Flickr dataset indicated satisfactory results in recommending groups to images with a mean precision of 0.62 and a mean recall of 0.69, compared to 0.49 and 0.59 respectively for a *k*-NN based baseline framework. We conclude that user tagging and communication based characterization of images helps improve recommendation performance significantly against image content alone. Our recommendation framework also captures social interactions among users through user communication history which is central to online social media.

As future work, elaborate understanding of communication among users can help provide better recommendations. Moreover exploiting the social network of users to understand their mutual coupling can also improve recommendation performance.

8. REFERENCES

[1] *Flickr* http://www.flickr.com.

[2] H.-M. CHEN, M.-H. CHANG, P.-C. CHANG, et al. (2008). *SheepDog: group and tag recommendation for flickr photos by automatic search-based learning*. Proceeding of the 16th ACM international conference on Multimedia. Vancouver, British Columbia, Canada, ACM**:** 737-740.

[3] N. GARG and I. WEBER (2008). *Personalized, interactive tag recommendation for flickr*. Proceedings of the 2008 ACM conference on Recommender systems. Lausanne, Switzerland, ACM**:** 67-74.

[4] T. HOFMANN (1999). *Probabilistic latent semantic indexing*. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. Berkeley, California, United States, ACM**:** 50-57.

[5] Z. LIU and R. LAGANIERE (2007). *Phase congruence measurement for image similarity assessment.* Pattern Recogn. Lett. **28**(1): 166-172.

[6] D. G. LOWE (2004). *Distinctive Image Features from Scale-Invariant Keypoints.* Int. J. Comput. Vision **60**(2): 91-110.

[7] G. LOY and A. ZELINSKY (2003). *Fast Radial Symmetry for Detecting Points of Interest.* IEEE Trans. Pattern Anal. Mach. Intell. **25**(8): 959-973.

[8] R. A. NEGOESCU and D. GATICA-PEREZ (2008). *Analyzing Flickr groups*. Proceedings of the 2008 international conference on Content-based image and video retrieval. Niagara Falls, Canada, ACM**:** 417-426.

[9] A. I. SCHEIN, A. POPESCUL, L. H. UNGAR, et al. (2002). *Methods and metrics for cold-start recommendations*. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. Tampere, Finland, ACM**:** 253-260.

[10] B. SIGURBJORNSSON and R. V. ZWOL (2008). *Flickr tag recommendation based on collective knowledge*. Proceeding of the 17th international conference on World Wide Web. Beijing, China, ACM**:** 327-336.

[11] Z. XIAO, Z. HOU, C. MIAO, et al. (2005). *Using phase information for symmetry detection.* Pattern Recogn. Lett. **26**(13): 1985-1994.