Prediction of Online Conversation Compatibility through Personality Clustering: Milestone

Alex Zamoshchin, Sam Beder, Jay Hack

I. INTRODUCTION

Chatous is a new online chat platform that has gained a large amount of traction since its inception a year ago. It seeks to use data gathered on users in order provide fulfilling IM-style interactions. Like Guo et al., we will attempt to predict user compatibility in this online chat setting. We intend to do the following:

- Designate a set of metrics that evaluate how compatible two users are given their interactions and shared history.
- Identify characteristics of *Chatous.com* users that could correlate with their compatability with other users, such as statistics on their activity and usage of natural language in different contexts.
- Design an algorithm, or set of algorithms, that will leverage these characteristics in order to map pairs of users to their predicted compatibility.

II. PRIOR WORK

Our primary inspiration is the initial analysis of the networks and interactions on *Chatous.com* performed by Guo et al. In their paper, they describe this novel social network and provide some level high-level analysis on its properties. Furthermore, they provide the reader with a description and evaluation of an algorithm they developed to predict unknown edge weights. Guo et al. define user compatibility between two users as the geometric mean between their respective ratings of a shared online interaction. They employ a linear classifier trained on a portion of their data set; then, in order to evaluate their classifier's performance, they measure the percentage of the time that it correctly predicts which of two unobserved edges is of higher weight. In their paper, they claim a 93% success rate.

While Guo et al.'s work applies directly to our chosen task, we also plan to use the findings of two other papers describing similar approaches to signed edge classification in different domains. Leskovec et al examine various predictive techniques in social networks composed of 'positive' and 'negative' edges. In particular, they describe an edge-weight classifier trained on features of neighboring relationships and neighboring-neighboring relationship, a model based upon theories of balance and status, and an approach to edge-weight prediction using heuristics. Chiang et al. also explore sign prediction, though they focus on the insights offered by of longer 'walks' about the network. Their treatment of longer walks can be considered an expansion upon or generalization of the idea of triads. One contribution of particular note is their description of

III. DATASET

Our dataset consists of metadata describing approximately 9 million conversations occurring between 80,000 users on *Chatous.com*, as well as the following information on the individual users: Geographic location, age, gender and hashed unigrams from a short

personal statement. In addition, for each conversation, we have access to a bag of (hashed) words representation of the users' dialogue.

III. INITIAL FINDINGS AND SUMMARY STATISTICS

Thus far, we have concentrated our efforts on (1) creating a code infrastructure that is capable of handling the large, sparse dataset we are provided, (2) developing unsupervised learning techniques in order to better characterize individual users, and (3) discerning the structure and underlying properties of the data we are provided.

III.1: Infrastructure

Our Dataset consists of several gigabytes of pure csv files containing primarily natural language data, in addition to a very large network structure with very few edges of significant weight. We are dealing with a set of very sparse datasets and, therefore, in order to make any sense of it, must deal with the challenge of sifting through enough of it to arrive at any conclusions at all. Currently, we initially loading the data into a set of Pandas data frames; after extracting a set of custom objects representing a single user's activity and history from the dataframes, we construct a network structure on top of it using networkx. For unsupervised feature learning, we have opted to use scipy's sparse matrix representations, which dramatically speeds up our unsupervised feature learning. At this point, we are capable of loading in 1,000,000 unique chats, extracting user information, and performing our unsupervised learning routines in under 20 minutes on a Macbook Pro.

III.2: Unsupervised Feature Learning

Our goal in unsupervised feature learning is to (1) arrive at a model for n different personality clusters and (2) to generate a response histogram with respect to the personality clusters for each user. Our baseline attempt consisted of treating each user as a vector of features based on their objective profile (age, sex and location, among a few other high-level features) and clustering based on k-means. We achieved the following cluster sizes when training on 10,000 unique chats:

Cluster ID	0	1	2	3	4	5	6	7	8	9
Size	221	2642	4	226	69	529	38	4332	1476	10

The relatively consistent size of clusters demonstrated here is a desirable property, though it is not clear how much correlation these 'personalities' would have with user compatibility. As a second, more sophisticated attempt, we created a 'tf.idf profile vector' for each user in order to represent the words that were characteristic of a certain user. We performed clustering via a Gaussian mixture model and the EM algorithm, and arrived at the following cluster sizes on a training set of one million users:

Cluster ID	0	1	2	3	4	5	6	7	8	9
Size	2764	2935	8583	2160	531	708	2813	1266	1931	904

III.3: Data Characteristics

During our initial investigations into the dataset, we attempted to establish key characteristics that could inform our future feature selection. We found that the network of users and their IM-style interactions is very sparse and, given an arbitrary pair of users, there tend to be very few paths between them. This is what lead us to the Chiang et al. paper on long walks - they report that their methods work particularly well on networks that exhibit such structures.

Therefore, drawing on the work of Chiang et al., we converged upon the following predictor for user compatibility:

1. We assign each edge a compatibility score of $(1 - \frac{1}{x})$, where x = min(lines spoken, lines received). (In the case of no verbal exchange, we set the compatibility score to zero.)
2. For a given path, we find it's 'score' by taking the product of all the edges that constitute it.
3. Finally, we define the compatibility score between two users to be the sum of the compatibility scores of all paths between them above a certain threshold.

We think this metric is quite promising both for its intuitive appeal and our observations suggesting its descriptive ability. Since edge weights range from zero to one, this metric weights shorter paths higher. In addition, in contrast to the social theories on balance, we observed that paths with many negative edges are not indicative of user compatibility in this context. That is, the idea of "the enemy of my enemy is my friend" does not necessarily hold true for conversation compatibility. Instead, a negative conversation is not useful for our purposes, since it provides little to no information on the compatibility of two distantly related individuals. Therefore, our predictor assigns a very low weight to paths that involve unfruitful (short) interactions.

Our initial implementation of using long walks proved largely unsuccessful. As discussed below, this network is not only very sparse, but also saturated with negative edges; since such a large percentage of the conversations are unfruitful, almost all of our path scores are miniscule. Therefore, even while maintaining relatively high precision, this predictor demonstrated a very low recall in predicting which users would be compatible.

Another unrelated insight we gained on this network was that the frequencies of conversation lengths exhibited a distribution like that of inverse power laws. Due to some irregularities and outliers in the data, we found that the least-squares linear regression estimator for the value of alpha was a much better estimator than the MLE. These results are demonstrated below, in Figures 1 and 2. Using a least-squares linear regression, we estimate a value of alpha of roughly 2.22, and with the MLE we approximate alpha to be 3.63.

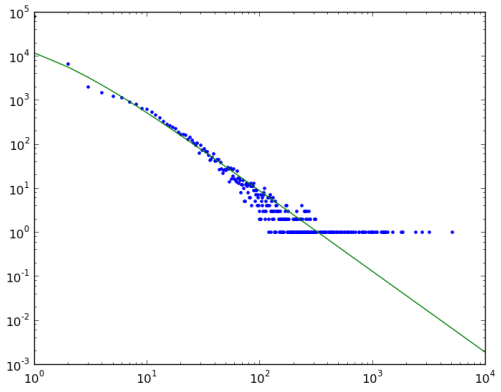

Figure 1: histogram of conversation length frequencies, plotted with power law parametrized by least-squares LR estimate for alpha

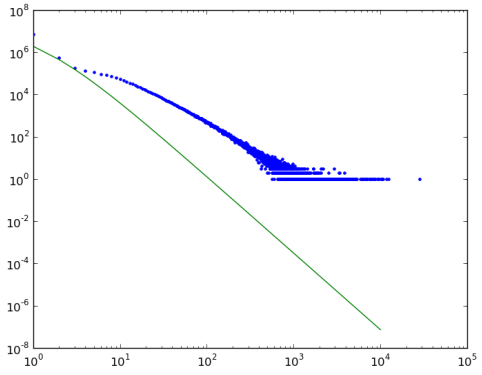

Figure 2: histogram of conversation length frequencies, plotted with power law parametrized by MLE estimate for alpha

This information provides us with the following insights: We see that the network follows common network properties and resembles a power law distribution. Moreover, a high value of alpha indicates there is an extreme drop-off in conversation length. This corresponds to a vast majority of conversations being of length 0, which will have a large impact on all future explorations. We discuss this 'heavy-tail' property below.

IV. MATHEMATICAL BACKGROUND AND ALGORITHMS[1]

One of the novel contributions of this project is our treatment of users as having unobserved 'personality-type' variables; we intend to both model individual users' personality types and, perhaps most significantly, the attraction between personality types. That is, we intend to introduce a latent variable for each user representing the personality cluster that they belong to, and then find parameters for the distributions $P(u|p_i)$, where p_i is the (latent) personality variable and *u* is a feature-vector representing a user, and additionally $P(c|p_i,\ p_j)$, where c is a level of compatibility and p_i and p_j are the two personality-types concerned.

We intend to parametrize these distributions using the EM algorithm, where $P(u|p_i)$ is parametrized as a multivariate Gaussian distribution, and $P(c|p_i,\ p_j)$ is parametrized as a binomial distribution. (Note: this is a very similar problem to parametrizing a Gaussian Hidden Markov Model for gesture recognition, as presented by Daphne Koller in CS 228, from which we derive our inspiration. The analogy is as follows: $P(u|p_i)$ is analogous to the observation probability, and $P(c|p_i,\ p_j)$ is analogous to a function of the transition probability in a gaussian HMM.)

By themselves, these distributions serve as a means to predict user compatibility.

[1] This section is very similar to the algorithms discussion in our last milestone. Our progress in the past few weeks has been primarily in the implementation of these algorithms, with just a few small adjustments made in the algorithms themselves due to practical concerns.

Let u_1, u_2 *be vector representations of users*
Let p_k = *the* k^{th} *value that the latent personality variable can take on*
Let c = *user compatibility*

$$\frac{P(u_1 \mid p_i) * P(p_i)}{\sum_j P(u_1 \mid p_j)} = P(p_i \mid u_1)$$

$$P(p_i \mid u_1) * P(p_j \mid u_2) * P(c \mid p_i, p_j) = P(c, p_i, p_j \mid u_1, u_2)$$

$$\sum_i \sum_j P(c, p_i, p_j \mid u_1, u_2) = P(c \mid u_1, u_2)$$

$$\textit{predicted user compatability } c* = argmax_c P(c \mid u_1, u_2)$$

We intend to use the full distribution $P(c|u_1, u_2)$ as features in our input vector to the linear algorithm, however, which will allow us to combine it with other, non-probabilistic features. We hypothesize that these features will have relatively high weights, though will not be the only significantly-weighted features.

One challenge that we will face in developing this module is picking the number of possible personality types; we intend to empirically determine this parameter using a held-out development set.

V. DIFFICULTIES

As mentioned numerous times above, our greatest challenge is overcoming the sparsity of chat content - an overwhelming majority of the chats that we are presented with have no verbal content whatsoever. Long walks, in particular, are damaged by this lack of information associated with edges. We plan to combat this sparsity by performing unsupervised feature learning, in order to learn a more abstract representation of users where little information is available, as well as through making scalable infrastructure and using the sheer size of our dataset.

VI. PLANNED WORK

At this point, we have implemented a baseline predictor for user compatibility - the long walks algorithm described above - in addition to multiple methods for learning higher-level user representations in an unsupervised manner. Our next step is to combine our more sophisticated user representations with our baseline predictor, in addition to other informative features that we can extract from the network, in order to create an integrated predictor that benefits from the strengths of each of its components. In particular, we plan to construct a linear classifier that combines a pair users' distributions over personality-types, the attraction between those types, and features extracted from 'long walks' in order to predict their compatibility. We expect to find that the unsupervised feature learning will help us to avoid overfitting our linear classifier on the data at hand and, in addition, allow us to infer more about a set of users when data concerning them is very sparse.