Automated Message Prioritization: Making Voicemail Retrieval More Efficient

Meredith Ringel
Stanford University
merrie@cs.stanford.edu

Julia Hirschberg
AT&T Labs -- Research
julia@research.att.com

ABSTRACT
Navigating through new voicemail messages to find messages of interest is a time-consuming task, particularly for high-volume users. When checking messages under a time constraint (e.g., during a brief meeting break), users need to identify those messages requiring urgent action, since not all messages can be processed in limited time. For these users, it would be useful if messages of greater urgency could be played first. For other users, distinguishing personal from business voicemail is a pressing need, to separate their home and business lives. We have successfully applied machine-learning techniques to lexical, acoustic, and contextual features of voicemail in order to sort messages based on urgency and on business-relevance.

Keywords
Voicemail, phone interface, speech recognition, machine learning

INTRODUCTION
Searching a full inbox of voicemail messages to find one specific item can be a very lengthy process. Messages must be played sequentially, and at least several seconds of each must be played before the user can decide whether or not they have found the target message [3].

Users checking their voicemail under time constraints (a meeting break, between flights at an airport, driving between work locations) may not have sufficient time to scan through all their incoming messages in sequence, and thus may miss critical messages at the end of their inbox. In such situations, users might be better served if they could play their messages in order of message urgency. Other users, seeking to separate work and office life, look for methods to distinguish personal messages from business messages, which may be difficult if they have a single voicemail inbox. To address both of these needs, we have developed techniques for re-ordering a voicemail inbox based on the urgency and business-relevance of its constituent messages.

METHOD
Our corpus consists of 3466 voicemail messages collected from 132 AT&T employees' inboxes over a period of several months. This represents 45 hours of a 100-hour corpus used to develop an audio browsing and retrieval system for voicemail access, SCANMail [2]. The messages were transcribed by hand and further annotated with information about the caller, including name, gender, and age, where available. Entities such as dates, names, and phone numbers were also identified in the messages.

From reading the hand transcription and listening to the audio, the authors rated each of these messages on a scale from 1 to 5 along two dimensions: First, we rated the message's urgency. Factors we noticed increased a message's urgency rating included the mention of pressing scheduling issues, the need to take speedy action (e.g., FedEx a document, pick up a child that afternoon, fix broken equipment), the importance of the person placing the call (e.g., the user's boss), and the number of calls received from a particular caller/on a particular topic. We also rated messages according to the degree to which they were 'personal' or not. Factors that we noted increased a message's 'personal' rating and decreased its likely business-relevance included callers' use of nicknames in the message greeting, callers' failure to identify themselves by name, and the discussion of personal topics such as children, pets, and cars in the message.

We then extracted 79 features from the message itself, from the message's transcription, and from the state of the user's inbox. Examples of extracted features include the number of dates mentioned in the message, the distance of mentioned dates from the date the message was sent, the number of previous messages the current caller had left for the user within the past *n* days, the caller's speaking rate, the time of day the message was sent, the duration of the message, and key words appearing in the body of the message.

We used the RankBoost machine-learning algorithm [1] to learn a ranking procedure for voicemail messages based on the utility of the aforementioned features for predicting the urgency and business-relevance ratings. First, we randomly partitioned the messages into training (90%) and test (10%) sets.[1] We then used the hypotheses output by RankBoost to re-order the test data from its original

[1] We also created a second, different random partition into training and test data. The statistical analysis that follows represents the averaging of the results from both training/test partitions.

chronological order in two ways: from the least to most urgent and from the least to most business-relevant messages.

RESULTS

We evaluated the orderings generated by the RankBoost algorithm in terms of their distance from the "correct" ordering (according to the ratings we had earlier assigned for urgency and for business-relevance) of the messages within each voicemail box. For example, messages we ranked as '5' for urgency should be ordered first within an inbox, '4' next, and so on. As baselines for comparison, we also evaluated the degree to which the chronological ordering and a random ordering of the messages in each box differed from our labeled ordering.

The algorithm's orderings produced significantly more perfectly ordered inboxes (i.e. inboxes in which no lower-ranked message was ordered above a higher-ranked one) than the baseline orderings (Figure 1). Furthermore, for the remaining boxes our algorithm ranked fewer items out of order than the baseline, and those misordered items were generally much closer to their proper location than were misordered items in the baseline orderings. Figures 2 and 3 summarize these results by illustrating the average amount of misordered items and the average distance of a message from its proper position, normalized for mailbox size.

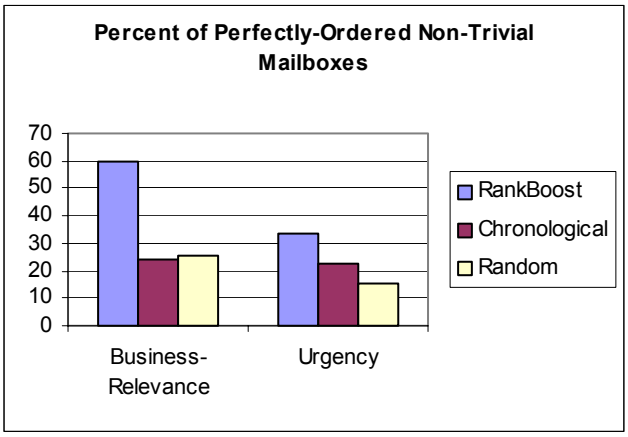

FIGURE 1

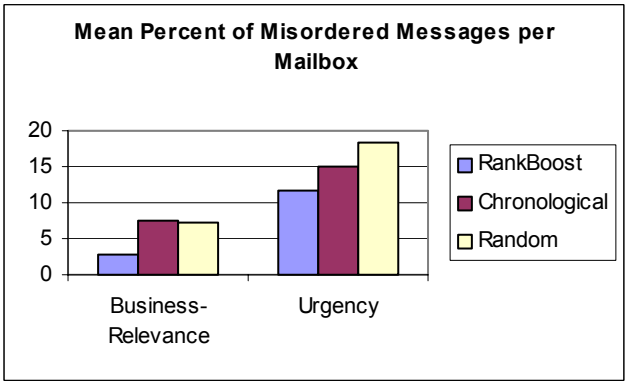

FIGURE 2

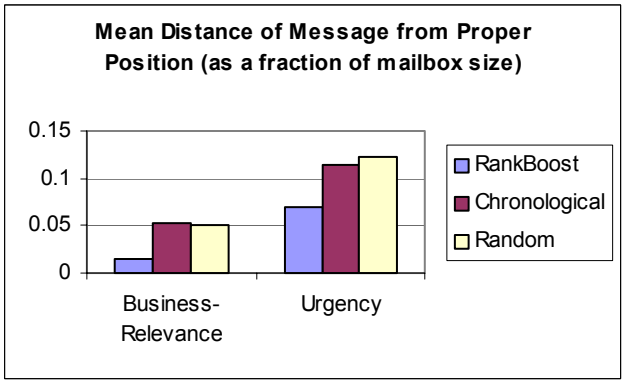

FIGURE 3

DISCUSSION

Our results are an encouraging demonstration that machine-learning techniques can be successfully used to create new capabilities for communications interfaces. While our technique is not always accurate, we conjecture that it will be sufficiently precise to improve access for heavy voicemail users. However, only actual user testing will tell us whether users will find this form of message ordering beneficial and how accurately our algorithm must rank messages in order to be useful.

However, before our procedure can be integrated into our voicemail browsing prototype and tested on real users, we must first train it using only automatically available data. Instead of relying on hand transcription of messages, we will substitute ASR (automatic speech recognition) transcripts. We expect that about half of our features may be impacted by this transition (although since the ASR transcription is consistent in its misrecognitions, a good cue in a hand transcription should still prove useful in an ASR transcription), and it remains to be seen how this will impact the overall performance of our technique.

Our initial success re-ordering voicemail messages based on criteria of urgency and business-relevance paves the way for a novel method for users to interface with their voicemail. We plan to continue to evaluate this method in order to determine the utility of priority-based, rather than chronology-based, message access.

REFERENCES

1. Freund, Y., et al. An Efficient Boosting Algorithm for Combining Preferences. *Proc of Machine Learning*, 1998.

2. Hirschberg, J., et al. SCANMail: Browsing and Searching Speech Data by Content. *Proc. of EuroSpeech*, 2001.

3. Whittaker, S., et al. All Talk and All Action: Strategies for Managing Voicemail Messages. *Proc. of CHI*, 1998.