Am I wasting my time organizing email? A study of email refinding

Steve Whittaker, Tara Matthews, Julian Cerruti, Hernan Badenes, John Tang
IBM Research - Almaden
San Jose, California, USA
{sjwhitta, tlmatthe}@us.ibm.com, {jcerruti, hbadenes}@ar.ibm.com, johntang@microsoft.com

ABSTRACT
We all spend time every day looking for information in our email, yet we know little about this refinding process. Some users expend considerable *preparatory* effort creating complex folder structures to promote effective refinding. However modern email clients provide alternative *opportunistic* methods for access, such as search and threading, that promise to reduce the need to manually prepare. To compare these different refinding strategies, we instrumented a modern email client that supports search, folders, tagging and threading. We carried out a field study of 345 long-term users who conducted over 85,000 refinding actions. Our data support opportunistic access. People who create complex folders indeed rely on these for retrieval, but these preparatory behaviors are inefficient and do not improve retrieval success. In contrast, both search and threading promote more effective finding. We present design implications: current search-based clients ignore scrolling, the most prevalent refinding behavior, and threading approaches need to be extended.

AUTHOR KEYWORDS
Email, refinding, management strategy, search, conversation threading, folders, usage logging, field study, PIM.

ACM Classification Keywords
H5.3 Group and Organization Interfaces: Asynchronous interaction, Web-based interaction.

INTRODUCTION
The last few years have seen the emergence of many new communication tools and media, including IM, status updates, and twitter. Nevertheless, in work settings email is still the most commonly used communication application with reported estimates of 2.8 million emails sent per second [15]. Despite people's reliance on email, fundamental aspects of its usage are still poorly understood. This is especially surprising because email critically affects productivity. People use email to manage everyday work tasks, using the inbox as a task manager and their archives for finding contacts and reference materials [2,7,23]. This paper looks at an important, under-examined aspect of task management, namely how people *refind* messages in email.

Refinding is important for task management because people often defer acting on email. Dabbish et al. [7] show that people defer responding to 37% of messages that need a reply. Deferral occurs because people have insufficient time to respond at once, or they need to gather input from colleagues [2,23]. Refinding also occurs when people return to older emails to access important contact details or reference materials.

Prior work identifies two main types of email management strategies that relate to different types of refinding behaviors [13,23]. The first management strategy is *preparatory* organization. Here the user deliberately creates manual folder structures or tags that anticipate the context of retrieval. Such preparation contrasts with *opportunistic* management that shifts the burden to the time of retrieval. Opportunistic refinding behaviors such as scrolling, sorting or searching do not require preparatory efforts. Previous research has noted the trade-offs between these management strategies. Preparation requires effort, which may not pay off, for example if folders do not match retrieval requirements. But relying on opportunistic methods can also compromise productivity. Active foldering reduces the complexity of the inbox. Without folders, important messages may be overlooked when huge numbers of unorganized messages accumulate in an overloaded inbox [2,7,23].

Choice of management strategy has important productivity implications since preparatory strategies are costly to enact. Other work has shown that people spend an average of 10% of their total email time filing messages [3]. On average, they create a new email folder every 5 days [5]. People assume that such preparatory actions will expedite future retrieval. However, we currently lack systematic data about the extent to which these folders are *actually used*, because none of these prior studies examined actual access behaviors. Such access data would allow us to determine whether time spent filing is time well spent. This is important because prior work suggests that organization can be maladaptive, with people creating many tiny 'failed folders' or duplicate folders concerning the same topic [23].

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Another important reason for reexamining how people manage and access email is the emergence of new *search-oriented* clients such as Gmail [12]. Such clients assume the benefits of the opportunistic approach as they do not directly support folders. A second novel characteristic is that they are *thread-based*. Building on much prior work on email visualization [2,19,20], Gmail offers *intrinsic* organization, where messages are automatically structured into threaded conversations. Threads potentially help people more easily access related messages. A thread-based inbox view is also more compact, enabling users to see more messages without scrolling, helping people who rely on leaving messages in the inbox to serve as 'todo' reminders. We therefore examine the utility of these new email client features by determining whether search and threads are useful for retrieval.

We extend approaches used in prior work that tried to identify email management strategies by analyzing single snapshots of email mailboxes for their structural properties, such as mailbox size, number of folders, and inbox size, [2,11,13,23]. We also know that users are highly invested in their management strategies [2,23] so it is important to collect objective data about their efficacy. We therefore logged *actual daily access behaviors* for 345 users enacting over 85,000 refinding operations, and looked at how access behavior relates to management strategy. Our method has the benefit of capturing systematic, large-scale data about refinding behaviors 'in the wild'. It complements smaller-scale observational studies of email organization [1,2,23], and lab experiments that attempt to simulate refinding ('find the following emails from your mailbox') [10]. Finally our study also extends the set of users studied. Unlike prior work, only 2% of our users are researchers.

To apply this logging approach we needed to implement and instrument a fully featured modern email client. Later, we describe the client used to collect this data, which supports efficient search, tags, and threading.

This paper looks at the main ways that people re-access email information, comparing the success of preparatory vs. opportunistic retrieval. We explore how two aspects of refinding interrelate. On one level we wish to *characterize basic refinding behaviors* to determine whether people typically search, scroll, access messages from folders, or sort when accessing emails. We also want to determine the efficiency and success of these different behaviors, as well as how behaviors interrelate. At the next level, we want to examine *the relationship between refinding behaviors and people's prior email management strategies*, to determine for example, whether people who have constructed complex folder organizations are indeed more reliant on these at retrieval. We therefore ask the following specific questions:

Access behaviors: What are people's most common email refinding behaviors, when provided with a modern client that supports search, tagging, and threads, as well as folders? Do people *opportunistically* refind emails by scanning their inbox, searching, or sorting via header data? Or instead do they use *preparatory* behaviors that exploit pre-constructed organization in the form of folders or tags? Also, what are the interrelations between behaviors? For example, are there people who rely exclusively on search and never use folders for access?

Relations between management strategy and access behaviors: Does prior organizational strategy influence actual retrieval? Are people who prepare for retrieval by actively filing, more likely to use these folders for access? In contrast, are people who make less effort to prepare for retrieval more reliant on search, scanning, and sorting?

Impact of threads on access: Do threads affect people's access behaviors? Are people with heavily threaded emails less reliant on folders for access?

Efficiency and success of management strategies and access behaviors: We also wanted to know whether access behaviors affect finding outcome. Which behaviors are more efficient and which lead to more successful finding? We might expect folder-access to be more successful than search, as people have made deliberate efforts to organize messages into specific memorable categories. On the other hand, search may be more efficient as it might take users longer to access complex folder hierarchies. Finally, are people who create many folders more successful and efficient at retrieval?

RELATED WORK

Studies of email use have documented how people use email in diverse ways, including for task management and personal archiving [2,13,23]. Foldering behaviors are the most commonly studied email management practice. Whittaker and Sidner [23] characterized three common management strategies: no filers (forego using email folders, relying on browsing and search), frequent filers (minimize the number of messages in their email inbox by frequently filing into many folders and relying on folders for access), and spring cleaners (periodically clean their inbox into many folders). Fisher et al. [11] also added a fourth management strategy: users who kept their inboxes trim by filing into a small set of folders. Other studies [1,2] discovered similar management strategies, but also found that users did not exclusively fall into one category. Rather, users employ a combination of strategies over time [1,11].

Grouping messages together according to conversational threads (i.e., a reply chain of messages on a common topic) has been explored in prior research [2,3,19,20]. Gmail [12] uses threads (rather than individual messages) as the basic organizing unit for email management, although a more recent version also combines the functionality of folders and labels [16]. A thread-based inbox view is more compact, enabling users to see more messages without scrolling, helping those who rely on leaving messages in the inbox to serve as 'todo' reminders. Collecting messages into threads also gives users the context for interpreting an individual message [19]. While Venolia and Neustaedter

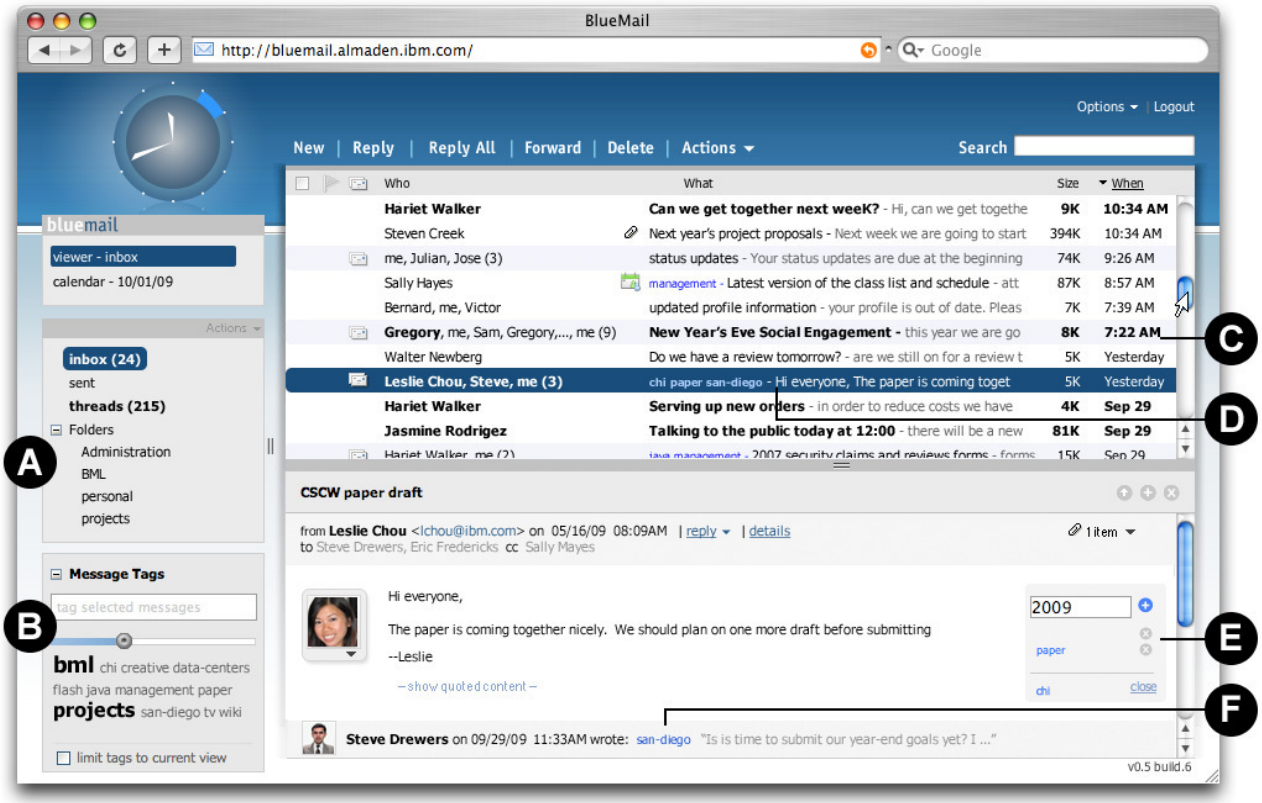

Figure 1. User interface design for Bluemail, showing panes for foldering (A) and tagging (B), on the left, a message list area in the top center showing a threaded message (C) and a selected thread (D) which is displayed in the message preview below showing an interface to add tags to a message (E) and display tags already added to a message (F).

[19] and Bellotti et al. [2] conducted studies of threading with small groups of users, there has not been a large-scale study of thread usage.

One might think that the emergence of effective search would lead users to reduce preparatory foldering. Yet Teevan et al. [18] observed for web access that even a perfect search engine could not fully satisfy users' needs for managing their information. Instead, their users employed a mix of preparatory and opportunistic refinding behaviors. We explore if this result holds for email refinding as well.

Other work has examined how people refind personal files on their personal computers, showing that people are more reliant on folder access than search. In addition, search and navigation are used in different situations: search is only used where users have forgotten where they stored a file, otherwise they rely on folders [4]. Dumais et al. [9] found that refinding emails was more prevalent than files or web documents, and that refinding tended to focus on recent emails. However, that study focused on search and did not compare it to other access methods, e.g. folders or scrolling.

Elsweiler, et al. [10] looked at memory for email messages. Participants were usually able to remember whether a particular message was in their mailbox. Also, memory for specific information about each message was generally good; people remembered content, purpose, or task related information best, correctly recalling over 80% of this type of information, even when items were months old. However, frequent filers tended to remember less about their email messages. Filing information too quickly sometimes led to the creation of archives containing spurious information; premature filing also meant that users were not exposed to the information frequently in the inbox, making it hard to remember its properties or even its existence.

THE BLUEMAIL SYSTEM

Bluemail is the email client used for this study. It is a web-based client that includes both traditional email management features such as folders, and modern attributes such as efficient search, tagging, and threads. This combination of features allowed us to directly compare the benefits of preparatory retrieval behaviors that rely on folders/tags, with opportunistic search and threading. We could not have made this direct comparison if we had used a client such as Gmail that does not directly support folders separately from tags. Also, Bluemail could be used to access existing Lotus Notes emails, making the transition to Bluemail very straightforward. For a full description of the design see [17].

Figure 1 shows the main Bluemail interface. The layout follows a common email pattern with navigation panes on

the left for views and foldering (to which Bluemail adds an interface for tagging), a central content area with a message list on top, and a message preview panel at the bottom. Messages are filed into folders by drag and drop from the message list into a folder in the left pane. One novel feature of Bluemail that enhances scrolling is the Scroll Hint. As the user engages in sustained scrolling (> 1 second) the interface overlays currently visible messages with metadata such as date/author of the message currently in view. This hint provides orienteering information about visible messages without interrupting scrolling.

Bluemail also supports efficient search (shown in the upper right of Figure 1) based on a full content index of all emails, with the search index being incrementally updated as new messages arrive. As in standard email clients, and unlike Gmail, messages can also be sorted by metadata fields such as sender ('who'), or date ('when'). The default view is by thread, which we now describe.

Message Threads

A message *thread* is defined as the set of messages that result from the natural reply-to chain in email. In Bluemail, threads are calculated against all the messages in a user's email database, i.e., threads include messages even if they have been filed into different folders. This design contrasts with clients that do not have true folders (like Gmail).

Bluemail uses the thread, not the individual message, as the fundamental organizing unit. Deleting, foldering, or tagging a thread acts on all the messages in the thread, even messages already foldered out of view. Figure 1C shows how threads are represented in the message list view. Each thread is gathered and collapsed into a single entry in the list. Users can toggle the view in the interface between the default threaded view and the traditional flat list of messages by clicking on the icon in the thread column header. The 'what' column for a thread shows the subject field corresponding to the most recently received message. After the subject text, we show in gray text as much of the message that space allows. User-applied tags are also shown pre-pended to the subject in a smaller blue font, as will be described in the tagging section below.

Tagging Messages

The interface for message tagging comprises four elements: a tag entry and display panel in the message, pre-pended tags in the list view's 'what' column, a tag cloud, and a view of the message list filtered by tag. As a user tags messages, the tags are aggregated into a tag cloud as shown in Figure 1B. Clicking on a tag (anywhere a tag appears) filters the message list to show only messages across a user's email (including other folders) with that tag. If any of those messages are part of a thread, the whole thread is shown in threaded view. Toggling to the unthreaded view shows only the individual messages marked with the tag.

METHOD

Users

The Bluemail prototype was released in our organization and used long term by many people. For our analyses, we focused on *frequent users*, i.e., people who used our system for at least a month, with an average of 64 days usage. As our main focus was on access behaviors, a criterion for inclusion was that a user had to have used each retrieval feature (folder-access, scroll, search, sort, tag-access) at least once. This assured us that users were aware of that feature's existence. Overall 345 people satisfied these criteria. Users included people from many different job roles (marketing, executives, assistants, sales, engineers, communications) and organizational levels (managers and non-managers). Unlike many prior email studies there were few researchers (just 2% of our frequent users).

Measures

Many prior studies of email have taken a snapshot of a user at a single point in time. This approach has the disadvantage that it may capture the email system in an atypical state. To prevent this, we therefore recorded longitudinal *daily system use*, averaging measures across the entire period that each person used the system.

General usage statistics

For each user, we collected and averaged the following usage statistics over each day they used the system:

- *Days of system usage*. We only included people with more than 30 days of usage.
- *Total messages stored* - number of messages included in all folders and the inbox.
- *Inbox size* - number of inbox messages.
- *Number of folders.*
- *Messages per thread* - number of messages in each thread, excluding messages without replies.
- *Daily change in mailbox size.* Other work notes that it is hard to determine the exact numbers of received messages because users delete messages [11]. We therefore recorded the *daily change* in mailbox size, i.e., the number of additional messages added or, in some cases, removed from the total archive each day. From a refinding perspective this is a better measure as it represents the set of messages users *potentially access* longer term.

Access behaviors

We also recorded various daily *access* behaviors. We logged each instance when the behavior was invoked.

- *Sort* - whenever the user clicked the various header fields such as sender, subject, date, time, attachments, etc.
- *Folder-access* - whenever a user opened a folder.
- *Scroll* - whenever users scrolled for more than one second (a conservative criterion adopted to identify when scrolling is used for refinding).

Table 1: Overall Usage Statistics.

	Mean	Std. Deviation
Days Used	**63.97**	42.61
Total Messages Stored	**2568.79**	3107.77
Inbox Size	**870.28**	1422.96
Number Folders	**46.89**	91.65
Messages/Thread	**3.61**	1.54
Daily Change in Size	**24.24**	58.07

- *Tag-access* - whenever a user clicked on a tag.
- *Search* - whenever the user conducted a search.
- *Open Message* - whenever the user opened a message.
- *Operation duration* - measured by subtracting the timestamp of each operation from the timestamp of the subsequent operation.

To preserve user privacy we did not record search terms or the names of folders and tags. We initially recorded other access operations, e.g., filter by flag (filtering for messages users had marked as important), or filter by unread messages (selecting the interface view which showed only unread messages). However, these behaviors accounted for less than 1% of all access behaviors and were only ever used by 8% and 17% of our users respectively. We therefore do not discuss them further.

We also recorded the *success* and *duration* of finding sequences. We define a finding sequence as a set of access behaviors containing one or more sort, scroll, search, tag-access, or folder-access. Each finding operation was treated separately, so that opening a folder followed by a sort was treated as two separate operations. Searching followed by sorting was treated the same way. Our analysis is quantitative and relied on parsing large numbers of logfiles, so we aimed to define an automatically implementable definition of success and duration.

Success: People usually want to find a target message to process the information it contains. We began by defining as successful an unbroken sequence of finding operations that terminated in a message being opened. Opening a message did not always indicate success, however. Observations of finding sequences revealed that users sometimes opened a message briefly, discovered that it was not the target, and then immediately resumed their finding operations. To determine the upper bound for this unsuccessful message opening interval, we timed 12 pilot users opening and reading two standard paragraphs from an email message that we felt would be sufficient for message identification. We found this took 29s. Any 'open message' operation lasting less than 29s and followed by subsequent finding operations was therefore treated as a non-terminal part of the finding sequence. 23% of sequences contained such unsuccessful opening of messages. Note that a user briefly opening a message and hitting 'reply' would be classified as a 'success' because the operation after 'open message' is not a finding operation.

Failure: We classified as *failures*, sequences of finding operations that did not terminate in a message being opened, e.g., when the sequence was followed by the user closing their browser, or composing a new message.

We acknowledge that finding success may also be influenced by subjective factors such as urgency or message importance. However our large-scale quantitative approach requires clearly definable success criteria, and it is hard to see how to operationalize these contextual factors in a working logfile parser.

Duration: The finding sequence duration was the sum of the finding operation durations it comprised. For one specific case, we excluded final operation time: when people abandoned an unsuccessful finding sequence, there were sometimes long intervals, lasting tens of minutes before the subsequent operation. We could not assume that the user was actively engaged in that operation for the entire interval, so we excluded it.

One potential limitation of this study is that we observed behavior for people who have been using our system for an average of two months. This may not be sufficient time for people to modify long-term email behaviors. To qualitatively profile our population however, we interviewed 32 users. We found that 60% regularly used Gmail, indicating that features such as tagging and search were highly familiar. Furthermore, we ensured that all users had used all access features at least once and found that certain features such as threading were immediately used ubiquitously—suggesting that people will readily change access strategy if they see the value of new technology.

RESULTS

Overall Statistics

Table 1 shows overall usage statistics, derived from daily samples. These are consistent with prior work (see Whittaker et al. [21] for a review), showing that users tend to build up large archives. However, the proportion (33%) of messages we observed being kept in the inbox is smaller than that reported in prior work. This may be due to different sampling methods, i.e., that we were sampling daily rather than relying on a single snapshot. Also, there may be over-representation of researchers in prior samples, and others [11] have speculated that researchers tend to hoard more than other types of workers. Finally threads did not tend to have a complex structure, with an average of 3.61 messages per thread, after we exclude singleton messages (i.e., messages without replies). As with all prior email research, there is high variability in most aspects of usage, as shown by the large standard deviations.

Access Behaviors

We next examined people's access behaviors, which have not been systematically studied before.

Table 2. Daily Usage, Distributions and Durations for Each Access Behavior. (Opportunistic behaviors are shaded.)

	Mean (SD) Accesses/day	% of All Accesses	Mean (SD) Duration in seconds
Folder-accesses	**0.21** (0.42)	**12**	**58.82** (30.22)
Tag-accesses	**0.02** (0.09)	**1**	not recorded
Searches	**0.71** (1.82)	**18**	**17.15** (36.86)
Sorts	**0.17** (0.33)	**7**	**13.96** (17.99)
Scrolls	**1.49** (1.76)	**62**	**25.77** (30.22)

Overall, each person had an average of 61.75 (SD 110.7) finding sequences, lasting a mean of 69.59s (SD 33.48). Overall 88% of finding sequences were successful, and the average number of finding operations per sequence was 3.85 (SD 3.57). As expected, successful sequences contained fewer operations than unsuccessful ones (M = 3.67, SD 1.83 and M = 5.03, SD 5.26), paired t-test $t(344) = 9.41$, $p < 0.001$, presumably because on encountering failure people persist in trying to find the target message.

Opportunistic behaviors dominate retrieval

Our first question concerned the overall frequencies of different access behaviors. We can distinguish between: (a) accesses based on *preparatory activity,* i.e., using folders and tags that users deliberately create in anticipation of retrieval, and (b) *opportunistic* accesses that do not rely on preparatory activity, i.e. sorting, scrolling and searching.

Table 2 shows that opportunistic behaviors dominate. These account for 87% of accesses. This is mainly explained by the predominance of scrolling which accounts for 62% of all accesses. Of course, scrolling might be used in preparatory contexts, e.g. scrolling through a large folder. However scrolling and folder-access are highly negatively correlated ($r(343)=-0.52$, $p<0.001$). Overall then, preparatory activities (folder- and tag-accesses combined) are not prevalent. They account for just 13% of all access operations overall.

A within sequence analysis indicated that specific behaviors tended to perseverate, with people relying on one or two strategies to find a specific message. Note too that there is enormous variability in individual usage for each of these behaviors (as indicated by their large standard deviations). The use of tagging was minimal, accounting for just 1% of all accesses. We therefore excluded it from subsequent analyses, removing finding sequences that included tags and relaxing the criterion that each user in the sample had to have used tag-access at least once. This added 13 users to our original user population.

Efficiency of Different Access Behaviors

Table 2 also indicates large differences in duration for the different access behaviors. Each individual folder-access took approximately one minute, more than twice as long as scrolls, with both searches and sorts being relatively short (around 15s). Paired t-tests show that folder-access operations are significantly longer than scrolls ($t(357) = 6.71$, $p<0.001$). Scrolls in turn are significantly longer than searches ($t(357) = 2.87$, $p<0.01$). However, there is no difference between the durations of searches and sorts ($t(357) = 0.51$, $p>0.05$).

Interrelations between Access Behaviors

We next explored the *interrelations* between access behaviors, where we anticipated specific patterns. Users who have made the effort to create folders should be more reliant on a preparatory behavior like folder-access, and avoid opportunistic behaviors like search, sort, and scroll. Others we expected to rely exclusively on these opportunistic behaviors, eschewing folder-access.

We found these expected combinations of access patterns. When we correlated behaviors per user, preparatory behavior, i.e. folder-access, was *negatively* correlated with search ($r(356) = -0.25$, $p<0.001$), and with scrolling ($r(356) = -0.52$, $p<0.001$). Thus, scrolling does not co-occur with folder-access, e.g. scrolling through large folders. As we expected, the various opportunistic behaviors positively intercorrelate, with searches correlating with sorts: $r(356) = 0.23$, $p<0.001$), and with scrolls $r(356) = 0.61$, $p<0.001$). This indicates that people tend to rely exclusively on either preparatory or opportunistic behaviors, but not a mixture of the two. This is an important result because it suggests that email clients, that mainly support search, like Gmail, are unlikely to be optimal for all users.

The Relationship between Email Management Strategy, Threads and Access Behaviors

So far our analysis has only examined access behaviors. In this section, we examine the relationship between access behaviors, threads, and email *management strategy*. Are people who engage in preparatory activity by making folders more likely to rely on these for retrieval? We also explore the effects of the intrinsic organization afforded by threads. Do threads make people less likely to use folders for access?

People who create folders use them more often for access

To analyze access behaviors with respect to management strategy, we must first operationalize management strategy. Prior work [1,13,23] proposed strategy differences, based on a combination of inbox size, number of folders as well as 'large scale changes' in inbox size over time. However, recent work [11] critiques these definitions arguing that they are ad hoc and do not reliably identify distinct user types. To avoid both of these problems, we used a simple *propensity to organize* metric based on the percentage of the user's total mailbox that is stored in folders. People who are more committed to foldering should have a higher percentage of their information in folders, as opposed to the inbox. We conducted a median split on this percentage to divide users into *high* and *low* filers.

Table 3 shows the frequency of different access behaviors (folder-access, scroll, sort, search), depending on whether a user is a high or low filer. To control for the large variance

Table 3. Access behaviors for High and Low filers based on a median split of percent messages in folders.

Access Behavior	% Mailbox Foldered	% of Accesses in which Behavior is Used	SD	Significance
Folder-accesses	High	**16**	22	t(356) = 4.68, p < 0.001
	Low	**7**	16	
Searches	High	**16**	18	t(356) = 2.05, p < 0.05
	Low	**21**	26	
Sorts	High	**8**	10	t(356) = 1.95, p=0.05
	Low	**6**	9	
Scrolls	High	**59**	26	t(356) = 2.39, p < 0.02
	Low	**65**	25	

Table 4. Access behaviors for High and Low threading based on median split of percent messages in threads.

Access Behavior	% Mailbox that is Threaded	% of Accesses in which Behavior is Used	SD	Significance
Folder-accesses	High	**10**	18	t(356) = 2.05, p < 0.05
	Low	**14**	20	
Searches	High	**24**	26	t(356) = 4.73, p < 0.001
	Low	**13**	18	
Sorts	High	**7**	10	t(356) = 0.09, ns
	Low	**7**	9	
Scrolls	High	**58**	25	t(356) = 2.65, p < 0.01
	Low	**66**	26	

in absolute numbers of accesses across different users, we normalize by expressing each access behavior as a *proportion* of all access operations for that user.

We expected that high filers would be more likely to use folders at retrieval, and less likely to scroll, sort, or search. This was confirmed: independent t-tests showed high filers were more likely to use folder-access and less likely to search or scroll (see Table 3). Contrary to our expectations, filers were slightly more likely to sort, possibly after accessing a large folder to identify a message from a particular person, time or topic.

These are striking results because a median split is a conservative statistical approach, as users who are just above or below the median may be very similar in terms of their filing strategy. We therefore checked our approach by comparing upper and lower deciles (i.e., people who *almost always* file with those who *almost never* file). To check the validity of the normalization, we also compared absolute numbers of each access behavior for the high/low split. Both analyses replicated the basic findings reported above.

Intrinsic organization: Threading reduces reliance on folders

There are other factors, such as threading, which potentially affect access behaviors. As shown in Figure 1, our client automatically organizes and presents emails as threads. Threading imposes a structure on messages and potentially represents a way for people to access related messages, without the burden of manually organizing them.

The perceived utility of threading is shown by the fact that our users made almost exclusive use of the threaded view. Users were able to switch from this view to a more standard sequential message view, but seldom did so. For all system users, 56% *always* used the threaded view, and for those who switched to the unthreaded view, only 1% stayed there more than 40% of the time.

We explored the effect of thread structure on access behavior, using the same approach as for foldering strategy. We first identified the percentage of each person's messages that participated in threads. We again conducted a median split on this thread percentage to distinguish people who received mostly heavily threaded emails ('high threading') from those whose messages tended not to be threaded ('low threading'). We expected that those with high degrees of intrinsic organization would be less reliant on access methods that demanded manual preparation for retrieval, such as folder-access.

Table 4 shows that, as we expected, people with highly threaded emails are less likely to use folder-access. This effect may be reinforced by the fact that in Bluemail, threads include messages that have been foldered. People who had more threads were also more reliant on search, which is possibly a response to situations where threads provide insufficient organization to access the message people need. Finally, people with more threads were less likely to scroll suggesting that threads were indeed an effective way to compress information in the inbox. Threads allow people to see more messages, reducing the need to scroll.

How do Efficiency and Success Relate to Management Strategy and Threads?

Foldering is Less Efficient and No More Successful

We next explored the overall *efficiency* of the preparatory management strategy. It takes time and effort to manually organize emails into folders, but does this effort pay off? Do people who prepare find information more quickly and successfully? Do they find information in fewer operations?

Table 5 reveals that high filers managed to find messages using fewer operations in each finding sequence. However, this did not equate to faster overall finding sequences, as high filers took marginally longer in their finding sequences. There is a simple explanation for this: high filers are more reliant on folder-accesses, which Table 2 shows take much longer than the searches and sorts.

More important is how often people *successfully find* the target message. Again, to control for the fact that people had very different numbers of finding sequences, we evaluated what percentage of their finding sequences were

Table 5. Success and efficiency of finding sequences for High and Low filers based on median split of percent messages in folders.

Measure	% Mailbox Foldered	Mean	SD	Significance
% of All Sequences that are Successful	High	**0.88**	.12	t(356) = 0.98, p > 0.05
	Low	**0.88**	.11	
Sequence Duration (secs)	High	**72.87**	38.05	t(356) = 1.97, p < 0.05
	Low	**66.07**	26.64	
# Operations	High	**3.69**	1.46	t(356) = 2.17, p < 0.05
	Low	**4.16**	2.50	

Table 6. Success and efficiency of finding sequences for High and Low threads based on median split of percent messages in threads.

Measure	% Mailbox that is Threaded	Mean	SD	Significance
% of All Sequences that are Successful	High	**0.91**	.11	t(356) = 2.01, p < 0.05
	Low	**0.85**	.12	
Sequence Duration (secs)	High	**66.43**	27.20	t(356) = 1.71, p > 0.05
	Low	**72.35**	37.44	
# Operations	High	**4.03**	2.55	t(356) = 0.98, p > 0.05
	Low	**3.81**	1.44	

successful. We expected high filers to be more successful given their investment in preparing materials for retrieval ('*I know where that message is because I deliberately filed it*'). As Table 5 shows, contrary to our expectations, high filers were no more successful at finding messages than low filers. Again we checked whether high vs. low filers had a greater absolute success rate, but found no differences.

These analyses examine how *management strategy* affects efficiency and success. However to confirm our analyses we can also look directly at people's *access behaviors* (as opposed to their management strategy) to see how these behaviors affect both efficiency and success. Thus for *efficiency*, we would expect that people who were more reliant on folder-access behaviors would tend to have finding sequences that are longer in duration. To determine which access behaviors dominated for each user, we again calculated the frequency of each access behavior expressed as a proportion of all accesses. We then correlated this with the overall duration of their finding sequences. Consistent with the above results, a high reliance on folder-access was positively correlated with an increased sequence duration ($r(356)=0.33$, $p<0.001$). Scrolls, sorts and searches were all negatively correlated with sequence duration.

We next examined the relationship between retrieval behavior and *success*. Did a reliance on folder-access predict success, or was search a stronger predictor? We found that people who relied on search were more likely to have successful finding sequences ($r(356)=0.15$, $p<0.005$). None of the other behaviors was correlated with success.

Threads improve finding success

We also explored whether the intrinsic organization afforded by threads improved success and efficiency. Table 6 shows that people who had higher threading were more successful at finding messages. Threads did not seem to influence efficiency: there were no differences in either sequence duration or operations per sequence.

Explaining Management Strategy

Given that actively creating folders does not increase efficiency or success, why do some people use them? One possibility is that they are a task management strategy. Many people use the inbox as a 'todo' list, a function which is compromised by a high incoming message volume, causing them to folder. Foldering removes messages from the inbox, reduces its complexity, and allows users to see outstanding tasks at a glance. To explore whether foldering is used for task management, we compared incoming message volume to users' propensity to folder. Our measure of incoming volume was the daily change in inbox size, i.e., how many messages people kept each day. We correlated this with our standard measure of people's propensity to folder, i.e., proportion of the mailbox in folders. We found that people who kept more messages each day were more likely to folder ($r(356)=0.16$, $p< 0.01$). This suggests that foldering may be a reaction to incoming message volume.

To further understand this result, we asked our 32 interview participants about their email management and refinding practices. Though people used their folders to find messages, the predominant reason given for foldering in the first place was related to task management, as comments from four different folder users illustrate: "Generally everything sits in inbox until actioned… [I] attempt a daily run through to move inbox items to subfolders." "My inbox is a todo list." "I'm trying not to drown." "My inbox stays clean. It has things I need to respond to or do."

A related possibility is that assiduous filing is the result of a greater need to refind messages due to a user's job role, regardless of whether filing is less efficient. However our data do not support this hypothesis: there were no overall differences in the number of accesses conducted by filers and no filers ($t(356)=1.05$, $p > 0.05$). To confirm this we also directly explored whether job role affects strategy, but found few effects. Although managers get more messages ($t=3.77$, $p<0.001$), when compared with non-managers they show no overall differences in refinding behaviors (sorts ($t=1.49$), searches ($t=.14$), folder-access ($t=.80$), scrolls ($t=.08$)) or in their propensity to folder ($t=.25$) (all $df=356$).

DISCUSSION AND CONCLUSIONS

This is the first large-scale quantitative study of how people refind email. Our results show that opportunistic access behaviors dominate, mainly because of the prevalence of

scrolling. Nevertheless, folder-accesses account for 12% of overall accesses. Thus even with a client that supports effective search and threading, some users are still reliant on folder-accesses.

Prior work has argued that folders may be poorly organized and sometimes ill-suited for retrieval [23]. While people who manually organize more information into folders are more likely to rely on these for retrieval, high filers were no more successful at retrieval. Further, they were less efficient because folder-accesses took longer on average.

Why then, do users persist with manual foldering when it is known to be onerous to enact [3]? First, it was clear that foldering was not a response to increased demands for refinding emails; filers were no more likely to reaccess messages. Instead we found that filing seems to be a *reaction* to receiving many messages. Users receiving many messages were more likely to create folders, possibly because this serves to rationalize their inbox, allowing them to better see their 'todos'. Interview data confirms that people file to clean their inboxes to facilitate task management. This result contradicts prior work arguing that people who receive many messages do not have the time to create folders [1]. Further work should move beyond refinding to explore trade-offs between opportunistic and preparatory strategies for task management.

We also found that the *intrinsic* structure afforded by threads affected folder-access. People who received more threaded messages were less likely to rely on folders for access. Threads impose order on the mailbox, reducing the need for preparatory strategies. In part, this validates our design. Threading in Bluemail draws messages out of folders and into relevant inbox threads, making people less reliant on folders for access. Threads also serve to compress the inbox, reducing the amount that users need to scroll. As a result, people who received more threaded emails were more successful in their retrievals.

There are direct technical implications of our results. Search was both efficient and led to more successful retrieval, in part supporting the search-based approach of clients like Gmail. However in our study, other behaviors, especially scrolling, were prevalent. Gmail, which mainly supports search at the expense of scrolling, foldering, and sorting may be suboptimal. Even with a threaded client, scrolling was by far the most common access mechanism. However, scrolling is not well supported in Gmail, which breaks the mailbox into multiple pages, each of which has to be accessed and viewed separately. Gmail also does not support sorting, although this was a less frequent access behavior. Finally, folder-access was a preference for a minority of users, accounting for 12% of accesses (compared with 18% that were searches). Recent versions of Gmail attempt to *combine* folders and search [16]. However our data argue for the opposite: users employ either preparatory or opportunistic approaches, suggesting we need to design different features or mailbox views that optimize for each population tendency.

Another important design implication concerns *threading* which proved very useful. People who received more densely threaded emails created fewer folders and relied less on folder-accesses. They were also more successful at accessing emails. Threading can be improved, however. At the moment, threading imposes a very low level of organization [21]. The average thread length we observed was just 3.61 messages, and only 16% of messages participate in threads. This suggests there may be room for different intrinsic organizational tools that collate larger numbers of messages around a task.

How might we impose higher-level intrinsic organization on email? One possibility is to re-organize the inbox according to 'semantic topics'. One could use clustering techniques from machine learning to organize the inbox into 'superthreads' by combining multiple threads with overlapping topics, using techniques similar to [8].

In addition to 'superthreads', there may be other opportunities to exploit intrinsic organization to reduce the burden of manual organization. Several systems organize emails on the basis of *social* information, such as key contacts and social networks [14,22]. This approach has led to new commercial clients that include these features, such as Xobni and newer versions of Outlook. However, we currently lack systematic data about the utility of these new socially organized clients.

Finally, there are important empirical and theoretical links to other areas of PIM. Our findings that people resist using tags to manage emails are consistent with PIM studies showing people are unwilling to use tags for organizing personal files [6]. However recent studies also demonstrate that powerful new search features do not cause people to abandon manual navigation to desktop files [4] or web documents [18]. In contrast, we found a preference for newer automatic methods, such as search and threading, and that these were more effective than manual techniques. This may be because email data is more structured than personal files and webpages, leading to more effective searches. Another possibility is that the volume of email messages received is high compared with files created or web pages visited, making manual organization too onerous for email. Future work needs to explore this more.

In conclusion, we have presented a study that contributes a deeper understanding of email message refinding, a topic that has not yet been systematically studied. We have extended prior studies that focused on snapshots of email data. We have provided new data about the relations between management strategy, intrinsic threading structure, and actual access behaviors. We have also shown the value of threading and search tools. These data also offer direct design implications for current and future clients, including improved scrolling and threading. In future work, we will

explore new forms of intrinsic organization that our results here suggest.

REFERENCES

1. Bälter O. Keystroke Level Analysis of Email Message Organization. Proc. of CHI 2000, 105-112.
2. Bellotti, V., Ducheneaut, N., Howard, M., & Smith, I. Taking email to task: The design and evaluation of a task management centered email tool. Proc. of CHI 2003, 345-352.
3. Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., & Grinter, R. Quality vs. quantity: Email-centric task-management and its relationship with overload. Human-Computer Interaction, 20(1-2), 2005, 89-138.
4. Bergman, O., Beyth-Marom, R., Nachmias, R., Gradovitch, N., & Whittaker, S. Advanced search engines and navigation preference in personal information management. TOIS, 26(4), 2008, 1-24.
5. Boardman, R., Sasse, M.A. 'Stuff goes into the computer and doesn't come out': A cross-tool study of personal information management. Proc. of CHI 2004, 583-590.
6. Civan, A., Jones, W., Klasnja, P., Bruce, H. Better to organize personal information by folders or by tags?: The devil is in the details. Proc. of ASIS&T, 45(1) 2008, 1-13.
7. Dabbish, L.A., Kraut, R.E., Fussell, S., Kiesler, S. Understanding email use: Predicting action on a message. Proc. of CHI 2005, 691-700.
8. Dredze, M., Lau, T., Kushmerick, N. Automatically classifying emails into activities. Proc. of IUI 2006, 70-77.
9. Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., Robbins, D. Stuff I've seen: A system for personal information retrieval and re-use. Proc. of SIGIR 2003, 72-79.
10. Elsweiler, D., Baillie, M., Ruthven, I. Exploring memory in email refinding. TOIS, 26(4), 2008, 1-36.
11. Fisher, D., Brush, A.J., Gleave, E., Smith, M.A. Revisiting Whittaker & Sidner's 'email overload' ten years later. Proc. of CSCW 2006, 309-312.
12. Gmail, About Gmail. http://mail.google.com/mail/help/about.html, verified Jan. 11, 2011.
13. Mackay, W., Diversity in the use of electronic mail: A preliminary inquiry. TOIS, 6(4), 1988, 380-397.
14. Neustaedter, C., Brush, A.J., Smith, M., Fisher, D. The social network and relationship finder: Social sorting for email triage. CEAS, 2005.
15. Radicati Group, Email statistics report 2010-2014. http://www.radicati.com/?p=5282, verified Jan. 11, 2011.
16. Rodden K., Leggett, M. Best of both worlds: Improving Gmail labels with the affordances of folders. Extended Abstracts of CHI 2010, 4587-4596.
17. Tang, J.C., Wilcox, E., Cerruti, J.A., Badenes, H., Schoudt, J., Nusser, S. Tag-it, snag-it, or bag-it: Combining tags, threads, and folders in e-mail, Extended Abstracts of CHI 2008, 2179-2194.
18. Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.R. The perfect search engine is not enough: A study of orienteering behavior in directed search. Proc. of CHI 2004, 415-422.
19. Venolia, G., Neustaedter, C. Understanding sequence and reply relationships within email conversations: A mixed-model visualization. Proc. of CHI 2003, 361-368.
20. Wattenberg, M., Rohall, S.L., Gruen, D., Kerr, B. Email Research: Targeting the enterprise. Human-Computer Interaction, 20, 1 & 2, 2005, 139-162.
21. Whittaker, S., Bellotti, V., & Gwizdka, J. Everything through email. In W. Jones, J. Teevan (Eds.), Personal information management. UW Press, 2007, 167-189.
22. Whittaker, S., Jones, Q., Nardi, B., Creech, M., Terveen, L., Isaacs, E., Hainsworth, J. ContactMap: Organizing communication in a social desktop. TOCHI 11, 4, 2004, 445-471.
23. Whittaker, S., Sidner, C. Email overload: Exploring personal information management of email, Proc. of CHI 1996, 276-283.