Email-Set Visualization: Facilitating Re-Finding in Email Archives

Doug Gorton, Uma Murthy, Srinivas Vemuri, Manuel A. Pérez-Quiñones
{dogorton, umurthy, nvemuri}@vt.edu, perez@cs.vt.edu

ABSTRACT

In this paper we describe ESVT – EmailSet Visualization Tool, an email archive tool that provides users a visualization to re-find and discover information in their email archive. ESVT is an end-to-end email archive tool that can be used from archiving a user's email messages to visualizing queries on the email archive. We address email archiving by allowing import of email messages from an email server or from a standard existing email client. The central idea in ESVT's visualization, an "email-set", is the set of emails that are the result of a query on a user's email archive. ESVT provides a multiple email-set view - visualization of multiple email-sets on a time axis. In addition, each email set can be individually visualized based on person and time axis, using the single email-set view. Query logs, individual email visualization, multiple email set visualization provide rich contextual cues, thus enabling end users to deal with email overload and re-find past email which otherwise wouldn't be discovered easily.

Author Keywords

Email archive, Visualization.

ACM Classification Keywords

H5.m. Information interfaces and presentation

INTRODUCTION

Email is one of the primary communication tools in our daily life. We use it for communication, archiving personal information, task management and contact management. Whittaker et al. discuss the importance of email in daily life, and its active role as a personal information management (PIM) tool [19]. Despite its wide use as a PIM tool, the traditional focus of email clients is merely on retrieval and simple organization and searching of emails, not necessarily all encompassing information management suites [17]. Even worse, the sheer volume of information being pushed to users via email each day leads to an excess of data commonly referred to as email overload [19]. Individual emails lack context when simply listed with others textually thus being able to view emails visually amongst one another is critical for full understanding. Based on how email users combat this abundance of email they can be categorized into three groups based on folder usage – no filers, frequent filers, and spring cleaners [19]. Each category uses a particular strategy to solve this overload problem.

These different strategies to handle email, however, fail after a certain point where volume becomes too great or users attempt to use email for purposes it wasn't designed such as massive archiving. We believe that it might be useful to separate archiving functions of email, from an email client to a different application, in order to focus on the archiving PIM issues more effectively.

The current state of art tools for email archiving is entirely focused on archiving enterprise mail servers. IDC, a premier global market intelligence firm, predicted that email archiving application revenue will reach $310 million globally by the year 2005 [12]. Most commercial email archiving solutions focus on mail servers since it is a lucrative market with far more impact and potential revenue than the personal archiving market. There have been some partial solutions proposed such as forwarding email to a database [6], using a separate email client like Eudora or Mail.app for archiving [13], using a combined tool as a client, server and an archival application, like ZOE [3]. Related to archival visualizations, Li et al [16] developed email mining kit for visualizing email archives of groups of users.

We observe that there is little work done in addressing visualizations of email archives from a personal perspective. In order to help ease the issues associated with email overload and with information re-finding, we have developed a personal email archive tool, ESVT. ESVT allows users to work with email archives, while focusing on re-finding information through more visual and directed searching. Our hypothesis is that an "email-set" (described in detail below) and its visualization will help users work with archived email, especially in their information re-finding activities.

RELATED WORK

Email is a critical communication and personal information organization tool and thus a fair sized body of work has focused on its impact and user behaviors. For our email visualization tool, we have extended the email thread visualization architecture mentioned developed by Perer and Shneiderman [11], to support visualization for sets of emails resulting from searches across email archives. Choices that lead to our visualization interface is supported by other findings as well. Donath describes the importance of representing memories and conversations in email visualization [8]. She focuses on showing ties and context of emails with regard to others in visualizing collections of communications. In their multi-scale email interface that ties together intimacy and chronology-based visualization of past conversations, Mandic et. al. explains email history as a vibrant, clear record of one's past [10]. There, they suggest multiple icons to represent outgoing and incoming emails in order to clearly identify those qualities to the user in visualizations. In our tool we draw upon this concept of a continual history in emails to better show users their email conversations over time.

Viegas et al. [14], through their investigations on visualizing past email archives, observed that these visualizations motivated retelling stories from the user's past to others. They created two kinds of visualizations for the past email archives. One approach highlighted the use of social networks, while the other approach depicted temporal rhythms of interactions, thus emphasizing the importance of people and time in email archive visualization. In Viegas et. al.'s Themail email visualization tool, [15], the authors focus on visualization based on content of email messages. The tool uses two main interaction modes, a "haystack" mode showing overall evolution in conversations, based on keywords in email message content. A "needle" mode allows users to see more details and locate specific pieces of information, individual emails for example. Cole et al. have proposed a lattice structure for navigating through emails versus a hierarchical tree structure that is based on folder organization [7]. Whittaker et al discuss remembering past communications with the help of associative retrieval cues - remembering items/people based on other items/people related to the target [18].

In the multi-store model of human memory, Atkinson and Shiffrin make a distinction among sensory, working/short-term and long-term memory [4]. Analogous to this notion is the idea information types (ephemeral, working, archived) described by Nardi et al. [5]. Atkinson and Shiffrin describe long-term memory (LTM) as being responsible for our persistent ("crystallized") memories of procedural, episodic and semantic information. Archived information, to a large extent, is what an individual makes persistent for later use. Wiseman and Tulving [20] originally formulated the encoding specificity principle, which maintains that information items are encoded in our LTM with respect to their context and that retrieval is a function of similarity to that encoding context [9]. Thus, retrieval cues, which we use to retrieve information from out LTM, serve as context for information we are trying to recall.

THE EMAIL-SET APPROACH

The design of ESVT revolves around the idea of an email-set. We use the term *email-set* to denote the result of a querying activity on a user's email archive. An email-set simply refers to a set of emails. It can be a result of a complex search query, a thread conversation, a set of emails from a mailing list, etc. Our tool focuses on visualization of email-sets within archives with the goal of helping a user in the information re-finding activity. We believe that the visual representation of the email-set provides context, defined by two dimensions – time and people.

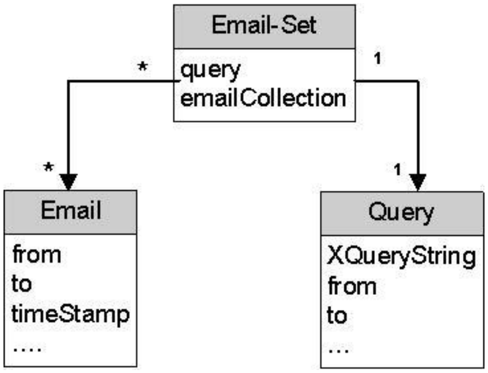


Figure 1. ESVT information model.

ESVT's information model is shown in Figure 1. An email-set is a collection of emails and a query. Each email object may belong to one or more email-sets, and contains various fields in an email, like to, from, time information, etc. Each query contains the XQuery string that specifies the query and may contain other details like to, from, etc incase of an advanced query.

ESVT is a complete email archive tool that supports keeping, organization, and re-finding activities. We discuss all three activities here, with focus on re-finding.

Keeping and Organizing

ESVT allows a user to add email from her existing email client using an import feature. Each email essentially is an XML serialization of the Mime message format. Currently, one can import email either from an email server or from a standard email client such as Thunderbird. The process of importing from an email server is automatic, while importing from an email client is semi-automatic. These XML formatted email messages are then stored in eXist - a native XML database [1].

While exporting emails into the archive, the folder organization of emails is retained. This ability to retain the structure of emails allows our application to support both filers and pilers. Our loose *email-set* concept can relate to the habits of both pilers and filers. Based on our definition of email sets, for a filer, a set could refer to the contents of a folder, whereas for a piler, it could be the result of a query.

Re-finding

As mentioned in the Related Work section, retrieval cues, which we use to retrieve information from out LTM, serve as context for information we are trying to recall. Context in the case of an email in an archive, could be the date, persons involved, keywords in the email subject or content, etc. In querying an email archive, we believe that the results provide us with additional cues, and eventually, enough context to recall information (that we initially set out to recall).

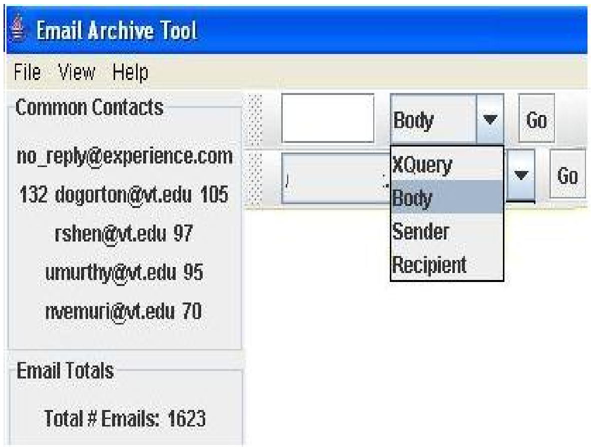
The ESVT visualization aims to help in this process of information recall. In order to assist users in finding other emails in a given message or context retrieved from a query, we provide a way to re-find relevant email from the email-set visualization. The visualization also aims to help users discover new links among information items.

The re-finding process using ESVT involves querying the email archive, visualizing the result email-set, and then discovering relevant email from the context provided by the email-set visualization.

Querying

ESVT allows users to formulate and run queries over archived emails. In the current prototype, a user has the option to search in the “sender” (from), “recipient” (to), and “body” fields of an email. In addition, the user can formulate an XQuery expression to represent any kind of advanced search query. This query functionality may be replaced by a more user-friendly interface along with advanced functions in the future. A built-in query log feature logs all user queries. We believe that this might serve as a reminder of past searches. It also saves the user the trouble of repeatedly entering recurring queries.

Figure 2 shows the main interface of ESVT (the figure has been adjusted to show details in the given space). The left two panes are used to show the user’s email archive statistics. For example, total number of archived emails and most common contacts (people from whom most emails were received). The top right pane is the query pane, where the user can specify a query. Below the query pane is the history pane, which the user may use to go back to past queries.



Common Contacts

no_reply@experience.com
132 dogorton@vt.edu 105
rshen@vt.edu 97
umurthy@vt.edu 95
nvemuri@vt.edu 70

Email Totals

Total # Emails: 1623

Figure 2. ESVT main interface.

Visualization

Our main emphasis is on visualizing email-sets and enabling the user to discover connections and relevant email using this visualization. We extended the Beyond Threads architecture by Perer and Shneiderman [11], to support visualization of email-sets as well as groups of email-sets. ESVT supports two related views: a multiple email-set view and a single email-set view.

Figure 3 is a snapshot of the multiple email-set view, with time on the X-axis and the XQuery expression (representing the associated email-set) on the Y-axis. The outlined red squares signify that the user was the sender of the email (part of that email-set) and the blue filled circles signify that the user was a receiver in the email (part of the email-set).

A snapshot of the single email-set view is shown in Figure 4. As with the multiple email-set view, the X-axis represents time. The Y-axis shows the people (represented by their email addresses) involved in the email-set. Each vertical line parallel to the Y-axis denotes an email. The

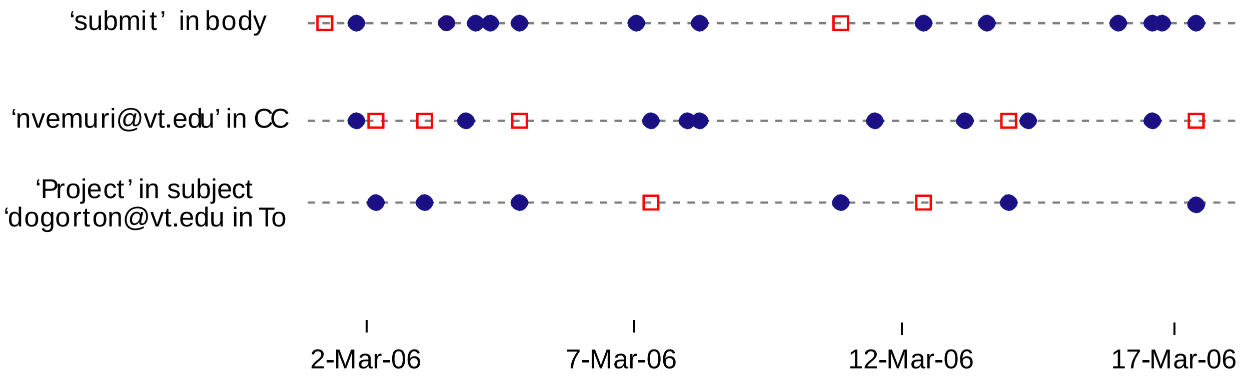


Figure 3. Multiple email-set view.

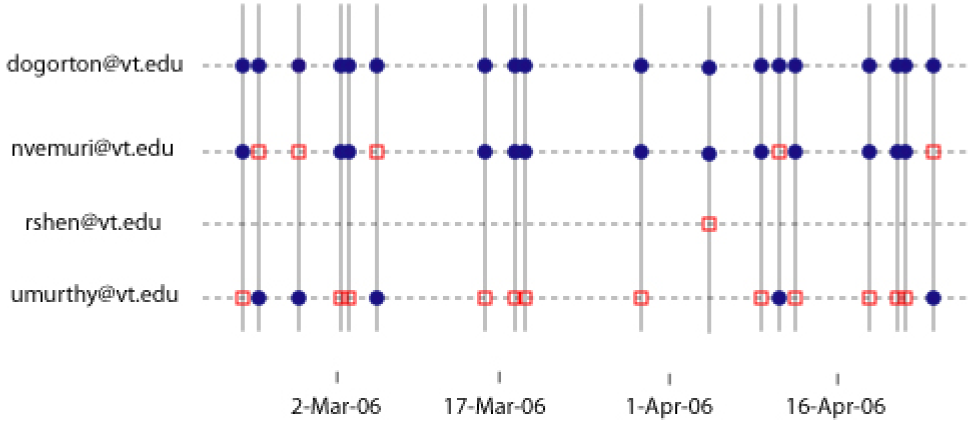


Figure 4. Single email-set view.

outlined red squares signify that the person (on the Y-axis) was the sender of the email and the blue filled circles signify that the person (on the Y-axis) was a receiver in the email.

In each of these visualizations, tooltips help a user to gain a quick overview of an email. Currently, only the timestamp of an email is enabled. In the future, we envision a user being able to see other email information, like the subject. Also, on clicking an email, the contents of the email will be displayed.

These visualizations have been developed using JFreeChart [2].

Discovering new connections/relevant emails

According to Viegas et. al., the two main dimensions of email archives are people and time [14]. Our email-set visualization method keeps this in mind and lets the user discover relevant email, by using one of two options:

1. Discovery using time scale: ESVT allows the user to zoom in and zoom out the time axis (X-axis) of the visualization. By zooming in the user can avoid visual clutter and focus on a shorter duration within the time-frame of the email-set. Another feature that can be provided (not present in the current prototype), is allowing the user expand the time scale (zoom-out), thus facilitating discovery of new email involving at least one person present in the original email-set. This might result in showing new persons who are additionally involved in new emails.

2. Discovery by person dimension (not present in the current prototype): This feature will allow the user to remove and add people (email addresses) from and to an existing email-set view. For example, a user should be able to remove a person from the Y-axis, which would result in filtering away all email sent/received by that person. Alternatively, by adding a person, all email sent/received by that person during the specified time frame are displayed. In many cases, a user may be interested in email from a certain sender, and the ability to show only emails from that individual would greatly cut down the degree of information overload.

EVALUATION

Our hypothesis is that the idea of an email-set and its visualization will help users work with archived email, especially in their information re-finding activities. It will help users discover new connections based on content (information contained within an email) and metadata (people, time, and subject).

We performed an informal evaluation to test our hypotheses. The evaluation involved three participants consisting of two Graduate students and one Undergraduate student. All three participants used their email client to archive emails. For each participant, we imported their email and used it to give a demonstration of ESVT and its features. Then participants formulated sample queries and tested ESVT. They tested both, the multiple email-set view and the single email-set view. All three participants felt that ESVT would be a useful tool to someone who received and archived large volumes of email. Among other things, the participants said that ESVT could be used for performing complex searches over email archives and for viewing trends in email exchanges. Since ESVT is a prototype, there is considerable scope for improvement in terms of robustness of the tool as well as additional features. The participants suggested adding features like being able to view the content of an email through the visualization, indicating common emails among email-sets, etc.

This evaluation was meant to be information and its objective was to get initial feedback on the use and usability of ESVT. A formal evaluation would involve more structured tasks or a longitudinal study of use of ESVT. Along with qualitative measures, quantitative measures from logs of user-system activity could be studied to make inferences about its use.

CONCLUSION AND FUTURE WORK

We developed ESVT – an email archive tool that can archive a user's email in a seamless manner. The user can query the archived email, visualize each email set- result of a query, in a multiple email set visualization panel. Also, the user can visualize individual email set on a person and time axis. The entire tool is written in Java, thus is portable across various platforms. The XML formatted emails can be transformed into any proprietary format using XSL-FO, XSLT style sheets.

These visualizations provide rich information, and patterns that otherwise are difficult to discover by looking at a flat list of emails. These visualizations allow the user to navigate to different portions of an email by interacting with the person and time axes.

ESVT is a prototype that was developed to demonstrate an idea. We see possible future work in different aspects. For example, the multiple and single email-set visualizations may be improved in order to provide more manipulative interactions on person, subject and time dimensions. This will allow the user to discover other email from this context. Although we performed an informal evaluation, a more traditional evaluation could be performed as mentioned in the previous section.

XML formatted email messages are stored in, with their folder hierarchy preserved. Another future improvement would be letting users refine queries based on folder structure in addition to current and advanced methods. In addition, XQuery's capability to access content at element level (subject, from, to etc) rather than document level (individual email) can be exploited further to build other data mining tools on top of an email collection.

ACKNOWLEDGMENTS

We thank Johnny Livingstone for suggesting visualization package, JFreeChart, and Dr Perez and our fellow classmates in the Personal Information Management class for feedback and suggestions.

REFERENCES

1. eXist: Open Source Native XML Database, 2006, http://exist.sourceforge.net/.
2. JFreeChart, 2006, http://www.jfree.org/jfreechart/index.php.
3. ZOË Email Archiving, 2005, http://www.cdegroot.com/blog/2005/12/05/zoe-email-archiving/.
4. Atkinson, R.C. and Shiffrin, R.M. Human memory: A proposed system and its control processes. in Spence, K.W. and Spence, J.T. eds. *The psychology of learning and motivation (Vol. 2)*, Academic Press, London, 1968.
5. Barreau, D. and Nardi, B.A. Finding and Reminding: File Organization from the Desktop. *SIGCHI Bull.*, *27* (3). 39-43.
6. BORNSTEIN, N. Personal Email Archiving, http://monkeyfist.com/articles/508.
7. Cole, R., Eklund, P. and Stumme, G., CEM - A Program for Visualization and Discovery in Email. In *Proceedings of Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases*, (2000), Springer, 367-374.
8. Donath, J. Visualizing Email Archives, 2004, http://smg.media.mit.edu/papers/Donath/EmailArchives.draft.pdf.
9. Eysenck, M.W. and Keane, M.T. *Cognitive Psychology: A student's handbook*. Psychology Press, Philadelphia, PA, 2000.
10. Mandic, M. and Kerne, A., Using intimacy, chronology and zooming to visualize rhythms in email experience. In *Proceedings of CHI '05 extended abstracts on Human factors in computing systems*, (Portland, OR, USA, 2005), ACM Press.
11. Perer, A. and Shneiderman, B., Beyond Threads: Identifying Discussions in Email Archives. In *Proceedings of InfoVis 2005, the eleventh annual IEEE Symposium on Information Visualization*, (Minneapolis, MN, 2005).
12. Thomas, B.D. MailArchiva: Open Source Email Archiving Server, 2006, http://www.linuxsecurity.com/content/view/121268/65/.
13. Tsai, M. E-Mail Archiving with Eudora and Mail.app.
14. Viegas, F., Boyd, D., Nguyen, D., Potter, J. and Donath, J., Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments. In *Proceedings of 37th Hawaii International Conference on System Sciences*, (2004).
15. Viegas, F.B., Golder, S. and Donath, J., Visualizing Email Content: Portraying Relationships from Conversational Histories. In *Proceedings of CHI 2006*, (2006), To Appear.
16. Wei-Jen, L., Shlomo, H. and Salvatore, J.S., Email archive analysis through graphical visualization. In *Proceedings of 2004 ACM workshop on Visualization and data mining for computer security*, (Washington DC, USA, 2004), ACM Press.
17. Whittaker, S., Bellotti, V. and Gwizdka, J. Email in personal information management. *Commun. ACM*, *49* (1). 68-73.
18. Whittaker, S., Jones, Q., Nardi, B., Creech, M., Terveen, L., Isaacs, E. and Hainsworth, J. ContactMap: Organizing communication in a social desktop. *ACM Transactions on Computer-Human Interaction*, *11* (4). 445-471.
19. Whittaker, S. and Sidner, C., Email overload: exploring personal information management of email. In *Proceedings of CHI 96. Human Factors in Computing Systems.*, (Vancouver, BC, Canada, 1996), ACM, 3-18.
20. Wiseman, S. and Tulving, E. Encoding Specificity. *Journal of Experimental Psychology: Human Learning and Memory*, *2*. 349-361.