Beyond Threads: Identifying Discussions in Email Archives

Adam Perer*, Ben Shneiderman
Human-Computer Interaction Lab, Department of Computer Science
University of Maryland at College Park

ABSTRACT

Email archives have the promise of serving as great resources for historians and social scientists. However, making sense of the information in these archives is a challenge. Email messages are often not self-contained and are instead a part of an ongoing discussion. The process of determining when discussions commence and conclude is a difficult task to automate. Threading messages by common subject lines and reply-chain information in email headers has been a common way of assembling messages into discussions. However, even though email provides this structured information, it does not usually reflect the user's behavior. Our work helps email archive explorers interpret the archived messages by providing access to the full scope of discussions that stretch beyond the thread. We present an interactive visualization that allows explorers to perceive a discussion and navigate over time and people to gain the context they need.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces - Graphical User Interfaces (GUI)

Keywords: social information visualization, email archives

1 INTRODUCTION

Email archive explorers typically use keyword searching to find emails of interest. Advanced techniques also exist, such as rhythms of patterns [5], social networks [1], and statistical user behavior models [4]. However, these techniques also output a set of email messages, which then requires the explorers to make sense of them.

Past attempts to show an email's discussion focus on visualizing threads of messages [2,6]. Threads are typically generated by linking messages featuring common subject lines or reply-chain information contained within email headers. However, this information does not necessarily correlate with the user's behavior. Beyond these threads, email users continue discussions by forwarding other messages, modifying subject lines, or sending new emails to a subset (or superset) of the original participants. In order for historians and social scientists to be able to understand archived messages, they need tools to navigate through a full discussion, and not just the parts easily recovered from the structure of email.

In this paper, we present a tool for explorers to identify, visualize, and navigate discussions. When an explorer arrives at a potentially interesting message during a search, the message is shown along with all other messages connected by traditional threading techniques. Beyond the thread, explorers have quick access to all of the participant's messages during relevant time periods, as well as the messages of the people the participants talked to on the side. The visualization allows explorers to not only attempt to understand what a particular message means, but can also act as a point of departure for further exploration. We also present a novel interaction technique for exploring temporal data, with which the explorers can use to adjust the length of time shown in the visualization.

*email: adamp@cs.umd.edu

```
Email 1:  Date: Tue, 31 Jul 2000, 07:00 PST
From: gregg, To: sara
Subject: tomorrow's agenda
Message: lets do it!

Email 2: Date: Wed, 1 Aug 2000, 10:30 PST
From: sara,  To: gregg
Subject: RE: tomorrow's agenda
Message: ok. when?

Email 3: Date: Thu, 2 Aug 2000, 06:30 PST
From: gregg, To: sara, Cc: eric
Subject: RE: RE: tomorrow's agenda
Message: tonight.  meet me at my office.
```

Figure 1. The first email simulates a message retrieved by an explorer. The other two emails were recovered from traditional threading, but add little context.

2 IDENTIFYING DISCUSSIONS

Discussions in email archives are not limited to a single thread, and therefore analysis should not solely rely on data contained within the structure and headers of email. Our techniques allow explorers to locate the actual horizons of a discussion. As an example, we illustrate our approach on a simple discussion manufactured for this paper in Figure 1. Suppose an explorer retrieves the first email, which reads, "Lets do it!" If this message came from a corpus such as the Enron archive [3], an explorer might wonder if this message is a initiation of a lunch meeting or an impetus for committing fraud. The second and third emails listed in Figure 1 are those that were part of the same message thread. The minimal content in the body of the messages provide little useful information, although they convey a third participant is added to the discussion in the third email.

Figure 2 shows these three emails in the visualization we designed. Each vertical line represents an email, oriented along the horizontal axis according to the time it was sent. All participants in the discussion are listed on the left. For every message a participant is addressed, a colored circle is drawn aligned with their name, and connected to the email's linear representation. For clarity, a subtle gray line is drawn

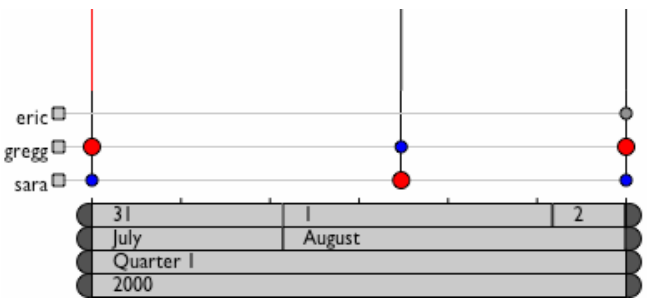

Figure 2. Visualization of the discussion presented in Figure 1.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2005		**00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Beyond Threads: Identifying Discussions in Email Archives	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Defense Advanced Research Projects Agency,3701 North Fairfax Drive,Arlington,VA,22203-1714**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

Email archives have the promise of serving as great resources for historians and social scientists. However, making sense of the information in these archives is a challenge. Email messages are often not self-contained and are instead a part of an ongoing discussion. The process of determining when discussions commence and conclude is a difficult task to automate. Threading messages by common subject lines and reply-chain information in email headers has been a common way of assembling messages into discussions. However, even though email provides this structured information, it does not usually reflect the user's behavior. Our work helps email archive explorers interpret the archived messages by providing access to the full scope of discussions that stretch beyond the thread. We present an interactive visualization that allows explorers to perceive a discussion and navigate over time and people to gain the context they need.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**2**	

horizontally from the participants name to each node that represents them. The circle's color also defines the role of the participant: senders are painted red, direct recipients are painted blue, and copied recipients are painted gray. The sender's circles have a slightly larger size, which emphasizes their importance as the author of the message.

The horizontal axis represents time and it is labeled by our time navigation tool described in Section 3. Moving the mouse over an email's nodes will present a brief summary of the message within the visualization, and clicking an email's node will highlight it in red, as well as show the full message in a separate panel.

The visualization's power shines when emails that are not a part of the thread are unveiled. At any time, explorers can enable messages to be shown that are temporally and socially relevant. These revealed messages include those during the active time period that feature at least one of the original participants somewhere on the address line. Figure 3 is a visualization of the same earlier discussion with the addition of these relevant nodes. Five additional emails are revealed and are notated by yellow lines. A new email (second from left) is identified that involves all three original participants. This email was not included in the previous and traditional visualizations because it was in a separate thread. A new participant, Matt, is now included in the discussion because he exchanged emails with several of the participants throughout the course of their original discussion. Explorers can then determine whether his messages contribute any new information to the discussion. By displaying all of the messages that are connected by time and people, explorers can become aware of messages they might not have easily reached.

Explorers also have the option of filtering out nodes that are not of interest. Irrelevant people and messages can be removed from the visualization, as well as entire classes of nodes, such as all sender or recipient nodes.

3 Navigating Time

In the visualization, the span of time shown on the horizontal axis defaults to the duration of time from the message's subject line thread. That is, the leftmost time point is the oldest message in the thread, and the rightmost time point is the most recent message in the thread. Since a discussion is not limited to the time span of a thread, an explorer needs a way to navigate through time easily. We have created a new interaction technique for temporal navigation, shown at the bottom of both Figures 2 and 3. The number of rows on the tool depends on the length of time being visualized. Each row on the tool represents a different granularity of time. In these figures, the discussion spans three days, so the navigation tool shows the year, season, month, and day details.

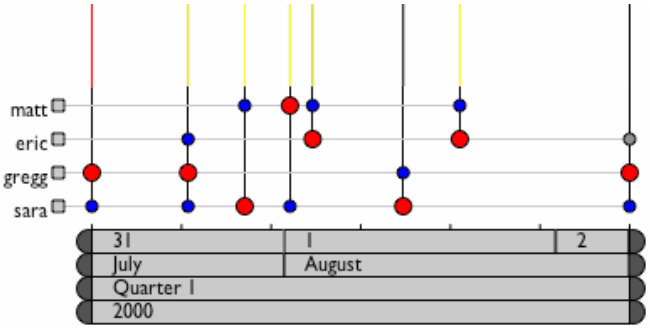

Figure 3. Visualization of the same discussion in Figure 2 after identifying additional relevant messages.

Explorers can customize the seasons depending on the archive. For instance, if the emails are from a company's archive, it might make sense to have the seasons represent fiscal quarters, whereas if the emails are from a professor's archive, it might make more sense to have them represent semesters from an academic calendar.

Each row is labeled with segments identifying the granular values of time from left to right. These segments also act as buttons that allow the users to zoom in to more specific time periods. For instance, if the user clicked the segment labeled '31', the visualization would zoom in and only messages from July 31, 2000 would be shown. A new row, representing hours, would also appear on the top of the navigation tool, allowing the user to get more detailed information on the time of day the messages where sent. Conversely, if the user selects 'Quarter 1', the visualization will zoom out to show the entire length of the quarter, and the day and hour rows would disappear, as the labels would be too small to be read. In general, rows disappear if the labels on the buttons cannot be displayed legibly. Each row also has left and right buttons on each side, which allow the explorer to expand the visualization easily with the granularity of their choice. For example, clicking the button on the right of 'August' segment in Figure 3 will add the full month of September to the visualization, whereas clicking on the button next to '2', will add only the next day, August 3.

4 Conclusion

Many email archives are currently being preserved with the intention that they will be a valuable resource in the future. Governments preserve some email archives as public records, businesses may keep their emails to comply with certain laws, and some individuals save their personal email as a record of their life. In order for historians and social scientists to make use of these archives, they need powerful tools to navigate through a vast number of messages. In this paper, we present a new technique to allow email archive explorers the ability to identify discussions. By presenting visualizations that go beyond the threads of email, explorers can become aware of messages that would have otherwise been hidden. Our legible visualizations present relevant temporal and social data that provides explorers with an enhanced opportunity to understand discussions evidenced in email archives.

Acknowledgments

We'd like to thank Douglas W. Oard for inspiration and advice for exploring email archives. Adam Perer has been supported in part by DARPA cooperative agreement N660010028910.

References

[1] Heer, J. (2005): 'Exploring Enron: Visual Data Mining', http://www.cs.berkeley.edu/~jheer/anlp/final/.

[2] Kerr, B. (2003): 'Thread Arcs: An Email Thread Visualization', *2003 IEEE Symposium on Information Visualization*.

[3] Klimt, B. and Yang Y. (2004): 'Introducing the Enron Corpus', *2004 Conference on Email and Anti-Spam*.

[4] Li, W., Hershkop, S., and Salvatore, S. (2004): 'Email Archive Analysis Through Graphical Visualization', *2004 ACM Workshop on Visualization and Data Mining for Computer Security*.

[5] Perer, A., Shneiderman, B., and Oard, D. W. (2005): 'Using Rhythms of Relationships to Understand Email Archives', *University of Maryland HCIL Tech Report HCIL-2005-08*

[6] Venolia, G. and Neustaedter, C. (2003): 'Understanding Sequence and Reply Relationships within Email Conversations: A Mixed-Model Visualization', *Proceedings of ACM CHI 2003*.