A Sentiment-Aware Approach to Community Formation in Social Media

Thin Nguyen[*], **Dinh Phung**[*], **Brett Adams**[†] **and Svetha Venkatesh**[*]
[*]School of Information Technology, Deakin University, Geelong Waurn Ponds Campus, Australia
[†]Department of Computing, Curtin University, Perth, Australia
{*thin.nguyen,dinh.phung,svetha.venkatesh*}*@deakin.edu.au,b.adams@curtin.edu.au*

Abstract

Participating in a community exemplifies the aspect of sharing, networking and interacting in a social media system. There has been extensive work on characterising on-line communities by their contents and tags using topic modelling tools. However, the role of sentiment and mood has not been studied. Arguably, mood is an integral feature of a text, and becomes more significant in the context of social media: two communities might discuss precisely the same topics, yet within an entirely different atmosphere. Such sentiment-related distinctions are important for many kinds of analysis and applications, such as community recommendation. We present a novel approach to identification of latent hyper-groups in social communities based on users' sentiment. The results show that a sentiment-based approach can yield useful insights into community formation and meta-communities, having potential applications in, for example, mental health—by targeting support or surveillance to communities with negative mood—or in marketing—by targeting customer communities having the same sentiment on similar topics.

Introduction

Unlike conventional broadcasting media, users in social media can exchange content and interact with others. A means enabling this practice is *communities*, through which people of common interest can join to discuss their preferred topics. To enable users to find suitable communities to join requires the categorisation of those communities into hyper-groups of communities. This task has been attempted by learning the link structures among communities (Kumar, Novak, and Tomkins 2006). A drawback of such an approach is the dynamic nature of media, meaning the link structures are not stable over time. Also, in many instances, these links are not explicit.

An alternative way to discover hyper-groups is to use the content itself and characterise communities by topic, including blog sub-communities (Adams, Phung, and Venkatesh 2010) and tagged media (Negoescu et al. 2009). However, there might be difference in sentiment between two communities discussing the same issues. E.g., where one forum might host conversations about politics in a cerebral, serious-minded and friendly fashion; another will discuss the same issues adversarially, with zest and tolerance of profanity.

Here, we explore the role of sentiment conveyed in the content. We will collate the sentiments for the blogs belonging to one community and perform clustering across these '*bag-of-sentiments*' to uncover meta-groups. These meta-groups produce similar mood or sentiment and thus serve as a barometer of meta-group mood. For comparison, two other conventional aspects of content are used. First, we consider the topics discussed by users in each community by performing a topic-based analysis of the blogs and then cluster these 'bag-of-topics' to discover meta-groups. The expectation is that these meta-groups will share similar topics. Second, we use psycholinguistic features to categorise communities alike in linguistic style into clusters.

Our contribution is to examine the potential for mood to reveal hyper-communities, or inter-community boundaries, not apparent from topical analyses. Our study includes results for community clustering using topical, sentiment-based and psycholinguistic-based features, in a comparative analysis that is the first of its kind. Our hyper-community formulation has direct application to any social media community applications with a textual component, and could serve as a useful feature in a domain whose lifeblood is product differentiation and rapid innovation.

Hyper-community Detection Framework

In this section, we present content-based, sentiment-based and psycholinguistic-based approaches to cluster blog communities into groups automatically—a problem we term *hyper-community detection*. The aim is to group communities that are related in either content, sentiment or both. In the content-based method, we extract topics from blog content using topic-modelling tools and measure content similarity using topic-based representations for clustering. For the sentiment-based case, we investigate the usefulness of including sentiment information in the clustering task. To achieve this, a mood or emotion-bearing lexicon is extracted from blog content and used as features. In the psycholinguistic-based approach, we use psycholinguistic features provided from psychological studies to cluster communities. In all cases we use data extracted from Livejournal for our investigations. However we note that the proposed method is directly applicable to similar datasets.

We crawled the communities listed in the Livejournal directory.[1] From the 579 communities obtained, we extracted a subset consisting of the top 100 communities having the most members across 10 categories (Advice-Support, Creative-Expression, Entertainment-Music, Fandom, Fashion-Style, Food-Travel, Gaming-Technology, Parenting-Pets, Politics-Culture and Television), resulting in a dataset of 211,740 posts by 59,496 users.

Community representation

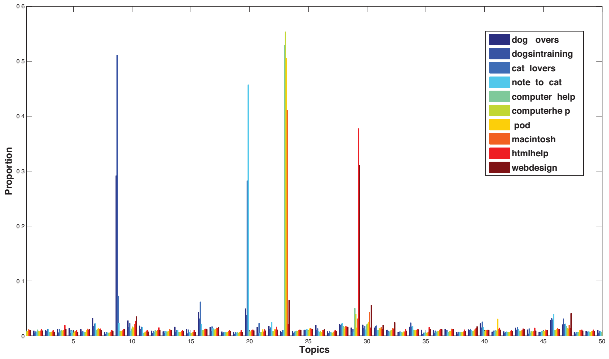
On-line communities come in many shapes and sizes and are affected by many factors, including the demographics of their members, reason for existence and facilities afforded by the hosting application. The Livejournal blog site includes a community feature. Each community is defined by the scope of topics it aims to host and comprises, among other things, members and posts.

Topic-based representation To represent what community members talk about, we apply Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003)—a Bayesian probabilistic topic model—to the corpus of blog posts. All posts for each community are aggregated to form the corpus input to LDA, wherein each post is considered as one document. LDA learns the probabilities $p\,(\text{vocabulary} \mid \text{topic})$, that are used to describe a topic and assigns a topic to each word in every document. Each post can then be represented as a mixture of topics using the probability $p\,(\text{topic} \mid \text{document})$. A topic-based representation for each community is then constructed based on topic mixtures for those blog posts belonging to that community. We expect similar communities to discuss a similar mix of topics, and hence to have a similar mixture of $p\,(\text{topic} \mid \text{document})$ aggregated from their posts.

Formally, let *J* be the number of communities, denote by $\mathbf{x}_j = \{x_{1j}, x_{2j}, \ldots, x_{n_j j}\}$ the set of posts in community *j* where n_j is the total number of posts by this community. Thus, the corpus to be modelled consists of $N = \sum_{j=1}^{J} n_j$ documents aggregated from all communities $D = \cup_{j=1}^{J} \mathbf{x}_j$. Finally, if θ_{ij} denotes the topic mixture for blog post x_{ij}, community *j* can be represented by $\theta_j = (1/n_j) \sum_{i=1}^{n_j} \theta_{ij}$. θ_j is a *K*-dimensional vector, where *K* is the number of topics used by LDA and the k^{th} element represents the mixture proportion of topic *k* for community *j*.

The topic distributions are well separated among some groups of communities. As can be seen in Figure 1, {*dog_lovers*, *dogsintraining*} could be inferred as a group of communities mainly talking about the character ***Dog*** (topic 9); similarly, {*cat_lovers*, *note_tocat*} about ***cat*** (topic 20); {*macintosh*, *computer_help*, *computerhelp*, *ipod*} about ***computer/ipod*** (topic 23); and {*webdesign*, *htmlhelp*} about ***web design*** (topic 29).

Sentiment-based representation Instead of grouping communities based on their topics as in the previous section, we group communities based on sentiment. Sentiment extracted from blog posts is analysed without considering topic. Two methods to extract sentiment from a community are used in this study. If a blog post was tagged with a mood when it was composed, we can use this information to compute an overall sentiment for the community based on moods aggregated from its posts. Otherwise, when mood is not available, we propose to use a sentiment bearing lexicon.



Topic	Top Topic Terms
9	dog dogs puppy training animal gets started ried visit outside pet mo ths house problem walk away ot y un big ov s
16	baby months sleep weeks month birth started start milk question thanks tried hospital doctor eat daughter breast couple pain weight
20	dear cat cats food stop mommy kitty thank bed vet water sleep mom big eat kitten litter glad room clean
23	computer ipod tried problem windows itunes using apple mac thanks drive files screen file running internet music fix open download
29	link table page thanks site code links click text journal change post website layout background thank picture entries box entry

Figure 1: Above: topic proportions of 10 communities. Below: example topics and most likely words sized by $p\,(\text{word} \mid \text{topic})$.

Using mood Livejournal offers 132 moods for users to tag their posts. We assume that there exists a difference in tagged moods among communities, supporting the intuition that such communities can be grouped by mood.

Let $\mathcal{M}$= {*sad, happy, ...*} again be the predefined set of moods where $|\mathcal{M}| = 132$ is the total number of moods provided by Livejournal. Using the notation in the previous section, each blog post x_{ij} in the j^{th} community is further tagged with a mood $m_{ij} \in \mathcal{M}$. For each community, a 132-dimension mood usage vector $\boldsymbol{m}_j$ is constructed whose k^{th} element is the number of times the k^{th} mood in $\mathcal{M}$ was tagged within this community.

Figure 2 shows a plot of the mood usage by eight different communities in Livejournal. It can be seen that the mood usage in one group of communities (*computer_help*, *computerhelp*, *htmlhelp*, *ipod*, *webdesign*) is well separated from another group (*ncisficfind*, *sgagenrefinders*, *sgastoryfinders*).

[1] http://www.livejournal.com/browse/, retrieved July 2011.

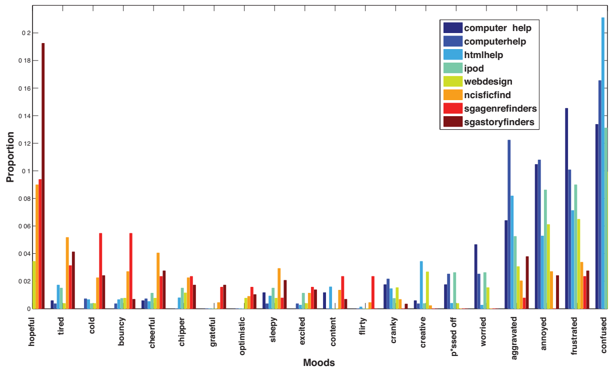


Figure 2: An illustration of mood usage proportions in two groups of communities: {*computer_help*, *computerhelp*, *htmlhelp*, *ipod*, *webdesign*} and {*ncisficfind*, *sgagenrefinders*, *sgastoryfinders*}.

The first group favours using moods having low valence (such as *p*ssed off*, *worried* and *confused*) while the second prefers high valence moods (for instance, *hopeful*), empirically suggesting that it is sensible to study grouping behaviour based on mood.

Using an emotion bearing lexicon When mood ground truth is not available, sentiment-based hyper-community detection can be performed using vectors of sentiment bearing words. In this paper we use the Affective Norms for English Words (ANEW) (Bradley and Lang 1999). ANEW is a set of 1034 sentiment-conveying words rated in terms of valence, arousal and dominance.

In this representation, each j^{th} community is now represented with a 1,034-dimension ANEW feature vector $\boldsymbol{a}_j$, whose k^{th} element is the number of times the k^{th} ANEW word is used in the content of the blog posts made by users belonging to the community.

Psycholinguistic-based representation As a final point of comparison that bridges pure topical and sentiment-based representation of communities, we use psycholinguistic features drawn from the Linguistic Inquiry and Word Count (LIWC) package (Pennebaker, Francis, and Booth 2007). The LIWC package assigns English words to one of 68 linguistic and psychological features.[2] These LIWC features are used to build a vector to provide a psycholinguistic representation of each community.

Community clustering

To group communities into hyper-communities we use affinity propagation algorithm (AP) (Frey and Dueck 2007)—a non-parametric clustering algorithm. AP can automatically discover the number of clusters as well as the cluster exemplars. This is crucial in our setting as the number of hyper-communities can be extremely difficult to know in advance. The algorithm requires the pairwise similarities between data points. In our case, it is the similarity computed between $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_l$ for the (j, l)-pair of communities.

For topical hyper-communities, we note that each $\boldsymbol{\theta}_j$ is a proper probability mass function over topics, summing up to 1. For sentiment and psycholinguistic-based hyper-communities, the feature vectors are normalised. Thus, any suitable probability distance functions can be employed to compute the similarities. In this work, we use the negative Kullback–Leibler divergence (Kullback and Leibler 1951).

Hyper-community Detection Results

Overall clustering performance for the different community representations—topic, mood, mood-proxy and psycholinguistic—is shown in Table 1. We report cluster purity and Normalised mutual information (NMI) (Manning, Raghavan, and Schütze 2008) using the Livejournal community classification, which is a topical classification, as the groundtruth. Therefore, it is expected that these metrics will be highest for the topic-based community representation. We are chiefly concerned with new knowledge discovered using mood-related representations, which will be analysed in more detail below for each type of representation.

	Topic	Mood	ANEW	LIWC
No. Clusters	20	9	15	12
Purity	70%	46%	63%	54%
NMI	62%	43%	59%	51%

Table 1: Cluster purity and NMI of the clusterings based on different community representations.

Topic-based hyper-groups

Using LDA with 50 topics yielded 20 hyper-communities. Clustering appears to have gathered topically similar communities together in a number of cases (e.g., {*ofmornings*, *bentolunch*, *picturing_food*, *trashy_eats*}), but also elucidated finer distinctions (such as in the cases of {*cat_lovers*, *note_to_cat*} and {*dog_lovers*, *dogsintraining*}). On further inspection, a number of its communities have a significant romance or relationships component. E.g., in addition to those communities with obvious topics, three are about particular fictional relationships: *house_cameron*, *sheldon_penny* and *time_and_chips*.

It is found that, in the feature space, the intra-category distances are much smaller than inter-category distances. The smallest distance is between hyper-communities {*cat_lovers*, *note_tocat*} and {*dog_lovers*, *dogsintraining*}, which indicates the high degree of topical commonality of discussion about pets, despite their being different animals. It is interesting to note how close Entertainment-Music and Politics-Culture are. The farthest distance is found between Food-Travel and Gaming-Technology or Fandom.

Mood-based hyper-groups

Clustering based on explicit mood labels yielded nine hyper-communities. In contrast to the topic-based clustering, only two hyper-communities have 100 per cent purity with respect to the topical ground truth, one of which (*htmlhelp*, *computer_help*, *computerhelp*, *ipod*, *webdesign*, from Gaming-Technology) is characterised by negative mood.

[2] http://www.liwc.net/descriptiontable1.php, retrieved Jan 2012.

Mood-based clustering reveals distinctions not apparent in the topic-based representation. E.g., the group including *behind_the_lens*, while having significant overlap with the group with *behind_the_lens* in the topic-based clustering, has some illuminating differences: gone are the communities *beatlepics, madradstalkers*, *ru_glamour*, *topmodel* and *worldtourist*; replacing them are *add_a_writer*, *just_good_music* and *ofmornings*.

From an appraisal of the content of these communities we find the distinctions to be nuanced. The topic-based hyper-community is loosely united by pictures and people, whereas the mood-based hyper-community is united by the desire to create and its outcomes—differences that are best explained by prevailing mood and intent. Indeed, these distinctions are captured by the predominant moods of the different hyper-communities, respectively *curious*, *cheerful* or *happy* versus *calm*, *accomplished* and *creative*.

ANEW-based hyper-groups

Clustering based on ANEW features as proxy mood yielded 15 hyper-communities. Of these, five consisted of communities with matching Livejournal categories (e.g., *curlyhair*, *beauty101*, *dyed_hair* and *vintagehair* all classified as Fashion-Style). Two hyper-communities are examples of the sub-category distinctions returned by the topic-based clustering: {*macintosh*, *computer_help*, *computerhelp*, *ipod*, *webdesign*} and {*worldofwarcraft*, *gamers*, *wow_ladies*} are both from Livejournal's Gaming-Technology category.

LIWC-based hyper-groups

Clustering based on psycholinguistic features yielded 12 hyper-communities. Three hyper-communities contain communities with the same Livejournal category and appear to have been associated topically. The top three LIWC categories for these hyper-communities are illuminating: for Fashion-Style, *feel*, *body* and *percept* (i.e., perceptual processes); for Food-Travel, *ingest*, *bio* (i.e., biological processes) and *percept* and for Parenting-Pets, *family*, *health* and *humans* (e.g., adult, baby and boy).

Other hyper-communities appear to exhibit a characteristic mixture of topic and style of discussion, which is in part captured by the psycholinguistic processes of LIWC. E.g., {*sheldon_penny*, *adayinmylife*, *house_cameron*, *miracle______*, *rpattz_kstew* and *time_and_chips*} aggregates all of the communities in the dataset about fictional relationships (plus one community about documenting a day in one's life). These communities are a kind of meta-genre not easily captured by topical features alone. Linguistic features, such as post length and extensive use of the third personal singular (i.e., *shehe*), appear to help associate these communities.

Discussion

It is not surprising that the different community representations lead to hyper-communities that reflect these varying emphases. Topic-based representation is the method of choice for recovering hierarchy within, and associations across, Livejournal's canonical topic categories. Likewise, the results for the mood-based representation indicate an ability to recover non-topical features of a community such as prevailing intent and atmosphere of discussion. However, contrary to expectations, ANEW does not appear to be a well-suited and cheap alternative to mood-based representation for the task of hyper-community detection.

The clustering results for LIWC's psycholinguistic representation are worthy of follow-up. LIWC offers a wide scope of classification—due to including topical, linguistic, stylistic and mood categories—yet is cheap to obtain. Some of the distinctions captured by the hyper-communities arising from LIWC representation are a kind of topic + atmosphere that seems relevant to the Web 2.0 denizen, who is faced with a surfeit of choice and whose decision as to which community they will invest in may turn on the presence of more than one characteristic of the content. Consequently, psycholinguistic analysis demonstrates potential for use in community recommendation (and analysis).

Conclusion

We have investigated the problem of discovering hyper-groups of communities by using topics, sentiment information and psycholinguistic properties of the posts of members. We presented an unsupervised approach based on a non-parametric algorithm to detect hyper-groups of communities in the blogosphere and to reveal interesting content-, sentiment- and psycholinguistic-based grouping behaviours.

We have proposed a novel approach for addressing hyper-community detection based on users' sentiment. The grouping of meta-communities based on sentiment information has potential applications in, e.g., mental health or in marketing. In addition, the psycholinguistic hyper-groups detected provide insight into the language styles of people in specific categories (e.g., Fashion-Style bloggers favour spoken language) while topical hyper-groups enable users to find suitable communities based on their interests.

References

Adams, B.; Phung, D.; and Venkatesh, S. 2010. Discovery of latent subcommunities in a blog's readership. *ACM Transactions on the Web* 4(3):1–30.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bradley, M., and Lang, P. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. *University of Florida*.

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

Kullback, S., and Leibler, R. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Kumar, R.; Novak, J.; and Tomkins, A. 2006. Structure and evolution of online social networks. In *Proc. of SIGKDD*, 617.

Manning, C.; Raghavan, P.; and Schütze, H. 2008. *Introduction to information retrieval*. Cambridge University Press.

Negoescu, R.; Adams, B.; Phung, D.; Venkatesh, S.; and Gatica-Perez, D. 2009. Flickr hypergroups. In *Proc. of ACMMM*, 813.

Pennebaker, J.; Francis, M.; and Booth, R. 2007. Linguistic inquiry and word count (LIWC) [computer software]. *LIWC Inc.*