Marked for Deletion: An Analysis of Email Data

Laura A. Dabbish
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA
dabbish@cs.cmu.edu

Gina Venolia, JJ Cadiz
Microsoft Research
Microsoft Corporation
One Microsoft Way, Redmond, WA 98052 USA
{ginav; jjcadiz}@microsoft.com

ABSTRACT
What characteristics of an email message make it more likely to be discarded? Statistical analyses of a set of deleted and non-deleted messages revealed several factors that were important in predicting the fate of a message. After controlling for the owner of the particular message, four factors turned out to be most important: history of communication with the sender (messages sent to and messages received from), intra-organizational vs. external sender, and size of the recipient group.

Keywords
Electronic mail, Email, filtering, messaging, CMC

INTRODUCTION
Several studies about email have focused on how people save their email, the purposes it serves for them, and its importance as a tool for coordination in the workplace [1,3]. In this paper we address the following question: What factors indicate that an email message is more likely to be deleted? For email that is not spam, what characteristics of a message affect how users choose to deal with it? To the best of our knowledge we are the first to conduct a close examination of the factors that affect the deletion of an email message. Identifying these factors could provide insight for the design of email systems. Intelligent agents could identify messages prime for deletion, or prioritize certain messages to receive attention first (as in [2]).

DATA COLLECTION
We employed two techniques to discover how people deal with incoming email: interviews and analysis of email stores. The reason for conducting the interviews was to be able to select intelligently which factors to examine from the email data collected.

E-mail Store Analysis Methodology
Six employees with a broad range of jobs within QSOFT[1], a software corporation, allowed a data collection program to be run on their email. This program collected information from their email, such as number of messages in their inbox and folders, message status (read or unread), and the thread structure for messages that were replied to and forwarded.

ANALYSIS OF E-MAIL STORE DATA
Each participant's E-mail store data was examined and a set of messages were obtained over a certain period of time. The time period was chosen so that some of the messages would be marked for deletion while some would not. Data from six participants were used in the analyses performed with a total of 16199 e-mail messages in the data set. Of these messages 1478 had been marked for deletion, while 14721 had not.

Based on insights gained from the interviews and a review of the literature, several characteristics of the messages were hypothesized to be important in predicting likelihood of deletion. The following list of factors were hypothesized to be influential[2]:

- **Owner of a message**
- **Importance of message**
- **Whether a message was read or unread**
- **Number of Recipients**
- **Is message part of a thread?**
- **Length of the Subject of message**
- **Number of Attachments**
- **Address Type**
- **Top Sender**
- **Does Have History**

Data Analysis Performed
A nominal logistic regression was performed on the factor 'IsDeleted', a binary response variable indicating whether a message was deleted or not. A total of 16199 observations were used in the model where each observation indicated an email message. All factors believed to be influential were included in the first model.

Results
Five out of the ten factors in the first model were actually significant. The significant factors included who the messages belonged to, the number of recipients of the message, whether the sender was internal or external to the organization, whether the sender of the message was one of the highest people sent to in the past, and whether the person had received a lot of mail from the sender of

[1] Company name was changed to protect the identity of study participants.

[2] Age of the message was not included as a factor because the messages analyzed for each participant spanned varying time periods.

CHI 2003, April 5-10, 2003, Ft. Lauderdale, Florida, USA.
ACM 1-58113-637-4/03/0004

the message in the past (indicating past communication histories with the sender).

A model was run with only the significant factors included in order to obtain reasonable parameter estimates and odds ratios for each of the factors. The effect on the probability of message retention determined using the odds ratios from the second model is shown in Figure 1.

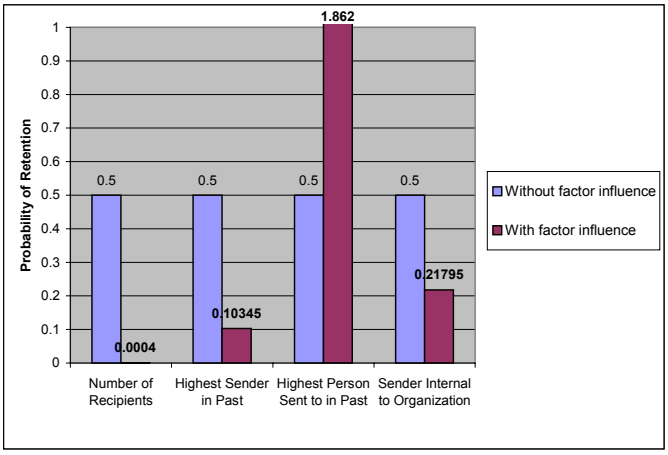

Figure 1 - Important factors and their effect on the odds of message retention

DISCUSSION OF RESULTS

It is worth noting that there was a large main effect for the owner of a message, which was expected due to the variation between individuals' email management strategies.

Highest Person Sent to in the Past

If a message was from one of the five people sent to the most in the past, this increased the probability of retaining the message, with a 3.724 times increase in the odds of message retention.

Sender Internal to Organization

Messages from addresses internal to the company were more likely to be deleted, with the odds of retention decreased by 0.26 times. Messages from addresses external to the company were more likely to be retained.

Number of Recipients

An increase in the number of recipients of a message caused a decrease in the probability of a message being retained. The more recipients on a message, the less likely it was to be personally directed to the user, therefore they were more likely to delete it.

Highest Sender in the Past

If the sender of a particular message was one of the top five senders in the past, this decreased the odds that the message would be retained by about 0.2069 times. It could be that the majority of the messages received from these kinds of senders are simply non-informational replies to previous messages sent, or continuations of previous conversations that do not need to be saved.

CONCLUSIONS

In our data set, the following factors affected the likelihood that a message would be deleted:

- *Past communications directed to the sender*
- *Internal communications vs. External*
- *Number of recipients*
- *Past communications received from the sender*

There are several possible reasons why these factors were most influential. One reason might be that these factors were the elements of the message that were made most visible in the interface for the email program these participants were using. A second reason might be because these factors typify how the participants internally categorized messages. These users may have used sender name and email address type, for example, as a cue of whether to retain a message or not. Because the data used in this study was 'observational' in nature and obtained from a population that was not controlled, the external contextual effects not accounted for could influence the outcome of the analysis. Thus, the results of the study must be considered in relation to the population of messages they came from. Though the results can not be directly generalized to all email messages at large, they do provide an interesting example of what factors were most important for retention or deletion for this particular set of messages.

FUTURE WORK

Future studies could involve more controlled sampling of messages deleted over a certain period of time from a random set of individuals within a specified population. A larger set of individuals could be used for the study, and messages could be randomly selected for consideration during the model creation.

ACKNOWLEDGMENTS

We are grateful to the participants of this study who so graciously allowed us access to their email data.

REFERENCES

1. Ducheneaut, N., and Belloti, V. (2001). Email as Habitat: an exploration of embedded personal information management. *Interactions of the ACM*, September/October 2001.
2. Horvitz, E., Jacobs, A., and Hovel, D. (1999). Attention-sensitive alerting. *Proceedings of UAI '99, Conference on Uncertainty and Artificial Intelligence.*
3. Whittaker, S., and Sidner, C. (1996) Email Overload: Exploring Peronal Information Management of Email. *Proceedings of CHI '96, the ACM Conference on Human Factors in Computing Systems.*