Navigation, Organization and Retrieval in Personal Collections of Email

Benjamin M. Gross (bgross@uiuc.edu)
University of Illinois Urbana-Champaign
University of California Berkeley

September 27, 2002

1 Introduction

Collections of personal information are continually growing in size and importance. Electronic mail remains the dominant application for the Internet and is the most ubiquitous type of personal collection. According to the Messaging Online "Year-End 2000 Mailbox Report" the number of electronic mailboxes rose 67 percent from the 533 million in use at the end of 1999. Globally, the number of electronic mailboxes has grown to more than 891 million.

Over the last decade, the primary improvements to email have been in groupware integration, rendering of multimedia content, remote access and security. Further research is needed in the areas of navigation, organization and retrieval within electronic mail collections. In this paper, I make the following contributions:

First, I conduct user interviews that reveal problems in current email systems, including role conflict, high cognitive overhead associated with organization and retrieval, inability to navigate conversations and difficulties in addressing messages. Next, I discuss a prototype system that addresses many of these problems through improvements in the message store, support for identity and roles, authority control and novel query interfaces.

2 Interviews

2.1 Role

I interviewed twelve users, who were an equal number of novice and expert computer users. Five users were male and seven users were female. All users had at least five years of experience with email. Their education ranged from high school diploma to Ph.D. candidate. The majority of users, seven, are employed in the information technology sector, while five users are in non-technical fields.

I found that most users maintain multiple email addresses in order to "act" in multiple "roles." For example, the number of email address ranged from at least two to dozens. The majority of users maintained separate email addresses for work and personal communications. Additional roles for which people had accounts included organizational affiliations, multiple work roles, online shopping identities and pseudononymous identities.

Even though many applications allow users to select from a number of accounts to send or receive mail, users expressed difficulty in managing multiple email addresses. The most typical coping mechanism was for users to forward multiple addresses to a smaller number of address. One difficulty in managing multiple email addresses was role conflict. For example, users consistently reported being embarrassed by mailing a professional contact with a personal address.

2.2 Organization, retrieval and navigation

Users manage their collections by creating categories and filing messages into them, moving messages from one category to another, duplicating messages and deleting messages. Most users categorize a small amount and sort or search a large amount. [2] [8]

Nearly all classification mechanisms require users to place messages into fixed categories. A message can not be located in more than one category unless it is duplicated. This creates a burden on the user to choose the "correct" categories to file messages under and to remember later in order to retrieve the message.

Most users categorize email, at least in part, by the sender of the message (e.g. a folder for John Smith). [2] The burden is on the user to either file all messages from an individual into a single category or to remember the name variants or email addresses in order to search for that individual later. Categories tend to change over time, leading to "category drift." Recategorization is time consuming as users must move each message to another location, and often old categories are never fully removed. [3] [6]

Users typically locate messages through sorting by name or date and then browsing to find the desired item. A smaller number of users rely on the built in search function to locate messages. Users reported that sorting columns was faster and easier than searching in most cases. Users want to search for messages by a person's name, not their email address. The problem is that there is no way to reference an individual consistently over time, as their email address and name may change. Using sorting alone to locate message by an individual's name is problematic as name forms are not standardized.

When we communicate with individuals, our interactions may be brief, where the conversation may consist of only one message in each direction, or it may be a sustained interaction lasting for years. Because sent mail is saved in a separate folder, a message and its response are hard to display together in most email clients. To reconstruct conversations, users typically must go back and forth between their sent mail, inbox and folders to correlate messages.

2.3 Addressing

All modern email applications have a mechanism to store and retrieve email addresses. I found a number of techniques used to address messages, including address books, aliases, typing addresses in by hand, relying on auto-complete and replying to previous messages.

Most users only placed their frequently used addresses in their address book. Many users reported confusion about how entries got into their address book. Once an address is entered in the address book, the recipients name may "auto complete," or expand after the first few characters are typed. In all cases where it was available, users relied heavily on auto-complete. Occasionally, auto-complete would choose an address that the user did not expect. Users also reported difficulty in using this feature to send email to recipients who had more than one email address.

There was little reliance on the address book aside from auto-completion. If an address did not auto-complete for a user, it was common to simply type the address in by hand. Many users reported that replying to an old message was faster than composing a new message and addressing it.

I found a temporal and geographic component to addressing. Users reported sending email to different addressees depending on the time of day and location of the recipient. Many users who had multiple accounts did not have access to all of their accounts from each location.

3 System design

3.1 An email message store

Many of the limitations of current email systems discussed in the previous sections can be traced to limitations in the data structure in which the messages are stored. The prototype system I am developing has an underlying message store that adds substantial improvements for organization, retrieval, addressing and

navigation. Messages are stored in a database with a full text index and allow user supplied metadata in the schema.

Rather than creating a folder on the file system, categories are created by queries. Most queries are handled by a simple query based interface. An advanced interface will be available for creating complex queries, as well as for editing queries. For example, if a user wants to create a category for a mailing list, she can query all messages sent to the mailing list address. If the user then wants to save this category, she can choose to make the query a standing query, and it will appear as a traditional folder in the interface. User studies will be needed to determine the appropriate interface and methods of user interaction.

The benefit of this system is that users no longer have to manually categorize messages in order to organize them. Instead of filing messages into fixed categories, users are able to add metadata to create additional categories. Categories are simply views of the collection, allowing messages to be in multiple and overlapping categories. Categories may contain other sub-categories which emulates the traditional folder hierarchy. The prototype includes full text indexing with boolean searching.

3.2 Role and identity management

Most modern email applications have "roles" or "personality" functionality. However, this functionality is limited and the personalities and roles cannot be used for organization or retrieval.

The prototype system includes a notion of an individual person, each with a locally unique identifier within the email collection. The unique identifier allows senders and recipients to have a persistent "identity" comprised of multiple facets: name forms, email addresses, roles, contact information, notes, etc. This allows authority control, which is useful for mapping multiple entries into one entry for the purposes of retrieval. [5]

The most common organization method is for users to file messages is by sender. For this reason, the prototype system automatically generates a category for each identity in the collection. This reduces the amount of categorization for most users and the cognitive overhead associated with remembering multiple name forms for a single person. A disadvantage is that users must associate every email address to an identity. New addresses are recognized automatically through similarities in names and email addresses.

By attaching role information to an identity, the system is able to perform "role matching." For example, if a user sends a message using a personal role to someone who is both a friend and a coworker, the application will use the recipient's personal email address by default. More complex matching can be achieved through the use of temporal and geographic facets.

Another advantage of using a canonical identity, rather than a series of email addresses is that it improves the reconstruction and display of threads. The system is able to display an entire conversation with any individual, including messages both sent to and received from that person.

3.3 Time based classification

One common user practice is to categorize messages by ranges of dates and time. In the prototype, simple time based categories are available by default, for example, email received today, this week, this month or this year. Other ranges can be added without difficulty. The advantage is that time based categories can be combined with and overlap with other categories (queries). The prototype provides a simple but powerful, interface for selecting time based queries. The user is able to select dates or date ranges on a calendar that are translated into queries.

4 Conclusions and future work

I conducted user interviews that revealed problems in current email systems, including role conflict, high cognitive overhead associated with organization and retrieval, inability to navigate conversations and difficulties in addressing messages. The prototype system addresses many of these problems through improvements in the message store, support for identity and roles, authority control and novel query interfaces.

In the future, I will conduct performance evaluations, including user studies, to test improvements made in the prototype system. Performance evaluations will include precision, recall, speed and efficiency comparisons. User studies will include an analysis of organization versus retrieval time, ease of use and effectiveness of various interfaces.

5 Previous work

A number of research and enterprise email systems have been built on top of databases or other indexed data structures, including Lotus Notes, Microsoft Exchange, Novel Groupwise, HP OpenMail and Compaq CRC Pachyderm. However, many of these systems require additional programming in order to expose or implement many of the features discussed in this paper. A few third party applications, such as Altavista Discovery, Enfish Onespace and Glimpse, can be used to index email collections. [7]

The Lifestreams system stores records in a database as a time ordered stream and all presentation is time based. [4] The TimeStore system is designed to help users locate messages in their email collection through a time based display. [1]

References

[1] Ron Baecker, Kellogg Booth, Sasha Jovicic, Joanna McGrenere, and Gale Moore. Reducing the gap between what users know and what they need to know. In *Proceedings of the 2000 International Conference on Intelligent User Interfaces*, Architecture and Experience, pages 17–23, 2000.

[2] Olle Bälter. *Electronic Mail in a Working Context*. PhD thesis, Royal Institute of Technology, IPLab, NADA, KTH, 10044 Stockholm, 1998.

[3] Hilary D. Burton. Famulus revisited: Ten years of personal information systems. *Journal of the American Society for Information Science*, 32(11):440–443, November 1981.

[4] Eric Freeman and Scott Fertig. Lifestreams: Organizing your electronic life. In *AAAI Fall Symposium: AI Applications in Knowledge Navigation and Retrieval*, Cambridge, MA, November 1995. MIT.

[5] F. W. Lancaster. *Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington, VA, second edition, 1986.

[6] Ann Lantz. Heavy users of electronic mail. *International Journal of Human-Computer Interaction*, 10(4):361–379, 1998.

[7] Udi Manber and Sun Wu. GLIMPSE: A tool to search through entire file systems. In USENIX Association, editor, *Proceedings of the Winter 1994 USENIX Conference: January 17–21, 1994, San Francisco, California, USA*, pages 23–32, Berkeley, CA, USA, 1994. USENIX.

[8] Steve Whittaker and Candace Sidner. Email overload: Exploring personal information management of email. In *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, volume 1 of *PAPERS: Collaborative Systems*, pages 276–283, 1996.