Which are the Representative Groups in a Community? Extracting and Characterizing *Key* Groups in Blogs

Munmun De Choudhury[1] Hari Sundaram Ajita John Dorée Duncan Seligmann
Arts Media & Engineering, Arizona State University Collaborative Applications Research, Avaya Labs
Email: {munmun.dechoudhury,hari.sundaram}@asu.edu, {ajita,doree}@avaya.com

1. Problem and Motivation

In this paper, we analyze online communities in blogs to extract and characterize their representative *key* groups. Tracking communication in online communities has been of interest to a wide spectrum of applications in recent times. E.g. corporations are interested in understanding share-holder sentiment with respect to product releases, as well as identify 'hot' points of resource consumption in a network of users. However, although communities usually exhibit coherence in communication, some groups might have properties deviated from the overall community. Hence to track the activities of the overall community, it is useful to identify a subset of groups within the community as *key* groups. Moreover, it is more cost effective to track a small set of representative groups over time, instead of the entire community.

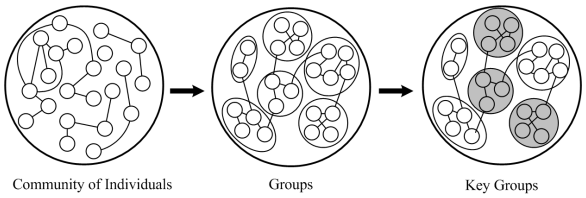


Figure 1: Conceptual representation of groups (white circles) and key groups (shaded circles) in a community of individuals.

We define key groups to be a set of representative groups in a community which capture its overall context and the content of communication. A key group therefore should have two characteristics: (a) at a certain time interval, it should be able to draw individuals from a large number of groups in the past, and (b) its communication should be aligned to the communication of the overall community. A conceptual representation of a community consisting of several individuals has been shown in Figure 1, along with their group representation. A subset of the groups (shaded circles) comprises the key groups.

In the following sections, we discuss our approach to tackle the problem of identifying key groups, followed by some experimental results and analysis.

2. Our Approach

There are two parts to our approach: extracting groups at each time interval, and identifying a subset of groups to be key, based on a set of characteristics that capture the context and content of communication in the community.

2.1 Group Extraction

We use a set of communication based features to represent each individual in the community, and then use an unsupervised method of group extraction.

Communication features of individuals should reflect either their communication activity in the past, or their intrinsic habits of communication. We consider a three-dimensional activity characterization per individual: responsivity, participation and impact. The second type of feature captures the intrinsic habits of the individual u in communication – u is a 'leader' if several others write comments and replies following her; while u is a 'follower' if she is observed to follow several others.

We then extract groups using the group extraction algorithm [2] known as 'mutual awareness expansion' (MAE) over each time interval t_i.

2.2 Key Group Characterization

To characterize key groups at a certain time interval, we need to compute two measures per group: (a) *composition entropy*, and (b) *topic divergence*.

The measure of subscription of the composing individuals in a group g at t_i with respect to their subscription to groups in the previous time interval t_{i-1} is given by its composition entropy:

$$e(g) = \frac{1}{|G_i|} \sum_{h \in G_{i-1}} \left(-p(g \mid h) \log p(g \mid h) \right), \quad (1)$$

where $p(g|h)$ gives the fraction of the number of individuals who are in group g at t_i given they had been in group h at time slice t_{i-1}. Obviously, when the composition entropy of individuals in a particular group at t_i is very high, with respect to all groups at t_{i-1}, the group can be considered to be a key group.

Next we characterize the topic distributions of groups and the topic distribution of the overall community at each time interval. To model topic distributions of groups and community as a whole, we represent the communication content i.e. comments and replies as a bag-of-words λ_g (for group g). We assume that the words in λ_g are generated from N multinomial topic models θ_1, θ_2, …, θ_N whose distributions are hidden to us. A word can be generated

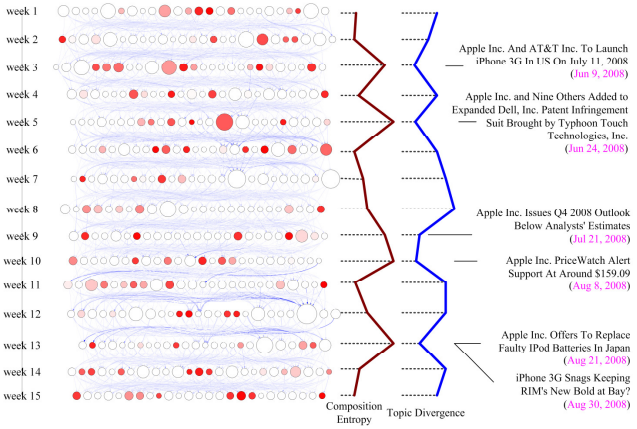
[1] The first author Munmun De Choudhury is a PhD student in Computer Science at Arizona State University, Tempe. She has played the primary role in this work, including the theoretical modeling of the framework as well as data collection and the experiments. Her advisor is Dr Hari Sundaram. This work is supported by a grant from Avaya Labs, New Jersey. Her collaborators (coauthors in this paper) helped her with valuable suggestions and guidance.

either due to the content of communication λ_g or the time indicator t_i:

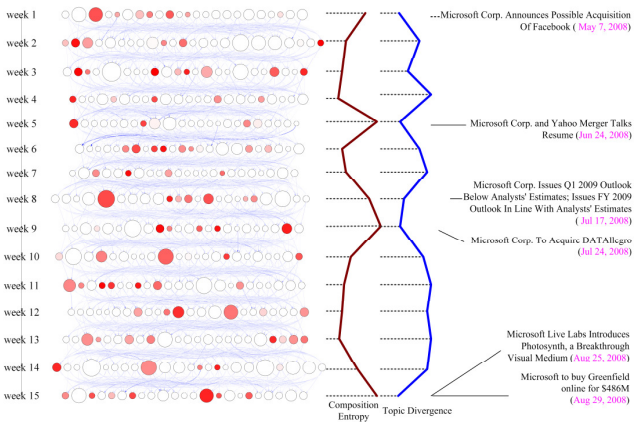
$$p(w:\lambda_g,t_i)=\sum_{j=1}^{N} p\left(w,\theta_j \mid \lambda_g,t_i\right), \quad (2)$$

where w is a word in λ_g and θ_j is the j^{th} topic. The unknown topic distribution parameters $\bullet_1$, $\bullet_2$, …, $\bullet_N$ can now be learnt by maximizing the log likelihood over the entire collection, based on the EM algorithm. We can now compute the topic distribution $\Theta_{g,i}$ of group g at t_i. Note, the topic distribution Θ_i of the overall communication set at t_i can be similarly determined. The divergence $d(g)$ of topic distribution of group g from that of the overall community is given by the KL-divergence of $\Theta_{g,i}$ from Θ_i. Finally in our framework, the groups with optimal measures of both the characteristics, composition entropy and topic divergence are considered key groups.

3. Experiments



(a) 'Apple'



(b) "Microsoft'

Figure 2: Visualization of groups (circles) and key groups (shaded circles) in two communities.

We tested our framework on a dataset crawled from the popular technology blog Engadget [1], comprising 78,740 individuals with 4,580,256 comments and replies between May 1 and August 31, 2008. The dataset was organized into communities along names of two technology companies – 'Apple' and 'Microsoft', based on the tags associated with the blog posts.

The results of experiments conducted on this dataset have been shown in Figure 2. The visualizations show the dynamics of key groups based on their mean composition entropy and topic divergence for the two communities, Apple and Microsoft over a period of 15 weeks.

We observe that the mean composition entropy increases with respect to significant company related events (based on the NY Times website); and topic divergence decreases due to such events. This is because during significant company-related happenings, individuals from different groups at a previous week are likely to get interested in discussing the event and thereby subscribe to the key groups at the current week, yielding high composition entropy. Low topic divergence during significant happenings is also meaningful because most individuals are likely to communicate on the event, yielding groups with substantially low topic divergence. Hence we observe that our method of extraction and characterization of key groups in communities is meaningful as they can capture overall dynamics of the communities over time.

4. Conclusions

We characterized online communities in blogs to extract and characterize their representative key groups. We conducted extensive experiments on a popular blog, Engadget with excellent qualitative results. We observed qualitatively that the key groups are able to capture the dynamics in the community. Our results are promising and sentiment detection in communication leveraging the properties of the key groups can be an interesting direction towards future work.

5. References

[1] *Engadget* http://www.engadget.com/.

[2] Y.-R. LIN, H. SUNDARAM, Y. CHI, et al. (2007). *Blog Community Discovery and Evolution Based on Mutual Awareness Expansion*. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society**:** 48-56.