Reply Expectation Prediction for Email Management

Mark Dredze, John Blitzer, and Fernando Pereira
Computer and Information Sciences Department
University of Pennsylvania
Philadelphia, PA 19104
{mdredze|blitzer|pereira}@cis.upenn.edu

Abstract

We reduce email overload by addressing the problem of waiting for a reply to one's email. We predict whether sent and received emails necessitate a reply, enabling the user to both better manage his inbox and to track mail sent to others. We discuss the features used to discriminate emails, show promising initial results with a logistic regression model, and outline future directions for this work.

1 Introduction

Email has evolved to encompass a plethora of work-related activity. Whittaker and Sidner [6] analyzed the use of email to perform task management, personal archiving, and asynchronous communication and referred to the three as "email overload". They concluded: *(1)* Users perform a large variety of work-related tasks with email. *(2)* As a result, users are overwhelmed with the amount of information in their mailbox. A quotation from interviews conducted by [6] characterizes some frustrations:

"Waiting to hear back from another ... employee can mean delays in accomplishing a particular task, which can ... have significant impact on our overall operations. ... it can be critical or just frustrating."

"One of my pet-peeves is when someone does not get back to me, but I am one of the worst offenders. I get so many emails ... that I cannot keep up."

In this work, we address the issue of waiting to hear back from others by learning to predict whether emails need replies. Our system identifies incoming messages that require a reply, providing another means of prioritizing emails in a cluttered mailbox. Similarly, the system tracks outgoing messages to which it thinks the user expects a reply so as to maintain a list of outstanding requests for follow-up.

2 System

Our system relies on the intuition that a user's previous patterns of communication are indicative of future behavior [5]. While reply prediction, like spam detection, is a binary classification problem, they are quite different. Nearly all agree on what is spam and thus it can be aggregated to obtain a large pool of (positive) training examples. By contrast, legitimate emails sent to a group may require only one person to reply. Additionally, keywords are less useful in reply prediction, while social network factors are very good predictors.

In addition to standard features such as word identity and message length, we designed a variety of features specifically tailored for reply prediction. *(1) Dates and times.* Emails containing dates and times are time sensitive and might require a reply. *(2) Salutations.* "Dear John" or "Hi John" directly address the recipient and might require personal attention. *(3) Questions.* Questions indicate requests. *(4) Header fields.* The sender (for received emails), as well as the TO and CC recipients are important fields for reply prediction.

The system's classifier uses a logistic regression model with a base set of features, including those above, together with feature induction [4]. The feature extraction components are integrated with the IRIS application framework as part of the CALO project [1].

3 Evaluation

We evaluated our predictors on spam-free inboxes and *sent mail* of two UPenn computer science graduate students. We detected replies by matching the `in-reply-to` and `references` fields of a message with the `Message-ID` field of potential parents. User 1 received 1218 messages and replied to 449 of them. He sent 637 messages, and received replies to 215 of those. User 2 received 596 messages and replied to 129 of them. He sent 323 messages and received replies to 91 of those.

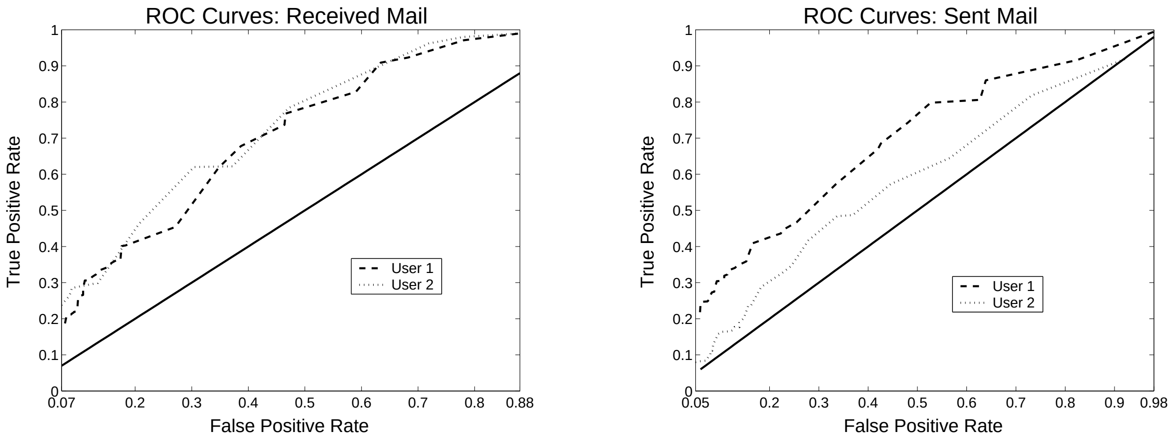

Figure 1: ROC curves for reply prediction on the received and *sent mail* of two UPenn graduate students

Figure 1 shows ROC curves for 2 users on both of our prediction tasks. The curves are generated by weighting negative (unreplied) instances. The false positive rate is the percentage of emails the classifier marked as replied, but were not actually replied to. The true positive rate is the percentage of replied emails which were correctly identified as replied. By tuning the unreplied weight we can effectively trade off a low false positive rate for a high true positive rate. For example, we see that in order to correctly find 80% of user 1's replied emails, 50% of the emails that we mark will be incorrect. Each point on the curves is an average over 10 9-1 random splits of the received and sent messages.

User 1 data performed better than user 2 data for *sent mail*, perhaps because of less *sent mail* data from user 2. Additionally, user 2 data represented mostly personal communications, while user 1 data were mostly work related. Work mail may be easier to predict because it may be more structured and contain more explicit requests. More analysis is needed to determine the performance differentials.

4 Future Work

Reply prediction is a difficult task, and while the initial results are promising, there is room for improvement. The context of an email is critical to predicting whether or not it will be replied to, and while some of the features we introduce serve as proxies for context, we believe that important information is still missing. One goal is to incorporate social network analysis such as that of [2]. Another is to incorporate a notion of thread activity, under the assumption that active threads are likely to remain active. Additionally, [3] presents a survey of email users that yield features for reply prediction. Finally, an analysis of the features is needed to determine the most effective predictors.

We also plan to develop a GUI in conjunction with the IRIS platform. A GUI should integrate user feedback and perhaps use reply prediction as a proxy for message priority. We intend to investigate the possibility of treating message priority prediction as an instance ranking problem. Priority may indicate email reply time, specifically what replies must be sent first.

5 Acknowledgments

We thank Kuzman Ganchev for data collection, and Colin Evans, Girish Acharya, and Rich Giuli from SRI for guidance. Dredze is supported by a NDSEG fellowship. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under Contract No. NBCHD030010.

References

[1] Cognitive agent that learns and organizes. http://ai.sri.com/project/CALO, 2003.

[2] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. *CEAS '04*, Mountain View, CA, 2004.

[3] L. Dabbish, R. Kraut, S. Fussell, and S. Kiesler. Understanding email usage: Predicting action on a message. *CHI '05*, Portland, OR, 2005.

[4] A. McCallum. Efficiently inducing features of conditional random fields. *UAI '03*, pages 403–410, San Francisco, CA, 2003. Morgan Kaufmann Publishers.

[5] J. Tyler and J. Tang. When can I expect an email response? *ECSCW '03*, 2003.

[6] S. Whittaker and C. Sidner. Email overload: exploring personal information management of email. *CHI '96*, pages 276–283. ACM Press, 1996.