Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

User Interest and Interaction Structure in Online Forums

Di Liu, Daniel Percival and **Stephen E. Fienberg**
Department of Statistics
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
diliu, dperciva, fienberg@stat.cmu.edu

Abstract

We present a new similarity measure tailored to posts in an online forum. Our measure takes into account all the available information about user interest and interaction — the content of posts, the threads in the forum, and the author of the posts. We use this post similarity to build a similarity between users, based on principal coordinate analysis. This allows easy visualization of the user activity as well. Similarity between users has numerous applications, such as clustering or classification. We show that including the author of a post in the post similarity has a smoothing effect on principal coordinate projections. We demonstrate our method on real data drawn from an internal corporate forum, and compare our results to those given by a standard document classification method. We conclude our method gives a more detailed picture of both the local and global network structure.

Introduction

Social network analysis has grown as a topic of interest with the growth of the internet as an interactive environment, especially in connection with online communities. The general goals of these approaches include characterizing user behaviors and interactions, as well as extracting information from actual user discussions. In this paper, we define a measure of similarity between users of an online forum, based on a modification of document classification, which takes into account both their interests and interactions.

Establishing a notion of distance or similarity between the people in a social network provides a useful way to illustrate the structure of the social network. For example, we might define similar people to represent friendship, shared interest, or similarity in skill. These interpretations give user similarity a wide variety of applications. For example, recovering friendship from another form of personal interaction data is useful in sociological studies. People with similar interests could be targeted with a certain advertisement or product suggestion. A company could assign people with similar skills to work together on a project.

We base our method on establishing a measure of similarity between all posts created by all users of an online forum. Our measure takes into account both the textual information and the particular context of the online forum. Usual approaches take only the textual information of the posts into account. From similarity between posts, we establish similarity between users. We then use this similarity to investigate the structure of the social network.

Our Data

Online Forums We examine user similarity in the context of an online forum. An online forum is a system designed for the discussion of topics, with each topic separated into its own area, called a thread. A thread is begun by a user writing a short document, called a post, which introduces the topic or asks a question about the topic. Typically, this user also writes a separate title for the thread, which summarizes or highlights the thread topic. Other users can then continue the discussion by adding their own posts to the thread. Thus each thread in the forum is a place where many users discuss a certain topic.

Corporate Forum Data

Our data come from a global IT company. The company created an internal forum in order to enhance information flow between employees. We have data collected from this forum over a one year period from August 2006 until August 2007. Over this period, 2,974 users wrote 79,128 posts in 20,090 threads. The users of this forum are skilled IT professionals, and so the topics discussed in this forum are very technical and specific. The company is interested in grouping employees in creative ways based on the employee's skills, areas of interest, and other strengths.

By using the available thread ID and user ID information, we can link posts to threads, and authors to posts. Table 1 gives a summary of the attributes of the forum data. We see that most posts only contain a few words. As we will see, this makes it difficult to apply traditional document classification methods, which treat each post as a document. We also see that most users write only a few posts, and each thread is only a few posts in length. All of this means that most posts have very little or no content, thread, and user information in common. Our method will seek to address these issues.

Table 1: A summary of the attributes of the corporate forum data. Note that the post word counts include stopwords.

Attribute	Min	1st Quartile	Median	3rd Quartile	Max
Words in a Post	1	14	28	59	8980
Posts made by a User	1	1	5	22	975
Posts in a Thread	1	2	3	5	265

Table 2: A sample thread from the corporate forum data set. Here the posts in the thread do not share many words in common. Traditional document classification methods would therefore consider these posts nearly unrelated.

	Thread Title: Madriva 2007 3D desktop
Post 1	Anybody tried mandriva 2007? Its cool with a XGL 3D desktop.. But is hungry for RAM..
Post 2	You should give ubuntu 6.10 (or the 7.04 dev) a try. You might also find this interesting: [HYPERLINK]
Post 3	And lookout for KDE Plasma. More info in : [HYPERLINK]
Post 4	Here are few resources on getting Beryl (beryl.. is extremely irresistable.. enter at your own risk :-)) [HYPERLINK] [HYPERLINK] [HYPERLINK] (best of all)

Method

Our method consists of two main steps. In the first step, we create a matrix which measures the similarity between all pairs of posts in the forum. In the next step, we build a similarity matrix for users by creating a coordinate system based on the similarity matrix from the first step. The results of this second step allow us to examine the structure of the relationships and activity of the forum users.

Measuring Similarity Between Posts

If we consider each post as a document, then our goal is to establish a notion of similarity between the documents. Methods such Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) and cosine similarity (Weiss et al. 2005) have been shown to be effective document classification techniques. Since we wish to establish a numerical measure of post similarity, we will modify cosine similarity.

Cosine similarity is insufficient for analyzing forum data. Cosine similarity is based on word overlap, and is most effective when applied to long documents. However, a typical forum post is only a few sentences long. Additionally, cosine similarity mistreats or ignores information available by considering the threads and the author of the posts. These two issues result in a very sparse similarity matrix — many documents intuitively related by thread or author have no relation at all.

To address these problems, we modify cosine similarity as follows to take into account all of the available information in forum posts:

- We append to each post the title of the thread in which it appears. This makes posts within the same thread more similar in word content and therefore closer in cosine similarity.

  Posts made in the same thread might share little or no words in common, even though they are on the same topic. Such posts would not be considered related under usual cosine similarity. Table 2 illustrates this problem via an example of a typical thread in our data set.

  We use the thread titles since they roughly represent the topic of the thread. Additionally, a user typically only reads the title of the thread before deciding to read the rest of the thread and then possibly making a response post. Therefore, the thread title captures both post topic and user interests.

- We modify the tf-idf(D, j) measure of word importance to take into account the thread in which a document D appears - see (Weiss et al. 2005) for a full description of tf-idf. tf-idf measures word importance only using the overall frequency of a word. However, if a word appears often in a particular thread, then it is likely to be of particular importance to the thread topic, whether or not it is a common word in an overall sense. Table 3 gives an example of a thread which illustrates this point.

  We define $T(D)$ to be the document consisting of the concatenation of all posts in the thread containing document D. We then define:

$$\text{df}_{T(D)}(j) = \frac{\text{df}(j)}{\text{tf}(T(D), j)}.$$

  Which gives us the following formula, where N is the total number of documents:

$$\text{tf-idf}_{T(D)}(D, j) = \text{tf}(D, j) \times \log_2(N/\text{df}_{T(D)}(j)). \quad (1)$$

  This new measure takes into account the importance of a word within a thread. Examining the original df(j) measure, we see that as df(j) increases, the importance of word j goes down. Dividing by the thread word frequency tf$(T(D), j)$ means df$_{T(D)}(j)$ decreases as a word becomes more common within a thread.

  Note that in combination with the previous point, we have that the words in the thread title are of great importance to the thread topic. Since the thread title usually represents the topic, this is a desirable effect.

- After computing the cosine similarity using the above modifications, we add an additional term to capture our belief that documents authored by the same user are similar. Since we want this term to be independent of both post content and the particular user, this term should be a universal constant.

Table 3: A sample thread from the corporate data set. Here the words "data" and "migration" appear frequently in the posts. Therefore, within this thread, these words should be given high importance. In the usual tf-idf framework, these words would be given high importance only if they were relatively rare throughout the forum.

	Thread Title: data migration
Post 1	Basically what is data migration?
Post 2	Data migration, basically means to porting data from one environment (format/OS/Database/Server etc) to other environment.
Post 3	The process of translating data from one format to another. Data migration is necessary when an organization decides to use new computing systems or database management system that is incompatible with the current system. Typically, data migration is performed by a set of customized programs or scripts that automatically transfer the data.
Post 4	Migrating to higher version also one of the part in data migration.

Our goal is not to cluster posts or to assign posts to users. Rather, we are interested in examining the relationships between users. This term does not affect the distance of posts written by different users. Therefore, the inclusion of an author term is not a circular step. This term has a smoothing effect – it reduces the within group variance for each user.

We therefore modify the cosine similarity equations by replacing $\mathrm{df}(j)$ with $\mathrm{df}_{T(D)}(j)$, and by replacing each post D with post D^*, which has the thread title appended. We define the function $U(D)$ to return the author of post D. We then define:

$$\mathrm{sim}(D_1, D_2) = \mathrm{cosine}_{T(D)}(D_1^*, D_2^*) + \lambda I_{\{U(D_1)=U(D_2)\}} \quad (2)$$

$$\mathrm{dist}(D_1, D_2) = \max\left(0, 1 - \mathrm{sim}(D_1, D_2)\right). \quad (3)$$

Here, λ is our universal author constant as discussed above and $\mathrm{cosine}_{T(D)}(D_1, D_2)$ represents the cosine distance, with the modified tf-idf measure given in equation 1 replacing tf-idf in both the distance and the norms. We then convert the similarity measure in equation 2 to a dissimilarity measure via equation 3. Note that the maximum of the cosine similarity measure is 1. This formula is applied to all pairs of posts, giving us a dissimilarity matrix between all posts.

Measuring Similarity Between Users

We now seek to create a dissimilarity matrix between all users in the forum, given the dissimilarity matrix between all the posts. We first seek to visualize the relative position of all the posts in some low dimensional space. Note that our dissimilarity matrix only gives us a function of the position of the posts, it does not give the coordinates directly.

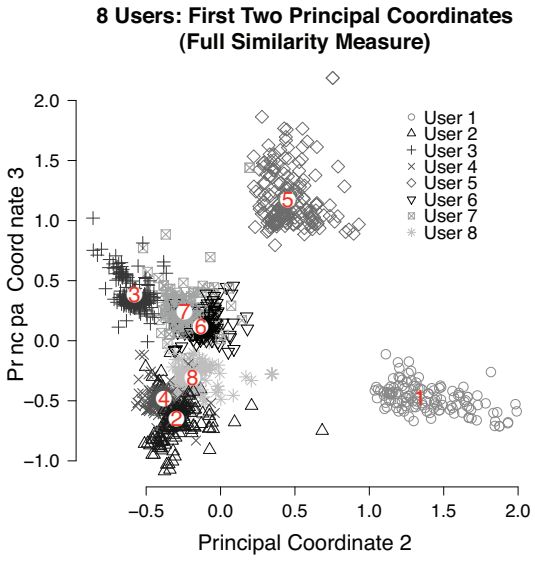

Figure 1: The second and third principal coordinates for the 8 user data set. The centroids are given by the numbers in the white circle. We can clearly see each user separately, but are still able to see spread information and the similarity between users.

Therefore, in order to visualize this result and to facilitate further computation, we find a low dimensional representation of the posts which preserves the geometry implied by the original dissimilarity matrix. We achieve this via principal coordinate analysis (Hastie, Tibshirani, and Friedman 2001).

Towards the main goal of this paper, we can also use this representation to characterize users. We can give each user a single set of coordinates by finding the centroid of all that user's posts in this low dimensional space. This roughly gives us a center which follows the areas of high density. The universal author constant λ reduces the variance of the centroid, which is the estimator of the position of the user in this space.

Using these centroids, we can simply make a distance matrix between users by taking the euclidean distance between all pairs of user centroids. This distance matrix can characterize the social network structure in a wide variety of ways, such as clustering, spanning trees, or nearest neighbor methods.

Results

Illustration: A Small Set of Active Users

In order to illustrate our method clearly and intuitively, we present results for a small set of users. We consider all users who wrote between 200 and 210 posts on the forum. In the full dataset, the set of users who wrote more than 200 posts accounts for about 50% of all posts. Therefore, this range represents users who are roughly in the middle in terms of posting activity. These users are also easier to compare since they wrote roughly the same number of posts. This range

gives us eight users in total.

Using only the subset of posts authored by these users, we apply our previously described method. Note that the dictionary is built from all of the posts, so the tf-idf measures take into account the overall importance of the words. Therefore, the tf-idf measures are not biased in this case, and so our results for these particular users will not differ greatly from those obtained when all of the posts are included in the analysis.

Figure 1 shows the second and third principal coordinates for the posts. We use $\lambda = 0.059$ for our universal author similarity constant. This is the 75th quantile of our nonzero similarities obtained without taking author into account. We can see a clear separation between users, as well as differing spreads for each user. Looking at the centroids, we see six users who are somewhat similar, and two users who are separated by these principal coordinates. These two users also seem to have larger spreads than the other six, which possibly indicates a broader interest in topics.

To illustrate an application of the user distance, we next build a hierarchical clustering tree using complete linkage (Hastie, Tibshirani, and Friedman 2001). From this tree we see that there are two main clusters: users $\{1, 5\}$ and users $\{2, 3, 4, 6, 7, 8\}$. This is consistent with our earlier display of the users in Figure 1. This clustering gives a picture of the network structure within this 8-user group.

Active Users

We now consider a large set of active users in the corporate forum data set. We take all users who wrote between 200 and 400 posts on the forums. This gives us 71 users and 18,682 total posts. We consider this subset for computational and interpretive reasons.

We apply our method to these 71 users, with $\lambda = .054$. This λ is obtained from the 75th quantile of the nonzero similarity matrix entries for the modified cosine similarity measure. Due to the large number of users in this data set, plots of the principal coordinates do not give clear pictures of the relationships between users. Note that plots of subsets of users can show individual user spread.

As mentioned before, there are many ways to look at the social network structure once we calculate the distance between all of the users. We present a well-known examples here: complete linkage clustering.

Figure 2 shows the complete linkage hierarchical cluster dendrogram for the 71 users. We see that a five cluster solution looks appropriate. The majority of the users are in two of these five clusters. The remaining three groups are small and separated from these two large groups. In particular, we see a group of three users located far away from the other four groups. This group may represent a collection of users who have the same specialized interest. The two large groups perhaps deal with general or popular topics.

Discussion

Our main contribution is a new similarity measure between posts in a forum. This measure effectively modifies document similarity to incorporate the special structure of forums. We discussed the properties of our modification, and presented some results on a real data set.

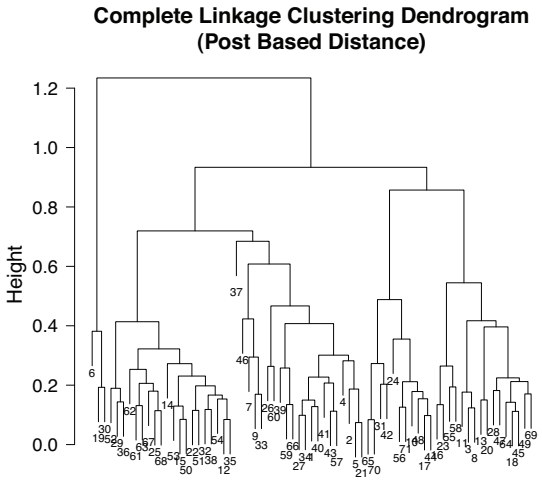

Figure 2: Cluster dendrogram for the 71 users. We can create a clustering by drawing a horizontal line at any height in the tree, and taking the clustering given by the links below the line. Overall, we see that five clusters seems most appropriate.

Users in a forum demonstrate their interests and interactions with other users in two ways. First, users write posts whose words can tell us in which topics they are interested. Second, users post in particular threads, indicating both topic interest and interaction with the other users who have already posted in the thread. By including the title of threads in each post, we view both types of information in a single unified context. User interactions within a thread are transformed into shared words.

In our database, thread titles tend to be very short once we remove stopwords. Therefore, the longer the post, the smaller the effect of the thread title on cosine similarity. Since longer posts contain more textual information, this is a desirable effect. Our modification give us information about posts which are otherwise hard to characterize.

We need a more systematic way to pick the author constant λ. We believe λ also has beneficial properties with regard to statistical testing. We would like to develop a framework to investigate these properties. Different estimators of user location besides the centroids proposed in this paper may lead to more rich estimators of user distance. We are also interested in additional validation of this method on data with some known and recoverable social structure.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Jornal of Machine Learning Research* 3:993–1022.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. *The Elements of Statistical Learning*. Springer.

Weiss, S. M.; Indurkhya, N.; Zhang, T.; and Damerau, F. J. 2005. *Text Mining*. Springer.