

The Lifetime of Email Messages: A Large-Scale Analysis of Email Revisitation

Tarfah Alrashed*

Massachusetts Institute of Technology
Cambridge, MA, USA
tarfah@mit.edu

Ahmed Hassan Awadallah

Susan Dumais
Microsoft Research
Redmond, WA, USA

{hassanam,sdumais}@microsoft.com

ABSTRACT

Email continues to be one of the most important means of online communication, leading to a number of challenges related to information overload and email management. To better understand email management practices in detail, we examine the distribution of visits to emails over time. During their lifetime, emails may be visited one or more times, and with each visit different actions may be taken. Emails that are revisited over time are especially interesting because they represent an opportunity to improve email management and search. In this paper, we present a large-scale log analysis of email revisitation, the activities that people perform on revisited email messages (e.g. responding to, organizing or deleting messages, and opening attachments), and the strategies they use to go back to these emails. We find that most emails have a short lifetime, with more than 33% having a lifetime of less than 5 minutes. We also find that deleting is the most common action taken on messages visited once, and that responding and organizing are more common for messages visited more than once. We complement the log analysis with a survey to understand the motivation behind revisits and the types of emails that are revisited. The survey results show that 73% of the visits are to find information (e.g. a link or document, instructions to perform a task, or answers to questions), while 20% of revisits are to respond to the email. Our findings have implications for designing email clients and intelligent agents that support both short- and long-term revisitation patterns.

KEYWORDS

Email interaction, email revisitation, email lifetime

ACM Reference Format:

Tarfah Alrashed, Ahmed Hassan Awadallah, and Susan Dumais. 2018. The Lifetime of Email Messages: A Large-Scale Analysis of Email Revisitation. In *Proceedings of 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3176349.3176398>

*Research was conducted at Microsoft Research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions.acm.org.

CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-4925-3/18/03...\$15.00

<https://doi.org/10.1145/3176349.3176398>

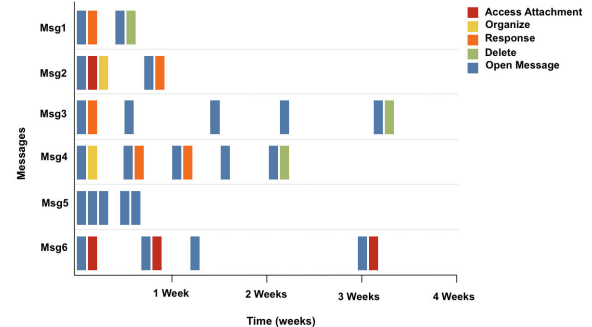


Figure 1: An illustration of activities on email messages over time.

1 INTRODUCTION

Email is one of the most familiar mediums of communication. It has evolved over time to also support task management and serve as a personal repository of information, leading to a number of challenges related to information overload and email management [11], in both work and personal settings. People use different strategies for managing personal and work accounts. They tend to create more structure in their work accounts, where they organize their emails into categories and folders, but they create fewer folders in personal account and rely to search to find emails [4].

Every message we receive in our mailbox goes through a journey. That journey starts with opening a message and ends with deleting, responding, archiving, or no further interaction. During its journey, the message undergoes a sequence of actions. This journey represents the lifetime of that email. Most emails have a short lifetime, never visited or visited once, and some have a longer lifetime, visited more than once (revisited). To better understand and characterize how and why people interact with emails over time, we investigate email revisitation behavior. Different users have distinct mailbox characteristics. For example, the distribution of the type of emails they receive may vary (e.g., instructions, information about an event, shared document or resource, etc.), and hence, their interactions with those emails vary. Users interact with their emails in different ways as well [7, 22]. Some users like to read all the email and reply to each immediately, others triage emails and return to important ones later. To go back to important emails, people use a variety of strategies - some users flag such emails or mark them as unread to facilitate quick visual scanning, others place them in folders or delete them, and some do nothing and use search to get back to important emails.

Figure 1 shows an illustration of email activities for different messages. Each row represents a unique email message and the x-axis shows a user's interaction with that message over time. Different colors represent different actions taken on the email (e.g., blue is "Open Message"), as discussed in more detail in section 3.3.2. The figure shows that different messages have distinct revisitation patterns. Some messages are revisited within short intervals and others remain relevant for a longer period of time. Some messages receive revisits in a small number of bursts, while others are revisited periodically. Some messages are revisited to retrieve information from the message body or attachments, other messages are revisited because the recipient needs to take an action against the message (e.g., reply, delete, etc.).

In this paper, we present a detailed study of the lifetime of an email through the lens of email revisiting behavior. We present a large-scale log analysis of one hundred thousand anonymized users of Outlook Web Access. We complement the log analysis with a survey to gain more insights into the motivation for the observed user behavior. Previous research has studied email and specifically email search [1, 10, 14, 17], however less work has focused on analyzing the journey that different emails go through by examining how and why people go back to emails. The research presented in this paper differs from previous work in several ways. We use large-scale log analysis to analyze interaction patterns in work email accounts. We examine the lifetime of emails by analyzing email revisitation patterns using two complementary methods: a large-scale log analysis, which captures the email activities and patterns, and a survey, which helps complement the log analysis. The research questions we address through our analysis are:

(1) *What are the characteristics of email revisitation?* We examine email revisitation using different dimensions: the number of visits, the inter-visit interval, the lifetime of revisited emails, and the distribution of actions on emails.

(2) *Why do users go back to emails?* We want to understand the type of emails that are most likely to be revisited, the intent behind these visits, and the actions performed on these emails.

(3) *How do users find emails they want to revisit?* In particular, we want to understand what strategies people use to re-find these emails (e.g., search or browse), and if this is influenced by the lifetime of emails.

2 RELATED WORK

Email is one of the most important means of online communication as well as a way to manage tasks and archive personal information. As the volume of email grows, challenges related to email management and retrieval increase [3]. In the research described in this paper, we examine how people interact with email messages over time, with a special focus on understanding how and why people revisit email messages. Two lines of prior work are especially relevant, one on email management and organization and the other on large-scale log analysis of email interaction.

2.1 Email Management and Organization

Most previous studies of email management strategies were based on small samples and used qualitative methods such as interviews, surveys, or diaries. In their pioneering work, Whittaker and Sidner

[22] investigated how people manage and organize their email. They identified three user types based on their strategies for managing email overload: no filers (search for emails), filers (put emails into folders), and spring cleaners (occasionally organize their emails). More than a decade after the original paper, Fisher et al. [7] and Grevet et al. [9] conducted qualitative studies based on Whittaker and Sidner strategies. They found that these strategies were still evident and that email overload was still an issue in both work and personal settings.

Increasingly people have multiple email accounts, and prior work has examined how people manage both work and personal email. Smith et al. [18] conducted a qualitative diary study with 16 participants, in which they found that more than half of their participants had two email accounts. A later study by Capra et al. [2] showed that 84% of participants had separate accounts to help manage boundaries between work and personal life. These two types of accounts had different characteristics: respondents reported more frequent use of keeping behaviors and larger mail boxes in their work accounts. Cecchinato et al. [4] used a diary study to investigate email management and search strategies. They found that people manage their personal and work accounts differently, and use different retrieval strategies. Work accounts were more structured, and email was generally retrieved through the folder structure, whereas people had fewer folders in personal accounts and relied more on search to find email.

Our research also examines email interaction behavior, but we focus on work email and use large-scale log analysis, as opposed to small studies. We also examine in greater detail the journey that individual emails go through by analyzing email revisitation behavior using data from both log analysis and a survey.

2.2 Large-Scale Log Analysis of Email Interaction

One of the earliest naturalistic log studies of re-finding in email was conducted by Elswiler et al. [6]. They examined a variety of email interactions such as selecting messages, opening folders, sorting or changing views and searching. Using a small amount of labeled data, they developed models to identify re-finding behavior from email logs and described how various interface features were used in re-finding. Whittaker et al. [21] carried out a larger-scale field study of 345 users using a web-based email client who conducted over 85,000 re-finding actions. They investigate different re-finding strategies, and found that although users who create complicated folders do use them for email retrieval, this approach did not improve retrieval success. On the other hand, both search and threading support more effective retrieval. Kalman and Ravid [13] conducted a study of email management strategies on thousands of users over a period of 8 months using a popular email web client add-in. They showed that people use a wide variety of strategies to manage their emails, many more than had been identified in earlier studies.

More recently, both Koren et al. [15] and Grbovic et al. [8] investigated email re-finding strategies using larger-scale log analyses of popular web email clients. They found that search is an increasingly important alternative to tagging emails and creating folders. Ai et al. [1] and Narang et al. [17] also used large-scale log analysis

to investigate interaction patterns in work email accounts. Ai et al. [1] examined the actions that people perform on emails after searches and compared re-finding in email search with web search. Narang et al. [17] also examined the activities performed on messages following searches, and how this related to the characteristics of people's mailboxes and email organization strategies. They found that people with larger mailboxes search more, and people who organize less tend to search more.

Re-finding previously seen information is a frequent activity that goes beyond email [5]. Re-finding has been studied in the context of Web search and browsing, and in desktop search. Teevan et al. [19] showed that about 80% of web-page visits are re-visits and 39% of all queries issued to a search engine are to re-find something seen before. Tyler and Teevan [20] provided key insights into the behavior people employ when re-finding. Dumais et al. [5] showed that email is the by far most common type of information that people re-find in a desktop search application, and that more than half of the items re-found using search are more than a month old.

Our research is similar to this line research in that we are also investigating user's email activities. We extend prior work by studying email revisitation, regardless of whether it is accomplished by searching or browsing (which we find to be much more common). We focus on investigating enterprise email behavior, since accessing email makes up a large proportion of users' information seeking efforts within enterprises [16]. We also look more broadly at the lifetime of email messages from the time they are first read to the time they are deleted, responded to, archived or abandoned. And, we use complementary techniques (large-scale log analysis and a survey) to provide a more complete picture of how and why people revisit emails.

3 METHODOLOGY

We studied email revisitation using two complementary techniques: a large-scale log analysis and a survey. The log analysis provides valuable information about general patterns in email usage. We used a sample of email action logs from a popular email client. We analyzed user behavior along different dimensions to understand email revisitation. We also developed a survey to better understand properties of email revisitation that are not possible to study using log analysis. The survey allowed us to examine the intent behind revisitation and the relation between email content and revisitation behavior. In this section, we describe these data sources, and the measures we considered through out our analysis.

3.1 Log Data

We analyzed a sample of the anonymized email logs from Outlook Web Access over a four months period from February 1, 2017 to May 31, 2017. The email web client can be used on both desktop and mobile with multiple browsers. The typical configuration has a folder list on the left and a search box is on the top left, the message list in the middle, and the message reading pane on the right.

Our sample included emails from enterprise users only (as opposed to consumer Web email users). The logs do not provide access to the text of the email message, headers or email search queries. The email log contains actions performed against messages with timestamps and other metadata. In order to capture revisitation

behavior in email logs, we only considered active users, specifically those who were active for more than 75% of the workdays within those four months. A user is considered active on a given day if she performs at least one action (e.g., reply, read) on that day. Our sample contained one hundred thousands active users who performed about 800 million actions during four months. The actions are described in section 3.3.2. Since our analysis focused on the lifetime of emails by analyzing email revisiting behavior, we focused on actions against revisited email messages.

3.2 Survey

We conducted a survey to investigate the intent behind email revisits, and types of information users were looking for in revisited emails. The survey was distributed to a random set of employees within Microsoft who were based in the USA. 395 respondents completed the entire survey, while 3 additional respondents provided partial responses (response rate: 13%, completion rate: 99%). In our analysis, we only consider the 395 who completed the survey in its entirety. 74% of the respondents were male, and were distributed across a wide age range ranging from under 20 to more than 60. Respondents came from a diverse set of roles within the company including: software development, program management, sales, marketing, legal, human resources, administrative assistance, IT support, finance, retail, etc.

The survey was structured into several sections, and like the log data, focused on work emails. In the first section, we asked our respondents about their general email behavior, such as the email client(s) that they use, and the number of emails that they typically receive during a work day.

In the second section, we asked our respondents to recall the last time they went back to an email they had seen before, and briefly describe what the email was about and why they went back to it. We also asked them whether they were successful in re-finding that email, and how long it took them to find it.

In the third section, we asked about characteristics of revisited emails. Specifically we asked respondents to estimate how long ago they received that email (e.g., in the last few hours, today, this week, this month, etc), and what the email was about (e.g., information about an event, instructions to perform a task, an email asking them to do something, etc).

In the fourth section, we asked respondents about their motivation in revisiting the email (e.g., did they go back to the email to find information, to respond, or to organize). We also asked them about the type of information they were looking for (e.g., an email that contained an attachment, instructions to perform a task, etc.).

In the fifth section, we asked respondents about the strategy they used to re-find the email they recalled (e.g., search, browse, or filter for flagged/unread emails). We also asked about attributes that used to find the message (e.g., sender name/email) and whether they had previously revisited the email. In addition, we asked whether they anticipated that they would need to go back to that mail when they first read it, and if so what technique they used to facilitate getting back to that email (e.g., flag/mark as unread, move to a folder, etc.). Lastly, we asked a few optional demographic questions.



Figure 2: A simple illustration of how we defined email visits and revisits. For a given user, each row represents a message that the user interacts with, and for every message, each set of consecutive actions represents a visit.

3.3 Measures

In order to understand the lifetime of emails, we analyze email revisitation behavior. We examine: (1) characteristics of emails that are revisited, (2) why users revisit emails by analyzing the distribution of email actions (from logs) and types of emails (from surveys), (3) how users get back to revisited emails (browse or search). We start by defining email visits/revisits, and then describing the actions we considered on email messages, the types of email messages, and email revisitation strategies. We describe each of these measures in more detail below.

3.3.1 Email visits and revisits.

Definition 1. An *email visit* is a sequence of consecutive actions performed by a user on an email, which are not interleaved with actions on other emails. An *email revisit* occurs when a user returns to the email after interacting with other messages.

Figure 2, which is a magnification of Figure 1, shows a simple illustration of this. Each row represents a unique message, and all messages belong to one user. The user “visits” Msg1, opens it and then responds to it. Then the user “visits” Msg2, opens it, accesses an attachment and then organizes it. The user then “revisits” Msg1 and opens, then deletes it. This is a revisit to Msg1 because there were intervening actions on another message, Msg2 in this case – that is, they went to another message and came back to (revisited) Msg1. Note that our definition of a visit requires an unbroken sequence of interactions on the same message. We realize that actions performed on one email could be interleaved with actions performed on another. We opted for treating these as multiple visits. One direction of future research could focus on organizing related visits into higher-level hierarchies similar to the work on segmenting search sessions into missions, tasks and goals [12].

The lifetime of an email starts with the user’s first interaction with a message (typically opening the message) and ends with their last interactions with the message. We divide users’ actions on messages into visits, and analyze the actions taken on messages within these visits, the interval between visits, the type of email messages, and the strategies used to revisit them.

3.3.2 Actions on email messages. In email, unlike web search, several actions of different types can be performed on a message (e.g., read, reply, move, etc.). In our analysis, we considered the ten most common actions present in the log data sample described in Section 3.1. Table 1 shows these ten actions grouped into higher level categories of actions – open, delete, response, and organize

Table 1: Groups of Message-Related User Actions

Group	Actions
Open Message	StartReadingDisplayPane
Delete	DeleteMessage
Response	Reply, ReplyAll, Forward
Organize	Flag, FlagComplete, MarkAsUnread
Access Attachment	OpenAttachment, EditAttachment

Table 2: Enterprise Email Classification Used for our Survey

Email Type	Description
Commitment	Email committing to do something
Communication	Informal communication with a colleague
Event	Information about an event
Instructions	Instructions to perform a task
Receipt	Receipt or confirmation (e.g., flight)
Request	Email asking someone to do something
Resources	Email sharing resources or documents
Status	Status update
Task	Email asking me to do something
Other	Other types of emails

message, and access attachment. Open message occurs when a message is clicked and its content is shown in the reading pane. Delete occurs when the user deletes an email message. Three actions are used to respond to email messages, and these actions are (Reply, ReplyAll, and Forward). Three actions are used to organize messages in different ways, these actions are (Flag, CompleteFlag, and MarkAsUnread). Delete can be considered as a type of organization activity, but we consider it separately because it is so common. Lastly, two actions are used to access email attachments (OpenAttachment and EditAttachment). We summarize the actions that people take on messages by computing the proportion of total email actions that each activity group accounts for. We further analyze the distribution of actions by the number of visits and the interval between visits.

3.3.3 Types of emails. To better understand email revisits, we study different revisitation patterns in light of the intent behind sending that email. To characterize the intent behind work emails, we classify work emails into several main types listed in Table 2. These types were derived from analyzing the free text responses to a survey question asking the respondents to describe the work emails that they needed to go back to. The types cover machine-to-human emails such as receipts, confirmations, events, etc. as well as human-to-human emails such as sharing a status update, assigning a task, requesting a resource, sharing instructions, etc.

This is not a complete taxonomy of work email types but it gives us some insights on how the content of the email could affect revisitation patterns. Analyzing the types of emails that are revisited enables us to answer questions like: are emails that contain requests, or attachments more likely to be revisited?

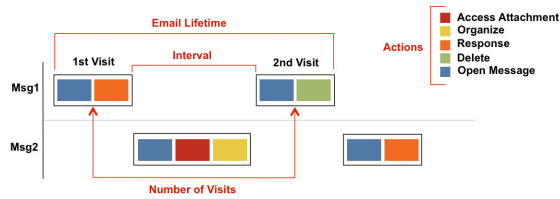


Figure 3: An illustration of the four dimensions we investigate when analyzing email revisits: number of visits, inter-visit interval, email lifetime, and actions taken on revisited emails.

3.3.4 Revisitation strategies. People use several strategies to go back to (revisit) emails that they have previously seen. Understanding the ways users revisit their emails, and how this might be affected by various email properties, is important in understanding email revisitation. We examine two revisitation strategies: searching and browsing. Searching occurs when a user explicitly types a query and accesses a message from the search results. Browsing occurs when a user browses messages in the inbox or navigates to a folder or a category to access the message. We also analyze revisitation strategies for emails with short and long lifetimes, and investigate the actions taken on messages that were found by searching or browsing.

4 RESULTS

In this section, we address our research questions. First, we characterize email revisits. Then, we address our second research question "why do users go back to their email?". Lastly, we address our third research question "how do users go back to their emails?".

4.1 Characterizing email revisits

In order to understand the lifetime of emails, we consider several dimensions that might influence general interaction and revisitation patterns. Figure 3 shows the four dimensions we investigate: number of visits, inter-visit interval, lifetime of revisited email, and actions taken on revisited email. Although we focus on revisited emails, we also examine the distribution of actions on emails that are visited only once and compare it with the distribution of actions on emails that are revisited.

4.1.1 Distribution of number of visits. As described in Section 3.3.1, we first divide users' interactions into visits and compute the number of visits each message receives. Most emails are visited only once, but about a third are visited two or more times. Figure 4 (top) shows the distribution of number of visits for messages that are revisited. The x-axis shows the number of visits and the y-axis shows the relative frequency of occurrence for each on a log-scale. The distribution of the number of visits for revisited emails is heavily skewed to the right. The average number of visits is 10.7 but the median is only 4. Of the revisited emails, 25% are visited only twice and 64% are visited five times or less. The proportion of messages drops quickly as the number of visits they received increases; leaving only a small percentage of messages receiving large number of visits. Now we move on to study the time between subsequent visits on the same message.

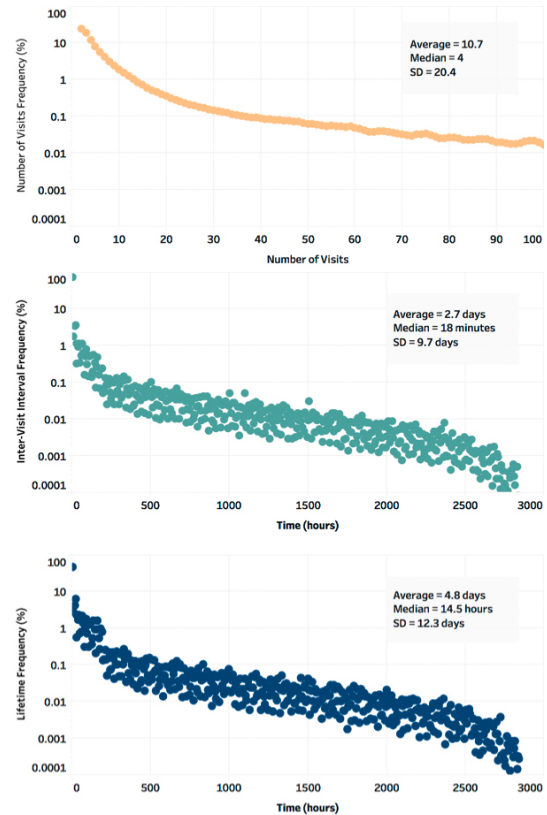


Figure 4: Distribution of number of visits (top), inter-visit interval for revisited emails (middle), and lifetime of revisited emails (bottom).

4.1.2 Distribution of inter-visit intervals. The distribution of time intervals between visits for revisited emails is shown in Figure 4 (middle). The x-axis shows the inter-visit intervals in hours and the y-axis shows the relative frequency of occurrence for each time interval on a log-scale. Again the distribution is heavy tailed – most revisits happen within a very short time from the previous visit (short-term revisits), while a much smaller number of revisits happen several months after the previous visit (long-term revisits). The average inter-visit interval is about 64 hours (2.7 days) and the median is 18 minutes. Further examination of the log data revealed that short-term revisits are more frequent and usually happen when the users are scanning their emails moving back and forth in the message list or when they are interleaving visits to different messages (e.g., read an incoming email and then go back to the previous email to respond or take other actions). The difference between short and long-term revisits has implications for designing methods to support revisiting behavior. Long-term revisits may be well supported by search or organizing into folders or marking; short-term revisitation may be better supported with quick access to recently viewed messages.

4.1.3 Distribution of email lifetimes. As described above, we define the lifetime of an email as the time between a user's first and last interactions with the email. Figure 4 (bottom) shows the

Table 3: Ratios of Actions for Emails Visited Once and Revisited

Action Group	Likelihood Ratios (Revisited/Visited once)
Open Message	3.1
Access Attachment	2.7
Response	1.1
Organize	1.1
Delete	0.4

distribution of lifetimes for revisited emails. Again, we see a heavy-tailed distribution. The average email lifetime is about 115 hours (4.8 days) and the median is 14.5 hours. In our analysis, we find that most emails have a short lifetime - 36% of emails have a lifetime of 5 minutes or less, 43% of emails have a lifetime of 1 hour or less, and 63% of emails have a lifetime of 5 days or less. This temporal distribution could be used as a feature to email ranking functions.

4.1.4 Distribution of actions. To analyze the distribution of actions, we computed the proportion of total email actions that each action group accounts for across all other actions in the dataset. Although we focus on revisited emails, we start by comparing the distribution of actions for emails that are visited once and those that are revisited.

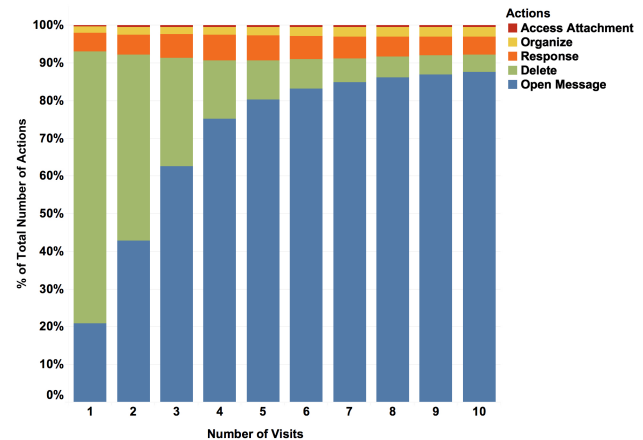
Table 3 shows the ratio of different actions for emails that are revisited compared to those that are only visited once. All actions except delete occur more in revisited emails than in emails visited once. Messages visited only once are 2.5 times more likely to be deleted (ratio $2.5 = 1/0.4$). This is expected, since when a message is deleted it is less likely to be visited again. On the other hand, messages that are visited more than once are 3.1 times more likely to be opened and 2.7 times more likely to be accessed for attachments compared with messages visited only once. Revisited messages are also somewhat more likely to be responded to or organized. If users know that they will go back to the emails when they receive them, then they are not likely to delete those emails.

4.2 Why do users revisit their emails?

To answer this question, we analyze the users' intent behind their email revisits and what actions they perform after they revisit the email. Log data provides insights about the actions performed on emails after they are revisited and survey data give us insights about the intent behind email revisits and the types of emails that are revisited. For the log data, we further partition the result by the dimensions described in the previous section – number of visits, inter-visit intervals, lifetime of emails, and type of email.

4.2.1 Actions on revisited emails / Intent. In this section, we examine both actions taken on revisited emails and the intent behind email revisitation. First, we analyze the actions taken on revisited emails broken down by number of visits, and the inter-visit interval. Finally, we investigate revisit intent using our survey data.

Actions vs. Number of Visits. To better understand why users revisit their emails, we analyze actions taken on revisited emails (e.g., did the user go back to a specific email to reply or to open an attachment?). Figure 5 shows the distribution of actions over

**Figure 5: Distribution of actions over number of visits.**

number of email visits (for number of visits between 1 and 10). Note that this figure shows the distribution of actions for email that were visited once and revisited emails (2 or more visits). The most common action for emails that are visited once or twice is delete, and the most common action for emails visited more than twice is open. The proportion of delete action decreases with the increase of the number of visits. On the other hand, the proportion of open message actions increases with the number of visits, which we expect to see since users need to open their emails every time they visit them. It is interesting to see that response actions increase with number of visits until the fourth visit, then decrease after that. We also observe a small increase in the proportions of both organizing a message and accessing attachments from emails as the number of visits increases. It appears that responding to an email is most likely to happen within the first four visits to that email, and after that organizing or access to attachments are more common.

Actions vs. Inter-visit Intervals. As we discussed earlier, different emails exhibit different revisitation patterns, some of the revisits happen within a short amount of time, and others occur over a longer period of time. Different actions might be taken for emails revisited within a short interval versus longer ones. We analyze the distribution of actions over the interval between visits. Figure 6 shows the distribution of actions over inter-visit interval (in days) for the first seven days. For emails that are revisited within one day delete and open message are both common. For emails revisited after more than one day, the most common action by far is open. We also find a slight increase in the response percentage with an increase of the inter-visit interval. Open message and access attachment decrease a bit with emails that are revisited more than four days apart.

In Figure 7, we zoom into the action distribution of actions over inter-visit interval (in hours) within six hours. Of the deletes that happened within one day (shown in Figure 6), more than half happened within an hour. We also notice that if emails are revisited after more than an hour users are less likely to delete these emails (10-15%). The other actions (response, organize, and access attachment) increase slightly as the inter-visit interval increases. Observe the overall similarity of patterns seen in Figure 6 and 7. They both

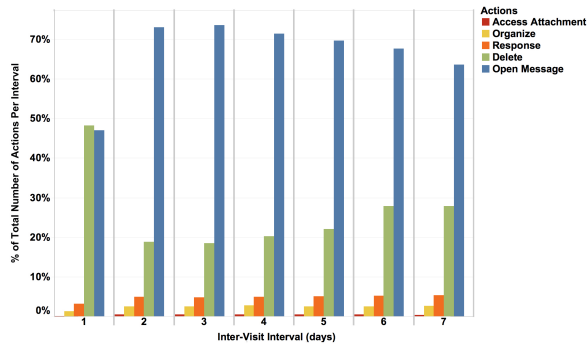


Figure 6: Distribution of actions over inter-visit interval in days (messages revisited within 7 days).

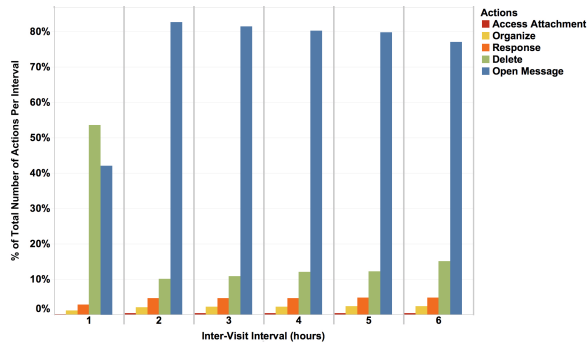


Figure 7: Distribution of actions over inter-visit interval in hours (for messages revisited within 6 hours).

show different proportions in the first day (Figure 6) or the first hour (Figure 7) compared to intervals above one day or one hour. That is because; the vast majority of deletes happen after very short revisits, which affects the proportions of the other actions.

Revisit Intent. In the previous section, we studied actions that are taken on a message during a revisit. In this section, we study the intents that lead a user to revisit a message. In the survey, we asked respondents why they went back to the email they recalled revisiting - was it to find some information, to do something with the email (e.g., reply, forward), to organize or clean up the email (e.g., flag, delete), or something else. Table 4 shows the distribution of the email revisit intents. Most of the emails (73.4%) were revisited to find some information, 20.6% to respond to the email, only 0.6% went back to the emails to delete it or organize it, and 5.4% revisited emails to do something else. For respondents who went back to their emails to find some information (73.4%), we further investigated the type of information they were looking for. Giving our respondents a list of types, as shown in Table 5, we found that 24.1% of the email revisits were to find instructions, 22% were to look for a document (e.g., attachment, link), 16.3% were to find an answer to a question, 10.2% were status updates, 9% were to find a solution to a problem, and other intents occurred less than 5% of the time.

4.2.2 Types of revisited emails. We are interested in understanding the influence that the type of email has on revisitation patterns.

Table 4: Distribution of Email Revisit Intents

Revisit Intent	Percent
Find information in the email	73.4%
Do something with the email (e.g., reply, forward)	20.6%
Organize or clean up their email (e.g., flag, delete)	0.6%
Other	5.4%

Table 5: Distribution of Types of Information Users were Looking for in Emails that were Revisited

Type of Information	Percent
Instructions to perform a certain task	24.1%
A document (e.g., attachment, link)	22.0%
An answer to a question that was previously asked	16.3%
status update	10.2%
A solution to a problem	9.0%
A task request to you	4.9%
A person/customer (e.g., contact information)	2.0%
An appointment/event	2.0%
Machine generated message (e.g., reservation)	0.8%
Other	8.6%

For example, do users go back to emails that contain instructions more often than emails about an event? In the survey, we used a general classification of work email types listed in Table 2, and asked respondents what the email that they went back to was about. Table 6 shows the distribution of the types of revisited email.

38.2% of revisited emails were about sharing resources or documents, 17.3% were emails containing instructions on how to perform a task, 11.9% were emails that asked users to do something, 9.9% were status updates, 5.1% were emails that asked someone else to do something, 3.6% contained information about an event, 2.4% were committing to do something, 1.8% were communicating with a colleague, 0.9% of them were containing receipt or a confirmation, and 7.5% did not fit into our classification. From the previous section we found that 73.4% of users go back to emails to find information, which explains why the two common types of emails are those sharing resources or documents (38.2%), and containing instructions on how to perform a task (17.3%).

4.3 How do users revisit their emails?

To better understand email revisitation behavior, we analyzed the strategies people used to revisit emails, and how effective those strategies were. Before we describe strategies for revisiting email, we briefly summarize the success and effort required to do so. 95.5% of our respondents said they found the email successfully, and most found it quickly (48.4% found it in less than a minute, 41.3% in less than 5 minutes, and 9% needed more than 5 minutes). Respondents indicated that they had previously revisited 62.9% of these messages.

The vast majority of respondents said they used search to revisit the email - 71.6% said they used search to go back to the email, 20.9% said they browsed through their emails (11% browsed in their inbox and 9.9% browsed in a folder/category), only 1.5% filtered for flagged

Table 6: The Distribution of Email Types from Table 2

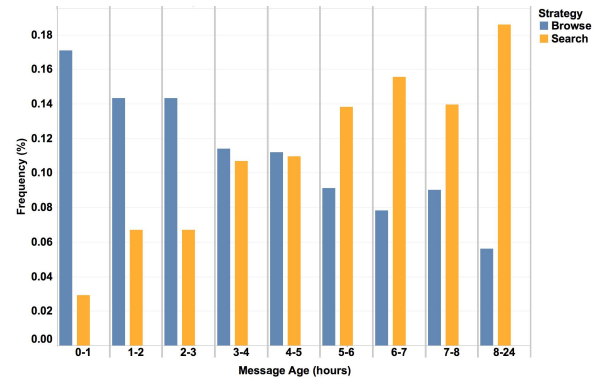
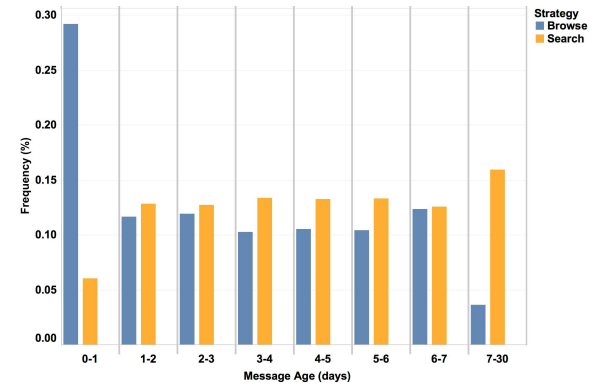
Email Type	Percent
Resources	38.2%
Instructions	17.3%
Task	11.9%
Status	9.9%
Request	5.1%
Event	3.6%
Commitment	2.4%
Communication	1.8%
Receipt	0.9%
Other	7.5%

or unread emails, and 6% used other strategies. For respondents who used search we also asked about the type of queries they issued - 75.3% said they used keywords that were in the message body or title, 66.1% of them used sender name or email, 14.2% used the recipient name or email, and 7.9% use date/time range that they believe the message was sent in.

Finally, we wanted to understand whether users anticipated that they would need to go back to emails when they first read them. Our survey showed that 62.3% of users anticipated they would revisit these emails. Although the majority of respondents knew that they would revisit these emails in the future, 38.5% of them did not do anything to help facilitate getting back to these emails, 27.4% moved them to folder, and 19.2% flagged/marked them as unread. Note, however, that deleting other messages is another way to make it easier to re-find content by reducing the size of the inbox.

4.3.1 Revisit Strategy vs. Message Lifetime. In our survey, we asked respondents how long ago they received the email that they revisited. 29% said they received it more than a month ago, 27.5% within the same month, 23% within the same week, 6% the day before, 6% the same day, 2.4% the last few hours, 1.5% the last few minutes, and 4.5% could not recall when they had first seen it. It is interesting to note that 56.5% of the revisited emails were received more than a week ago, and the main strategy our respondents said they used was search (71.6%). It appears that when we asked respondents to think of an email that they had revisited, they tended to think about older emails that they used search to get back to. They did not think of the many revisits that occur over short intervals (that are evident in the log data) as "revisits".

Using the log data, we examined the distribution of message lifetime (in hours), for messages received within 24 hours, for searching and browsing; the results are shown in Figure 8. We find that users are more likely to browse when the message was received within three to four hours, and more likely to search for messages received more than 4 hours ago. We also analyze the distribution of revisiting strategies over the lifetime of revisited emails over days. Observe that a large portion of browsing happens for messages received within the first day as shown in Figure 9. The portion of the messages revisited through browsing drops significantly for messages older than one day and remains constant for messages up to one week old, and drops again for messages older than one week.

**Figure 8: Distribution of message lifetime in hours for both revisit strategies (for messages received within 24 hours).****Figure 9: Distribution of message lifetime in days for both revisit strategies (for messages received within 30 days).**

4.3.2 Revisit Strategy vs. Action taken on Message. In this section, we examine the differences in the actions that people take on messages depending on whether the message was revisited by searching or browsing. The top actions taken on messages after browsing are in order: open message, delete, response, organize, access attachment. After searching, the top actions taken are: open message, response, delete, organize, and access attachment. For both strategies, open message is the most frequent action, and organize and access attachment are the least frequent actions. An important difference is that delete is proportionally more frequent after browsing while response is more frequent after searching.

Table 7 shows the likelihood ratios for different actions after browsing or searching. Organizing and deleting are more likely after browsing (ratio: 2.9 and 2.3 respectively). On the other hand, responding is significantly more likely after a search (ratio:0.5), and reading a message is slightly more likely after a search (ratio:0.8). Managing content is roughly the same regardless of revisit strategy (ratio:1.0). Accessing attachments occurs roughly the same proportion of the time regardless of the strategy used to revisit the message (only 3% more likely after browsing). This suggests that people browse through their inbox to triage mail, deleting some and filing others. On the other hand, when they search for a message

Table 7: Likelihood Ratios Of Actions After Browsing or Searching

Action	Likelihood ratio (Browse/Search)
Organize	2.9
Delete	2.3
Access Attachment	1.0
Open Message	0.8
Response	0.5

they have seen before, they are more likely to read or respond the messages.

5 DISCUSSION

In this paper, we examined the journey that emails go through from the time they are first read to the time they are last viewed using two complementary methods. We analyzed 800 million email actions from a sample of 100 thousand people, and conducted a survey of almost 400 people to provide insights about search intents, success and strategies. We *characterized* email interaction patterns for work emails with a focus on revisitation (regardless of whether emails are revisited via search or browsing).

The analysis presented in this paper is just a first step toward understanding the lifetime of email messages and revisitation behavior. There are several directions of future work to extend the work presented in this paper both in terms of developing a deeper understanding of the lifetime of different messages by further investigating techniques that people use to facilitate revisiting emails (e.g., flagging or tagging messages, marking messages as unread, moving messages to folders, etc.) and designing email clients and intelligent assistants that help users with going back to messages. In this section, we highlight some of the key findings and discuss their implications as well as directions for future work and conclude with a discussion of the limitations of this work.

Revisits and Lifetime: Most emails are visited only once. For these emails, the most common action is delete, accounting for 72% of actions. Most emails that are visited once are deleted without having been opened. Delete is 2.5 times more likely in emails that are only visited once. This suggests that, on average, people quickly triage their email by deleting emails. On the other hand, for emails that are revisited, the most common action is open/read message. Opening messages is 3 times more likely in revisited emails and opening attachments is 2.7 times more likely.

Although most emails receive only one visit, more than a third of emails are revisited at least once. The median number of visits is 4, while the average is 10.7. This reflects that some emails receive a much larger number of visits (see Figure 4). The proportion of emails are revisited and the steady increase in the the volume of messages make supporting people in getting back to emails increasingly important.

We also found out that the lifetime of most emails is very short. For example, 63% of emails had a lifetime of 5 days or less and 36% of emails had a lifetime of 5 minutes or less. The distribution of the email lifetimes was also skewed with an average email lifetime of 4.8 days and median of 14.5 hours. This temporal distribution could

be leveraged to provide a better ranking functions for email search or caching of most recently visited emails.

The emails that have longer lifetime exhibit interesting revisitation patterns. Figure 1 illustrates some of these patterns, where we see some messages receiving visits in a small number of bursts while other are revisited periodically. Additionally, some emails are revisited over short periods of time, while others are revisited over longer periods in regular intervals or irregular intervals. Deeper understanding of these patterns and how they can be leveraged to support revisiting and re-finding messages is an interestingness direction for future work. This temporal distribution (in combination with previous interaction patterns on emails) could be used as a signal to email ranking functions.

Types of Revisitation: Our analysis uncovered two distinct types of email revisitation: short- and long-term revisits. Most revisits occur shortly after an email is first read – this is salient in Figures 6 and 7. Other revisits are longer term and have longer inter-visit intervals. This distinction between types of revisitations is also evident in the actions that people take on the emails. Very short-term revisits often result in a message being deleted. This is similar to messages that receive a single visit where users quickly triage their messages by deleting some of them. Revisits that happen after more than an hour are less likely to result in a message being deleted. For revisits that occur after an interval of one day or more), the most frequent action, by far, is opening the message. The other actions taken on emails (response, organize, and access attachment) increase slightly as the inter-visit interval increases.

Short-term and long-term revisits also impact the strategy users employ to get back to the message. This is evident in Figures 8 and 9 which show that users are more likely to browse when the message was received within three to four hours, and more likely to search for messages received more than 4 hours ago. The difference between short and long-term revisits has implications for designing methods to support revisiting behavior. Short-term revisitation may be supported with UX elements that enable quick access to recently viewed messages. On the other hand, long-term revisiting might be supported by improving search ranking techniques to assist with retrieving previously seen messages.

Revisit Intent: From the survey, we find that most longer-term revisits are to find information (74%) or to take some action on the email (20%). Such information varied from a status update, information about an event, instructions to perform a task, etc. Our analysis also showed that although 62.3% of users knew that they would go back to emails when they first read them, 38.5% of users did nothing to help them re-find the items and relied on search instead. This suggests that there is an opportunity to provide better support for going back to previously seen messages. For example, we can build a model to predict which emails are most likely to be revisited and the actions that will be taken on these emails. Predicting such behavior can provide new means to support getting back to relevant emails.

Limitations: Finally, we should discuss some limitations of the work presented. We used two complementary techniques for this research: a large-scale log analysis and a survey. Log-based studies allow us to characterize how people interact with email, but provide limited insights into the intentions of the users as they interact with emails. As a result, our insights related to revisit intents were limited

to the survey data. Our log data sample comprises logs recorded for a hundred thousand active users over a 4-month period of time. The logs could be segmented along different dimensions (e.g., demographics, industry type, etc.) to better understand revisitation behavior varies depending on such segmentation. We leave the study of the impact of these segmentations on the user behavior to future work.

On the other hand, the tremendous scale of log-data allowed us to study the behavior of a large number of users from different industries and backgrounds, this is typically harder to achieve with survey-based or observational analysis. Our survey targeted a smaller number of subjects (relative to the log-data size) from one organization. To partially address this, we included participants from various job roles (e.g., sales, legal, HR, software development, etc.) in our survey. Additionally, our survey design required participants to recall the last time they revisited an email and answer questions about this particular event. While recollecting specific events is more accurate than asking about general patterns, the retrospective methodology can still run the risk of inaccurate or incomplete recollections. Since we asked respondents to recall the last occurrence of a frequent event, the vast majority of them (85.3%) reported no difficulty in recalling the last message they needed to go back to. However, they seemed to interpret "revisit" as referring to older emails (~80% of them reported receiving the email more than 24 hours ago). They did not think of the many short-term revisits that are evident in the log data as revisits. Despite the limitations, the log-based and survey-based studies provided complimentary insights on user behavior but additional research (e.g., involving an *in situ* survey) will be required to better align the findings from the two studies.

6 CONCLUSION

We sought to understand the lifetime of email messages by investigating email revisitation behavior, and understanding why people want to get back to messages and how they do so. We found that while most messages have a short lifetime, many messages remain relevant for a longer time period. We also studied the intents that lead people to go back to messages that they have seen before, and found that finding information and responding to email are the most common reasons. We also found that the strategy used to revisit (search or browse) depended on the age of the email and the intent behind the revisit. Understanding the lifetime of an email message could have implications on understanding how people interact with their email and designing email clients and intelligent agents to help people with managing and organizing their messages. Our future work will aim to develop deeper understanding of the lifetime of an email and explore applications that would better support email revisits.

REFERENCES

- [1] Q. Ai, S. Dumais, N. Craswell, and D. Liebling. 2017. Characterizing email search using large-scale behavioral logs and surveys. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. Perth, Australia, 1511–1520. <https://doi.org/10.1145/3038912.3052615>
- [2] R. Capra, J. Khanova, and S. Ramdeen. 2013. Work and personal e-mail use by university employees: PIM practices across domain boundaries. *Journal of the Association for Information Science and Technology* 64, 5, 1029–1044. <https://doi.org/10.1002/asi.22815>
- [3] D. Castro, Z. Karnin, L. Lewin-Eytan, and Y. Maarek. 2016. You've got mail, and Here is what you could do with it!: Analyzing and predicting actions on email messages. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. San Francisco, CA, USA, 307–316. <https://doi.org/10.1145/2835776.2835811>
- [4] M.E. Cecchinato, A. Sellen, M. Shokouhi, and G. Smyth. 2016. Finding email in a multi-account, multi-device world. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. 1200–1210. <https://doi.org/10.1145/2858036.2858473>
- [5] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. 2016. Stuff I've Seen: A system for personal information retrieval and re-use. In *ACM SIGIR Forum*, Vol. 49. 28–35. <https://doi.org/10.1145/2888422.2888425>
- [6] D. Elswiler, M. Harvey, and M. Hacker. 2011. Understanding re-finding behavior in naturalistic email interaction logs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 35–44. <https://doi.org/10.1145/2009916.2009925>
- [7] D. Fisher, A. J. Brush, E. Gleave, and M. Smith. 2006. Revisiting Whittaker & Sidner's "email overload" ten years later. In *Proceedings of the 2006 20th anniversary conference on Computer Supported Cooperative Work (CSCW '06)*. Banff, Alberta, Canada, 309–312. <https://doi.org/10.1145/1180875.1180922>
- [8] M. Grbovic, G. Halawi, Z. Karnin, and Y. Maarek. 2014. How many folders do you really need?: Classifying email into a handful of categories. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. Shanghai, China, 869–878. <https://doi.org/10.1145/2661829.2662018>
- [9] C. Greve, D. Choi, D. Kumar, and E. Gilbert. 2014. Overload is overloaded: Email in the age of Gmail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Toronto, Ontario, Canada, 793–802. <https://doi.org/10.1145/2556288.2557013>
- [10] M. Harvey and D. Elswiler. 2012. Exploring query patterns in email search. *Advances in information retrieval* (2012), 25–36.
- [11] A. Jerejian, C. Reid, and C. Rees. 2013. The contribution of email volume, email management strategies and propensity to worry in predicting email stress among academics. In *Computers in Human Behavior*, Vol. 29. Elsevier, 991–996. <https://doi.org/10.1145/2661829.2662018>
- [12] R. Jones and K.L. Klinkner. 2008. Beyond the Session Timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. 699–708. <https://doi.org/10.1145/1458082.1458176>
- [13] Y.M. Kalman and G. Ravid. 2015. Filing, piling, and everything in between: The dynamics of E-mail inbox management. In *Journal of the Association for Information Science and Technology*. Toronto, Ontario, Canada, 2540–2552. <https://doi.org/10.1002/asi.23337>
- [14] J.Y. Kim, N. Craswell, S. Dumais, F. Radlinski, and F. Liu. 2017. Understanding and modeling success in email search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Perth, Australia, 265–274. <https://doi.org/10.1145/3077136.3080837>
- [15] Y. Koren, E. Liberty, Y. Maarek, and R. Sandler. 2011. Automatically tagging email by leveraging other users' folders. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11)*. San Diego, CA, USA, 913–921. <https://doi.org/10.1145/2020408.2020560>
- [16] U. Kruschwitz and H. Charlie. 2017. Searching the Enterprise. *Foundations and Trends in Information Retrieval* 11, 1 (2017), 1–142.
- [17] K. Narang, S. Dumais, and Q. Ai. 2017. Large-scale analysis of email search and organizational strategies. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. 215–223. <https://doi.org/10.1145/3020165.3020175>
- [18] H. Smith, Y. Rogers, and M. Underwood. 2003. Managing personal and work email in the same box: overcoming the tensions through new metaphors. *Proceedings of the Home Oriented Informatics and Telematics (HOIT '03)* (2003).
- [19] J. Teevan, E. Adar, R. Jones, and M.A.S. Potts. 2007. Information re-retrieval: repeat queries in Yahoo's logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. 151–158. <https://doi.org/10.1145/1277741.1277770>
- [20] S.K. Tyler and J. Teevan. 2010. Large scale query log analysis of re-finding. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. 191–200. <https://doi.org/10.1145/1718487.1718512>
- [21] S. Whittaker, T. Matthews, J. Cerruti, H. Badenes, and J. Tang. 2011. Am I wasting my time organizing email?: A study of email refinding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Vancouver, BC, Canada, 3449–3458. <https://doi.org/10.1145/1978942.1979457>
- [22] S. Whittaker and C. Sidner. 1996. Email overload: Exploring personal information management of email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '96)*. Vancouver, British Columbia, Canada, 276–283. <https://doi.org/10.1145/238386.238530>