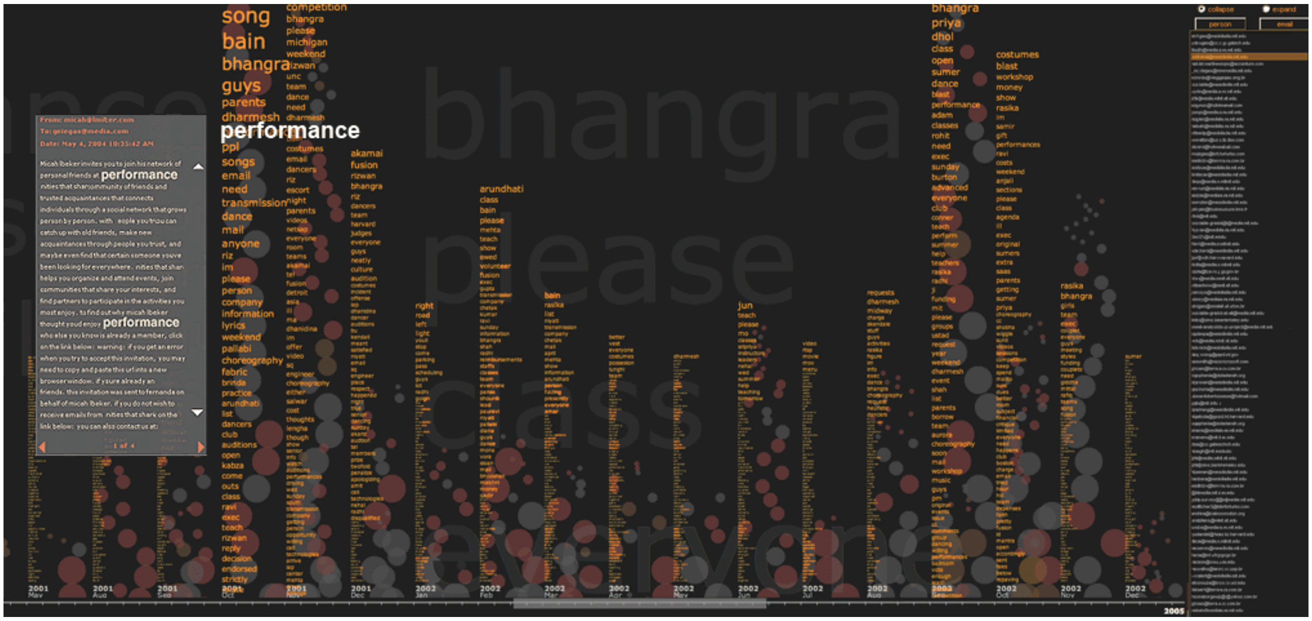
Visualizing Email Content: Portraying Relationships from Conversational Histories

Fernanda B. Viégas
IBM Research
Cambridge, MA 02142
viegasf@us.ibm.com

Scott Golder
HP Laboratories
Palo Alto, CA 94304
scott.golder@hp.com

Judith Donath
MIT Media Lab
Cambridge, MA 02139
judith@media.mit.edu



From: micah@limiter.com
To: geiegas@media.com
Date: May 4, 2004 10:35:42 AM

Micah lbeker invites you to join his network of personal friends at **performance** nities that sharcommunity of friends and trusted acquaintances that connects individuals through a social network that grows person by person. with eople you trou can catch up with old friends, make new acquaintances through people you trust, and maybe even find that certain someone youve been looking for everywhere. nities that shan helps you organize and attend events, join communities that share your interests, and find partners to participate in the activities you most enjoy. to find out why micah lbeker thought youd enjoy **performance** who else you know is already a member, click on the link below: warning: if you get an error when you try to accept this invitation, you may need to copy and paste this url into a new browser window. if youre already an friends. this invitation was sent to fernanda on behalf of micah lbeker. if you do not wish to receive emails from nities that shark on the link below: you can also contact us at:

1 of 4

Figure 1: *Screen shot of Themail showing a user's email exchange with a friend during 18 months.*

ABSTRACT

We present *Themail*, a visualization that portrays relationships using the interaction histories preserved in email archives. Using the content of exchanged messages, it shows the words that characterize one's correspondence with an individual and how they change over the period of the relationship.

This paper describes the interface and content-parsing algorithms in Themail. It also presents the results from a user study where two main interaction modes with the visualization emerged: exploration of "big picture" trends and themes in email (*haystack* mode) and more detail-oriented exploration (*needle* mode). Finally, the paper discusses the limitations of the content parsing approach in Themail and the implications for further research on email content visualization.

Author Keywords

Email archive, visualization, content

ACM Classification Keywords

H.5 Information Interfaces and Presentation; H.5.2 User Interfaces

INTRODUCTION

Email users tend to save the overwhelming majority of messages they receive [4]. In fact, email storage and retrieval have, early on, been identified by researchers as two of the main uses of this communication technology [18, 19]. It is not clear, however, why users save such large amounts of messages.

Users save "important" messages, ones that announce a new policy at work or the arrival of a friend's baby. They also save seemingly insignificant messages, ones that suggest we have lunch at 12:30 instead of at noon, ones that involve the logistics of trips long since taken, of meetings long since held [19, 4].

Most tools for handling email archives have focused on the task of finding either specific messages—commercial email clients have search functionalities that let users look for messages by subject or sender—or the "important" emails. Less attention has been paid to the overall patterns of communication that can be gleaned from the daily accumulation of messages over time. PostHistory [17], our first email visualization project, explored temporal patterns of email exchanges. The project revealed that users were quite fascinated by the ability to look back at overall patterns of exchange in their archived messages. This positive feedback prompted us to explore email visualization possibilities even further.

Our hypothesis is that the patterns of communication we build up over time are themselves significant. As email archives grow, they become valuable records of people's relationships. An increasing amount of our interaction with colleagues, friends, family members, etc. occurs via electronic media such as email.

Whereas face-to-face interactions are rich in sensory detail, online conversations, by contrast, are abstractly sterile. As archives of online interactions "pile up" on a daily basis, users are left with amorphous, homogeneous records of online interactions, little more than white noise. In the same way that we rely on pictures, videos, scrapbooks, and photo albums to remember people and events in real life, we need better ways to capture and depict the email-mediated relationships in which we engage.

In this paper, we describe Themail, a visualization of the contents of an email archive. We also discuss the results from a user study where participants visualized their own email archives. The primary contribution of this paper is the discussion of two main themes that emerged from users' reactions to Themail: appreciation of the overall picture ("the haystack") and seeking specific pieces of information ("the needle"). We describe how Themail's design supported these two themes, and what implications these findings have for the future of email studies in general.

VISUALIZING EMAIL ARCHIVES

The growth of email archives presents challenges not only to the end user but also to librarians, scholars, historians, forensics experts, and intelligence analysts. It is no surprise then, that research on these collections spans a wide variety of fields: from information management, retrieval and security, to spam detection, social network analysis, and user interface design. More recently, the information visualization community has also become interested in the idea of exploring email archives and the opportunities they provide for the visual discovery of patterns.

Roughly speaking, most of the work done on email archive visualizations falls into four main categories:

- thread-based visualizations [10, 16]

- social-network visualizations [6, 9, 17]

- temporal visualizations [7, 12, 17]

- contact-based visualizations [11, 15]

As a communication medium, email is inherently suitable for social network analysis (SNA). Moreover, given SNA's long history of visual exploration [8], it is only natural that researchers came to create email-based social network visualizations with increasing efficiency. Most of this work has been done so that experts can better understand communication patterns in third-party email archives. A good example of this approach is the work being done by team of researchers in Berkeley, which built an entire suite of visualizations for revealing social network patterns in the now public Enron archives [9].

There have also been a few projects that visualize social networks with the end user in mind, an approach that is more in line with the work presented in this paper. Both *Social Network Fragments* [17] and the work done by Fisher and Dourish [6] are ego-centric visualizations of the social network in an individual's email archive. Both projects were aimed at end users and were tested in a similar manner to Themail.

In addition to visualizing the structure of email networks, researchers have also started to look at different aspects of email chronemics[1] in the hope of finding meaningful patterns of behavior. In [12], researchers uncovered the temporal rhythms of email archives to create context for further email analysis. PostHistory [17], an ego-centric visualization of email archives based on frequency of exchanges, revealed that users could effectively map bursts of email exchange to events in their lives without having to rely on the content of the messages.

Finally, some systems have been developed for contact management – to allow users to keep track of the various people with whom they communicate over email [11, 15].

Themail differs from most email visualizations described here because it relies on the content of messages, instead of on header information, to build a visual display of interactions over time. Visualizations such as [7, 9, 10, 12, 15, 17] depict the patterns and structure of correspondence. Such visualizations are useful for showing the networks of acquaintanceships and the temporal rhythms of interactions, but they do not provide any clues about the topics people discuss or the type of language they use with

[1] Chronemics refers to the temporal dimension of communication.

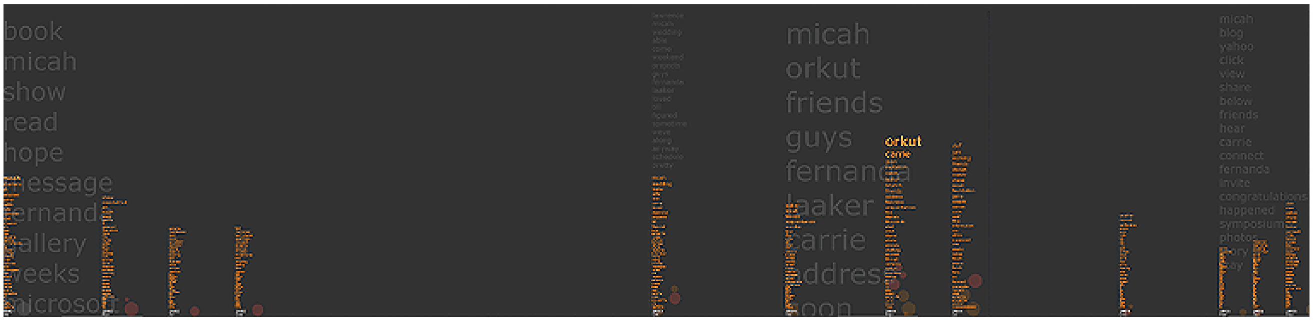


Figure 2: *Expanded view of Themail showing the sporadic nature of a relationship. "Blank" spaces between columns of words stand for months when no messages were exchanged between the user and the selected email contact.*

different members of their social circles. By visualizing the content of messages, Themail creates a more nuanced portrait of the mediated relationship.

Interestingly, outside of the email research area, projects like Conversation Map [13] have explored the validity of using content visualization to make online communities and large-scale conversations more legible.

We developed *Themail* with the working hypothesis that a visualization of email content constituted meaningful portraits of people's relationships. To test this claim, we needed to let users visualize the relationships encoded in their own email archives. Users' familiarity with the materials being visualized turned out to be critical in revealing both the successful and the problematic aspects of our content parsing mechanism. We believe that these are valuable insights for researchers working with email content and we discuss the broader implications of our findings to some of the related work happening in email visualization.

THEMAIL

Themail is a typographic visualization of an individual's email content over time. The interface shows a series of columns of keywords arranged along a timeline. Keywords are shown in different colors and sizes depending on their frequency and distinctiveness.

The application was designed to help the owners of email archives answer two main questions:

- What sorts of things do I (the owner of the archive) talk about with each of my email contacts?
- How do my email conversations with one person differ from those with other people?

As Themail is designed for the exploration of dyadic relationships, it visualizes one relationship at a time, between the owner of the mailbox, and one of her email contacts.

Themail displays multiple layers of information, each encoding a different content-parsing technique and aesthetic treatment. *Yearly words*—large faint words—show up in the background, whereas *monthly words*—columns of yellow words—appear in the foreground.

Yearly words: reveal the most used terms over an entire year of email exchange.

Monthly words: are the most distinctive and frequently used words in email conversations over a month. The selection and font size of words is based not only on frequency but also on how distinctive the word is to a specific relationship against the rest of the archive. For instance, if the owner of the email archive uses the word "environment" a lot with a friend but not with anyone else, the word will appear fairly large when one visualizes this relationship. If, on the other hand, the word "environment" is used a lot with other people in the archive, the word will not be nearly as large in the visualization. The more frequent and distinctive a word is, the bigger it appears in the monthly columns.

Why monthly and yearly words?

By displaying multiple layers of topical words in different colors and sizes, Themail creates a richly textured portrait of conversations over time. The yearly words in the background, being the most common words used over one year, function as broad brushstrokes in revealing the overall tone of the relationship. For instance, when users in our study visualized their conversational history with family members, several of the yearly words would be terms such as "love," "hug," "Thanksgiving," "family," etc. Other times, yearly words with a friend would reflect the social nature of the relationship: "dinner," "tomorrow," "lunch," "movie." Visualizing conversations with a colleague would, many times, generate fairly representative yearly words such as "meeting," "project," "deadline."

Monthly words, on the other hand, revealed a much more detailed portrait of a person's past email exchanges. Being bound by much shorter periods of time, monthly words successfully depicted the time-based, episodic nature of email conversations. These words clearly depicted the occurrence of events in a relationship. For instance, major events such as an individual's wedding or preparation for thesis defense were clearly represented in monthly keywords. In several cases a before-and-after-the-fact effect could be seen in the visualization: all along the duration of the event, monthly keywords would be larger in font size and columns would be taller, attesting to the significance of the conversations that took place during that time [Figure 4]. In many cases, keywords would change sharply from a month column to the next, exposing the ever-changing nature of people's conversations.

Unlike the background yearly words, which are static, monthly words are interactive. Selecting a monthly keyword meant users could retrieve the email messages that caused that term to appear in the visualization.

Interacting with Themail

Figure 1 shows a comprehensive screen shot of Themail. All email contacts are listed on the right of the interface and the user has selected one of these contacts to visualize. The visualization panel shows the conversational history between the owner of the email and the highlighted contact.

Yearly words—faint, gray words in the background—and *monthly* words—yellow words in the foreground—are arranged in columns. The month and year information is displayed at the bottom of each column. Each colored circle represents an email message exchanged during a given month. The size of the circles stands for the length of the message and the color of the circles stands for the direction of the message: incoming or outgoing.

In figure 1, the monthly word "performance" has been selected, which causes the email messages that contain this word to appear in an information box (on the left). Users can scroll through all monthly columns by using the scrolling bar at the bottom of the visualization. Users may also search for words in the visualization by typing in the term they are looking for [Figure 3].

Retrieving Email Messages

In addition to exploring topical keywords, users can bring up the original context of topical words, namely, the email messages from which the keywords were selected.

Whenever a user mouses over a word in a month column, that word is shown highlighted in white and in a larger font size; this allows even words set in undersized fonts to be easily read. When a user clicks on a month word, the email messages that have the selected word appear in an information box [Figure 1], allowing users to recall the context in which the word was used in the past.

The box that shows emails packs a lot of functionality. It displays the headers of the displayed message and it highlights all instances of the word that has been highlighted in the visualization. The email box also tells how many emails contain the selected word during the chosen month [Figure 1].

Adjusting the Time Scale

Themail displays content over time and so temporal rhythms are an important aspect of the visualization. A sporadic relationship, one where correspondents exchange a few messages every other month, should look different from one where users correspond every single week. At the same time, in some situations users may only wish to see only months where email exchanges occurred. To that end, Themail has two ways of displaying content over time: the *expanded* view and the *collapsed* view. In the expanded view, monthly columns are placed in their real position in time and months without exchange emails show up as blank spaces in the visualization. The "collapsed" view displays only the months with email correspondence. In this view there are no blank spaces, meaning that screen real-state is used to its fullest.

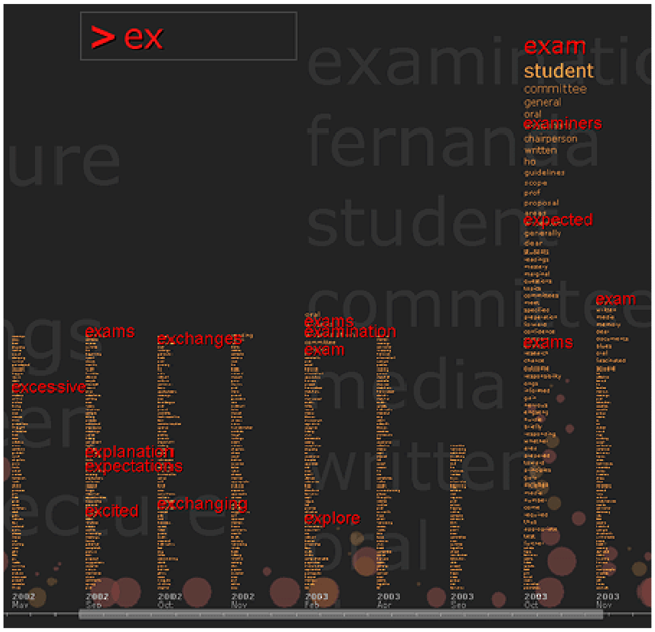


Figure 3: *Searching for words in Themail. Here the user has typed "ex" (at the top of the screen) and Themail has highlighted (in red) all the monthly words starting with these characters.*

PROCESSING THE CONTENT OF EMAIL MESSAGES

The data that Themail visualizes consist of processed email mailbox files. Themail begins with an email archive in the form of one or more mbox files, which are processed by applying a keyword-scoring algorithm. This application outputs a datafile that can be read by the visualization.

Several problems must be overcome in order to make email mailboxes more easily navigable. First, many people have multiple email addresses, and so can end up being represented repeatedly and fragmentally in the dataset. For example, "John Smith" may have email addresses *john@smith.com*, *smith@monkey.com* and *johnsmith@somethingelse.com*. The processing application in Themail allows users to identify all email addresses that belong to the same person and merge those addresses together so that the person in question has only one email address. In the case above, all of "John Smith's" email would appear as being from (or to) *john@smith.com*.

Secondly, spam plagues email users. Though many users attempt to filter their mail with spam filters, some spam may remain in their mail archive. Additionally, users who subscribe to many mailing lists will certainly have many mailing list-related messages from people they do not actually know. Because we wanted Themail to focus on *interactive* relationships—meaning, relationships in which the owner of the archive not only receives but also sends out messages—the processing application disregards any email addresses to whom the mailbox owner has not sent at least one email.

Calculating Topic Words

To generate the words that are at the heart of Themail's visualization, we use a measure of relative frequency. Salton's TFIDF algorithm [14], which scores words based on their relative frequency in one document out of a collection, served as a foundation for our process. However, some important differences apply. Namely, in processing monthly and yearly subsets of email, we compared subsets of documents against supersets, rather than one document against a collection.

For each person p, we compute scores for each word w in each month m and year y that the person exchanged email with the mailbox owner. We compute two scores:

$S_m(w,p,t) = F(w,p,t) * IF(w)$

score for word w in all messages to and from person p in timeslice t, where t represents one calendar month.

$S_y(w,p,t) = F(w,p,t)^3 * IF(w)$

score for word w in all messages to and from person p in timeslice t, where t represents one calendar year.

These two scoring functions are based on the following measure of frequency:

$F(w,p,t)$ = frequency of word w in all messages to and from person p in timeslice t

A word's inverse frequency is based on its raw count, where $C(w)$ is the frequency of word w in all emails in the email archive:

$IF(w) = \log(1 / C(w))$

Notice that S_m and S_y are the same, except that in S_y, $F(w,p,t)$ is cubed in order to increase its weight in the overall result.

The keywords shown in each month column in the main Themail visualization are selected, sorted, and sized according to $S_m(w,p,t)$. The gray words in the background of the visualization are chosen according to $S_y(w,p,t)$.

METHOD

Because we wanted to test Themail "in the wild," we decided against bringing users into our laboratory. Instead, we distributed the tool via email to participants in the study. Announcements were posted to several mailing lists within universities and research laboratories. No financial compensation was offered for taking the study. Participants were given both the processing and the visualization tools. This approach meant users were able to visualize their own email archives without having to be concerned about the privacy of their data. After having interacted with Themail, users were interviewed about their experience with the tool. These semi-structured interviews lasted 90 minutes each and were recorded for content coding (in the few cases where users were located in different states from where the researchers were, interviews were conducted over email).

Participants and their archives

Sixteen participants took part in an evaluation of Themail. The subjects ranged in age from 18 to 53; four were female and 12 were male. Participants came from two American universities as well as several technology and telecommunications companies; seven were graduate students and nine were professional researchers. Participants' email archives ranged in size from 90 MB to more than one GB, with the average size being 456 MB. The time span of these archives ranged from less than one year to over nine years of email activity.[2]

Participants were encouraged to upload multiple mailboxes to Themail – both incoming and outgoing mail. The number of mailboxes uploaded by each participant varied from two to over 55, with the average number being 19.

RESULTS

Overview

Overall, participants were quite excited to use Themail to look back at their email archives. When asked, on a scale from 1 to 5 (1 being the least and 5 being the most), how much they enjoyed looking at their email archives on Themail, participants responded, on average 3.9. When asked whether they would like to use the tool again if it were integrated in their email reader, 87% of participants responded yes.

Even though participants were, for the most part, impressed by the quality of the keywords shown in the visualization, they were also quick to point out critical shortcomings in our content parsing mechanism that merit note. We discuss these limitations in the section entitled *Limitations of content parsing in Themail.*

[2] Participants in this study were required to have archives that either covered three years of activity (at least) or were at least 100 MB.

Two main interaction modes emerged from the way participants related to Themail. In order to better explain the complimentary nature of these two forms of interaction, we have borrowed terms from the popular expression "looking for a needle in a haystack," and have called these modes "the haystack" and "the needle."[3] The former refers to gaining overall understanding and the latter refers to finding specific bits of information. In a sense, this distinction is reminiscent of the division in computer vision between trying to identify specific objects versus understanding scenes (vision for advanced robotics). About 80% of participants used Themail in the *haystack* mode whereas 20% utilized the visualization in the *needle* mode.

In the following sections we describe both the *haystack* and the *needle* interaction approaches and discuss some of the common usage patterns that emerged in each mode.

The *Haystack* mode

Users who interacted with Themail in the *haystack* mode, enjoyed using the visualization for the overall picture it presented of their relationships. They usually regarded the visualization as a portrait of their past conversations and frequently drew analogies between Themail and photo albums, in a manner that is reminiscent of our previous email visualization studies [17]. These users often put a premium on being able to see relationships with family members and friends. The more Themail confirmed their expectations—that is, their mental model of what their relationships were like—the more they enjoyed using the tool. This group of participants seemed more interested in overall patterns rather than in picking apart individual words that appeared in the visualization.

In the next subsections, we introduce the main usage patterns to have emerged from this group of users along with case studies of participants whose comments clearly illustrate the usage patterns being discussed.

Data as Portrait

Most haystack users appreciated having expectations of their relationships confirmed by the visualization while still being able to drill deeper and discover patterns they were not aware of. Several people enjoyed most looking at their families and friends in the visualization and comparing what they saw on screen with their impressions of these relationships.

> *The best "portrait" was for the mail with my mother... There have been all sorts of emotional things happening in the past few months (her mother/my grandmother passed away, she had surgery, etc.) and all of that comes through dramatically in the visualization.*

> *If you look at the ten first words of each monthly column, for instance, it's like you are following someone's life story.*

Case Study: Ann[4]

Ann is a graduate student in an American university. She is 26 years old and has recently gotten married. Her extended family lives in the south of the United States and she lives with her husband in New England. For Ann, one of the most exciting aspects of Themail was seeing all the correspondence that preceded her wedding:

> *It was funny going back both with [my husband] and my parents, there were these few months before our wedding… it's all about the wedding! There are all these words like "invitations," "tables," "drinks," guests' names. It's got all these words that are totally related to the wedding plans. It was all in October and November* [user gestures a peak] *and then the words completely changed after that. And the same happened with my friends that were bridesmaids. There are these few months where you can see that the words were related to our wedding theme but then, the month after the conversation it all switched back to normal. Yeah, it was like the before and after. You could definitely see the event.*

Ann thought it was important that Themail allowed her to look back at her relationships with loved ones, friends, and family. Even though she exchanges more emails with her coworkers on a daily basis, it was the personal facet of her email archive that she felt was the most exciting to explore.

> *Especially for my family, it was really exciting to see all the words and the things that we talk about for no reason other than to just reminisce; it was like looking through a photo album or something. For instance, I would never go back and search for the wedding planning emails, but it was fun to look at that! It's almost like this serves a different kind of purpose from regular email readers… It's more at a personal level… It's emotional, it's about reflecting and remembering.*

After looking at her correspondence with family members, Ann remarked that some "portraits" read very differently from others. With her brother, for instance, the themes ranged from talking about his kids to him asking Ann for help with his computer. With her grandmother, however, the words that came up on Themail referred to religious holidays and themes.

[Grandmother] *was interesting… I don't even remember, if you asked me, what kinds of emails I've exchanged with my grandmother; we don't write email all the time, and a lot of times the news flow through my Mom. But I felt like her Themail visualization really characterized her. It was probably because she was a whole lot different than anyone*

[3] We thank Martin Wattenberg for pointing out the "needle & haystack" metaphor in the context of our email research.

[4] All identifying information, including names of participants, their family members, acquaintances, and, where necessary, topical keywords, have been changed to preserve anonymity.

else in my email archive, so it makes her kind of a perfect person to get portrayed in a system like this and I felt like it really did a great job. It definitely brought out the things that were different about her than everyone else I talk to.

Evolution of Relationships

The sequence of keywords in Themail often reminded participants of how particular relationships had evolved over the years. Below is a list of evolution themes revealed in the change in keywords and communication patterns in Themail:

- from peer to boss
- from co-worker to social friend
- from classmates to lovers
- from spouses to former-spouses
- from child to adult (e.g. participants' offspring)
- from being co-located to moving away (and vice-versa)

> *During the past five years Ray has gone from being an acquaintance to a very good friend. Looking at this visualization I can see that it actually takes a while for the words to be dominated by social topics like bar names, beer and cinema! There are a couple of things that come out in the visualization, like a holiday when we all went to Sri Lanka and when Ray went to work in another town for a few months.*

> *This person was on my master's thesis committee so we emailed a lot during my masters about research topics and then we lost touch. She had a baby and we had a short interchange then. We exchanged some email this past March because I was defending and here she was* [user points at screen] *having another baby. To see different phases of a relationship is the best thing about this visualization.*

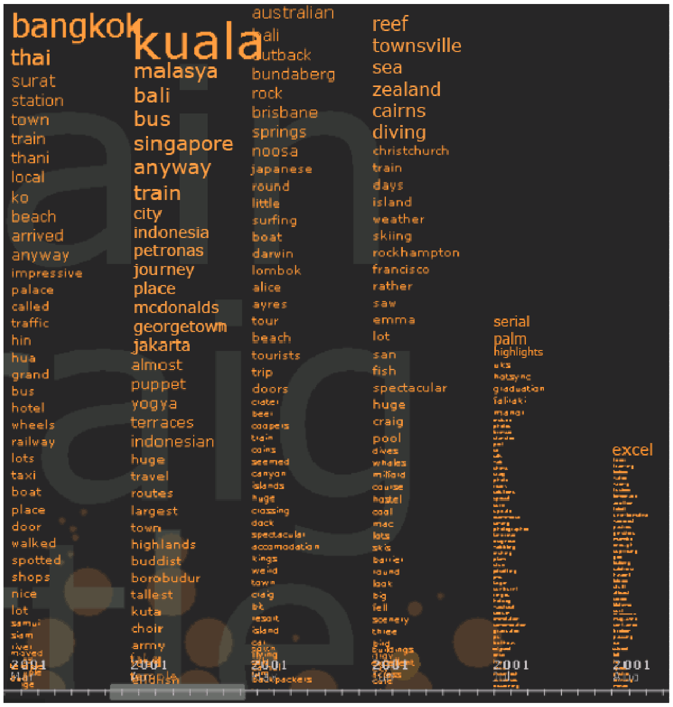


Figure 4: *Screen shot showing a user's email exchange with a friend over six months. There is a striking change between the first four months (four columns of words on the left) and the last two: the former are tall and contain several words set in big font whereas the latter are much shorter. The reason for this difference lies in the fact that, during the first four months, this user's friend was on a trip around the world. Therefore, the columns are full of unique words that are specific to the trip and not customary in their exchanges. By the time the friend returns to work—last two columns on the right—the conversation switches to programming and other usual themes in their conversation.*

Case Study: Jeff

Jeff is a researcher in his twenties, working for a European telecommunications company. He has recently spent some time in the US, working with researchers at a major university. He is single and his entire extended family lives in Europe. To Jeff, some of the most interesting information in Themail was related to his recent stay in the US and the realization of how much this change of environment was reflected in the history of his recent emails conversations:

> *During the time I've been [in the US], my friend Simon and I seem to have exchanged a lot of large emails - I suppose we were compensating for not just chatting ideas through face to face. It's also interesting to see how the content of the emails has changed and how long it took to go from very day-to-day issues, to more conceptual ideas.*

> *My mother: this is nice; it shows that during my time in the US I've used email a lot more with my Mom – and there are some good words coming out here too all about New York and Boston, when they came over for a week: "arrive," "Heathrow," "Logan," "JFK," "harbor," "staying," "itinerary," etc.*

The contrast between daily routine and extraordinary events in one's life was nicely illustrated for Jeff in his visualization of emails with a friend:

> *This is a nice view of my friend Chris. He went on a round-the-world trip and you can see in the first four month columns all the places he went and the order too. We were sending lots of long emails then – when he gets back in July the conversation switches to configuring Palm Pilots!* [Figure 4]

Finally, Jeff was able to see the evolution of his long-term relationships reflected in the visualization. One of these showed a colleague of his who went from being a peer to becoming his boss over the years. The other showed his interactions with a student who became his intern for one summer:

> *My internship student: this is very interesting (especially in the expanded view) – it shows the period for arranging the interview, day to day work emails and recent contact re-establishing the link.*

New Perspectives on Relationships

The sheer collection of words exchanged with a person made the texture of different relationships quickly obvious to users. By looking at these compilations of words, users were able to gain a new perspective on their relationships:

> *This is the one I got a little chuckle out of... this is the [ethnic dance] mailing list; this is a group of us who help manage a dance club at [our university]. And the thing that sort of stood out to me here was the fact that, just about every one of these columns has the word 'please' which is a reflection of everyone begging each other to do something! It's like, 'you guys, please do this, please do that…" and so, I thought it was really funny that this was sort of a predominant word that we're all just begging each other to do stuff! That's really what this is about.* [Figure 1]

> *This one reminded me of the fact that I was a slacker for the first couple of years [of my PhD program] and stuff like that…* [Interviewer asks: how did the visualization remind you of that?] *Well, because the name of our baseball team appears here! I'm talking more about baseball with [my advisor] than I am about work. I should probably have been working a little harder back then.*

The *Needle* mode

About 20% of participants in the study were more interested in finding specific bits of information rather than focusing on the overall patterns of the visualization (*haystack* users). Participants in this category displayed little interest in looking at the visualization of their family members, being more concerned with visualizing work-related relationships:

> *Instead of seeing my daily conversations with my wife I would rather be able to see that in 2002 I wrote a paper with Bob and this is what we talked about at that point and we haven't talked about any of that anymore, so on and so forth. That is why I don't think there are any big surprises [in Themail] because this is what I know without even having to look at a visualization.*

The last sentence in the quote above illustrates the main difference between *haystack* and *needle* users: the former take pleasure in seeing a tangible depiction of what they already know whereas the latter are more interested in discovering what they did not know or remember.

At first, *needle* users seemed a bit underwhelmed by Themail, however, when prompted to talk in detail about what the visualization showed of their relationship with top email contacts, most of these users became surprised by the richness of detail in the Themail keywords:

> *I'm not sure where the word 'femur' came from; why would I be talking with my dad about 'femur'?* [The user clicks on the word and reads the messages that contain 'femur'] *Ah…my grandmother got hurt; that's right. This is her name here* [pointing at the visualization].

> *I saw the word "horse" in the collection of correspondence with a family member and assumed that the email would be about the horses my brother has on his small farm in Minnesota. As it turns out, the email was one from my daughter (using my email account) describing the horse riding lessons she had just begun. Nice turn of events, as her email was written in the voice of a small child (~ 10 years old).*

> *I clicked on the work "decision" in an email from a friend of mine who worked with me on a local school technology planning committee. I was curious about what we might have had to "decide" about, and sure enough, it was an email about the MAC vs. Windows platform "decision" for the school. This brought back memories of many long, and quite heated discussions on the topic among parents, teachers and members of the committee.*

Such discovery episodes demonstrate that, even though Themail was not designed for querying data, its abundance of easy-to-get-to information let users find interesting bits of data quite effortlessly. Whenever users saw words that seemed out of place given the context of their relationships, they would invariably click on them to find out whether the system had made a mistake. Almost always, like in the above quote about the word "femur," users would be reminded of events they had forgotten. In fact, the ability to select keywords and see the email messages that caused those words to appear in the visualization was, more than any other aspect of Themail, the single feature that succeeded in strengthening users' trust in the visualization.

Limitations of content parsing in Themail

Other than the messages disregarded by the processing tool[5], the content analysis algorithm in Themail treats every message in the same way. This means that there are no messages with special weights or attributes in the keyword scoring mechanism. As it became clear from the user study, not all messages are created equal, and our egalitarian approach presents some serious limitations in terms of keyword output.

One of the main problems that participants identified on their Themail visualizations was the inadvertently high weight given to topical words in forwarded messages. For instance, sometimes the unique words in jokes that had

[5] Messages originating from people to whom the owner of the archive has never sent a message are automatically deleted from the dataset processed by the Themail content parsing tool.

been forwarded to the owner of the email ended up having too much weight, becoming the focus of the visualization. Whenever this was the case, participants remarked that Themail did a poor job of representing their email conversations. This phenomenon was not limited to forwarded messages, having also happened with sent-out announcements and pieces of code inserted in the body of email messages. In one extreme case, Themail displayed the relationship between an administrative assistant and the owner of the email as being studded with some of the most academic and highbrow words in that person's email archive. In fact, the administrative assistant had the task of sending out announcements for every thesis defense in that university department. The focal words in the visualization came straight from students' dissertation abstracts rather than having been produced by email exchanges between the owner of the email and the administrative assistant.

Unrepresentative keywords also came out of people's email signatures. Methods for removing signatures from emails include identifying email addresses and other contact information, as well as a large amount of non-alphanumeric characters, such as punctuation [2]. Themail accomplished some of this by ignoring URLs, email addresses and numeric strings. Signature content persisted in Themail most often when the signature (1) contained quotations, e.g. from famous people or song lyrics, (2) changed over time, or (3) varied according to multiple addresses the person held. In such cases, Themail would treat the words in the signature as content.

The second major limitation in our content parsing mechanism is granularity. Themail has no notion of expressions; it only knows individual words. This approach imposes clear limits to the output depicted in Themail as the --tone and texture of messages is more fully expressed by phrases rather than by separate words. A desirable next step for the work presented here would be to have it analyze phrases.

Implications for HCI

Methodology

Information visualization is generally used for understanding unfamiliar, complex data spaces. In effectively displaying overviews of large datasets, visualizations quickly reveal unknown patterns in the data. In our study of Themail, however, we distributed the tool to users who were *already familiar* with the datasets being visualized. Participants had some idea of what to expect in the visualization: they anticipated they would see the names of the people with whom they emailed the most, different kinds of words reflecting different kinds of relationships, etc. Even though building a tool to visualize familiar data is not the conventional course of action in information visualization, we feel that it is a valuable approach in email research and one that can lead to important advances in the field.

By testing Themail with users who were familiar with the datasets being visualized, we were able to learn about important limitations of our content parsing algorithm. Because users knew their email archives well, they were able to quickly point out some key problems that would have taken us much longer to discern. By distributing Themail to users, we made sure that it was tested in a natural setting—as opposed to having been tested at a laboratory—and that users did not have to worry about the privacy of their data.

We believe that this approach can be of value to other work being done in this area as well. For example, email visualizations that are built for the discovery of patterns by third-party experts, can also take advantage of insights from users who are familiar with the email archives being visualized. If, for instance, a visualization designer tests her expert system with owners of email archives—not her target audience but, rather, a "testing unit"—she might quickly learn about the kinds of patterns her tool displays and any potentially problematic artifacts the system generates. Most current studies of email visualization would be greatly complemented by this kind of approach.

Content Analysis

Text analysis is a vast and complex area of research. Our contribution is not a particular algorithm but rather a preliminary understanding of what is distinctive in the analysis of email content.

Results from the Themail user study made it clear that content parsing of email exchanges is inherently different from other kinds of text analysis. Email messages are not created equal and their differences need to be taken into account when exploring content. For instance, we may want to consider the content of forwarded messages and announcements differently from the content of messages exchanged between friends. Our user study revealed that participants were highly aware of and sensitive to these differences.

Some of these dimensions can be extracted from traffic patterns, chronemics, and the symmetry of exchanges. If such patterns were to be integrated with email content analysis, we could likely generate much more representative portrayals of email relationships. For instance, it would be useful to utilize traffic patterns to decide when an email exchange should be considered a conversation as opposed to a broadcast. If we can generate conversational models that take these structural elements into account, we will be in a better position to meaningfully analyze the content of email exchanges.

CONCLUSION

We have presented Themail, an original approach to email archive visualization that uses content to portray individual relationships. Our user study revealed two main interaction modes with the visualization, exploration of "big picture" trends and themes ("haystack") and more detail-oriented exploration ("needle"). The overwhelming majority of participants utilized the system in the "haystack" mode and were especially fond of looking at their relationships with family members and loved ones. These users often remarked on the photo-album quality of Themail and said that they would like to share the visualization with others. Users in the "needle" mode were more interested in finding specific bits of information in the visualization; they were especially interested in being able to identify information that was work related.

By displaying large collections of keywords that reflected the evolution of relationships over time, Themail placed users' past email exchanges within a meaningful context. Given that email is a habitat [5] and that an increasing amount of our daily interactions with others occur via email, the supplementary contextual cues presented in Themail can greatly improve users' utilization of email archives. We do not, however, expect users to utilize applications such as Themail on a daily basis. Given the analogy with photographs, it seems more likely that users might engage in sporadic explorations of the display to reminisce.

We relied on participants' familiarity with the data for effectively learning about some of the shortcomings of our content analysis algorithm. We propose that recognizing the importance of personal identification with the data is a key contribution to email content studies in particular and email research in general.

ACKNOWLEDGMENTS

We are grateful to Shreyes Seshasai and Daniel Martines for their help in implementing Themail. We would also like to thank the participants in our study for allowing us to share their thoughts on Themail.

REFERENCES

1. Bellotti, V., Ducheneaut, N., Howard, M., & Smith, I. (2003). *Taking Email to Task: The Design and Evaluation of a Task Management Centered Email Tool*. In Proc CHI.
2. Carvalho, V. & Cohen, W. (2004) *Learning to Extract Signature and Reply Lines from Email*. Conference on Email and Anti-Spam.
3. Csikszentmihalyi, M., and Rochberg-Halton, E. 1981. *The Meaning of Things*. Cambridge University Press, Cambridge, UK.
4. Dabbish, L., Kraut, R., Fussell, S. & Kiesler, S. (2005) *Understanding email use: predicting action on a message*. In SIGCHI, ACM Press.
5. Ducheneaut, N., & Bellotti, V. (2001). *Email as habitat: an exploration of embedded personal information management*. Interactions, 8(5), pp. 30-38.
6. Fisher, D., & Dourish, P. (2004). *Social and Temporal Structures in Everyday Collaboration*. In Proc. CHI.
7. Frau, S., Roberts, J., & Boukhelifa, N. (2005). *Dynamic Coordinated Email Visualization*. In WSCG.
8. Freeman. L. (2000). *Visualizing Social Networks*. Journal of Social Structure, 1 (1).
9. Heer, J. *Exploring Enron: Visual Data Mining of E-mail*. Available online at http://jheer.org/enron/
10. Kerr, B. (2003) *Thread Arcs: An Email Thread Visualization*. IBM Research Report.
11. Nardi, B., Whittaker, S., Isaacs, E., Creech, M., Johnson, J., & Hainsworth, J. (2002). *ContactMap: Integrating Communication and Information Through Visualizing Personal Social Networks*. Communications of the ACM.
12. Perer, A., Shneiderman, B., & Oard, D. (2005) *Using Rhythms of Relationships to Understand Email Archives*. In Review.
13. Sack, W. (2000) Conversation Map: An Interface for Very-Large-Scale Conversations. Journal of Management Information Systems, Vol 17, No. 3.
14. Salton, G. (1989) *Automatic Text Processing - The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.
15. Sudarsky, S., & Hjelsvold, R. (2002) *Visualizing Electronic Email*. In International Conference on Information Visualization.
16. Venolia, G. & Neustaedter, C. (2003) Understanding sequence and reply relationships within email conversations: a mixed-model visualization. In SIGCHI.
17. Viégas, F., boyd, d., Nguyen, D., Potter, J. & Donath, J. (2004) *Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments*. In HICSS-37.
18. Whittaker, S. & Hirschberg, J. (2001) *The character, value, and management of personal paper archives*. In ACM TOCHI.
19. Whittaker, S. & C. Sidner Year (1996) *Email overload: Exploring personal information management of email*. In SIGCHI.