Automatically Identifying Groups Based on Content and Collective Behavioral Patterns of Group Members

Michelle Gregory, Dave Engel, Eric Bell, Andy Piatt, Scott Dowson, and Andrew Cowell
Pacific Northwest National Laboratory
Richland, WA, USA
{michelle, dave.engel, eric.bell, andrew.piatt, scott.dowson, andrew}@pnl.gov

Abstract

Online communities, or groups, have largely been defined based on links, page rank, and eigenvalues. In this paper we explore identifying abstract groups, groups where member's interests and online footprints are similar but they are not necessarily connected to one another explicitly. We use a combination of structural information and content information from posts and their comments to build a footprint for groups. We find that these variables do a good job at identifying groups, placing members within a group, and help determine the appropriate granularity for group boundaries.

Introduction

The explosion of popularity in social media, such as internet forums, weblogs (blogs), wikis, etc., in the past decade has created a new opportunity to measure public opinion, attitude, and social structures (Agichtein et al. 2008, Qualman 2010). A very common social structure investigated is online communities, or groups. There are a number of motivations for studying online groups and communities; increasing online community involvement (Ling et al. 2005); recommender systems (Passant et al. 2009); collaborative filtering (Groh 2007); identifying authoritative or influential sources (Kleinberg 1999).

The majority of methods described to identify online communities tend to rely on, however, varying methods for link analysis (Kleinberg 1999, Wang & Kaban 2008, Chua & Xu 2007, Chin & Chignell 2007). In other words, an online community is defined as the amount of interconnectedness of individuals. While these approaches have been shown to be effective in some contexts, we argue that relying on interconnectedness for identifying a community misses many potential opportunities to identify groups of people that are very similar to one another, but may never actually interact online.

While the traditional definition of groups includes face-to-face interaction, it is well recognized that online groups need not have met face-to-face, but rather interact in some manner (for example, comment on a blog, email, etc.). We argue, however, that there is also great value to defining abstract groups (Groh 2007). An abstract group is one in which the members need not interact explicitly, but the members of the group still demonstrate cohesiveness in some way (Groh 2007, Groh 2009). In fact, the whole notion of compiling a focus group in marketing is based on the premise that one can make generalities about abstract groups: Marketers target demographic groups, for instance, females in the 18-25 age range. However, abstract online groups go beyond demographics. For example, on Live Journal1, there are a number of categories, gaming, for example, that one can categorize themselves and their blogs. While a number of those that self select that category may interact, there is no explicit requirement to do so. If one is interested in marketing to a gaming crowd, for instance, knowing all persons interested in gaming would be useful, even if they do not interact directly with one another.

Link type analyses are virtually ineffective at identifying abstract groups since the members are not connected. So we must look for other methods to identify groups, such as user behavior, which has been shown can help to identify online groups (Maia et al. 2008). The contribution that we add to this work is the inclusion of content-based features—those where the actual text of the posts and comments are analyzed—to help define groups. Because members of abstract groups do not interact, the content-based features enhance our knowledge of group composition. We also expand on previous work by using the same features that define a group footprint to investigate how well members' match up to the group and how similar a set of groups are.

[1] http://www.livejournal.com

Data

In order to validate our hypothesis that online abstract groups can be identified through the behavior of their members, we created a gold-standard dataset in which we know the groups that individuals belong to. From this set we extracted a number of features from the members' posts and comments. The features include both structural (metadata) and content-based information.

The data was harvested from LiveJournal from four different online groups that were active (Kramer & Rodden 2007) during the 2010 calendar year. LiveJournal allows blog authors to identify self-interests. We randomly selected 75 bloggers (individuals) that categorized themselves into a gaming group (2552 posts) and 75 that self categorized into a sewing group (2191 posts). In addition, to get an understanding of whether strictly online abstract groups behave different than online groups that correspond to physical groups, we harvested LiveJournal posts from members of a church (6 authors; 713 posts) and posts from members of a university program (4 authors; 104 posts).

The features that we use to describe member behavior are divided into two categories: structural and content-based. Features for clustering social media tend to be based on structure alone, not content (e.g., Maia et al. 2008). Content-based features require processing the text of all of the posts and comments to identify salient linguistics features. The content-based features were selected because these features have been demonstrated to be beneficial for other blog analyses or technically related tasks (Gregory et al. 2006, Orebaugh & Allnutt 2009, Webb et al. 2005). Table 1 shows the parameters that were extracted for each post.

Because there is generally more than one comment per post, the distribution (minimum, median, and maximum) for several features was included in our parameter space.

Methods

We use these features in a clustering algorithm to define groups. We have identified, developed, and implemented several algorithms to identify and analyze groups. For this paper, we will focus our analysis and discussion on a single algorithm to illustrate the process. The grouping was accomplished using the partitioning around medoids (PAM) clustering algorithm. This method is similar to the K-means method, except the representative member of the cluster is a medoid, which is an actual data point (observation) within the cluster (Kaufman and Rousseeuw 1990). Input into the clustering algorithm is a dissimilarity matrix that measures the distance between each observation. The selection of this (distance) algorithm is one of the key choices of the process, since it directly affects the clustering results and is dependent on the type of data to be clustered. As shown in the previous section, the data is of mixed type (numerical and categorical). For our analysis the Gower's General Similarity Coefficient was used (Gower 1971) as it is useful for measuring proximity of mixed data types. The Gower's General Similarity Coefficient Sij compares two cases i and j, as shown in Equation 1.

$$S_{ij} = \frac{\sum_k W_{ijk} S_{ijk}}{\sum_k W_{ijk}} \qquad (1)$$

Table 1 *Features extracted (metadata) and calculated (content based) from the training dataset*

Metadata	Content-Based
Comment number of Comments word count* time lag from post total time duration	**Comment** % comments that agree (disagree) with post positive (negative) sentiment* strong (weak) content*
Post word count average word length number of quoted words author	**Post** positive (negative) sentiment strong (weak) content theme

*minimum, median, and maximum values

Where Sijk denotes the contribution provided by the kth variable, Wijk is usually 1 or 0 depending upon whether or not the comparison is valid for the kth variable. It should be noted that the effect of the denominator ΣWijk is to divide the sum of the similarity scores by the number of variables.

The clustering methods used in this paper have been shown to produce decent partitioning of many different types of data (Meila 2007, Park et al. 2007) including online social media (Maia et al. 2008).

From the grouping process, a characteristic footprint for each group is produced. This footprint is used to define the boundary of each group. The algorithm for calculating this footprint is very dependent on the distribution of the numerical parameters that were used to define (cluster) each group. For this paper, we use the mean value of each parameter to define this characteristic footprint.

Identifying and characterizing groups enables us to test how accurately we can place new observations (post and comments) within the appropriate group by placing each new observation into its "closest" group. The numerical parameters of the new observation and the numerical footprint are used to calculate the distance between the observation and each group. A proportionality metric for themes and authors that reside in each group are also calculated. The closest group is defined by the minimum distance between the observation and each group. The themes and authors proportionality metrics are used to resolve which group to be placed when the numerical distance is very similar for 2 or more groups.

The final step in our analysis identifies outlier observations. Outliers will be defined as those observations or group of observations that don't fit well into any of the defined groups. Statistical inference and order statistics can provide a model for describing outliers with statistical confidence..

Experiments

We conducted 3 experiments to identify groups. The goal of Experiment 1 was simply to see if the algorithms and methods cluster the gold standard datasets into appropriate bins. We combined data from all four groups for this experiment. Clustering analysis was performed on a training set of 90% of the data, testing on 10%.

Experiment 2 was designed to see how well we can bin individual users into their appropriate group, in other words, how do the footprints of individuals fall within the footprint of the entire group? This method can be used to identify outliers to the groups.

One aspect of automatically identifying groups is to be able to define what appropriate group boundaries are. A group that is too large or diffuse is not much help. For *Experiment 3* we tested our methods and algorithms at identifying subgroups within our most diffuse group, the *gaming* group. Although we collected 75 members of each group for our test data, the overall gaming community was actually much larger (thousands of members). The number of members coupled with the results from *Experiment 1* suggests that this group should be further divided.

Results

Table 2 provides the results of how well we can cluster blog posts to the correct group (*Experiment 1*) and individuals to a group (*Experiment 2*). The first portion of the table (*Experiment 1*) shows the results from the clustering analysis for the four groups. Two separate analyses were performed and are illustrated in the table. First, the *gaming* and *sewing* datasets were combined and clustered, as well as the *church* and *school* datasets (2 group comparison). An analysis of the clustering results showed that 67% of the true *gaming* observations were clustered together (shown in column two), while 78% of the true *sewing* observations, 83% of the true *church* and 62% of the true *school* observations were clustered correctly. The highlighted column in Table 2 (column three) shows the results of the analysis when all four datasets were combined and then analyzed (4 group comparison). The last column represents the baseline; the expected values for each group if the whole process was entirely random. These values are based on the distribution of observations within the combined datasets.

***Table 2** Comparison of cluster results (% correct) to the self-rated grouping (gold-standard)*

Group	2 group comparison	4 group comparison	Baseline
Experiment 1			
gaming	67%	52%	40%
sewing	78%	76%	43%
church	83%	38%	14%
school	62%	10%	2%
Experiment 2			
gaming	73%	58%	25%
sewing	72%	70%	25%
church	70%	70%	25%
school	86%	86%	25%
Experiment 2 (themes and authors not used)			
gaming	54%	12%	25%
sewing	69%	69%	25%
church	27%	11%	25%
school	86%	86%	25%

In *Experiment 2*, we placed the posts from individual group members from the test dataset (10% of the total data) into one of the pre-defined groups. Each observation of the test dataset was placed into one of the groups based on the minimum distance to the different group footprints, with the theme and author being used to resolve minimum distances that were very similar. As in *Experiment* 1, the analysis was split into two parts; the two datasets analyzed separately (2 group comparison) and then all four datasets combined (4 group comparison).

The final analysis shown in Table 2 repeats the analysis for *Experiment* 2, except in this analysis the themes and authors were not used in order to ensure the grouping effects were not solely due to this information. In other words, the placement of an observation was made entirely based on the group with the minimum distance using only the numerical parameters from the observation and the numerical footprint for each group.

For *Experiment 3* we clustered only the *gaming* dataset to identify any subgroups that might be present. We found that the *gaming* posts fall into four main subgroups, each having a tighter distribution than the whole.

The distance between groups was defined using only the numerical parameters and the Euclidean distance. The results of calculating the distance between each group is shown in Figure 1.

Conclusion

We have defined methods to identify online groups automatically using both content-based and structural-based data. For this paper, we have selected only one grouping

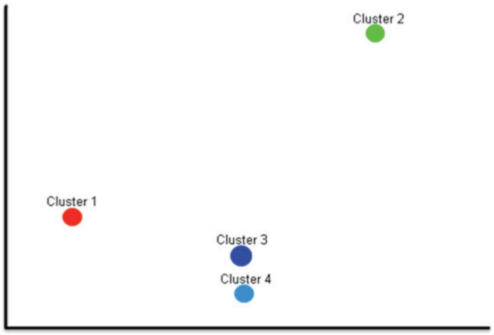

Figure 1 *Spatial distribution of the four gaming clusters, as defined by the Euclidean distance*

method to illustrate our process. Our initial experiments demonstrate this method can be used to distinguish different group types through defining their online footprints. Furthermore, we have shown that once a footprint for a group is identified, we can accurately place individuals within a group. Lastly, we can use the same features to identify outliers within a set of groups.

In our experiments we used two different group types to investigate whether the methods perform the same for abstract groups. This comparison is important in order to demonstrate that groups in which the members do not interact are still cohesive in important ways. We found no evidence to suggest that performance is better for groups whose members interact.

Defining abstract online groups allows one to identify like-minded individuals who may not interact directly, either online or physically. These findings have importance for applications such as targeted marketing and recommender systems. In addition, these methods can help to identify individuals who have overlapping interests but do not know each other. In addition, being able to compare the online footprint of an individual to those of known groups may have important intelligence applications as well.

The ability to identify groups that do not behave like (supposedly) similar groups might suggest the emergence of a new group. While this hypothesis will have to be explored in more detail, we did see that there was something in the content of the posts in one of the *gaming* subgroups that differs in significant ways from the others (i.e., Cluster 2 in Figure 1). We suspect this divide would only get greater over time to where characterizing it as a subgroup no longer is warranted. Additionally, these methods may be used to identify when two separate groups begin to merge into a single group. Imagine, if you will, if a new sewing game caught on, there might be much more overlap between the *sewing* and *gaming* groups.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. 2008. Finding High-Quality Content in Social Media. In *Proceedings of the International Conference on Web Search and Web Data Mining* (WSDM '08), 183–194. Palo Alto, Calif.: ACM Press.

Chaua, M. and Xu, J. 2007. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies.* 65(1):57-70.

Chin, A. and Chignell, M. 2007. Identifying communities in blogs: roles for social network analysis and survey instruments*. International Journal of Web Based Communities.* 3(3):345-363.

Gower, J. 1971. General coefficient of similarity and some of its properties. *Biometrics* 27:857-874.

Gregory, M., Chinchor, N., Whitney, P., Carter, R., Hetzler, E., and Turner, A. 2006. User-Directed Sentiment Analysis: Visualizing the Affective Content of Documents. *ACL, Association for Computational Linguistics Workshop on Sentiment and Subjectivity in Text*, ACL, pp. 23–30.

Groh, G. 2007. Groups and group-instantiations in mobile communities – detection, modeling and applications. In *Proceedings of the International Conference on Weblogs and Social Media*. http://www.icwsm.org/papers/paper7.html.

Groh, G. and Rappel, V. 2009. Towards Demarcation and Modeling of Small Sub-Communities/Groups in P2P Social Networks. In *International Conference on Computational Science and Engineering*, 304-311. Vancouver, BC: IEEE.

Kaufman, L. and Rousseeuw.P. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley series in Probability and Mathematical Statistics, John Wiley and Sons Inc.

Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46 (5):604–632.

Kramer, A.D. and Rodden, K. 2007. Applying a user-centered metric to identify active blogs. In *Proceedings of the CHI '07 extended abstracts on Human factors in computing systems.* (CHI '07), 2525-2530. New York, NY: ACM Press.

Ling, K. et al. 2005. Using Social Psychology to Motivate Contributions to Online Communities, *Journal of Computer-Mediated Communication*, Volume 10, Issue 4.

Maia, M. and J. Almeida, V. Almeida. 2008. Identifying User Behavior in Online Social Networks. *SocialNets'08,* Glasgow, Scotland, UK.

Meila, M. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 98:873 – 895.

Orebaugh, A. and Allnutt, H.. 2009. Classification of Instant Messaging Communications for Forensics Analysis. *The International Journal of FORENSIC COMPUTER SCIENCE*, 1:22-28.

Park, H., Lee, J., and Jun, C. 2006. A K-means-like Algorithm for Kmedoids Clustering and Its Performance*. Proceedings of the 36th CIE Conference on Computers & Industrial Engineering.* pp.1222-1231. Taipei. Taiwan.

Passant, A., Heitmann, B., and Hayes, C. 2009. *Using Linked Data to build Recommender Systems*. RecSys New-York, NY USA.

Qualman, E. 2010. *Socialnomics*. New York, NY: Wiley.

Wang X, and Kaban A. 2008. A dynamic bibliometric model for identifying online communities*. Data Mining and Knowledge Discovery*. 16:67-107.

Webb, N., Hepple, M., and Wilks, Y. 2005. Dialogue Act Classification using Intra-Utterance Features. *Proceedings of the AAAI Workshop on Spoken Language Understanding, Pittsburgh.*