Modeling and Predicting Group Activity over Time in Online Social Media

Munmun De Choudhury
Arts Media & Engineering, Arizona State University
Email: munmun.dechoudhury@asu.edu

ABSTRACT

This paper develops a probabilistic framework that can model and predict group activity over time on online social media. Users of social media sites such as Flickr often face the enormous challenge of *which group to choose*, due to the presence of numerous competing groups of similar content. Determining an empirical measure of significance of a group can help tackle this problem. The proposed framework therefore determines an optimal measure per group based on past user participation and interaction as well as likely future activity in the group. The framework is tested on a Flickr dataset and the results show that this method can yield satisfactory predictions of group activity. This implies that the computed measure of significance of a group can be used by end users to choose groups with rich activity.

Categories and Subject Descriptors

J.4 **[Social and Behavioral Sciences]**: Sociology.

1. INTRODUCTION

This paper develops a probabilistic framework that can model and predict group activity over time on online social media. The exponential increase in the number of social media sites has given the Web users variegated ways to share media content (e.g. images, videos etc) very easily. They can join media pools to engage in rich interactions with other users. The popular social media site Flickr [1] features such media pools in the form of 'groups'. Flickr users can share their images on groups enabling them to get critical feedback from others in the form of comments and voting feature (i.e. 'favorites') on them.

However, when several competing groups on a certain topic are available via the Flickr search tool, which group should a user choose? For example, a simple group search on Flickr using the keyword 'Arizona' reveals 4,967 relevant groups. Although users can filter their search spaces by ranking groups based on their sizes, date of creation or most recent activity; there is no efficient way to understand the measure of activity of the groups over time – users certainly would want to choose groups which are likely to have sustained activity in the past, as well as in the future. How can we help the end user choose groups based on rich activity dynamics? Can we predict group activity to ease choice of groups relating to a topic? This paper seeks to answer these questions.

There has been extensive prior work [2,3] on understanding the dynamics of groups on social media sites; but none of them attempts to determine measures of significance of groups. The framework proposed in this paper can let an end user identify groups that have rich activity in the past, as well as likely to have high activity in the future. The primary contributions are threefold: (a) a modeling method of group activity over time, (b) a prediction method to identify groups of high activity in the future and (c) an optimization method to compute an empirical measure of significance per group. These are discussed in the next section.

2. PROPOSED FRAMEWORK

Modeling Group Activity. Modeling activity in a group over time involves two aspects – modeling participation of different users around the shared media elements in the group, and modeling interaction among users. This paper considers that user participation in a group manifests in the form of (a) upload of media elements (e.g. images on Flickr) and (b) votes on the shared media elements (e.g. 'favorites' on an image on Flickr). User-user interaction can be gauged through the comments posted on the shared media elements. Hence the probability of a group g's degree of participation $P_{g,1:i-1}$ over the past i.e. the time prior to time interval t_i is given by the normalized sum of the number of media elements uploaded and voted by all users belonging to g:

$$p\left(P_{g,1:i-1}\right) = \frac{1}{|g|}\sum_{u\in g}\frac{\left(m_{u,1:i-1;g} + f_{u,1:i-1;g}\right)}{\sum\limits_{g'\exists g'\neq g}\left(m_{u,1:i-1;g'} + f_{u,1:i-1;g'}\right)}, \tag{1}$$

where $m_{u,1:i-1;g}$ and $f_{u,1:i-1;g}$ are respectively the number of media elements uploaded and number of media elements voted by user u on group g between time t_1 to t_{i-1}. Similarly the probability of a group g's degree of interaction $I_{g,1:i-1}$ prior to time interval t_i is given by:

$$p\left(I_{g,1:i-1}\right) = \frac{1}{|g|}\sum_{u\in g}\frac{c_{u,1:i-1;g}}{\sum\limits_{g'\exists g'\neq g} c_{u,1:i-1;g'}}, \tag{2}$$

where $c_{u,1:i-1;g}$ is the number of comments posted by user u on group g between time t_1 to t_{i-1}. Finally, the effective probability of the measure of activity $p(A_{g,1:i-1})$ of group g over time t_1 to t_{i-1} is given by the mean of measures of participation and interaction.

Prediction of Group Activity. Now we discuss the method of predicting group activity at a future time interval.

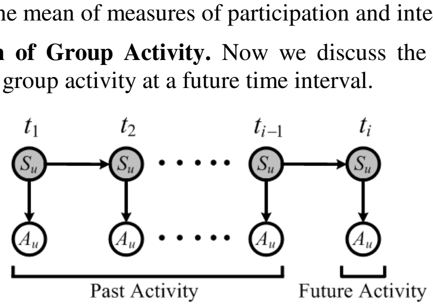

Figure 1: The structure of the HMM used in prediction of user activity. The HMM learns model parameters over past activity and uses them to determine probability of future activity of a user u, given the state at time t_i.

We assume that a particular user involves herself in some activity with respect to a group depending on a *latent* state reflecting her intrinsic intent. This latent state can take two values – *active*, indicating she is likely to perform some activity on the group and *dormant*, indicating otherwise. This idea can be modeled in the form a Hidden Markov Model [4] where the user's intrinsic intent are the hidden states and her activities on the group (i.e. participation through media uploads, votes and interaction in the form of comments) are the observed data (Figure 1). We therefore train a continuous HMM with the observed measures of activity $p(A_{u,1;g})$, $p(A_{u,2;g})$, …, $p(A_{u,i-1;g})$ for each user u over t_1 to t_{i-1} to learn the optimal model parameters (the emission probability matrix **B**) using the Baum-Welch algorithm [4]. Based on **B** learnt over t_1 to t_{i-1}, we can now easily determine the probability of activity $p(A_{u,i})$ of user u at time interval t_i given she is in the 'active' state (denoted by '1'):

$$p\left(A_{u,i}\right) = \mathbf{B}\left(A_u \mid \hat{S}_u = 1\right),$$

$$\text{where } \hat{S}_u = \begin{cases} 1 & \text{if } n\left(S_{u,1:i-1} = 1\right) \ge n\left(S_{u,1:i-1} = 0\right) \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

The above equation determines the mean likelihood of active state ($\hat{S}_u$=1) of the user u based on the number of times she was in it in the past from t_1 to t_{i-1}. Hence the predicted probability of activity of group g at t_i ($p(A_{g,i})$) is the mean of predicted activities over all users u in g ($p(A_{u,i})$).

Optimization. We now compute an optimal empirical measure of significance per group at t_i which implies whether the group is suitable to be chosen by a user at t_i. To compute this optimal measure, we define the following objective function:

$$\varphi\left(g,t_i\right) = \alpha_1.p\left(A_{g,1:i-1}\right) + \alpha_2.p\left(A_{g,i}\right), \qquad (4)$$

where α_1 and α_2 are the weights that respectively determine the impact of past activity and future activity. Maximizing this linear optimization function would yield optimal values of weights, as α_1^* and α_2^*, and also an empirical measure of activity $\varphi^*(g,t_i)$.

3. EXPERIMENTAL RESULTS

To test the framework, experiments have been conducted based on a dataset crawled from the popular image-sharing social media Flickr. A set of 200 groups comprising a mean number of 2,628 users *per group* were used for the experiments.

We conjecture that groups with high measures of optimal activity, predicted by our framework, would incur four consequences. They would have (a) increased participation of users through image uploads (b) increased number of favorites, (c) extensive user-user interaction in the form of comments, and (d) they would trigger new users to join in the future. A qualitative evaluation of our framework using these consequence metrics is presented in Figure 2. Four different groups of various sizes and characteristics, 'Gallery Photography', 'The Sky's The Limit', 'Surreal Nature' and 'Travelers' Eye' are shown, and over a period of 30 days in the month of November 2008. The predicted measure of group activity is observed to follow the four metrics closely in all the four cases. 'Gallery Photography' is observed to be a very small focused group comprising a mean number of 34 users over the time period of the experiment. Hence it is observed to have a decreasing trend of activity over time. Our framework captures this satisfactorily by yielding decreasing measures of predicted activity. On the other hand, 'The Sky's The Limit' is observed to be a highly dynamic group with an increasing trend of new users (mean 4464 users). It also shows high activity over time. Our framework captures the increasing trend well. Similar observations can be made for the other two groups.

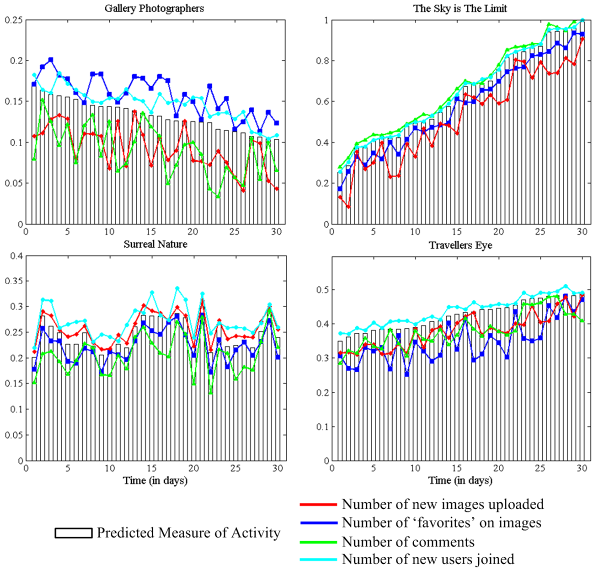

Figure 2: Experimental results of evaluation of the proposed framework based on four consequences – number of new images, new 'favorites' featured on the images, new comments and new users who join the groups. The predicted measure of activity for the four groups is observed to follow the four metrics closely.

Quantitatively, over all the 200 Flickr groups, correlation of our predicted group activity with the four metrics yields low mean error rates of 16.35%, 21.94%, 14.53% and 12.76% respectively. These results suggest our proposed framework is meaningful.

4. CONCLUSIONS

This paper proposed a probabilistic framework that can model and predict group activity over time on online social media. An empirical measure of significance per group was determined based on past user participation and interaction as well as likely future activity in the group. Experiments on a Flickr dataset demonstrated that this framework can yield satisfactory predictions of group activity; thereby implying that such measure of significance of a group can be used by end users to choose groups with rich activity in the past and future, when several competing groups of similar content are available on a certain topic.

5. REFERENCES

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

[2] L. BACKSTROM, D. HUTTENLOCHER, J. KLEINBERG, et al. (2006). *Group formation in large social networks: membership, growth, and evolution*. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA, ACM**:** 44-54.

[3] 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.

[4] L. R. RABINER (1990). *A tutorial on hidden Markov models and selected applications in speech recognition*. Readings in speech recognition, Morgan Kaufmann Publishers Inc.**:** 267-296.