Predicting User Quality in Anonymous Chat Networks

Claire Negiar and Dhruv Amin

Stanford University, Department of Computer Science

1. Introduction and Prior Work

Online social networks like Facebook and Twitter have been multiplying over the past few years, along with the amount of data that these sites handle and process. A lot of work has been done to optimize the quality and informativity of the data that reaches their users, such as Facebook's EdgeRank newsfeed calculation, or friend and follower suggestions. However, little such analysis and optimization has been done for less conventional platforms such as anonymous chat networks, which have much to gain from this kind of work. In particular, being able to predict users of poor "quality" is of paramount importance as they rarely have long conversations and are therefore over represented in the matching queue, which in turn affects the conversation quality of a large set of users. Past attempts at anonymous social networks (i.e. Chatroulette, etc.) have failed because of this very problem, as users of high quality leave the network after repeated interactions with suboptimal users. Offering a better filter of users would allow anonymous networks to create sub queues, or queues populated almost entirely of one type of user, enabling a better user experience and higher activation rate.

While many attempts have been made to predict optimal matching between pairs of users in network settings, little work has been done on classification of users based upon their own features. Most work on the topic has focused on the edges of the graph; in this project we hope to classify users simply by examining features of the nodes. On most networking platforms, the process of identifying malignant users or content is tedious, requiring user participation to report those that abuse the community. Our goal in this project is therefore to explore different representations of the chat network's users that are effective at separating users, or subsets of users. Ultimately, we wish to isolate groups of 'good' users so that they can be placed in a prioritized sub queue. These 'good' users would talk exclusively with other users from this sub queue, yielding higher average conversation lengths, better matching, and ultimately better retention rates on the site.

2. Data

Chatous is a text-based, 1-on-1 anonymous chat network that was used as the basis for exploration into user quality. Users can create a profile that contains a screen name, age, gender, location, and a short free-form "about me" field. After clicking the "new chat" button, users are matched up with one another in a text-based conversation. Interactions on Chatous include exchanging messages, sending/accepting a friend request, reporting an abusive user, and ending a conversation.

In order to extract relevant features for our work, we operated on all data collected from two weeks of user activity on the Chatous platform, which consists of approximately 80,000 users and 8 million conversations. The data was initially formatted in a graph structure, with users as the nodes and conversations as weighted edges with length as weight. We were also given access to profile data relating to these nodes (screen name, age, gender, location and "about me") and meta-data surrounding the edges (person to disconnect conversation, time started, time ended, friendship status, and word frequency vectors of the conversation with the underlying words anonymized by numeric ids). From this conversation data for two weeks, we transformed the graph structure into a schema with the user as opposed to the conversation as the atomic unit. Users were mapped to their entire conversation history, profile information, and meta-data surrounding their conversations. In order to increase efficiency, we select at random 10,000 of these users to serve as a training set for our exploration. Finally, a 'gold' set of 1034 users was labeled as belonging to the categories 'clean', 'dirty' or 'bot' based on a human examining their conversation history and making an intuitive decision on the nature of the user.

3. Quality Metric

In defining what it meant to be classified as a 'good' or 'bad' user, we decided that two key metrics could be

used to evaluate the performance of our algorithms: average conversation length of the user across all conversations and accuracy predicting the labels of the gold set. In past versions of the Chatous platform, users had been required to rate each conversation and this rating was collected as conversation meta-data; this rating requirement was dropped when it became clear that it directly correlated with length of conversation. Thus, we could consider one definition of suboptimal users as those for whom the average lengths of the conversations they are a part of is lower than the mean conversation length (2.58 lines per user). Yet another definition of user quality relies not upon the length of conversations, but on the type of conversations the user engages in. Users that harass or use a higher proportion of vulgar words are labeled as 'dirty' in the human created gold set. While there is some dependency between the two metrics, monitoring how our algorithms perform on both metrics provides a qualitative measure on how successful it is at determining which users are optimal. Ultimate choice of algorithm depends on which metric the reader desires to improve.

4. Data Model

Our method of representing a user evolved as we began to understand the problem more. Initially, users were treated as a vector of behavioral features that tried to capture how the user interacted on Chatous generally. When this approach ultimately failed to capture the same information as our human labeled gold set, we modified our approach to defining users based upon the terms in their conversation history.

4.1 Behavioral Feature Set

Originally, we hypothesized that 'clean' and 'dirty' users would interact with the Chatous platform in fundamentally different ways. Therefore, we started with a set of features that we believed would capture these differences in behaviors and inform proper labelings.

Feature	Motivation
Num of 'Long' conversations	Better users will have longer conversations
Num of 'Short' conversations	
Avg conversation ratio	Better users speak evenly with partner
Num reports generated against user	Harassing users are reported
Num reports generated by user	Users may abuse reporting system
Num of 'Friendship'	Better users have more friendship establishing convos
Num conversation zero length	Sign the user has high churn
Num conversation user finishes	User is the cause of churn
Num profiles used	Users changing their representation flagged
Num times gender changed	Gender change should be rare
Num times location changed	Location change should be rare
Num times age changed	Age should not change if true
Num times username changed	User changing representation
Variance in word history[1]	Bots and harassing users tend to say similar phrases in all convos

4.11 Logistic Regression

Training a random selection of 10,000 users from the Chatous dataset, we attempted logistic regression on the behavior features in order to see if these could accurately predict on the gold, human labeled data set of 1034 'dirty' and 'clean' users. In training our logistic regression, we initially used the user's average conversation length as the deciding factor on whether or not they were 'clean'. However, using the average length of conversation as a proxy training label ends up being suboptimal (74.2% as best accuracy rate), suggesting that the human labeling cannot be reduced to average conversation length despite the fact that the metric holds for judging individual conversation quality.

We therefore performed holdout cross validation using the smaller, hand-labeled dataset with 1034 users. The decision boundary created by the logistic regression procedure ends up labeling all of the users as 'clean' since this is the value that maximizes the accuracy for these features, yielding a 79.35% accuracy. This boundary occurs because the actual ratio of 'clean' to 'dirty' users is in fact approximately 78.5: 21.5. We performed tests on every possible subset of features to find the set of features that best separates our data. However, no subset of our features was able to achieve over 79.35% accuracy, with the lowest barely achieving 21%.

4.12 K – Means Clustering

Given that logistic regression failed to create a separating bound, we decided to move to an unsupervised algorithm to detect patterns in our data. Our hypothesis was once again that through clustering our data on similar behaviors, we could isolate pockets of users that were in fact of the same type. The ultimate goal changed to create cluster centers that grouped users of the same type together, or at the very least, improved upon the true distribution of 78.5:21.5 'clean' to 'dirty' users. This result would allow us to realize our goal of sub queuing on the live site.

[1] Average squared Euclidean distance between all unique combinations of a user's

Our gold-labeled dataset contains 230 'dirty' users and 804 'clean' users (22% dirty). We ran a k-means clustering algorithm for all k between 2 and 8 and achieved the following results:

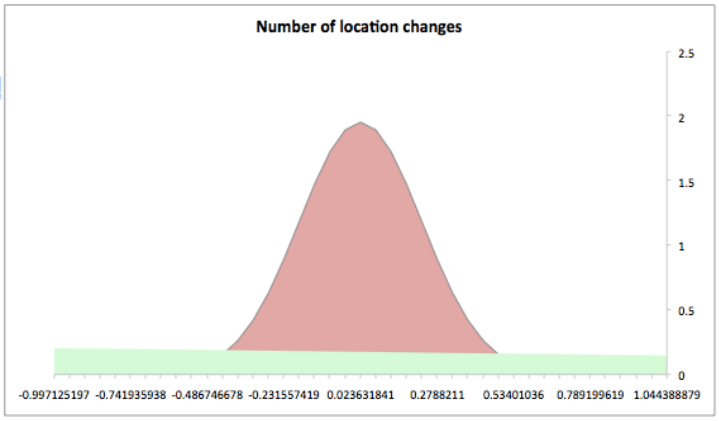
K = 2			
Cluser 1 size	64 users	Cluster 2 size	970 users
% dirty users	22%	% dirty users	26.56%
K = 4			
Cluser 1 size	46 users	Cluster 2 size	134 users
% dirty users	21.74%	% dirty users	26.87%
Cluser 3 size	33 users	Cluster 4 size	821 users
% dirty users	24.24%	% dirty users	21.4%
K = 8			
Cluser 1 size	12 users	Cluster 2 size	532 users
% dirty users	16.67%	% dirty users	22.74%
Cluser 3 size	41 users	Cluster 4 size	189 users
% dirty users	19.51%	% dirty users	21.70%
Cluser 5 size	10 users	Cluster 5 size	129 users
% dirty users	20%	% dirty users	20.16%
Cluser 7 size	40 users	Cluster 7 size	81 users
% dirty users	30%	% dirty users	22.22%

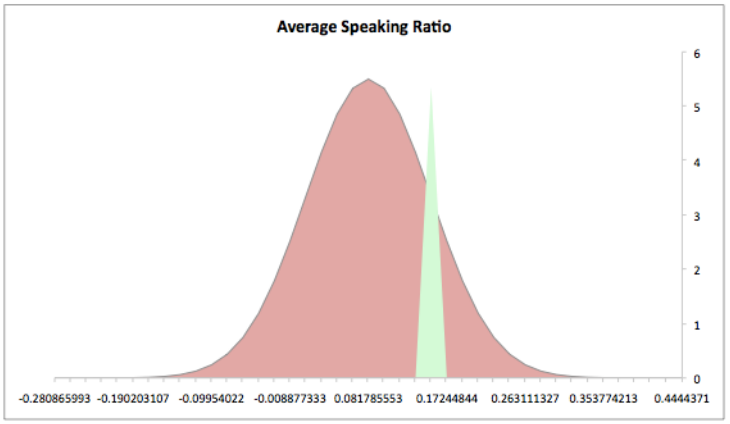
Given that our goal was to find clusters with more heavily skewed distributions of 'dirty' and 'clean' users than in the labeled data, our algorithm achieves moderate success. In the k=2 scenario, one cluster attains 26.5% of users 'dirty' (a 19.4% increase in the ratio of 'dirty' users from the original dataset). With higher numbers of clusters, the algorithm continues to 'polarize' the distribution, indicating that behavior captured by many of the clusters is more extreme. For instance, in the k = 8 scenario, the best cluster posts a 35% increase in the proportion of 'dirty' users, as well as two clusters with a 28% and a 14% increase in the proportion 'clean' users, respectively. However, these three clusters only represent 17.5% of our data, and the remaining 82.5% of the data is not differentiated. This observation led us to reconsider whether behavioral features were in fact differentiating.

For each of our behavioral features, we examined the potential distribution of the feature as if it were drawn from a Gaussian distribution[2]. The following means and variances were plotted for each behavior:

Avg Length Conversation			
Clean Mean	4.08	Dirty Mean	.34
Clean Var	39.4	Dirty Var	8.9
Percent Long Conversation			
Clean Mean	0	Dirty Mean	0
Clean Var	0	Dirty Var	0
Percent Short Conversation			
Clean Mean	.009	Dirty Mean	0
Clean Var	.009	Dirty Var	0
Avg Speaking Ratio			
Clean Mean	1.1	Dirty Mean	.008
Clean Var	.045	Dirty Var	.07
Avg Num Reports Against			
Clean Mean	3.0E-3	Dirty Mean	5.8E-5
Clean Var	1.4E-4	Dirty Var	7.3E-7
Avg Num Reports Filed			
Clean Mean	2.7E-3	Dirty Mean	1.4E-4
Clean Var	1.3E-3	Dirty Var	8.0E-6
Percent Friendship Conversation			
Clean Mean	.6	Dirty Mean	.05
Clean Var	.26	Dirty Var	.04
Percent Zero Length			
Clean Mean	.74	Dirty Mean	.05
Clean Var	.07	Dirty Var	.04
Percent User Finished			
Clean Mean	.44	Dirty Mean	.04
Clean Var	.06	Dirty Var	.02
Number of Profiles Used			
Clean Mean	4.0	Dirty Mean	.2
Clean Var	48	Dirty Var	1.3
Number of Gender Changes			
Clean Mean	0.24	Dirty Mean	.01
Clean Var	1.1	Dirty Var	.04
Number of Age Changes			
Clean Mean	.3	Dirty Mean	.02
Clean Var	1.6	Dirty Var	.2
Number of Location Changes			
Clean Mean	.2	Dirty Mean	.01
Clean Var	.49	Dirty Var	.04
Number of Username Changes			
Clean Mean	.55	Dirty Mean	.03
Clean Var	2.8	Dirty Var	.2
Conversation History Variance			
Clean Mean	.7	Dirty Mean	.05
Clean Var	.12	Dirty Var	.04

A few examples of the Gaussian plots:

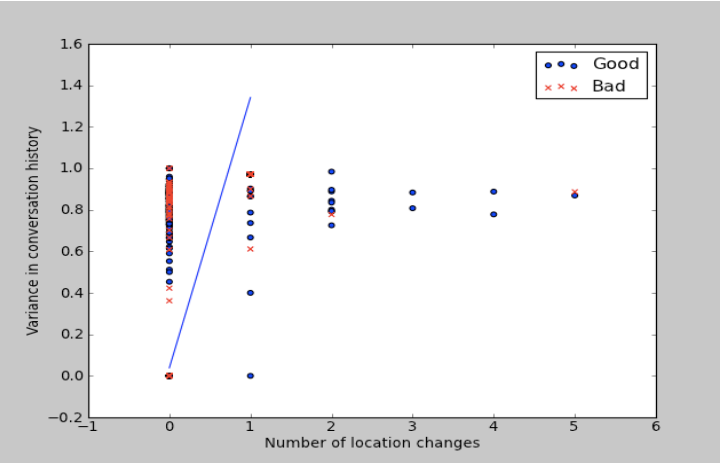




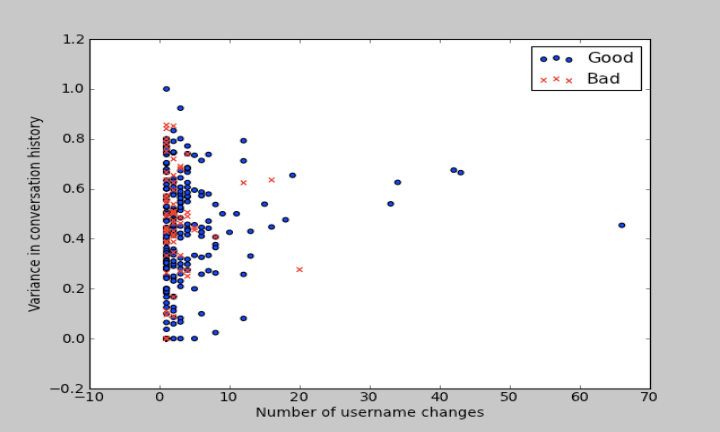
[2] This is a simplifying assumption that is only useful for comparison. True distribution across all users for each feature is of course unknown.

The striking aspect of these results is that the variance in the set for the labeled 'clean' users is almost always relatively large compared to the 'dirty' users. In a few select cases (average conversation length, speaking ratio), the variance in the 'dirty' users encompasses the range of 'clean' behaviors. As a result, the two spectrums of behavior almost always overlap in some form to the point that they are almost indistinguishable. Thus, few features in the behavioral feature set carry informative value when trying to separate the two labels.

Finally, in order to confirm that no subset of features would be more informative than the aggregate, we tested K-means and Logistic Regression on all subsets. Plotting features against one another yielded the same conclusion as above; it is difficult to separate users based upon behavioral features into the labels 'clean' and 'dirty'.



Logistic regression accuracy (69%) suffers since data is inseparable.



Best performance when all points labeled the same is an indication that few data axes spread the data.

4.2 Conversation History Feature Set

Given that the 'gold' set of data was created based upon a human examining users' conversations to assess whether a user's conversations tended to be 'dirty' or 'clean', the final approach we considered was clustering based on upon the textual similarity of users' conversation histories. As the proportion of vulgar words is much higher in dirty users' conversations, the assumption in this model is that it should be possible to separate the two into clusters based upon this difference in their word usage.

4.21 TF-IDF Document Clustering

For each user, we generated a 'document' of the users' entire conversation history. TF-IDF, which is defined as,

$$w_{i,j} = tf_{i,j} * \log (N / df_i)$$

where $tf_{i,j}$ is the number of occurrences of the word i in document j, df_i is the number of documents containing i, and N is the total number of documents, is used to generate a representation of each term w_i's importance to d_j, the document it is contained within. In this situation, the document is the users' entire conversation history. Due to the sparse nature of the data, the TF is normalized by dividing by the max frequency of any term in a given document. Using TF-IDF allows us to vectorize a user's conversation history and then use the cosine similarity, which is defined as,

$$\cos sim = \frac{v \bullet w}{\|v\|\|w\|} = \frac{\sum_{i=1}^{n} v_i \times w_i}{\sqrt{\sum_{i=1}^{n}(v_i)^2} \times \sqrt{\sum_{i=1}^{n}(w_i)^2}},$$

to determine how similar two users are in terms of the words they've used[3]. In order to normalize the lengths of the vector representations of the two conversations being compared, elements of the vocabulary that were not observed in a document were set to 0. Finally, in order to further de-emphasize words that did not create differences between user conversations, we identified the 670 most frequently used words, or stop words, and removed them from every user's conversation history.

Given that our objective function is non-convex, we are subject to finding local optima depending upon our starting conditions. In order to account for this, we ran the algorithm over many iterations with random starting points until we attained a sense of the global clustering. In order to determine K, we determined the median distribution over 10 runs of the algorithm for each value of K from 2 to 10. The median clustering is determined by ranking the outputs of the algorithm in terms of which increase the skewed nature of the data the most.

[3] We also briefly experimented using squared Euclidean distance as our distance metric. Square had the benefit of emphasizing any difference between the two histories but it underperformed cosine similarity as it over emphasized word frequency differences as opposed to word type differences.

Ultimately, the clustering that achieved the best performance had k = 2 cluster centers and created a slight improvement over the typical ratio of 'clean' to 'dirty' users. Since some users' conversation histories consisted entirely of stop words, these users were not considered in determining the cluster centers (cosine similarity undefined for 0 length vector). Therefore, these users were separated out and placed in their own undetermined cluster. Median results of k = 2 below:

k = 2					
Cluster 1		Cluster 2		Cluster undet.	
Size	159	Size	725	Size	150
% clean	74.84%	% clean	78.34%	% clean	78%
% dirty	25.15%	% dirty	21.65%	% dirty	22%

4.22 Expanded Stop Word Set

During the first attempt of k-means on history document vectors, we eliminated 667 of the most commonly used words on the Chatous network as these words were not informative as to the difference between any two documents. Seeing how this improved performance, we tried to find more words to eliminate so as to increase the separation between 'dirty' and 'clean' users. The finding that in our data, on average, 72% of the words in conversations of above average length are said only once further validated this potential approach as it indicates that a word form occurring at all is more important than differences in frequency. We were able to construct vocabularies V_c and V_d for each individual set of users, where $| V_c | = 40495$ and $| V_d | = 14491$ over the 1034 users. We hypothesized that variations in the word forms used is more informative of the difference between two users than the variation in the frequencies of words used, we considered expanding our stop word set to all words found in the intersection of V_c and V_d, which was $V_i = 8751$. With this dramatically increased stop word set, the median clustering of 10 iterations of k-means (at the optimal value of k = 2) yielded the following results:

k = 2, expanded stop words					
Cluster 1		Cluster 2		Cluster undet.	
Size	633	Size	146	Size	255
% clean	74.41%	% clean	100%	% clean	73.33%
% dirty	25.59%	% dirty	0%	% dirty	26.67%

The most encouraging result is that this approach is able to identify clusters of entirely one type of user no matter the value of k (we tested from k = 2 to k = 10) or starting initialization of the clustering. Although the number of users thrown out in the undetermined cluster has increased by 70%, the number of users that this algorithm can classify correctly with a high degree of confidence represented 14% of our test data, which could be used to satisfy our ultimate goal of sub queuing.

5. Analyses and Further Work

Ultimately, our approach of clustering was in fact able to make modest progress towards our goals of creating clusters of one type of user. However, by expanding the stop word set, there is a potential impact on the generalizability of this approach, especially if words in the intersection of 'clean' and 'dirty' users are not found in the training data. Therefore, we investigated the percentage of our vocabulary we were capturing and found that the size of the vocabulary of the cluster training set of 1034 human labeled users was 81% of the size of the vocabulary of the larger group of 10,000 users (37484 vs. 46277 distinct word, respectively). This result indicates that the approach has the potential to scale but more research needs to be done on the scenarios under which these conditions hold on the live chat network.

As a next step, we intend to run our results on the live Chatous platform and analyze the performance. Given that roughly 20% of our users are placed in the undetermined cluster due to stop word conversations, one suggestion would be to look into ways to classify these users. We were surprised by the conclusion that behavioral features failed to separate between our labeling, even in the clustering scenario. We recognize that one assumption we made was to use a hard clustering algorithm; users either belonged to one cluster that represented a labeling or they did not. Given the variance in behavioral features, our next step will be to change the underlying model to a softer notion of clustering, such as the Gaussian Mixture Model optimized with Expectation Maximization, in order to see if we can achieve better performance when we simply try to compare probabilities some user X belongs to some cluster Y.

In our dataset, we did not have access to the underlying words of our word frequency vectors; given that the conversation history approach has had limited success, yet another next step is to do more natural language processing to factor in considerations such as parsing, grammatical processing, relating concepts, extracting topics, etc. Ultimately, the initial success realized through our project in finding clusters of similarly typed users suggests sub queues can be established automatically using machine learning in order to dramatically improve the anonymous chat network experience.