Speak Little and Well: Recommending Conversations in Online Social Streams

Jilin Chen[*], Rowan Nairn[†], Ed H. Chi[†]

[*] University of Minnesota
200 Union Street SE, Minneapolis, MN 55455
jilin@cs.umn.edu

[†] Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304
{rnairn, echi}@parc.com

ABSTRACT
Conversation is a key element in online social streams such as Twitter and Facebook. However, finding interesting conversations to read is often a challenge, due to information overload and differing user preferences. In this work we explored five algorithms that recommend conversations to Twitter users, utilizing thread length, topic and tie-strength as factors. We compared the algorithms through an online user study and gathered feedback from real Twitter users. In particular, we investigated how users' purposes of using Twitter affect user preferences for different types of conversations and the performance of different algorithms. Compared to a random baseline, all algorithms recommended more interesting conversations. Further, tie-strength based algorithms performed significantly better for people who use Twitter for social purposes than for people who use Twitter for informational purpose only.

Author Keywords
Social stream, recommender system, conversation, user preference.

ACM Classification Keywords
H.5.3: Group and Organization Interfaces.

General Terms
Algorithms, Experimentation

INTRODUCTION

Online social streams such as Facebook news feeds, Google Buzz and Twitter streams have emerged as important channels of online information. Millions of people are reading statuses, tweets and alike to learn breaking news, useful tips, fun stories, keep up with friends' everyday lives, and engage in random chatters.

Conversation is a key element in the social stream experience. Prior research has suggested informal chatting as a major reason for using social streams [12, 13], and empirically demonstrated the prevalence of conversations in social streams [11, 12]. These conversations facilitate information exchange and social awareness, as well as help build common ground among users [19, 20]. The importance of conversations in social streams is also demonstrated by the interest within industry. For example, several startups have been formed to support the threading of Twitter conversations (e.g. *bettween.com*, *twonvo.com*).

However, not all conversations in social streams are interesting to read. Active Facebook users may receive over 100 conversations in their full feed per day, which they often have neither time nor desire to read completely. As a result, to avoid flooding users with boring conversations, service providers often selectively display conversations to users. Facebook used the proprietary EdgeRank algorithm, particularly favoring recent, long conversations related to close friends [14]. Instead, Twitter adopts simple rules to aggressively filter correspondences between users. In fact, the filter on Twitter was so aggressive that some users have wished Twitter to be more inclusive. To date, little research has been published on techniques for finding interesting conversations, despite their prevalence in social streams.

A great challenge in finding interesting conversations is the mixture of informational and social purposes in using social stream, and thus the potential diversity in user preferences of conversations. Between a short exchange about Java programming and a lengthy discussion around Alice's recent trip to Japan, Bob may prefer the former because of his interest in programming, while Charles may prefer the latter because he cares more about Alice. As a more nuanced example, David may skip the exchange about programming despite his interest in the topic, because the exchange itself is too short to be meaningful to him. Due to this potential diversity in preferences and shifts in contexts, a single model of users' interest might be doomed to fail.

Our research has three high level research questions:

RQ1: How do users differ on their preferences of conversations? Do their preferences correlate with their usage purposes, i.e., whether they use Twitter as an information medium or a social medium?

RQ2: How effective are different algorithms in selecting interesting conversations for recommendation?

RQ3: Do usage purposes of Twitter and preferences of conversations affect algorithm performance?

To answer these questions, we explored the design of a recommender system that recommends potentially interesting conversations on Twitter. We chose Twitter over

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

many other social stream platforms due to its popularity. Specifically, we chose Twitter over Facebook because Twitter's open APIs provide us greater flexibility in collecting data, designing conversation selection algorithms, and deploying these algorithms.

We deployed our recommender system online to conduct a user study, where Twitter users rated the interestingness of conversations produced by different algorithms. The user study allowed comparison of algorithm performances across different Twitter users.

The rest of the paper is structured as follows. First, we discuss how existing research relates to our work. We then provide an overview of Twitter and intricacies of Twitter conversations. We describe the design of recommender, and then detail our studies and the results. We conclude with discussions of our findings and design implications.

RELATED WORK

Social streams have attracted a fair amount of attention in the research community recently. Our work is built on two areas within social stream research: 1) characterization and analysis, and 2) information filtering and recommendation.

Characterization and Analysis

Research has demonstrated social streams to be used for a multitude of purposes. Java et al. [12] found major usage on Twitter to be daily chatter, conversations, information sharing and news reporting. Naaman et al. [15] manually coded Twitter messages, and suggested that people post on Twitter for both informational and social purposes. In a study of the enterprise social stream site Yammer, Zhang et al. [19] found that people used the site for different purposes, and have different preferences as a result. They also found conversations to be a large part of activity on Yammer, and suggested difficulty in finding relevant content being the greatest challenge.

Several works studied the conversational aspect of social streams. boyd et al. [4] analyzed the conversational usage of retweets (RT) in Twitter, revealing great variety in Twitter conversational practices. Honeycutt et al. [11] investigated Twitter conversations in form of @replies, and discussed the challenge in finding relevant conversations on Twitter. Our exploration in recommending conversations is an important step in addressing this challenge.

Our work is also informed by prior analytical research on Facebook. Joinson [13] characterized the motives on the use of Facebook and discussed the use of Facebook news feeds. Gilbert et al. [7] modeled tie-strength – the strength of social relationships – on Facebook, and suggested that tie-strength may be a useful factor for filtering messages in Facebook news feeds.

Information Filtering and Recommendation

Several recent works have addressed the information overload problem in social streams by utilizing topic as a key factor. Ramage et al. [17] applied LDA (Latent Dirichlet Allocation) to characterize topics in Twitter and to find messages worthwhile to read. Bernstein et al. [2] utilized search engines to support a topic-based browsing interface of Twitter. Both works were built on prior research on information retrieval and topic modeling, including Salton et al. [18] and Blei et al. [3].

Studies on filtering and recommendation in social streams have indicated the potential diversity in user preferences. Chen et al. [6] recommended news URLs in Twitter using topic relevance and social voting, and suggested that a single recommender may not be able to satisfy users' differing needs. Paek et al. [16] trained support vector machines to predict the importance of posts in Facebook feeds, and found that many posts were considered important to one user but worthless to another user. They indicated personalization as a promising solution to the problem.

Our exploration in recommending interesting conversations contributes to this existing body of research in two aspects: 1) While prior research focused on individual messages, our work is focused on conversations, and each conversation is a coherent thread of multiple messages; 2) While prior works concentrated on supporting news finding and information gathering in social streams, instead we explore conversation recommendation while facing the diversity in usage purposes and preferences.

In this work, we in particular explore three factors for recommending conversations: thread length, topic relevance, and tie-strength. We include thread length because it is a simple measure of the sustainability of a conversation [9]. Topic relevance is included due to its prior success on social streams [6, 17]. Tie-strength is included due to the suggestions by Gilbert et al. [7].

BACKGROUND: TWITTER CONVERSATIONS

Twitter is a popular social stream service with millions of registered users. Twitter users can post short messages, or *tweets*, each up to 140 characters long. By default these tweets are publicly available and can be viewed by anyone; in this work, we only consider such public tweets.

Users can direct a tweet to a particular user by adding an @ symbol and a user name (e.g. @Alice) in front of the tweet. Such directed tweets are usually referred as @replies. An @reply can be further replied, constituting a chain of tweets in a conversation.

Since Twitter does not otherwise have a separate feature for conversations, most Twitter users post a series of @replies to interested parties so as to engage in conversations [11, 12]. As a result, throughout this paper, whenever we mention *conversation* in context of Twitter, we mean a tweet followed by a series of interconnected @replies, which looks like the example below:

Alice: worked till the last minute before CHI deadline.

Bob: @Alice you submitted a paper to CHI?

Alice: @Bob yeah, I'll tell you details some time.

The social network on Twitter is constructed by "*follow*" relationships. This relationship is asymmetric – Alice can follow Bob without Bob following Alice back. Throughout

this paper, whenever Alice follows Bob, we refer to Alice as Bob's *follower*, and Bob as Alice's *followee*. If Alice and Bob follow each other, we refer them as *bi-directional friends* to each other.

Twitter users consume tweets mostly by reading their *home timeline*, a stream that contains tweets posted by all their followees. As a result, a typical Twitter user's followees are users that she is interested in reading from, while her followers are users who are interested in her posted tweets.

Twitter has gone through several changes on how @replies should be included in the home timeline. Originally Twitter did not treat @replies any differently in the home timeline. An @reply posted from Alice to Bob would be included in Charles' home timeline if Charles follows Alice, just like any other tweets posted by Alice. This practice appeared to cause confusion when Charles follows only Alice but not Bob, because he would only receive half of the conversation, seeing replies from Alice but missing replies from Bob. As a result, after several iterations of changes, in May 2009, Twitter added a mandatory filter on @replies – an @reply would be included in the home timeline only if the user follow both ends of the @reply. In case that Charles follows Alice but not Bob, he would miss all @replies between Alice and Bob.

Many believed such filtering is too aggressive. Popular IT blog TechCrunch and ReadWriteWeb criticized the filter immediately after Twitter's latest change. TechCrunch articulated the benefit for not having this filter: "[disabling the filter would have] led to an increase in noise, but it [would have] also exposed you to new Twitter users and conversations that you might have otherwise missed out on"[1]. ReadWriteWeb compared this filter to Facebook: "it's more fundamentally closed than Facebook is; on that site I may not be able to view the profiles of strangers talking to my friends, but I can see that the conversations are happening and I can read the comments."[2]

Indeed, Facebook is more inclusive on conversations compared to Twitter. On Facebook, all conversations that a friend participates in, including these involving strangers, are included in the full "Most Recent" feed. Facebook then applies its EdgeRank algorithm to selectively filter the full feed into the "Top News" feed on users' homepages [14]. Table 1 summarizes how Twitter and Facebook have been filtering conversations in their streams.

Facebook and Twitter also differ on how they display conversations in the stream. Facebook groups all posts in one conversation together in its feed, allowing users to see the whole conversation in one place. Twitter, in contrast, always displays tweets in reverse chronological order. As a result, in a Twitter home timeline, two interconnected

[1]http://techcrunch.com/2009/05/12/twitter-decides-were-not-smart-enough-for-replies-changes-them-again/

[2]http://www.readwriteweb.com/archives/twitter_puts_a_muzzle_on_your_friends_goodbye_peop.php

Message From	Message To	Would the Message Be Shown in the Stream?		
		Early Twitter	Current Twitter	Facebook
Followee or Friend	Followee or Friend	Yes	Yes	Maybe (determined by EdgeRank algorithm)
	Stranger		No	
Stranger	Followee or Friend	No		
	Stranger			No

Table 1. Conversation Filtering on Twitter and Facebook

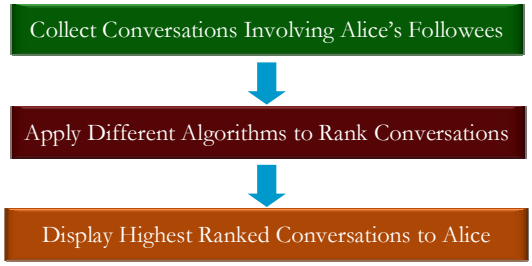

Figure 1. Recommending Conversations For Alice

@replies in one conversation may be separated by tens of other tweets that were posted between them. This design has made keeping track a long conversation particularly challenging on Twitter. In response, several third party services (e.g. *bettween.com*, *twonvo.com*, *tweetconvo.com*) have provided threading displays of Twitter conversations.

DESIGNING CONVERSATION RECOMMENDER

The high level workflow of our conversation recommender is illustrated in Figure 1. To recommend conversations for Alice, we begin by collecting all conversations that Alice's followees have participated in as candidate conversations, including conversations between followees and non-followees. This candidate set is similar to Facebook's candidate set for EdgeRank, and is broader than what Twitter currently shows to users (Table 1).

We then rank candidate conversations using different algorithms. We explore three factors of conversations in our algorithm design: thread length, topic relevance, and tie-strength. Below in this section we will discuss extensively the reason we choose each factor, how we approach the factor computationally, and how we use these factors to construct ranking algorithms.

Finally, after ranking, we present the highest ranked conversations as recommendations. We display each conversation as a thread (Figure 2), similar to Facebook and many third party Twitter conversation threading services.

Factor 1: Thread Length

Thread length of a conversation is the number of tweets that the conversation contains. This simple measure has been used to reflect how well sustained a discussion is in newsgroups and online forums [9]. In this work we apply the same intuition on Twitter conversations; that is, longer conversation threads are likely to be more sustained, of

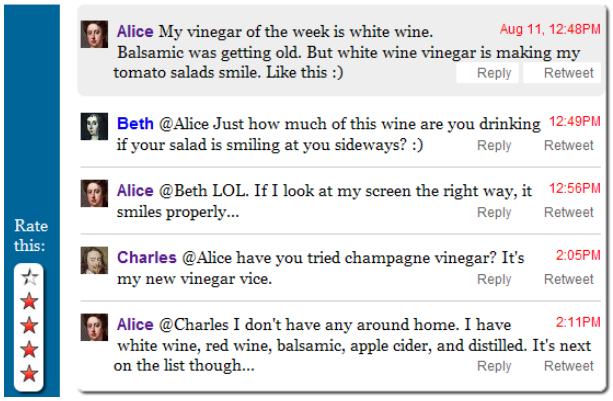

Figure 2. Conversation Displayed as a Thread

richer content, and therefore more promising as interesting conversations.

Factor 2: Topic Relevance

Topic relevance has been widely used to rank candidate items in recommender systems (See Adomavicius et al. [1] for a review) and has been successfully utilized to rank messages in social streams as well [6, 17].

Similar to the approach in Chen et al [6] and Ramage et al [17], in order to measure the topic relevance of a conversation to a user, we first build a topic profile vector for the user, then represent the content of the conversation in a similar vector, and finally match the user topic profile against the content of the conversation to compute a topic relevance score.

We build the topic profile of a Twitter user in the form of a bag-of-words vector, following the *Self-Profile* approach in Chen et al. [6]. Each dimension in the vector corresponds to a word used on Twitter and is calculated based on how many times the user has posted the specific word in her tweets, using a TF-IDF weighting scheme [18]. At a high level, Alice's topic profile constructed as such would represent her topic interest by capturing her interest on every word she has posted. That is, the more Alice posted the word in her tweets, the fewer other users posted the same word, the more confident we are that the word uniquely identifies Alice's topic interest.

The content of a conversation is represented in a similar TF-IDF weighted bag-of-words vector, using all words that have appeared in the conversation. However, since often a conversation contains only a handful of tweets and each tweet contains at most 140 characters, the constructed bag-of-words vector is often too sparse to be matched against user topic profiles. As a result, we also incorporate additional topic keywords into the vector so as to enrich the content of the vector. We identify these additional topic keywords through the TweeTopic technique [2]. That is, we feed each tweet in the conversation to the Yahoo! BOSS search engine, and extract the most salient keywords from the results returned from the search engine. At a high level, this approach allows us to discover keywords relevant to the conversation from the whole web corpus behind the search engine.

With both the user topic profile and the content of a conversation represented as vectors, we use a standard cosine similarity between the two vectors as the final *topic relevance score* for the conversation.

In preliminary explorations we have also represented topic profiles and content of conversations using LDA [3]. However, we found LDA-based representation by itself inferior to TF-IDF representation, concurring with Ramage et al.'s findings [17]. We also found the combination of TF-IDF and TweeTopic outperformed the combination of TF-IDF and LDA. We therefore decided to adopt our current approach to measure topic relevance of conversations. Note that the result we report here is merely meant to clarify our design choice; rigorous comparison of many different topic models is beyond the scope of this work.

Factor 3: Tie-Strength

Tie-strength is a characterization of social relationships between people [8]. Under this characterization, people in an individual's social network can be roughly divided into *strong-ties* and *weak-ties*. Strong-ties are families and close friends, with highly overlapped social circles. Weak-ties are, conversely, merely acquaintances. Weak-ties often provide access to novel information, information not circulating among the strong-ties.

In social streams, tie-strength of conversation participants can greatly affect the interestingness of a conversation. A discussion of Alice's recent trip to Japan would likely be much more interesting to her close friends than to her mere acquaintances.

As a result, for recommending conversations to a user, we give higher priority to conversations that happen among the strong-ties of the given user. When estimating tie-strength, we only assign non-zero tie-strength between bi-directional friends, because two users not following each other often means that the two do not know each other in person. For example, millions follow the user *BarackObama* on Twitter, but few are followed back or can claim to have a real social relationship with him.

Our approach of estimating tie-strength is largely inspired by Gilbert et al. [7], who have modeled tie-strength on Facebook in a regression model. In particular we take insight from three of their findings: 1) the existence of direct communications between two users is the strongest factor in predicting tie-strength between the two; 2) the frequency of such direct communications is another strong predictor; and 3) the tie strength between two users depends on the tie-strength between the two and their mutual friends.

We estimate tie-strength between a pair of users, Alice and Bob, using the following procedure:

1) For Alice, we first define the *communication score* for her bi-directional friend Bob as the logarithm of the

number of @replies that Alice and Bob have posted between each other;

2) Further, if Charles, David and Edward are all the mutual bi-directional friends between Alice and Bob, we compute the communication scores for these three people, and refer the average of their communication scores as the *mutual friend score* for Bob.
3) The final *tie-strength score* for Bob is the sum of communication score and mutual friend score for him.

As a result, Bob would be considered a strong-tie of Alice if he frequently exchanges @replies with Alice, or if a majority of his mutual friends with Alice (i.e. Charles, David, Edward) frequently exchange @replies with Alice.

We have two methods to associate the tie-strength scores of users to a conversation with several tweets. In one, we define the tie-strength score of a tweet as the tie-strength of its author, and then sum the tie-strength scores of all tweets within a conversation. We call this sum the *total tie-strength score* of the conversation. Intuitively, this score favors long conversations involving strong-ties.

In another, we divide the total tie-strength score of a conversation by the number of tweets in the conversation, and call the result *average tie-strength score*. Note that it is not equivalent to the average tie-strength score of all conversation participants. That is, if Alice has tie-strength score 1 and Bob has tie-strength score 0, a conversation containing 2 tweets from Alice and 1 tweet from Bob would have a final score 2/3 instead of 1/2. We pick this design so that people who participate more in a conversation will be weighted stronger in the tie-strength calculation.

Ranking Algorithms

From the above three factors, we construct five ranking algorithms for further exploration. We also included a random baseline. We will therefore compare the following six algorithms in our user study.

Random: Recommend random conversations. We use this approach as a baseline for other algorithms.

Length: Recommend conversations with the highest thread length.

Topic: Recommend conversations with the highest topic relevance scores.

Tie: Recommend conversations with the highest average tie-strength scores.

Tie-Sum: Recommend conversations with the highest total tie-strength scores. This approach is closely related to Facebook's EdgeRank algorithm [14]. Among other factors, EdgeRank attaches an "affinity score" to each post, and then decides whether to display a conversation based on the total score of all posts in the conversation. This approach can also be viewed as a hybrid of *Length* and *Tie*, because it essentially ranks conversations based on the product of thread length and the average tie-strength in *Tie*.

Topic-Tie-Sum: Recommend conversations with the highest product of the topic relevance score in *Topic* and the total tie-strength score in *Tie-Sum*. This approach can be viewed as a combination of all three factors that we have proposed.

Among the algorithms above, *Length*, *Topic* and *Tie* each directly correspond to one of the three factors. Performance of these algorithms would demonstrate the effectiveness of individual factors for selecting interesting conversations.

We include *Tie-Sum* as the forth algorithm due to its relationship to Facebook's EdgeRank, and include *Topic-Tie-Sum* as the fifth because it combines all three factors.

We did not include more algorithms, such as other ways of combining factors, because we do not want to burden our subjects with too many tasks in the user study: evaluating conversations from six algorithms already takes 20~30 minutes for a user in our pilot tests.

USER STUDY

We conducted the user study online at *zerozero88.com*, our living laboratory website that provides Twitter-based news recommendation. We recruited subjects for the study by sending Twitter private messages to existing users of *zerozero88.com*. We also posted tweets about the user study and let the information propagate through word-of-mouth on Twitter. As such, all subjects were already Twitter users before our study. Because completing the user study requires substantial effort, we offered two $50 Amazon certificates as raffle prizes.

As discussed below, the whole study consists of three parts: a pre-survey, the main study, and finally a post-survey.

Pre-Survey

In the pre-survey we sought direct inputs from the subjects for the following two purposes:

1) Confirming key design choices of the recommender, e.g. whether the threading UI (Figure 2) is appropriate;
2) Obtaining self-reports on subjects' Twitter usage purposes and their preferences of conversations, so as to understand the relationship between the two (RQ1).

We will detail the questions and responses from the subjects when we report the results of the pre-survey.

Main Study

In the main study, we collected ratings from subjects so as to compare the effectiveness of algorithms (RQ2). By relating the ratings to the self-reported usage purposes in the pre-survey, we also examined if usage purposes affect quantitatively the performance of the algorithms (RQ3).

Subjects were asked to rate the interestingness of a list of conversations on a 5-point Likert scale, where 1 is completely boring and 5 is the most interesting. Each of the conversations is displayed in a widget as shown in Figure 2.

We compile the list of conversations for a given subject as follows: First, we collect all conversations that the subject's followees have participated in during the last 7 days. Each of the 6 candidate algorithms then independently ranks this conversation collection and recommends the top 10 according to its ranking. The 6 algorithms generate a total of 60 recommendations. We then combine these 60

recommendations into a single list in random order. When algorithm A and B both recommend the same conversation, we only show one copy of the conversation to the subject. The subject's rating for the conversation is then reflected in the evaluation for both algorithm A and B, to ensure a fair comparison of all algorithms.

Post-Survey

We expect some recommended conversations to be rated as boring. One possible reason is model inaccuracy, e.g. the tie-strength score is high but in reality the conversation is all among weak-ties. However, subjects may still dislike a conversation even when the model is consistent, possibly due to their personal preferences. For example, the *Tie* algorithm may correctly recommend a conversation from strong-ties, but the subject then dislikes the conversation because she cares more about topic. Therefore, in the post-survey we asked subjects to clarify the reason for not liking a recommendation, to see qualitatively whether their preferences affect algorithm performance (RQ3).

After all conversations are rated, if any conversations generated by *Length*, *Topic*, *Tie* are rated 1, i.e., completely boring, subjects are asked to explain their ratings. In the questions we revealed why we thought the conversation should have been interesting: "we thought this conversation was in-depth" (for conversations from *Length*), "we thought this conversation was of your topic of interest" (for conversations from *Topic*), "we thought this conversation was among your close friends" (for conversations from *Tie*). In this way, we were able to qualitatively understand how much low ratings were due to model inaccuracy or due to other reasons, including personal preference.

We also ask subjects if they would be interested in receiving conversation recommendations as a service, and if they have any additional comment for this potential service.

RESULTS

We ran the user study live for three weeks and collected results from 38 subjects. We removed 3 subjects from the analysis because they follow fewer than 20 people and have fewer than 100 conversations to be considered as candidates for recommendation. The result we report is therefore based on responses from 35 subjects, who on average follow 248 people on Twitter. Below we will describe the results from the pre-survey, the main study, and the post-survey in order.

Pre-Survey

User Input for Conversation Recommender Design

In the pre-survey we asked questions related to two major design choices that we have made for the recommender: 1) displaying conversations in threads; and 2) including conversations involving non-followees, which are currently filtered away by Twitter (refer to Table 1 for a reminder).

For 1), we showed Figure 2 to subjects and asked if they think such threading display would be useful for them to track Twitter conversations. Among the 35 subjects, 29 subjects thought threading would indeed be useful. Among the other 6 subjects who did not give a positive answer, 2 subjects answered that individual @replies as in current Twitter are just fine; 3 subjects believed threading would not help much because they rarely see @replies on Twitter anyway; and 1 subject answered he/she is not sure if such display would be helpful.

Purpose	How much do you use Twitter for the following purposes?				
	1	**2**	**3**	**4**	**5**
	No		Kind of		A lot
Read useful or fun information	0 (0%)	1 (3%)	1 (3%)	15 (43%)	**18 (51%)**
Share useful or fun information	0 (0%)	3 (9%)	6 (17%)	12 (34%)	**14 (40%)**
Keep updated with people's lives	7 (20%)	5 (14%)	**12 (34%)**	7 (20%)	4 (11%)
Chat with people	6 (17%)	7 (20%)	**12 (34%)**	6 (16%)	4 (11%)

Table 2. Different Purposes of Using Twitter

Number in each cell is the number of subjects who have given the corresponding answer to the given purpose. The highest number in each row is bolded.

For 2), we briefly explained the fact that Twitter does not show a conversation if the conversation involves a person whom the user does not follow. We then asked subjects if they would rather see these conversations instead. Among the 35 subjects, 8 subjects answered a definite yes, 19 subjects said that it depends on how interesting the conversation is. The rest 8 subjects answered no, believing that the current filtering in Twitter is the right design.

We believe feedback from the subjects is supportive of both of our design choices, given that a majority of subjects believed threading as useful and showed interest in conversations involving non-followees.

Purposes of Using Twitter

In the pre-survey, subjects were asked about their purposes of using Twitter. We identified four potential purposes by reviewing prior research [12, 20]: *reading useful or fun information*, *sharing useful or fun information*, *keeping updated with people's lives*, and *chatting with people*. We consider the former two purposes as more "informational" and the latter two purposes as more "social".

We asked subjects to identify how much they use Twitter for each of the four purposes on a 5-point Likert scale, where 1 means "no", 3 means "kind of", and 5 means "a lot". The results are shown in Table 2.

An overwhelming majority of the subjects use Twitter for informational purposes. As shown in the first two rows of Table 2, 34 of the 35 subjects gave 3+ points to *reading useful or fun information*, and 32 of the 35 subjects gave 3+ points to *sharing useful or fun information.*

The answers to the two social purposes – *keeping updated with people's lives* and *chatting with people* – are more varied. As shown in the last two rows of Table 2, some subjects use Twitter for these purposes a lot, while some subjects do not use Twitter for these purposes at all. The answers to the two social purposes are significantly correlated ($r = 0.53$, $p < .01$), indicating that subjects who

use Twitter to keep updated with people's lives tend to also use Twitter for chatting, and vice versa. We found no significant correlation between answers to the two social purposes and answers to the two informational purposes.

In summary, we found that most of the subjects use Twitter for informational purposes, while social purpose usage is varied.

Grouping Subjects by Twitter Usage Purpose

To contrast the difference on social purposes, we divide subjects into two groups: The *Info-Only* group consists of subjects who gave 5 points or less in total for the two social purposes; the *Info-Social* group consists of subjects who gave 6 points or more in total for the two social purposes. Intuitively, subjects in the *Info-Only* group use Twitter mainly for informational purposes, while subjects in the *Info-Social* group use Twitter for both informational and social purposes. Under this designation, 16 subjects belong to the *Info-Only* group, and the other 19 subjects belong to the *Info-Social* group.

Self-Reported Preferences of Conversations

In the pre-survey subjects were also asked to report their preferences for Twitter conversations. We presented three statements about preferences, and asked subjects if they agree with each of them. The three statements directly correspond to the three factors in our algorithm design: thread length, topic relevance, and tie-strength. Agreement with each statement is measured on a 5-point Likert scale, where 1 means "strongly disagree", 3 means "neutral", and 5 means "strongly agree". We show the preference statements and the answers from the two subject groups in Table 3.

The two subject groups have shown similar preferences regarding thread length: most subjects gave either 3 points (neutral) or 4 points (agree) to the statement of preferring thread length.

As for preferences regarding topic, both subject groups mostly gave either 4 points (agree) or 5 points (strongly agree). The *Info-Only* gave averagely 4.7 points while the *Info-Social* group gave averagely 4.4 points. However, this difference is non-significant ($T[33] = 1.79$, $p = .08$)

The *Info-Social* group seems to prefer conversations with greater tie-strength more than the *Info-Only* group, since the former gave an average of 4.0 points (agree) while the latter gave an average of 2.9 (between disagree and neutral). This difference is significant ($T[33] = 3.34$, $p < .01$).

Main Study

In the main study, each of the 35 subjects rated recommendations from each of the 6 candidate algorithms. Each algorithm recommended 10 conversations. The whole dataset therefore contains 2100 ratings. Each rating is on a 5-point scale, with 5 representing the most interesting.

For the purpose of data analysis, *subject group* (*Info-Only*, *Info-Social*) is a between-subject factor, while *algorithm* (*Random*, *Length*, *Topic*, *Tie*, *Tie-Sum*, *Topic-Tie-Sum*) is a within-subject factor.

Our dataset is nested in nature – ratings given by a single subject are nested within each subject and may therefore be correlated. We thus employed a Hierarchical Linear Model (HLM) [5] for the analysis. HLM is an advanced form of linear regression, allowing us to model the correlation among ratings nested within the same subject. Compared to traditional ANOVA, HLM is known to provide a better fit for nested user study data like ours [5].

In the HLM model, *subject group* and *algorithm* are treated as fixed effects on the subject level, while correlations among a single subject's ratings are modeled via a covariance structure. We perform pair-wise comparisons between algorithms using post-hoc analysis with Tukey-Kramer adjustment of p-values [10].

Overall Performance of Algorithms

The overall performance of the 6 algorithms among all 35 subjects is illustrated in the top portion of Figure 3. The performance of an algorithm is measured by its mean interesting rating, i.e. on average how interesting subjects have rated recommendations from that algorithm.

In the HLM model, factor *algorithm* has a significant effect on the interestingness ratings ($F[5,2055]=36.69$, $p<.001$). Post-hoc analysis showed that the mean rating of *Random* is significantly lower than the other five algorithms ($p<.001$). In other words, the five non-baseline algorithms are all significantly better than the random baseline.

Among the five non-baseline algorithms, *Topic-Tie-Sum* performed the best overall. Subjects gave a mean interesting rating of 3.08 to its recommendations. *Topic-Tie-Sum* has been found significantly better than all other algorithms except *Tie-Sum* in the post-hoc analysis

Preference Statement	Subject Group	Do you agree with the following statements about your preferences of Twitter conversations?				
		1	2	3	4	5
		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Thread Length: "I prefer longer and more in-depth conversations over short exchanges."	Info-Only	0 (0%)	3 (19%)	5 (31%)	**6 (38%)**	2 (12%)
	Info-Social	0 (0%)	1 (5%)	7 (37%)	**10 (53%)**	1 (5%)
Topic: "I prefer conversations whose topic is close to my interest area."	Info-Only	0 (0%)	0 (0%)	0 (0%)	4 (25%)	**12 (75%)**
	Info-Social	0 (0%)	0 (0%)	1 (5%)	**9 (47%)**	**9 (47%)**
Tie-Strength: "I prefer conversations involving close friends or people I know personally"	Info-Only	1 (6%)	4 (25%)	**7 (44%)**	3 (19%)	1 (6%)
	Info-Social	0 (0%)	1 (5%)	4 (21%)	**8 (42%)**	6 (32%)

Table 3. Self-Reported Preferences of Twitter Conversations

Number in each cell is the number of subjects who have given the corresponding answer to the given preference statement. The highest number in each row is bolded.

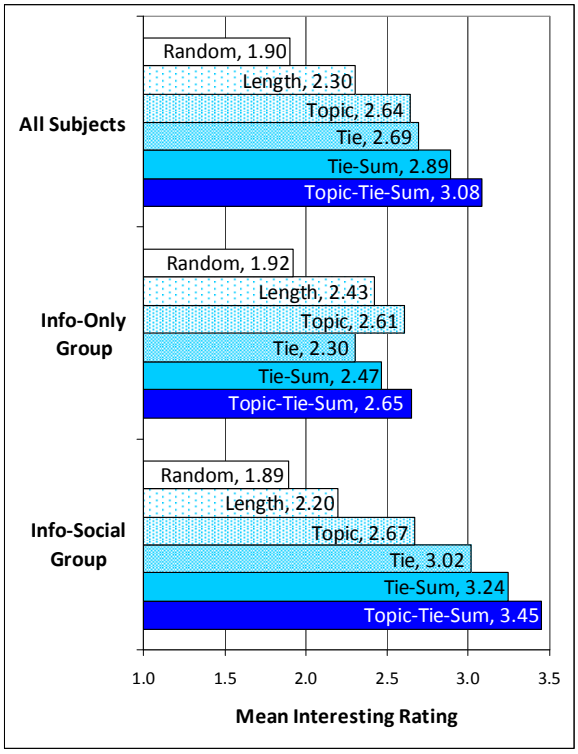

Figure 3. Mean Interesting Ratings by Algorithms and Subject Groups

(p<.001). *Tie-Sum* is the second-best with a mean interesting rating of 2.89, though it is only significantly better than *Length* and *Random* (p<.001), and is statistically indistinguishable from *Topic* and *Tie*.

Comparison of Algorithms between Subject Groups

There is a significant *algorithm x subject group* interaction effect in the HLM model ($F[5,2055] = 11.83$, $p<.001$). In other words, algorithm performance differs significantly between the two subject groups.

Compared between the two subject groups, the three algorithms that utilized tie-strength (*Tie*, *Tie-Sum* and *Topic-Tie-Sum*) have performed significantly better for the *Info-Social* group than for the *Info-Only* group:

- The mean rating of *Tie* is 3.02 for *Info-Social* vs. 2.30 for *Info-Only* ($p < .05$);
- The mean rating of *Tie-Sum* is 3.24 for *Info-Social* vs. 2.47 for *Info-Only* ($p < .05$);
- The mean rating for *Topic-Tie-Sum* is 3.45 for *Info-Social* vs. 2.65 for *Info-Only* ($p < .05$).

The performance of the other three algorithms, *Random*, *Length* and *Topic*, is statistically indistinguishable between the two groups.

We show the performance of algorithms for the two subject groups in the middle and bottom portion of Figure 3. As shown in the middle portion of Figure 3, due to the degraded effectiveness of tie-strength for the *Info-Only* group, the differences between the five non-baseline algorithms were small. In fact, the differences were so small that these five algorithms were found statistically indistinguishable for the *Info-Only* group.

Post-Survey

In the post-survey we gathered comments from subjects on lowly rated conversations recommended by *Length*, *Topic*, *Tie*. We analyzed the comments to qualitatively understand whether these low ratings were due to model inaccuracy or due to other reasons, including personal preference.

We also asked subjects if they would be interested in receiving conversation recommendations as a service, and if they have any additional comment for this potential service.

Why Are Recommendations from Length Algorithm Boring?

A majority of comments acknowledged that conversations recommended by the *Length* algorithm were more in-depth, and attributed the low ratings to other reasons.

One such reason is lack of topic relevance: "It's a detailed conversation about a topic I don't care about at all. I could care less about the details of StarCraft patches." "What the hell is Ithaca brute? I don't care." "This was a long conversation about a topic in which I have little interest."

Another reason is lack of social relationship: "I hardly know the people involved…" "I don't know any of these people personally, and only follow @[username]."

Several comments even indicated that thread length is actually an *undesirable* feature when other desirable features are absent: "Too long for the lack of thought represented in the conversation." "Yeah, is in-depth about crap. However, this is valuable in that I can skip the whole thread as one uninteresting blob - i LIKE that!!"

Why Are Recommendations from Topic Algorithm Boring?

For recommendations from the *Topic* algorithm, a number of comments suggested inaccuracy of our topic modeling. In other words, what the algorithm presented as topically relevant was not really relevant: "I'm not interested in the subject of the video and I don't use .wmv." "It is a very personal conversation that is not related to anything that I am interested in. It is very tech-support related." "Not interested in Formula one or why [username] is finally getting selective about what they tweet."

Several comments acknowledged topic relevance and explained low ratings by lack of substance: "The topic was interesting (state fair), but there was no content. 'Excited!' Tons of that junk on twitter." "The original tweet is interesting; the following conversation is not. and there is no answer to the important question in the second tweet!"

Why Are Recommendations from Tie Algorithm Boring?

For recommendations from the *Tie* algorithm, only one comment suggested that the people involved were not close friends: "These are not among my close friends. Cell phone stats do not interest me."

The rest of the comments largely attribute the low ratings to lack of relevance or substance: "Lack of useful content" "Idle chit-chat and not very interesting." "It is just small talk between two people and pointless/not relevant to me."

User Comments on Conversation Recommendation

At the end of the study, 32 of the 35 subjects indicated that they are interested in getting conversation recommendations as a new service on *zerozero88.com*. The other 3 subjects rejected conversation recommendations, 2 of which have given explanations: "I don't think I want to see the conversations. I like the brief 140 characters and that is enough for me." "In most cases, I was interested in the original tweet, but the follow-up conversation was not relevant to me. In fact, a lot of my followers often RT [retweet] part of a conversation if it becomes relevant to a broader audience, so the best parts of conversations seem to be coming into my stream already."

Several subjects summarized their experiences with the recommendations and explicitly explained their preferences of conversations: "Any conversation over 10 tweets or so is just a big space-waster unless it involves people that I really care about (i.e. people that I would hang out with and talk about stupid stuff with) or topics that I really care about (i.e. topics that are so interesting that I want to hear all I can about them)." "It seemed that the recommendations placed more emphasis on network rather than content. Many of the conversations I disliked came from people in my network who were complaining about some new phone or another product I don't care about."

Summary of Results

Usage Purpose and Preferences of Conversations

We found a large variation of social purpose usage among our subjects: some subjects use Twitter for social purposes a lot, while some other subjects have little social purpose usage (Table 2). In contrast, most subjects reported high informational usage of Twitter.

We found a link between the variation in social purpose usage and the reported preferences of conversations: subjects of high social purpose usage reported a high preference of tie-strength, while subjects of low social purpose usage reported a significantly weaker preference of tie-strength (Table 3). The above results answer RQ1.

Algorithm Performance and Comparison

Overall the five non-baseline algorithms performed significantly better than the random baseline. The most sophisticated algorithm, *Topic-Tie-Sum*, performed the best (top portion of Figure 3). This result answers RQ2.

The variation of social purpose usage among subjects greatly affected algorithm performance: algorithms utilizing tie-strength performed significantly better for subjects of high social purpose than for subjects of low social purpose in Twitter usage. Due to ineffectiveness of tie-strength, for subjects of low social purpose usage, the best algorithm only improved on average 0.73 on the 5-point scale rating (middle portion of Figure 3), and the five non-baseline algorithms were statistically indistinguishable from each other. In contrast, for subjects of high social purpose usage, the best algorithm was able to improve averagely 1.56 on the rating, and the difference among algorithms was more drastic than the overall case (bottom portion of Figure 3). This result answers RQ3.

DISCUSSION

Purpose, Preference, Performance, Personalization

At a high level, we have demonstrated in this work that the variation in usage purposes among users *matters* for conversation recommender design. More concretely, while several of our algorithms performed well for people of high social purpose usage (much thanks to the tie-strength factor), the same algorithms performed poorly for people of low social purpose usage.

This result has a straightforward explanation. Because some people do not use Twitter for social purpose as much, they do not have a strong preference on conversations involving strong-ties. Therefore, algorithms that assume importance of tie-strength would have a poor performance for these people, because the assumption of the algorithms is invalid.

There is also a less obvious alternative explanation. That is, because these people use Twitter mainly for informational purpose and not for social purpose, their Twitter social network is more focused toward weak-ties and strangers. After all, when the purpose is purely information gathering, whether the source is a friend or not is not always relevant. As a result, for these people, tie-strength based algorithms may fail simply because there are not many real strong-ties to begin with. One subject in the post-survey hinted directly at this possibility: "The biggest issue is that the signal/noise ratio is incredibly low. It'll be hard to pick out a conversation I'm actually interested in. As you probably know, this is a fundamental problem: most of my Twitter followees are not personal friends, and most of my personal friends don't converse much. So there's not a lot of signal to begin with."

Fortunately, both explanations lead to the same design implication: more *personalization*. More specifically, while the *Topic-Tie-Sum* algorithm may already be good enough for people of high social purpose usage, something different should be designed for people of low social purpose usage.

In retrospect, for people who view Twitter mainly as a medium for information gathering, a better design may have been ranking by thread length and topic, but not by tie-strength. Further, the recommender may benefit from considering all public conversations as candidates, instead of only considering conversations involving followees. By considering all conversations, the recommender can pick the longest topically relevant conversations from, say, all the 500,000 recent conversations on Twitter, instead of limiting itself on the 500 recent conversations in a user's local social network. Related to this argument, one subject has suggested in the post-survey "finding conversations not only around me but around hashtags or search keywords".

Limitation and Generalization

Our result is limited by our small subject population. Nevertheless, our subjects have demonstrated a decent level of preference diversity in their post-survey comments.

One concern comes from the fact that most subjects reported high informational usage of Twitter. This bias may be because many subjects are recruited from our news recommender *zerozero88.com*, and thus hold a prior interest in informational usage. Alternatively, Naaman et al [15] has found that information sharing on Twitter is correlated with high conversation activities, so perhaps people interested in a conversation recommender are likely also interested in informational usage of Twitter anyway. Further research is needed to clarify between the two possibilities.

Readers trying to generalize our results beyond Twitter should also note that certain designs of the recommender, such as our way to estimate tie-strength in algorithms, were made particularly for Twitter and may therefore need adaption for other platforms. Moreover, as the tie-strength estimate depends solely on online interactions, it cannot always reflect true offline relationships. Such potential online vs. offline mismatch is a general limitation of these estimates [7].

The effect of usage purpose on preference and algorithm performance in our study is likely present in information filtering and recommendation for social streams in general. Many social stream platforms, including Facebook, Twitter, Yammer and Google Buzz, are used for information gathering and social awareness with various degrees. For predicting the interestingness of a piece of information, topic relevance is likely effective whenever the user cares about information gathering, and tie-strength is likely effective whenever the user cares about social awareness. In light of this argument, the emphasis on social relationship in Facebook's EdgeRank is justified, as Facebook is known to support social awareness more than other purposes [13]. Further improvement may be possible if Facebook can identify subgroups of users who also care about information gathering and incorporate topic relevance into the formula.

CONCLUSION AND FUTURE WORK

We have studied conversation recommendation on Twitter. We implemented five algorithms, and evaluated the algorithms in an online study of real Twitter users. Further, in the study we explicitly explored the diversity in usage purpose and preference among users, and found that the performance of the same algorithms can be significantly different for users of differing usage purposes and preferences.

A promising future direction is inferring usage purpose and incorporating the inference into recommendation. For example, we may implement a conversation recommender for informational purpose and another for social purpose, and mix the two recommenders depending on the inferred usage purpose of each individual user.

REFERENCES

1. Adomavicius, G. and Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734-749.
2. Bernstein, M., Suh, B., Hong, L., et al. 2010. Eddi: Interactive topic-based browsing of social status streams. Proc. UIST'10.
3. Blei, D.M., Ng, A.Y., and Jordan, M.I. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 4-5, 993-1022.
4. boyd, d., Golder, S., and Lotan, G. 2010. Tweet tweet retweet: Conversational aspects of retweeting on Twitter. Proc. HICSS '10.
5. Bryk, A.S., and Raudenbush, S. W. 1992. Hierarchical linear models: Applications and data analysis methods. Sage Publications.
6. Chen, J., Nairn, R., Nelson, L., et al. 2010. Short and tweet: Experiments on recommending content from information streams. Proc. CHI '10.
7. Gilbert, E. and Karrahalios, K. 2009. Predicting tie strength with social media. Proc. CHI '09.
8. Granovetter, M.S. 1973. The strength of weak ties. The American Journal of Sociology, 78(6), 1360–1380.
9. Guzdial, M. and Turns, J. 2000. Effective discussion through a computer-mediated anchored forum. Journal of the Learning Sciences, 9(4), 437–469.
10. Hochberg, Y. and Tamhane, A. C. 1987. Multiple Comparison Procedures. Wiley.
11. Honeycutt, C. and Herring, S.C. 2009. Beyond microblogging: Conversation and collaboration via Twitter. Proc. HICSS '09.
12. Java, A., Song, X., Finin, T., et al. 2007. Why we twitter: Understanding microblogging usage and communities. Proc. WebKDD '07, 56-65.
13. Joinson, A.N. 2008. Looking at, looking up or keeping up with people?: Motives on the use of Facebook. Proc CHI '08.
14. Kincaid, J. 2010. EdgeRank: The secret sauce that makes Facebook's news feed tick. Retrieved from http://techcrunch.com/2010/04/22/facebook-edgerank
15. Naaman, M., Boase, J., and Lai, C. 2010. Is it really about me? Message content in social awareness streams. Proc. CSCW '10.
16. Paek, T., Gamon, M., Counts, S., et al. 2010. Predicting the importance of newsfeed posts and social network friends. Proc. AAAI '10.
17. Ramage, D., Dumais, S., and Liebling, D. 2010. Characterizing microblogs with topic models. Proc. ICWSM '10.
18. Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513-523.
19. Zhang, J., Qu, Y., Cody, J., et al. 2010. A case study of micro-blogging in the enterprise: use, value, and related issues. Proc. CHI '10.
20. Zhao, D. and Rosson, M. B. 2009. How and why people Twitter: the role that micro-blogging plays in informal communication at work. Proc. GROUP '09.