

On Understanding the Relationship Between Recollection and Refinding

DAVID ELSWEILER¹, MARK BAILLIE² and IAN RUTHVEN²

¹david.elsweiler@i8.informatik.uni-erlangen.de

Department of Computer Science 8 (AI), Univeristy of Erlangen-Nuremberg,
Haberstrasse 2, 91058 Erlangen

²{ian.ruthven,mark.baillie}@cis.strath.ac.uk

Department of Computer and Information Sciences, University of Strathclyde
Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH

Abstract

Memory has long since been acknowledged to be important to the processes involved in Personal Information Management, especially to re-finding previously accessed information. Nevertheless, relatively little is known about the role that memory plays in these processes or how the user's recollections should be supported. Focusing, on email re-finding, this article investigates the relationship between recollection and information re-finding performance. A study is presented that examines the attributes participants remembered about email messages they were asked to re-find. The recollections are analysed statistically to learn if they influenced the participants' re-finding performance. We discover that a relationship exists, although it is more complicated than researchers have previously suggested, and that specific attributes appear to influence the performance when they are remembered. We discuss our findings with respect to past and future work and also to the design of new re-finding tools.

1 Introduction

Personal Information Management (PIM) refers to the activities, by which an individual manages his personal information. These include finding new information, organising any information found into a collection, re-finding information that has been previously seen or possessed, and the activities relating to the maintenance of a personal information collection (Barreau, 1995). These activities are largely psychological in nature and involve psychological processes, such as categorisation, recognition and recollection (Lansdale, 1988). Information re-finding – perhaps the most important PIM activity, as it is the end goal of all PIM behaviour (Jones and Teevan, 2007) – principally involves recollection. Recollection is crucial to information re-finding, not only because people re-find based on what they remember (Capra and Perez-Quinones, 2005), but because it is problems with memory or a failure to recollect that represents the main challenge to successfully re-finding information when it is needed (Elsweiler et al., 2007).

The importance of memory in the activity of information re-finding has meant that, in recent years, this has become a popular research focus. Several PIM systems have been designed based on principles of memory e.g. (Dourish et al., 2000; Ringel et al., 2003; Cutrell et al., 2006), and a number of investigations have attempted to learn more about what people remember about their information and the contexts in which information was created, discovered or used e.g. (Gonçalves and Jorge, 2004; Elswailer et al., 2005; Blanc-Brude and Scapin, 2007).

Despite these efforts, we lack a clear understanding of the role that memory plays in PIM and know little about what people tend to remember about their information, how they use those recollections when re-finding, or how important the quality of recollections are with respect to how people are able to re-find. Understanding the role of memory is vital in order to better understand PIM behaviour, to learn more effective methods of managing information and to allow the creation of better tools to assist the user in these tasks. In this article, we aim to add to our understanding by investigating the final issue of the three listed above.

We describe a email system evaluation designed to provide insight into how to design PIM tools so that they better support the characteristics of the user’s memory. We recorded data about what the participants remembered about emails they were asked to re-find and determined the features of interfaces which supported those recollections. We have reported some of the study’s findings in a previous article (Elswailer et al., 2008), which examined what the participants remembered and the factors that influenced their recollections. In this article, however, we focus on another aspect of the findings. Here, we examine the relationship between the recollected attributes and the user’s re-finding performance.

The main outcome of the analyses described in this article is the discovery that the relationship between recollection and re-finding is more complex than previous research has suggested. Although we found that a relationship does exist, our data show that remembering more does not necessarily equate with better re-finding performance. We also examine and identify the effect that remembering particular features of email messages has on the re-finding performance. Our findings feed into existing knowledge of how people manage their emails, what they tend to remember and the difficulties they have. We discuss what our findings mean in the context of designing more effective management and re-finding tools and outline future research directions.

The remainder of the article is structured as follows: in Section 2, we provide a summary of the relevant background literature and motivate our research aims; in Section 3, we outline our methodology and study design; Section 4 presents our data analyses and details our main findings; Section 5 addresses the limitations of our study; Section 6 provides a discussion of what our findings mean in the context of previous and future research. Finally, in Section 7, we present our conclusions and explain our future research directions.

2 Related Work

This section describes the background literature for the primary themes of this article. Section 2.1 explains our decision to study memory in the context of email re-finding. It introduces the problem of email management and summarises previous work in this area; Section 2.2 presents work that has related memory and PIM, first by summarising studies of PIM behaviour where the findings provide insight to the role that memory plays in PIM activities and second, by reviewing the numerous PIM systems that have been designed to support particular attributes of memory. The described background literature motivates both our study and our approach to investigating the problem.

2.1 Email Management

Email has been subject to a large amount of research attention. It has been shown, for example, that its uses as a practical tool go far beyond its original intended purpose as a means of asynchronous communication. People have been found to use email for collaborative working (Ducheneaut and Bellotti, 2001), data archiving (Mackay, 1988; Whittaker and Sidner, 1996; Bälter, 2000), as well as for managing tasks (Whittaker and Sidner, 1996) and contacts (Whittaker et al., 2002). This means that people tend to keep their messages, rather than deleting them as they are received and read and, as a result, many users have email collections consisting of thousands of messages (Whittaker and Sidner, 1996).

One popular line of email research has been to investigate how people organise their messages for these different purposes. Mackay (1988), for example, distinguished between users (she called them “*prioritizers*”) who prioritise particular messages that they believe require the most attention and users (“*archivers*”) who are mainly concerned with keeping information in case of future need. When Whittaker and Sidner (1996) looked at this issue, they identified three user strategies based on the use of folders and the how often users put effort into maintaining their collection. Some users (*no-filers*) make no use of folders and instead leave all of their messages in the inbox; other users (*frequent filers*) use folders as a means of placing an organisation on the collection and make efforts to regularly sort messages in their inbox into appropriate folders; and a final group of users (*spring-cleaners*) also make use of folders, but only clean up their inbox periodically. The evidence suggests that a tension exists between the way people organise their messages for the different email activities they perform. For example, when using email for task management purposes, a common strategy is to leave an email message in the inbox to act as a reminder to perform a particular task (Mackay, 1988; Whittaker and Sidner, 1996). This strategy may be less successful for “no-filers” who will have many messages in the inbox, or may hinder the re-finding attempts of “filers” who normally files messages into folders.

The combination of all of these factors – large inboxes with thousands of messages, multiple uses of email and multiple and conflicting filing strategies – places a huge burden on the user’s memory when re-finding. This makes email a particularly important media to study with respect to our research aims. Email overload is also a problem that affects the productivity of information workers and as a result costs industry billions of dollars¹.

Another reason for our decision to study email was that this medium has been suggested as means to unify different PIM collections, such as visited webpages, stored computer files, personal contacts etc. in order to reduce information fragmentation (Whittaker et al., 2007). Email is a good candidate for unification because previous research has shown that email is already used by many people as a mechanism for storing documents and web pages for future retrieval (Mackay, 1988; Whittaker and Sidner, 1996; Bälter, 2000; Jones et al., 2001). Email has been discovered to be by far the most dominant digital tool for recording information scraps – small pieces of information that tend to slip through the net of traditional PIM tools (Bernstein et al., 2008). Email is also an activity that people perform daily, and as such, it is a medium which is bound to the context of surrounding events and often the user can relate these events to individual messages in his collection (Ringel et al., 2003). However, before email can be established as a good method of unification and, indeed, to determine how email tools can support this function for which they were not designed,

¹As an example of the cost information overload, a recent study showed that in the UK, IT managers spend 5 million hours per year searching for lost email messages. This equates to 140M in staff costs. Computing Magazine, 24/9/07 (<http://www.computing.co.uk/computing/news/2199363/five-million-hours-wasted>) last accessed on 13/10/07

we need a better understanding of the cognitive processes involved in managing email. We add to understanding with the study described in this article.

2.2 Memory and PIM

Previous research has highlighted the connection between the activities involved in PIM and the processes and workings of human memory. Carroll (1982) was perhaps the first to note the link when he demonstrated that simple eight character file names can trigger a detailed recollection of a file's content. Lansdale (1988) was also interested in the psychological aspects of information management, describing office organisational problems as problems of categorisation, recognition and recollection; while Case (1991) proposed that memory and metaphor impact the way historians manage their resources. More recently, it has been observed that memory problems and the limitations of human memory hinder PIM (Jones et al., 2005; Czerwinski and Horvitz, 2002; Elsweiler et al., 2007).

The connection between memory and re-finding has inspired many groups to design systems that support known characteristics of memory. For example, the systems designed by Freeman and Gelernter (1996) and Ringel et al. (2003), amongst others, attempt to leverage episodic memories and the fact that events are framed temporally with respect to the times of other events; the systems designed by Kaptelinin (2003) and Jones et al. (2005) exploit strong human abilities to relate information objects to contexts in which they were created or used; the Placeless Documents system (Dourish et al., 2000) exploits the fact that attributes of documents may be remembered better than their storage location, and the systems designed by Dumais et al. (2003) and Cutrell et al. (2006) exploit the fact that people usually find it easier to recognise than to remember. All of these projects are credible attempts to leverage psychological research to improve PIM tools. There have been some PIM studies that shed some light on this problem by revealing clues about memory in the context of PIM. For example, by understanding that users prefer to locate their documents spatially rather than using keyword search (Barreau and Nardi, 1995) we can infer that spatial memory can be useful for re-finding. There is also evidence that spatial memory can be utilised more effectively when the document space is three-dimensional (Robertson et al., 1998). However, the work of Jones and Dumais (1986) warns against relying too heavily on a spatial organisation. Their findings indicate that semantic labels provide stronger retrieval cues than spatial organisation alone, although enhanced performance can be achieved by combining semantic and spatial organisations. Lansdale and Simpson (1990) built on this work by discovering that both semantic and spatial cues are improved when the user selects the cues themselves, rather than having them selected by an external party, reflecting similar findings in psychology e.g. (Cohen, 1981). There is also evidence in the PIM literature for the utility of temporal and episodic memories in re-finding. Ringel et al. (2003) discovered that people can relate documents to events that happened around the time that the documents were created or used and that this can be utilised in re-finding interfaces to improve access to information. Further, examining the interaction log files of re-finding tools shows that users often remember that documents are connected to particular people and use these memories when creating re-finding queries (Dumais et al., 2003; Cutrell et al., 2006).

Building on this work, there have been studies that have focused specifically on what people tend to remember about their information. Gonçalves and Jorge (2004) and Blanc-Brude and Scapin (2007) examined the memories people had for different kinds of computer files and Elsweiler et al. (2005) looked at what people remember about their personal photographs. These studies highlighted the importance of context to recollection: not only did the participants tend to remember fragments of the context in which the information is created or used, but the recollections tended to be context

dependent i.e. what exactly was remembered changed depending on a number of factors, including the type of information being recalled. Despite these three studies, relatively little is known about user’s recollections for their personal information.

Even less is known about relationship between memory and PIM i.e. how memories are used and what they mean in terms of re-finding performance. However, there have been some relevant studies. Kalnikaité and Whittaker (2007) studied the difference between a user’s organic memory and various prosthetic memory devices, including pen and paper notes, a dictaphone, and a prosthetic that combined written and verbal notes on a handheld computer. The main outcome of this work was the discovery that in order for a prosthetic memory device – as any PIM collection is – to function effectively, it needs to be aligned with the user’s memory. Kelly et al. (2008) examined what one individual remembered about the context in which his information was used and by using theoretical performance metrics, examined the benefit that using these contextual recollections had on re-finding performance. They discovered that using recollection data could potentially improve re-finding performance.

The work of both Kalnikaité and Whittaker (2007) and Kelly et al. (2008) hint at the benefits that can be achieved by aligning PIM tools with the functionality of human memory, but as yet we know little about how users use their recollections to re-find, if current systems effectively support their recollections or even if a relationship exists between recollection and re-finding. Most of the work described in this section was performed based on the premise that there is a direct and close connection between recollection and re-finding, yet there is no concrete evidence of this in the literature.

In this article, we attempt to learn more about the relationship between a user’s recollections and re-finding performance. We have three main research aims: First, to determine if there is indeed a relationship between the recollection of the participants and their re-finding performance. Second, to determine if quantity of recollection is important i.e. is it important for the participants to remember lots of attributes about the mails they were looking for in order to achieve good performance? Third, we are interested to know if there are any specific attributes that benefit the participants when they are remembered.

3 Methodology

We conducted a user study which examined the participants’ recollections while they performed email re-finding tasks using three experimental systems: a browse-based system similar to the folder-based interface of Mozilla Thunderbird ², a search-based interface similar to the search-based interface of the same Mozilla Thunderbird email client³, and a third interface which was designed specifically to support memory based on previous investigatory work (Elsweiler et al., 2005, 2006, 2007). The third interface offered the user a more visual form of interaction based on thumbnail images of the sender of messages and provided a means of interaction that was a mixture of browsing and searching. Our study, therefore, incorporates both the current standard – the browsing and searching interfaces are what most people use to manage their emails – and a novel interface. We omit specific details on the interface designs here because the focus of this article is not on evaluating the performance of

²available from <http://www.mozilla.com/en-US/thunderbird/>

³the search interface can be found in Thunderbird from the menu: edit>find>search messages.

the individual systems used per se⁴. Instead, we investigate the relationship between the recollection data and the performance data for all three systems. We account for and isolate the influence of the experimental systems in our statistical analyses and show that the system used did influence the data. We discuss this further in Section 6.

The difficulties involved in performing controlled studies of re-finding behaviour are well documented (Boardman, 2004; Capra and Perez-Quinones, 2006; Elswailer and Ruthven, 2007; Kelly and Teevan, 2007). These include sourcing collections, creating experimental tasks, balancing the experimental design and protecting the privacy of users. The methodology we employed to overcome these challenges was that proposed by Elswailer and Ruthven (2007), which allows the behaviour and performance of the participants to be evaluated while they complete **realistic** re-finding tasks, on **real** (their own) collections, in the **controlled** environment of the laboratory. This is achieved by first performing a series of investigatory steps (diary studies, interviews and tours) to establish why the participants use email and learn about the contents of their collections. The knowledge gained can then be used to devise pools of experimental tasks that are possible to solve by re-finding information within the collections of individual participants. We have described the task creation process and experimental design for this study in great detail in a previous peer-reviewed article (Elswailer et al., 2008). Therefore, to save space here, we only summarise the design and highlight the important aspects required to understand and interpret the findings described in this article.

Using the approaches recommended by Elswailer et al. (2007), we created 3 pools of experimental tasks, one for each of the three participant groups who took part in the evaluation. The task pools for each group of users can be found in [Appendix A]. The pools consisted of a mixture of the task types (lookup, item and multi-item tasks) described by Elswailer et al. (2007). According to the definitions given by Elswailer et al. (2007), lookup tasks involve searching for specific pieces of information from within an email e.g. a place name, a time or a phone number; item tasks involve looking for a particular message, perhaps to pass on to someone else or when the entire contents are needed to complete the task; and multi-item tasks require multiple messages to be found and often require the user to process or collate the information from different messages in order to solve the task. Some of the tasks in the created pools, such as task B6 were applicable only to other participants in the same group, while others, such as task A1, were applicable to all three groups. Nevertheless, great care was taken to ensure that overall, the tasks in the three pools reflected the emails in the participants' collections and the kinds of tasks that participants might need to perform based on the purposes for which the different groups of users use email.

3.1 Participants

21 participants participated in our evaluation, 7 from 3 distinct user groups: undergraduate students, postgraduate students and research and academic staff members from the Department of Computer and Information Sciences at the University of Strathclyde, Glasgow.

We structured the experiment around the user groups because of the differences between the email behaviours exhibited between the groups. The three groups of participants were very different. They had different numbers of email messages, used email for different purposes and had different levels of experience with using email. Descriptive statistics between participant groups are reported in Table 1.

⁴Details of the three interfaces as well as detailed evidence for how they supported the participants' attempts to re-find can be found in (Elswailer and Ruthven, 2009)

Property	Postgraduate	Undergraduate	Researchers
Number of emails (median)	106 (min=95, max=228)	187 (min=76, max=1165)	5132 (min=1097, max=8954)
Age of oldest email (days)(mean)	76.15 (SD =2.44)	634 (SD=314.65)	941 (SD=546.08)
Number of filers	0	2	2
Number of No-filers	5	3	2
Number of spring-cleaners	2	2	3
Emails received per day (mean)	1.78 (SD = 0.70)	0.57 (SD =0.60)	8.03 (SD = 4.51)
Experience with using email	3 (IQR=1.25)	4 (IQR=0.25)	4 (IQR=0.00)

Table 1: Descriptive statistics of the email properties for the three participant groups.

The postgraduate group had not been enrolled at university for long and therefore had very few email messages. Similarly, the undergraduate group had low numbers of messages. This was because even though the undergraduates were recruited from the 3rd and 4th academic years and had been using their accounts for some time, the accounts had only recently been upgraded to the IMAP standard (where messages are left on the server). This meant that although the undergrad collections contained a few important older messages, the majority of the messages in the collection had been received more recently (in the previous year). The undergraduate participants did have on average more emails than the postgraduate students. The difference in collection sizes between the two groups was not significant ($t=1.87$, $df=7$, $p=0.052$). Participants in the researcher group, on the other hand, had significantly more emails than both the postgraduate ($t=5.76$, $df=7$, $p<0.01$) and undergraduate participants ($t=5.194$, $df=7$, $p<0.01$).

Many of the postgraduate students came from non-computer science backgrounds and had less technical experience with computers than the other groups. They also had limited experience with email re-finding. This was evident during the evaluation with participants in this group exhibiting far fewer pre-defined strategies for the benchmark tools. From informal interviews it was discovered that the participants in this group generally used email for class announcements and reported that they had less need to re-find information. The undergraduate students were recruited from 3rd and 4th year classes and therefore, had much more experience with computers and using email. The main uses of their departmental email for the undergraduate participants were class announcements, university related task management, collaborative work, and social communication with university colleagues. Nearly all of the participants remarked that they also had other email accounts that they used for personal and non-university purposes. The participants from the researchers group were also very experienced email users, had been using their accounts for longer periods of time and therefore had large numbers of messages, used email as a way to manage their activities and their documents and as a result they reported having to re-find information often.

A pre-study questionnaire asked the participants to rate their experience with using email browsing and search facilities to re-find emails. The participants answered on a scale from 1 to 5 where 1 meant no experience, 2 meant limited experience, 3 meant average experience, 4 meant reasonably experienced and 5 meant very experienced. The questionnaire data show that the postgraduate participants had significantly less experience with email re-finding than the other groups. However, there was no evidence to suggest a difference in the experience levels between the participants in the undergraduate or the researcher groups.

The undergraduate participants received and kept the fewest emails of the groups. This was significantly fewer than the postgraduate group ($t=2.97$, $df=6$, $p=0.01$) and the researcher group ($t=3.92$, $df=6$, $p<0.01$). Further investigation explained this by revealing that the undergraduate participants tended to exert more effort in collection maintenance than the other groups. The researchers received and kept the most emails of the three groups. However, all three groups in our study processed considerably lower quantities of messages than reported in other studies. For

example, Fisher et al. (2006), Whittaker and Sidner (1996), and Mackay (1988) all found that their participants received between 40 and 60 emails per day. The differences can be partially explained by the way the figures were calculated. Both Mackay (1988) and Whittaker and Sidner (1996) asked participants to estimate the volumes of email they receive. We, on the other hand, calculated our figures based on the collections themselves by dividing the total number of messages by the total number of days passed since the date of the oldest email. Therefore our figures do not include any emails that were deleted, but do include holidays and weekends where the volume of email would likely have been much lower. Although we feel that the method of calculation accounts for some of the differences, it is fair to say that overall, our population had different characteristics to those of previous studies. We discuss this further in Section 5.

To summarise, the three groups of participants in our study had very different characteristics. The postgraduates had low expertise, few messages and used email as a simple communication tool. The undergraduates had high expertise, few messages and used email for keeping track of class assignments and other university related tasks. Finally, the researchers had high expertise, lots of messages, but used their email for many purposes including managing content, scheduling, and task management.

3.2 Tasks

The tasks used in the evaluation were taken from the pools created in the first stage of the experiment. Each participant performed 9 tasks, with 3 tasks (1 lookup, 1 item, and 1 multi-item) being performed on each of the three experimental systems. The task types and systems were rotated to create a balanced experimental design [see Figure 1]. The tasks were read aloud to the participants so that they were not assisted with spelling, nor could they refer back to the text again during the task.

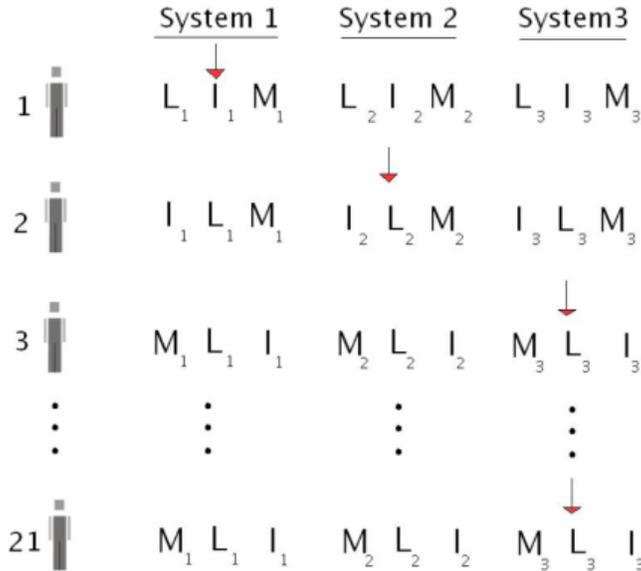


Figure 1: A diagram showing how task types (lookup, item, multi-item) and systems were rotated. The arrows demonstrate the starting system for each participant

Before completing each task the participants answered questions about the task, including how clear they felt the task description was and how difficult they perceived the task to be. Both questions

were answered on a scale from 1 to 5. The tasks were mostly rated as very clear (median = 5, IQR = 0)⁵ showing that the descriptions of tasks were understood, but there was a good mix of task difficulties (median = 2, IQR =1). Some of the tasks were perceived to be quite easy, while others were considered challenging. Similar ratings were applied across the different user groups and filing groups.

To determine how the recollection data changed over time, before each task we asked participants to specify roughly how long it had been since they last accessed the information the task required them to re-find. Participants could choose between the following options: In the last day, in the last week, in the last month, in the last year, or over a year. To simplify the data analysis process and to align our work with previous PIM research, the scale was simplified to the temperature metaphor proposed by Sellen and Harper (2003). Tasks that required information accessed in the last day or week were classified as hot, tasks that required information accessed in the last month were classified as warm and tasks requiring information last accessed over a month ago were classified as cold.

Of the 189 performed tasks, 45 were hot, 45 were warm and 60 were cold. The remainder of the tasks (mainly multi-item tasks) were classified as a temperature range i.e. different pieces of the sought-after information had been accessed more recently than others. Thus, although the way tasks were created and issued meant that we only had limited control over the temperature of the tasks, the issued tasks represented a reasonably balanced mix across the temperature range. The tasks were also balanced across the groups. See Table 2 below.

Temperature	Postgraduate	%	Undergraduate	%	Researcher	%
Hot	13	20.6	23	36.5	9	14.2
warm	26	41.3	12	19.1	7	11.1
Cold	8	12.7	25	39.7	27	42.9
Range	16	25.4	3	5.8	20	31.8
Total	63	100%	63	100%	63	100%

Table 2: The distribution of task temperatures across the groups of participants (Range indicates that more than one message was sought-after and these messages were of different temperatures)

3.3 Examining Recollections

We examined the participant’s recollections by employing a memory questionnaire. This is a technique that has been used often in the field of cognitive psychology to examine memory [see (Herrmann, 1982) for a review]. After each task, the participant was questioned on what they were able to remember about the information they had just looked for, i.e. the information they had available to help them with the search. Firstly, the participants were asked about four attributes common to every email. They were asked if, before completing the task, they *correctly* remembered (participants could answer “yes” or “no”):

- Roughly when the email was sent

⁵In descriptive statistics, the interquartile range (IQR) is the range between the third and first quartiles used to measure statistical dispersion. The interquartile range is a more stable statistic than the (total) range, and is often preferred to the latter statistic. A larger IQR means a larger range of data and IQR of zero would indicate no variance at all.

- The sender of the email
- What the email was about (we were interested in the topic of the email not exact syntax of the subject line)
- The reason why the email was sent

Secondly, the participants were asked if they remembered four attributes that are only applicable to some emails. We refer to these as additional attributes. The participants were asked if, before completing the task, they *correctly* remembered:

- Any other person(s) who may have received the email (this could include both individual recipients as well as organised mailing lists)
- If the email had any attachments
- If the email contained an image
- If the email contained a link or url

To clarify, these questions related to what the participants remembered before completing the task, i.e. information that could have been used to guide their re-finding strategy. Naturally, there were differences in the quality of recollections for the various attributes and sometimes the participants were unsure if what they remembered justified a “yes” response. A single experimenter⁶ applied the rule of thumb that for the recollection to count it had to be potentially useful to the re-finding task. For example, remembering that an email was sent on “Tuesday 23rd January because it is my birthday” is different from remembering that it was sent “around Christmas time last year”. Although, both of these recollections would be useful in a re-finding context. A recollection such as “I remember that it must have been some time in the last 3 years” is less useful and would probably not have been counted. However, the decision of whether or not a recollection should count was taken by the experimenter based on the context of the task and on the information provided by participants. It should be noted however that in the vast majority of cases it was clear whether or not the participant remembered a useful attribute.

We asked about recollections retrospectively to ensure that no bias was exerted on the participants’ behaviour while completing the tasks. There are advantages and disadvantages to employing this technique. We discuss these in Section 5.

The information collected was analysed by establishing the percentages of tasks for which each attribute was remembered. Firstly, all of the tasks were analysed to determine an overall picture of the participants’ recollection for email messages. Then, the data were analysed more closely by counting specific groups of tasks. This offered the opportunity to discover differences in the memories the participants had for different types of task, user and filing strategy.

3.4 Examining Re-finding Performance

Three metrics were used to gauge the participants’ level of performance: their ability to complete tasks, the time taken to complete tasks and the participants’ satisfaction with the information that he found.

⁶the lead author

A post task questionnaire surveyed the participants on how able they thought they were to complete each task on a 5-point scale, where 1 was not at all, 2 was not very, 3 was not sure, 4 was quite close and 5 was exactly. Responses of >3 were considered complete and all other responses deemed incomplete.

Each task was manually timed. In the *analyses of time taken* below only the times for completed tasks were considered. The reason for this is that when the participants failed to complete tasks they mostly used all of the 3 minutes allocated time in an attempt to complete the task. This means that if we had analysed time for both completed and incomplete tasks “time taken” would have been a proxy for “task completed”. By considering only time we were able to gain useful insights in addition to the factors that influenced task completion rates.

On completing each task, the participants were asked to rate their satisfaction with the information they found on a 5 point scale (1 = Not at all, 2 = not very, 3= not sure, 4 = satisfied and 5 = very satisfied). Responses of >3 were considered as satisfied and all other responses deemed unsatisfied.

3.5 Examining the Relationship between Recollection and Performance

We analysed the statistical relationship between performance and recollection by developing a number of generalized linear models for each performance metric (i.e. satisfaction, complete and time to complete) (McCullagh and Nelder, 1989). Given the large number of potential combinations between task, system, user group and user specific attributes for each performance metric, a multi-model approach was adopted for model development. That is, a set of pre-defined candidate models were evaluated, each reflecting a competing hypothesis of the relationship between recollection and performance. The most plausible model(s) out of this set were then selected for further inspection. This approach was chosen over other methods because we wanted to evaluate a set of potential hypotheses based on previous literature and domain knowledge, as well as minimise the selection of erroneous models through (semi-)automated techniques to model selection such as stepwise regression. We also measured memory from a number of different aspects: directly including memory attributes remembered or not, the interaction between memory attributes, and also in the form of a count of the total number of memory attributes remembered. To avoid problems with confounding factors, these viewpoints had to be developed in separate models.

To allow for a multi-model approach to inference, we used the Akaike information criterion (AIC) (Burnham and Anderson, 2002) – a measure of the goodness of fit of a model given observed data – to evaluate each candidate model. AIC penalises complex models with a large number of parameters in order to minimise overfitting (i.e. spurious relationships or artefacts of the data), providing a principled balance between bias and variance. The maximum likelihood of the model given the data is penalised by the number of free parameters (variables) in the model. As a result, the simplest, but most informative model, which reflects the observed data is selected. We report both the AIC and a scaled ΔAIC score, which is a measure of the relative difference between the top ranked and the remaining models. The larger the ΔAIC_i score, the greater the difference between the i^{th} and top ranked model. Therefore, models with a low ΔAIC_i score relative to the top ranked model could be assumed to be as informative. Using such an approach, therefore, enables different viewpoints to be included in the analysis rather than a strict acceptance or rejection of models.

The performance metrics were modelled depending on the underlying distribution. For example, time was found to be normally distributed and was, therefore, modelled as a general linear model using least squares regression. Both complete and satisfied were binary outcome variables and,

therefore, modelled using binary logistic regression. For all of the top ranked models, we report the associated model coefficients, F-test, fitted least squares mean or odds ratio for all factors included in the selected model(s). Given the sample size and the number of potential factors, and the exploratory nature of this analysis, statistical significance for each factor in the model was determined at the 10% level. All statistical procedures were performed using the the R GLM library⁷.

4 Findings

We describe our findings in three parts. Section 4.1 provides a summary of the recollection data, Section 4.2 describes the performance data, and finally, Section 4.3 details the findings of our modelling work, which investigated the relationship between what was remembered and how the participants performed. We also include a section describing the non-memory variables that our models show had an influence on the participants’ re-finding performance. These are discussed in Section 4.4.

4.1 Recollections

In this section, we review the findings with respect to the participants’ recollections. Even though these findings have been published before [see (Elsweiler et al., 2008)], we felt that it would be of benefit to summarise the findings in order to ease the understanding of this article’s contributions.

Overall the data indicate that for most tasks the participants remembered quite a lot about the email(s) they were looking for. Table 3 shows the percentages of tasks for which participants remembered the various attributes. The percentages are given for the different types of task (lookup, item and multi-item) and temperatures (hot, warm, and cold). In the vast majority of tasks the participants remembered what the email was about, why it was sent, as well as who sent it. Additionally, for many tasks these recollections were supplemented with additional temporal information, information about other recipients of email, as well as other attributes. In 42.33% of tasks, the participants remembered all of the common attributes, in 74.07% of the tasks 3 or more were remembered and in 85.9% of tasks participants remembered 2 or more common attributes. Therefore, if, as past research suggests, people re-find based on what they can remember, our data indicate that for the majority of the assigned tasks, the participants had options regarding which attributes to utilise when re-finding.

Task Type	When sent	Sender	Topic	Reason sent	Other recip.	Attach.	Image	Link	#Tasks
All tasks	57.45	77.13	85.11	80.85	46.81	12.77	2.13	21.81	188
Lookup	60.32	57.14	74.6	76.19	36.51	3.17	1.59	33.33	63
Item	58.73	68.25	85.71	87.3	42.86	6.35	0	6.35	63
Multi-item	34.92	90.48	82.54	66.67	46.03	14.29	4.76	11.11	63
Hot	71.11	95.56	91.11	88.89	53.33	22.22	2.22	17.78	45
Warm	68.89	71.11	93.33	91.11	44.44	17.78	2.22	37.78	45
Cold	56.67	56.67	80	81.67	41.67	6.67	3.33	26.67	60

Table 3: The percentages of all tasks in which the attributes were remembered

The most frequently remembered attribute was the topic of the email, which was remembered in 85.11% of tasks. This was followed by the reason the email was sent (80.85%), the sender of the email (77.13%), and temporal information (57.45%). Other recipients were remembered in 46.81% of the tasks, links or URLs in 21.81% of the tasks, and attachments were remembered in 12.77%

⁷<http://www.r-project.org/>

of the tasks. Images were reported as being remembered least often - only in 2.13% of the tasks did participants report remembering that the email contained an image. Nevertheless, it is likely that very few emails had attached images, and as there is no way of knowing what percentage of the tasks actually required the participants to retrieve emails containing images, links or URLs, or attachments, it is difficult to determine the importance of these additional attributes. However, what can be said is that for some tasks, and in the case of other recipients, a fairly large percentage of tasks, the participants had access to these extra pieces of information to help them search. Further, because not all of the emails have the additional attributes, such recollections could be useful for re-finding because of their discriminative power.

Despite good overall recollections, there were situations when the participants demonstrated poorer recollection. Using regression modelling techniques, we discovered, statistically, that certain factors influenced whether or not the attributes were remembered. We discovered that the temperature of the task, the filing strategy of the participant, the participant’s experience, the size of their collection, and their preferred method of re-finding all had an influence on the attributes that were remembered. The analyses of the recollection data are described in detail in (Elsweiler et al., 2008).

4.2 Re-finding Performance

Overall, the participants demonstrated reasonably good re-finding performance in the study. They were able to successfully complete 80.4% of the tasks assigned in an average time of 79.34 seconds. This is less than half of the 3 minutes time allocated to the participants to complete each task. The participants also seemed to be satisfied with the information that they found. The median of the satisfaction ratings was 5 with a very short inter-quartile range of 2.

Thus, the performance data seem to mirror the recollection data. Like the recollection data, which show that the participants generally remembered quite a lot, the participants demonstrated good performance overall. However, there were a number of tasks that were not completed and there were tasks that took longer to complete than others. At first glance, the situations for which the re-finding performance was poorest seem to align with the situations in which the participants remembered the least. For example, hot tasks, which were associated with better recollections than warm and cold tasks, were also associated with better performance. Similar observations were made when considering the user group, task type and filing strategy variables. The categories associated with better recollection were also associated with better performance.

The problem with these anecdotal observations is that there are many variables that may be influencing the data and it could be purely coincidental that memory appears to have an influence. In the following section, we analyse the data statistically to account for these variables to confirm if a relationship exists between what was remembered and how the participants performed. We also examine the factors that had an influence on this relationship.

4.3 Relationship Between Recollection and Performance

To formally analyse the relationship between recollection and performance, a number of models were developed for each performance metric (i.e. time to complete, task completion, and task satisfaction) using a multi-model approach to model selection (see Section 3.5). The final models were selected from a set of candidate models. All candidate models were ranked using the Akaike Information Criterion (AIC) (Burnham and Anderson, 2002), with the top ranked models chosen for further analysis. For brevity, we present only top six ranked models for each performance metric, Table 4.

Table 4: The six top ranked models for each performance metric with AIC scores. The combination of two attributes (i.e. WhenTopic) indicates an interaction between the two terms in the model. That is, an interaction analyses the combined effect of two or more attributes within the model (Burnham and Anderson, 2002).

Rank	Model	AIC	ΔAIC_i
1	Time = # Emails + System + Experience + When + Topic	-2096.4	0
2	Time = # Emails + System + Experience + WhenTopicSender*	-2097.7	1.3
3	Time = # Emails + System + Experience + WhenTopic*	-2098.0	1.6
4	Time = # Emails + System + Experience + Topic	-2098.4	2.0
5	Time = # Emails + System + Experience + When	-2099.9	3.5
6	Time = # Emails + System + Experience + #Memory Attributes	-2101.1	4.7
1	Complete = Pre-difficultly + Task + When + Topic	-184.05	0
2	Complete = Pre-difficultly + Task + WhenTopicSender	-184.18	0.13
3	Complete = Pre-difficultly + Task + Topic	-184.05	0.69
4	Complete = Pre-difficultly + Task + WhenTopic	-184.89	0.84
5	Complete = Pre-difficultly + Task + When	-185.22	1.17
6	Complete = Pre-difficultly + Task + # Memory Attributes	-189.78	5.73
1	Satisfied = Pre-difficultly + System + Task + When	-221.49	0
2	Satisfied = Pre-difficultly + System + Task + When + Topic	-223.1	1.83
3	Satisfied = Pre-difficultly + System + # Memory Attributes + Task	-223.3	2.03
4	Satisfied = Pre-difficultly + System + Task + Topic	-223.6	2.09
5	Satisfied = Pre-difficultly + System + Task + WhenTopic	-224.7	3.43
6	Satisfied = Pre-difficultly + System + Task + WhenTopicSender	-225.99	4.72

Table 5: Time to complete model (Rank 6: # Emails + System + Experience + #Memory Attributes) † # emails is a covariate i.e. increases linearly with time to complete

Predictor	Level	p-value	LS mean	95% CI
# emails†		0.001		
System				
	Novel	0.023	104.09	[93.72, 114.46]
	Browse-based		78.38	[68.29, 88.47]
	Search-based		78.83	[68.77, 88.89]
Experience				
	High	0.098	76.93	[64.42, 89.44]
	Low		97.27	[90.13, 104.41]
# Memory attributes				
	0	0.095	114.48	[86.05, 142.91]
	1		57.54	[43.78, 71.30]
	2		95.63	[82.55, 108.71]
	3		85.13	[76.69, 93.57]
	4		82.72	[74.71, 90.73]

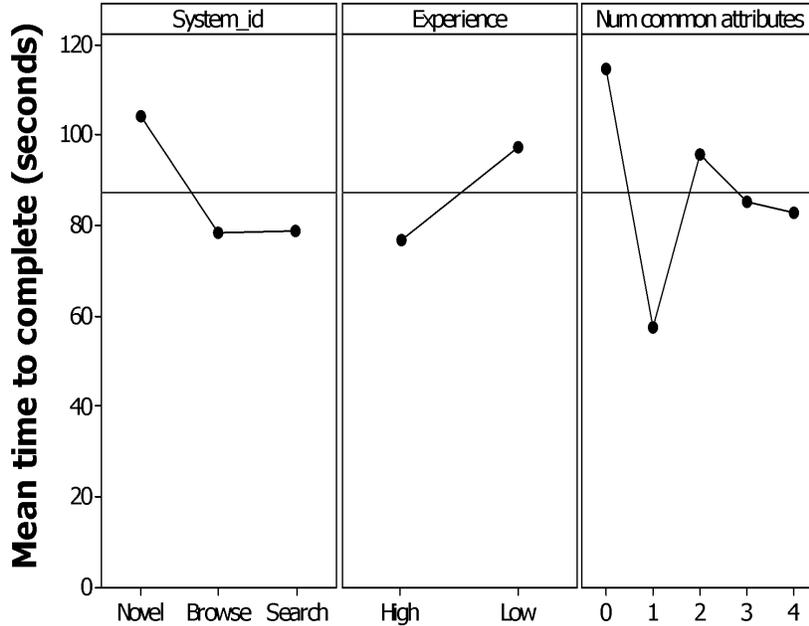


Figure 2: Model for time to complete (Rank 6: # Emails + System + Experience + #Memory Attributes)

A common trend emerged where similar attributes, specifically memory attributes, featured in the final top ranked models (Table 4). The modelling process highlighted the influence of memory on re-finding performance. In this section, we focus on the memory attributes included in the final models and further discuss the other user and task related attributes in the following section.

One important factor was the number of memory attributes remembered (“# Memory Attributes”), which measures the influence of memory recollection by counting the number of specific memory attributes remembered i.e. the recollection of when the email was received (When), who sent the email (Sender), what the email was about (Topic), and why the email was sent (Reason). The 6th ranked model for “time to complete” represented memory in this way. Further analysis of the effect of memory in this model (see Table 5 for model factors and Figure 2 for main effects plot) suggested that this relationship was not exactly what we expected, nor what previous research had suggested. We would have expected the relationship to depict that the more the participants were able to remember, the less time it would have taken them to complete the tasks or, if the systems were performing ideally, we would have expected that when the participants remembered something about the information, their performance would have been good and just as good if the participants remembered more information. However, the model did not infer either of these situations. Instead, it reported that when the participants did not remember any of the common attributes, they took a relatively long time to solve the tasks (between 86 to 143 seconds). The time taken was reduced significantly when one of the common attributes was remembered (44 to 71 seconds). However, when 2, 3, or 4 of the common attributes were remembered (i.e. when the participants remembered a lot about the information they were looking for), the time taken to complete tasks increased significantly in comparison. A possible conjecture is a situation of confusion: when participants remembered many details about the information they were looking for, they were unsure of which recollection to utilise

in their re-finding strategy.

A similar trend was identified in the task completed and satisfied models which were also ranked highly in the set of potential candidate models (see Table 4), although further inspection of both models indicated that there was insufficient evidence to suggest that the number of memory attributes remembered did have a significant effect on either model.

Table 6: Model for complete (Rank 1: Pre-difficultly + Task + When + Topic)

Predictor	Level	Coef	SE Coef	p-value	Odds ratio	95% CI
Constant		1.95	0.66	< 0.01		
Pre-difficultly		-0.73	0.45	0.02	0.48	[0.11 , 0.80]
Task type						
	Item	0.53	0.47	0.27	1.70	[0.67, 4.29]
	Multi-item	1.62	0.55	< 0.01	5.07	[1.72, 14.94]
When		0.82	0.44	0.06	2.27	[0.95, 5.41]
Topic		-1.15	0.66	0.08	0.32	[0.09, 1.16]

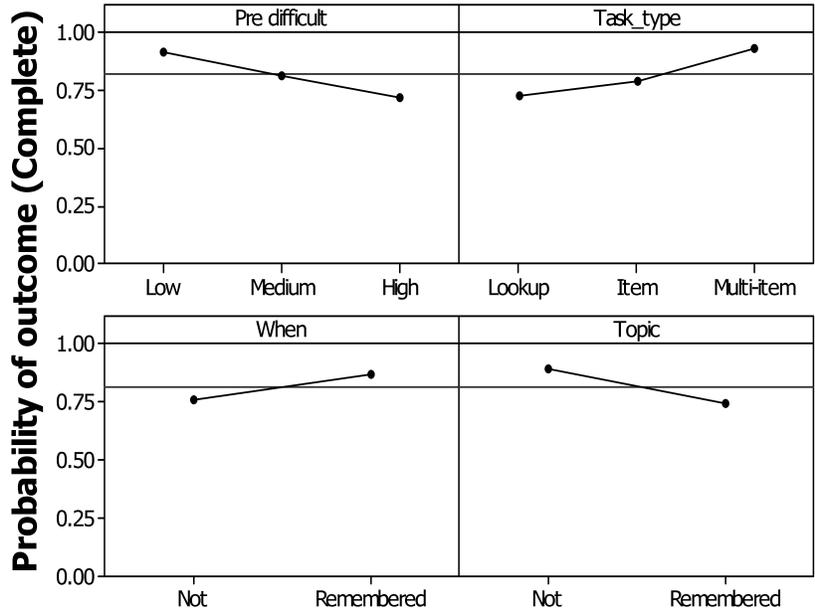


Figure 3: Model for complete (Rank 1: Pre-difficultly + Task + When + Topic)

Memory was also measured directly, where we were interested in identifying what memory attributes in particular influenced performance. Instead of modelling recollection as a count of the attributes remembered, we included each memory attribute as an individual factor to learn about how each attribute influenced the model. Measuring memory in this way resulted in improved models with respect to AIC, with models including “When” being the top ranked overall for all performance metrics i.e. temporal information had an important influence on the how the users were able to

Table 7: Model for time (Rank 1: # Emails + System + Experience + When + Topic) † # emails is a covariate i.e. increases linearly with time to complete

Predictor	Level	p-value	LS mean	95% CI
Number of emails†		< 0.01		
System		< 0.01		
	Novel		88.73	[67.57, 223.87]
	Browsed-based		63.6	[43.57, 150.73]
	Searched-based		63.75	[45.00, 153.75]
Experience		0.025		
	High		60.36	[35.26, 130.89]
	Low		83.69	[70.59, 224.86]
When		0.056		
	Not remembered		79.89	[63.27, 206.43]
	Remembered		64.16	[44.18, 152.52]
Topic		0.089		
	Not remembered		57.14	[30.00, 117.14]
	Remembered		86.91	[74.73, 236.34]

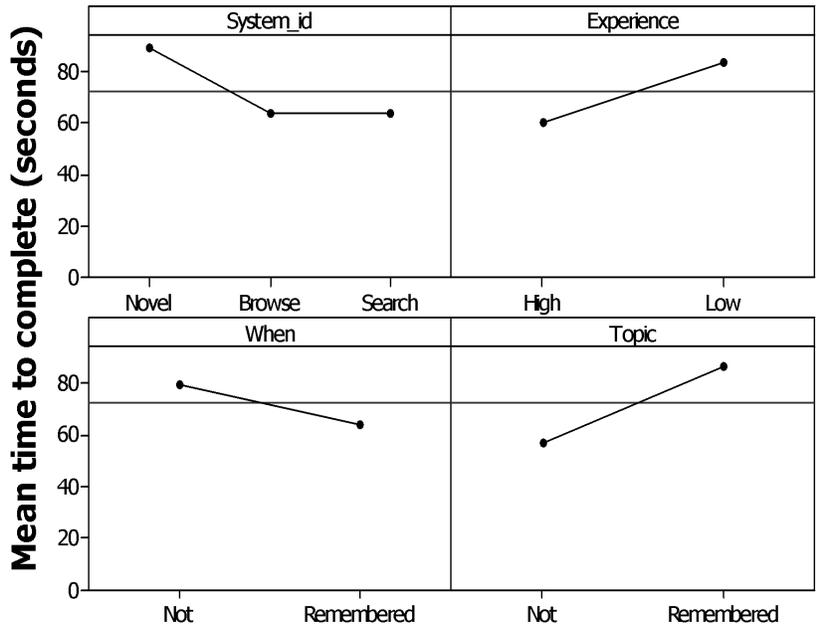


Figure 4: Model for time to complete (Rank 1: # Emails + System + Experience + When + Topic)

Table 8: Model for satisfaction (Rank 1: Pre-difficultly + System + Task + When)

Predictor	Level	Coef	SE Coef	P-value	Odds ratio	95% CI
Constant		2.23	0.78	< 0.01		
Pre-difficultly		-0.51	0.24	0.04	0.60	[0.37, 0.97]
System	2	0.90	0.44	0.04	2.47	[1.03, 5.90]
	3	0.44	0.42	0.30	1.55	[0.68, 3.53]
Task	2	0.17	0.45	0.69	1.19	[0.50, 2.86]
	3	1.02	0.47	0.03	2.77	[1.09, 7.00]
When		0.94	0.44	0.03	2.55	[1.08, 4.91]

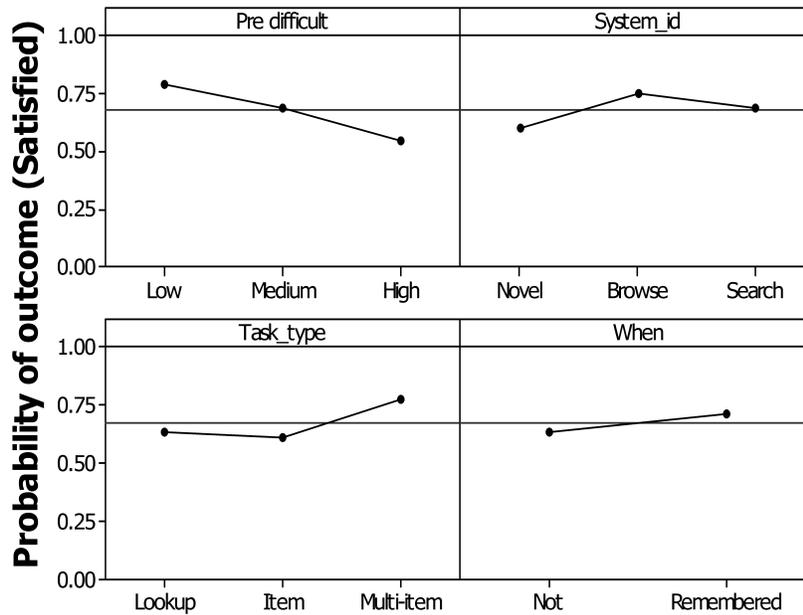


Figure 5: Model for satisfied. (Rank 1: Pre-difficultly + System + Task + When)

perform. Inspecting the models further (Tables 6, 7 and 8 and Figures 3, 4 and 5), highlighted the relationship between temporal information and performance. This finding demonstrated that when the participants remembered information regarding when the email they needed to find was sent, they tended to complete the task more often, tended to require less time to complete the task, and tended to be more satisfied with the information that they found.

The “topic” attribute featured in both the top-ranked complete and time to complete models. Inspecting the models indicated that remembering “topic” resulted in a negative trend in performance (see Tables 6, 7 and Figures 3, 4). As described in Section 4.1, these were the attributes remembered most frequently by the participants in our experiment.

From examining lower ranked models, we were able to hypothesize why having access to extra information, i.e. remembering what the sought after email was about, would negatively influence re-finding performance. A number of models measured the interaction between memory attributes, such as “When”, “Topic” and “Sender”. These models provide an explanation for the negative influence of the “topic” attribute, which may have been caused by the positive influence of the “when” attribute. In other words, if the participants were less likely to remember “when” – which we know positively influenced their performance – when they remembered “topic”, then this could make remembering “topic” appear to have a negative effect. To illustrate, we use the 3rd rank time to complete model as a case study of the interaction between memory attributes, examining the relationship between the “when” and “topic” attributes and the time taken to complete tasks.

Table 9: Model for time. (Rank 3 # Emails + System + Experience + WhenTopic*) † # emails is a covariate i.e. increases linearly with time to complete

Predictor	Level	p-value	LS mean	95% CI
Number of emails†			< 0.01	
System		0.03		
	Novel		85.75	[73.34, 98.16]
	Browse-based		60.74	[48.95, 72.54]
	Search-based		60.96	[49.79, 72.13]
Experience		0.06		
	High		57.57	[43.62, 71.52]
	Low		80.72	[71.54, 89.90]
When*Topic		0.09		
	None		67.00	[53.34, 80.66]
	When		35.93	[3.13, 68.73]
	Topic		94.11	[85.87, 102.35]
	All		79.55	[72.31, 86.79]

This model (shown in Table 9, Figure 6) shows the effect that remembering these attributes had on the time taken to complete the assigned tasks and disproves the theory that remembering the “topic” only had a negative influence because the “when” attribute was not remembered. When the participants had access to temporal information i.e. when they remembered the “when” attribute, much less time was taken to solve the tasks than when the “topic” attribute was remembered. When both attributes were remembered the tasks took less time than when “topic” was remembered alone, but longer than when “when” was remembered alone. In other words, according to these models,

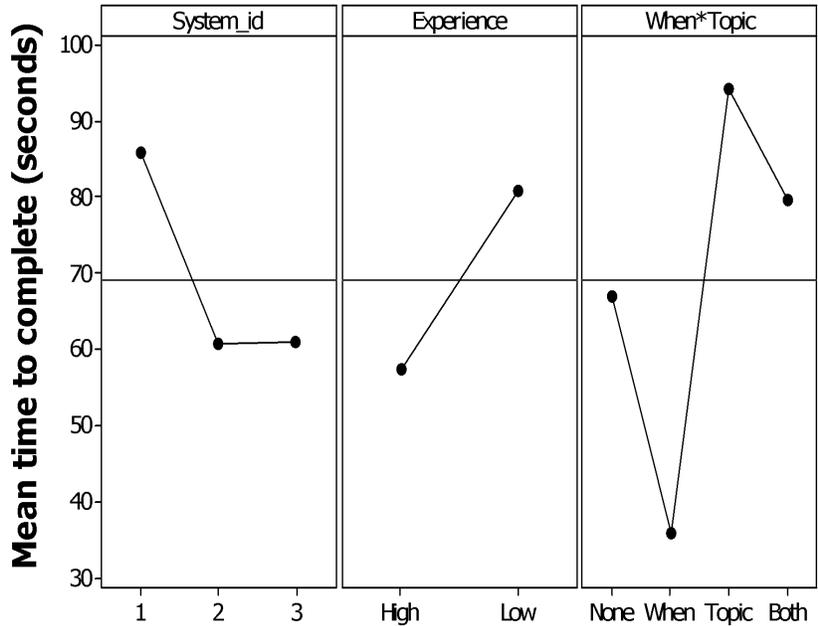


Figure 6: Model for time to complete. (Rank 3: # Emails + System + Experience + WhenTopic*)

it appears that remembering what the email was about had a detrimental effect when temporal information was also remembered.

To examine this situation more closely, we also examined the interaction between whether the “when”, “topic” and “sender” attributes were remembered, Table 10 and Figure 7. This model provided a relatively small increase in the ΔAIC_i score, however, it indicates an interesting trend. For both the “when” and “sender” attributes, better performance was evident when the attributes were remembered alone – fitting with the simpler model that modelled the recollection data as a count (Figure 2). Nevertheless, the “topic” attribute does not fit this pattern. Rather, the performance seems to improve when another attribute is remembered along with “topic”. One explanation for this could be that when the participants remembered semantic attributes, but not others, they were often unsure how to translate the recollection into a particular action or strategy using the three experimental systems. Our qualitative data seems to support this explanation. The experimenters regularly noted that the participants had difficulty with creating queries or deciding how to sort or browse through folders, even though they reported remembering what the email was about. Two participants even mentioned this situation during post-study interviews. However, the quantitative data presented here seem to indicate that the situation was much more widespread and not restricted to only a few participants.

4.4 Other Factors

The models presented above (Tables 5-10) also show that a number of other factors had an influence on the participants’ re-finding performance (see Table 4 for an overview). In the top-ranked tasks complete models, the participant’s pre-difficulty rating and type of task performed both had a significant effect. In the top-ranked time to complete models, the system used and the experience of

Table 10: Model for time (Rank 2: # Emails + System + Experience + WhenTopicSender*) † # emails is a covariate i.e. increases linearly with time to complete

Predictor	Level	p-value	LS mean	95% CI
Number of emails†			< 0.01	
System		0.02		
	Novel		90.93	[78.10, 103.76]
	Browse-based		63.72	[51.28, 76.16]
	Search-based		67.84	[56.07, 79.61]
Experience		0.08		
	High		63.27	[48.76, 77.78]
	Low		85.06	[75.36, 94.76]
When*Topic*Sender		0.07		
	None		116.46	[91.08, 141.84]
	When		15.24	[0.00, 54.97]
	Topic		92.65	[76.66, 108.64]
	Sender		49.96	[34.57, 65.35]
	When*Topic		67.59	[54.41, 80.77]
	Topic*Sender		95.44	[86.40, 104.48]
	When*Sender		72.36	[16.72, 128.00]
	All		83.61	[75.80, 91.42]

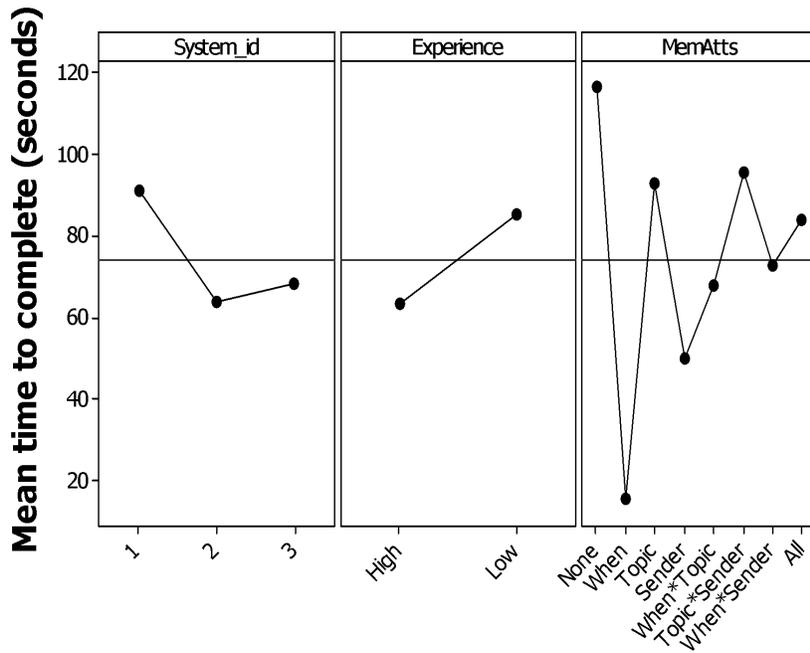


Figure 7: Model for time to complete (Rank 2: # Emails + System + Experience + WhenTopicSender*)

the user had an effect. Finally, in the top-ranked satisfied with found information model, the participant’s pre-difficulty rating, the re-finding system used, the type of task performed, and the number of emails in the participant’s collection all had a significant influence. We discuss these factors in section 6.

5 Limitations

Before discussing the implications of our findings, it is important to acknowledge the limitations of the work. Our study relates to the relationship between recollection and *email re-finding performance* and the findings should only be considered in this context. A limitation of the work is that our study population consisted of computer scientists and students taking computer science classes and investigated their recollections and performance for tasks that revolved around work-based email re-finding activities. Although computer scientists may not be representative of all email users, we argue that our results are generalisable, at least to some extent, for a number of reasons. Firstly, we included a group of users (the postgraduates) who had only recently started their course and who did not have a computer science background; indeed they had wide-ranging educational backgrounds and originate from different countries. These participants mainly had low levels of computer literacy and limited experience with email. Secondly, although we only examined re-finding tasks that were based on work and not leisure activities, we did examine each of the three categories of email re-finding tasks that previous work had shown users to complete in both work and leisure scenarios (Elsweiler and Ruthven, 2007). Also relating to the demographics of the study population, as mentioned in Section 3.1, the participants in our study had far fewer emails in their collections than had been reported in previous studies. Nevertheless, we do not feel this detracts from the usefulness of our findings because regardless of the quantities of emails in the collections and previous findings, our study analysed **real** users, performing **realistic** re-finding tasks (based on empirical work), on their **own** collections.

Regarding the creation of experimental tasks, great care was taken to learn about the contents of the participants’ collections as well as the kind of re-finding tasks they perform. The process involved recording real tasks that users in these groups performed and using these tasks as a template to create experimental tasks. It should be noted, however, that when asking a participant to perform a re-finding task it is necessary to tell the participant something about the information he should find before he can re-find it. This will undoubtedly affect the findings. However, again, great care was taken in the wording of the experimental tasks to minimise the effect. For example, rather than using phraseology that formed part of the textual content of the email, we chose wordings that would accurately communicate the information need, without providing keywords in the email text. For example, in task A2, rather than ask the participants to find information about the “MSDN academic alliance”, we asked them to find information about how they would go about getting free software from Microsoft through the university. Of course we were not always privy to the wording of emails, so in many cases we created a context where information would be required without mentioning details about the email e.g. task A3. We also tried to limit the amount of named entities in the task descriptions. Further, as the tasks were read aloud to the participants so that they were not assisted with spelling, nor could they refer back to the text again during the task.

Another limitation of our work is the number of variables present in the study. As noted above, other researchers have acknowledged the difficulties in performing PIM evaluations and one of the main challenges is controlling the variables present in experimental designs. In this study we have

attempted to control the variables as much as possible through rotating the types of task performed and the experimental system used around the participants. However, we concede that there are many variables that we were not able to control including the temperature of the task, the frequency with which the users perform that kind of task, the filing strategy of the users, the collection size of the user, and the difficulty of the task. Nevertheless, the way that the data were analysed, i.e. by using statistical modelling techniques, accounted for the uncontrolled variables, allowing us to isolate the factors that influenced how the participants performed.

Regarding the methodology used to discover the attributes of email messages that were remembered, we asked the participants what they were able to remember retrospectively, after they completed each task. The reason for this was that our experiment was primarily designed to be a system evaluation and we did not wish to influence the participants' re-finding behaviour by asking them about their memories before they performed each task. However, this means that we did not record what the participants remembered, rather what they remembered remembering. It also means that the process of re-finding and the information that they saw during the completion of the task may have subconsciously influenced what the participants believed they remembered. Again, however, we took steps to address this potential bias. During the completion of the task the participants often voiced their thoughts aloud. This gave the experimenter some idea of what was remembered before the task and what was learned (or cued) during the task. By retrospectively questioning, the experimenter was able to check with the participant when doubt arose. Nevertheless, the participants were generally good at determining what information they remembered and what information was cued and this information was communicated as part of the flow of conversation between the experimenter and the participant.

6 Discussion

In this article we have presented a study of the recollections people have when re-finding email messages and analysed the relationship between these recollections and how they performed when re-finding. We analysed the relationship by constructing statistical models from the data.

There were several outcomes to our analyses. We discovered that:

- there was a relationship between what the participants remembered and their re-finding performance
- the quantity of recollection was important and seemed to influence how the participants performed
- there were particular attributes that when remembered affected the re-finding performance
- and there were a number of non-memory related attributes that had an influence on the performance

Below we summarise our findings and discuss what they mean in the context of previous work, in terms of improving the design of email re-finding tools and in terms of future research directions.

6.1 Memory Factors

The models generated from our data indicate that there was indeed a relationship between how much and what the participants remembered about the emails they were trying to re-find and their performance when attempting to re-find the emails. Although, as we mentioned above, many research projects have been conducted based on this assumption, until now, to our knowledge, no study has evidenced this relationship empirically. The relationship we discovered, however, was not the one we expected to find. Instead of discovering that the more the participants remembered the better they performed, we found, in fact, the opposite situation. Our data suggest that remembering 1 common attribute was the optimal pattern of recollection and remembering further details tended only to make tasks require longer to solve. In Section 4, we suggested that one reason for this may be that the participants were unsure which recollection to use when they remembered more than one attribute. If this is true it may be beneficial for re-finding tools to prompt the user to use particular attributes, perhaps by highlighting highly discriminative fields or fields that when queried on facilitate good performance. In (Elsweiler et al., 2008), we proposed this as a means of helping the user remember more about their data, but our findings here suggest that encouraging (not restricting) the user to use one particular attribute may reduce the time needed to re-find information by reducing the uncertainty concerning which attribute to use. This technique could be implemented in many ways, although it would require proper user tests to establish the most effective approach. Possible methods could involve making the interface widget associated with a particular recollection more prominent by using bold text, making it especially large or by placing it in the line of sight of the user.

Our data show that the recollection of particular features of email messages seemed to influence the re-finding performance. Remembering temporal information was shown to have a particularly positive effect on all three of our performance metrics. This demonstrates that the experimental systems supported this kind of recollection particularly well and perhaps also suggests that, in-line with our previous suggestion, users should be encouraged by re-finding interfaces to re-find based on temporal recollections when they are available. It should be noted, however, that temporal recollections were not remembered very often. In fact, the participants only had access to temporal recollections in just over half of the tasks performed, highlighting the importance of additionally supporting other kinds of recollections.

Our models also seem to indicate that remembering what an email was about tended to have a negative influence on re-finding performance, suggesting that the experimental systems supported this kind of recollection particularly poorly. This is especially problematic as this was the email attribute remembered most frequently by the participants (85.11% of tasks). As such, it is vital that re-finding tools provide better support for semantic recollections. Each of the 3 experimental systems had features designed to leverage semantic recollection, although these were either based around keyword searching facilities or folder organisation. The evidence from our study suggests that these facilities are not adequate. Similar to studies of search behaviour, which have shown that users have difficulties expressing information needs as search queries (Rocchio, 1971; Belkin, 2000), there were many occasions when the participants in our study struggled to construct an appropriate search query that exploited a semantic recollection. The effectiveness of the other major alternative for exploiting semantic recollections – folder organisations – largely depended on the participant being able to create and maintain an appropriate organisation. Very few of our participants were able or willing to achieve this and this was evident during the study. Of those participants who made some effort to organise their emails into folders, i.e. the participants in the filing and spring-cleaning groups, the majority had no clear strategy for filing and often their strategy had changed

over time to suit developing needs and uses, resulting in uncertainty regarding the organisation of the collection. Further, many of the participants who used folders tended to organise based on a mixture of strategies. It was common to organise some emails based on the sender of the email and others based on the email’s semantic content. It was evident in our study that often this lack of a clearly defined strategy negatively affected their ability to re-find using folders, and in particular their ability to exploit semantic recollections.

It is possible that enhanced support for semantic recollections could be provided through relatively straightforward facilities, such as encouraging a purely semantic organisation for documents. In other words by providing the opportunity for the user to organise their emails in a way that supports their semantic memory. Our findings seem to endorse the task-based approaches suggested by Gwizdka (2002) and Bellotti et al. (2003), as well as the project-based approach of (Jones et al., 2005) rather than organisations based on other attributes. Nevertheless, as described in Section 2.1, previous research has shown that users tend to apply different filing approaches to facilitate activities such as task management, contact management and content management and this would seem to counter our suggestion of a purely semantic organisation. We suggest that email tools should have integrated task and contact management features so that the user can structure his organisation to optimise the re-finding of content.

We must also add a note of caution regarding forcing an organisation on the user. Our analyses of the recollection data [see (Elsweiler et al., 2008)], show that filing, i.e. the process of the user placing an organisation on his the email messages, results in poorer recollection, particularly for temporal recollections – those we have shown to be most beneficial to re-finding performance. Our analyses here do not provide any evidence of the participant’s filing strategy influencing performance. Nevertheless, it may be of benefit to investigate other methods of supporting semantic recollections.

Several potential alternatives that have been applied in the neighbouring research fields of human computer interaction and information retrieval, may offer a solution to the problem of supporting semantic recollections in PIM. Faceted browsing (Yee et al., 2003; Schraefel et al., 2005) and tagging (Golder and Huberman, 2006) are both ideas that could potentially address this problem and some groups have started to look at these in a PIM context (Cutrell et al., 2006; Weiland and Dachsel, 2008; Wilson et al., 2008). However, much more work is required to understand the benefits that these solutions can offer and how they might be used to manage personal information. We are particularly interested in how tag clouds might support PIM behaviour. Tag clouds are “visual presentations of a set of words, typically a set of ‘tags’ selected by some rationale, in which attributes of the text such as size, weight, or colour are used to represent features, such as frequency, of the associated terms.” (Rivadeneira et al., 2007). Tag clouds offer an overview of the overall theme (or gist) and content of the resource or collection being described. We feel that tag clouds may be useful, not only because they have been shown improve classification by making the process of selecting appropriate keywords easier (Harvey et al., 2009), but they could be a means of cueing recollection in the re-finding process. Our findings reveal the importance of exploring techniques that may allow users to better explore their semantic recollections.

In addition to the recollection attributes, the models in Section 4 demonstrated that a number of other factors had an influence on the participants’ performance. In the following sub-section, we describe these factors and discuss what they could mean.

6.2 Non-memory factors

The experience of the user was shown to influence how long the tasks took to perform, with the experienced users taking less time than inexperienced users to complete tasks. However, neither the probability of completing a task nor the satisfaction with the information found were affected by user experience. In our study the more experienced users were able to use techniques not available to those with less experience, such as the ability to utilise advanced query features, which helped solve some tasks particularly quickly. These users had more confidence in their re-finding strategies and were not only more knowledgeable regarding the browse- and search-based systems, but also demonstrated greater awareness of the content and organisation of their collections. We believe that the main reason that experience was not found to have affected the completion of tasks is that the less experienced users tended to have far fewer emails than their more experienced counterparts. The number of emails variable featured in the top-ranking information satisfied models, demonstrating that despite greater awareness of their collections and better re-finding skills, the participants with the most emails (i.e. those who had the most experience) tended to be less satisfied with the information that they found.

Another variable that featured in both the task complete and information satisfied models was the pre-difficulty rating assigned to the tasks by the participants before they attempted to perform the task. Our models show that the difficulty rating applied by the participants was a good indicator of how they would perform, suggesting that if researchers were able to understand more about what influences the user's perception of a re-finding task, it would provide greater insight into the difficulties they have with re-finding and how assistance can be provided. We are currently examining our data with this goal in mind.

The experimental system used was another variable that was shown to influence the performance, featuring in both the time and information satisfied models. Although a thorough examination of the experimental systems and how their individual features supported the participants' recollections is beyond the scope of this article, it is important to acknowledge that the system used had an influence and to discuss briefly some of the reasons for this. As shown in the models presented in Section 4, when the participants used the novel interface designed specifically to support their recollections, they tended to take longer to complete the task and were less satisfied with the information found than when the browse- or search-based interfaces were used. There was little overall difference in the performance achieved by the two familiar systems. This finding can be largely explained by the participants' lack of experience with the novel system. Due to time restraints, the participants received limited practise time on the novel system and several participants commented that this lack of experience influenced their performance. Nevertheless, our evaluation revealed that all three systems had both positive and negative design features. There were aspects of each that clearly supported the users' recollections and others that hindered re-finding. Detailed evidence for how the three experimental systems supported the participants' attempts to re-find can be found in (Elsweiler and Ruthven, 2009).

6.3 Unfeatured Factors

It is also interesting to look at the variables that did not feature in the final models that were selected in Section 4.3. Based on previous research we might have expected some extra variables to have featured. We were surprised, for example, that remembering the sender of an email did not seem to influence the re-finding performance. All three systems had features to support remembering the

sender and these were used often during the evaluation. Further, many of the participants commented on the usefulness of these features. Nevertheless, this qualitative evidence was not reflected in the statistical models that were constructed from the quantitative data. Further experimentation on a new sample of users would help investigate this conflict in greater detail. Given the sample size of the study, larger more complex models containing a greater number of attributes were penalised using AIC in order to avoid overfitting. Therefore, the models discussed in the paper were considered the most informative with respect to the observed data. However, a further study pooling the results of both this and a new sample would enable the investigation of more complex models which could include those attributes, analysing whether they did have an important effect on performance which was reflected in the qualitative data.

Another variable that we were surprised to discover did not have an influence was the participants' filing strategy. Our previous analyses had shown that how the participants file their emails influences what and how much they remembered about this information (Elsweiler et al., 2008). One attribute that was particularly affected by filing strategy was "when". We discovered that participants who tend to file their emails were statistically less likely to remember when a sought-after email was sent – an attribute that our analyses here showed had a powerful effect on re-finding performance. However, as mentioned, filing strategy did not feature in the top-ranking models described above.

7 Conclusions and Future Work

This article has explored the relationship between what a user remembers about email messages he is trying to re-find and how they perform when re-finding. We demonstrated that such a relationship exists, but it is more complicated than previous researchers have suggested. We discovered that remembering more does not generally equate to better performance and can, in fact, have the opposite effect. Our findings also suggest that remembering particular attributes of an email can affect the re-finding performance. We discovered that remembering when an email was sent had a positive influence while remembering what an email was about tended to have a negative influence. Other factors that influenced how the participants were able to re-find were the experience of the user, how difficult the user perceived the task to be before completing it and the experimental system being used. We discussed what our findings mean with respect to previous work and also to the design of future re-finding tools.

In our discussion we suggested many possible avenues for future exploration. Our work thus far has been concerned with building up a fuller understanding of PIM behaviour by examining our collected data from different perspectives. We are in the process of building on this work in a number of ways. Studying recollections for information objects is a challenging research question and the methodology we have used here has its limitations. To counter this and to add to our understanding, we are currently attempting to replicate our findings using other methods of investigation. We are also analysing our collected data to examine the features of email re-finding tasks that led the participants to perceive them as difficult. Further, we are exploring new interface features to determine whether they can support the user's recollections while they are re-finding.

References

Bälter, O., 2000. Keystroke level analysis of email message organization. In: CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press, New York, NY,

- USA, pp. 105–112.
- Barreau, D., June 1995. Context as a factor in personal information management systems. *Journal of the American Society for Information Science* 46 (5), 327–339.
- Barreau, D. K., Nardi, B., 1995. Finding and reminding: File organization from the desktop. *ACM SIGCHI Bulletin* 27 (3), 39–43.
- Belkin, N. J., 2000. Helping people find what they don't know. *Commun. ACM* 43 (8), 58–61.
- Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., 2003. Taking email to task: the design and evaluation of a task management centered email tool. In: *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, New York, NY, USA, pp. 345–352.
- Bernstein, M., Kleek, M. V., Karger, D., Schraefel, M. C., 2008. Information scraps: How and why information eludes our personal information management tools. *ACM Trans. Inf. Syst.* 26 (4), 1–46.
- Blanc-Brude, T., Scapin, D. L., 2007. What do people recall about their documents?: implications for desktop search tools. In: *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, New York, NY, USA, pp. 102–111.
- Boardman, R., 2004. Improving tool support for personal information management. Ph.D. thesis, Imperial College London.
- Burnham, K. P., Anderson, D. R., 2002. *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach*, 2nd Edition. Springer.
- Capra, R. G., Perez-Quinones, M. A., 2005. Using web search engines to find and refind information. *Computer* 38 (10), 36–42.
- Capra, R. G., Perez-Quinones, M. A., 2006. Factors and evaluation of refinding behaviors. In: *SIGIR 2006 Workshop on Personal Information Management*, August 10-11, 2006, Seattle, Washington.
- Carroll, J., 1982. Creative names for personal files in an interactive computing environment. *International Journal of Man-Machine Studies* 16, 405–438.
- Case, D. O., 1991. Conceptual organization and retrieval of text by historians: The role of memory and metaphor. *JASIST* 42 (9), 657–668.
- Cohen, R. L., 1981. On the generality of some memory laws. *Scandinavian Journal of Psychology* 22, 267–281.
- Cutrell, E., D.Robbins, S.Dumais, R.Sarin, 2006. Fast, flexible filtering with phlat. In: *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM Press, New York, NY, USA, pp. 261–270.
- Czerwinski, M., Horvitz, E., 2002. An Investigation of Memory for Daily Computing Events. In: *Brewster, S., Zajicek, M. (Eds.), Proceedings of HCI 2002*. pp. 230–245.

- Dourish, P., Edwards, W. K., LaMarca, A., Lamping, J., Petersen, K., Salisbury, M., Terry, D. B., Thornton, J., 2000. Extending document management systems with user-specific active properties. *ACM Trans. Inf. Syst.* 18 (2), 140–170.
- Ducheneaut, N., Bellotti, V., 2001. E-mail as habitat: an exploration of embedded personal information management. *interactions* 8 (5), 30–38.
- Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., Robbins, D., 2003. Stuff i've seen: a system for personal information retrieval and re-use. In: Sanderson, M. (Ed.), *SIGIR '03: Proc. 26th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, New York, NY, USA, pp. 72–79.
- Elsweiler, D., Baillie, M., Ruthven, I., 2008. Exploring memory in email refinding. *ACM Trans. Inf. Syst.* 26 (4), 1–36.
- Elsweiler, D., Ruthven, I., 2007. Towards task based pim evaluations. In: *Proceedings of SIGIR*.
- Elsweiler, D., Ruthven, I., 2009. Supporting memory in email re-finding. Submitted for publication in *Journal of American Society of Information Science and Technology*.
- Elsweiler, D., Ruthven, I., Jones, C., 2005. Dealing with fragmented recollection of context in information management. In: Doan, B. (Ed.), *Context-Based Information Retrieval (CIR-05) Workshop in Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05)*, Paris, France. Vol. 151. CEUR-WS.org.
URL <http://wwwsi.supelec.fr/bld/CIR-2005/>
- Elsweiler, D., Ruthven, I., Jones, C., 2007. Towards memory supporting personal information management tools. *J. Am. Soc. Inf. Sci. Technol.* 58 (7), 924–946.
- Elsweiler, D., Ruthven, I., Ma, L., 2006. Considering human memory in pim. In: *SIGIR 2006 Workshop on Personal Information Management*, August 10-11, 2006, Seattle, Washington.
URL <http://pim.ischool.washington.edu/pim06/files/elsweiler-paper.pdf#search=%22elsweiler%22>
- Fisher, D., Brush, A. J., Gleave, E., Smith, M. A., 2006. Revisiting whittaker & sidner's "email overload" ten years later. In: *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, New York, NY, USA, pp. 309–312.
- Freeman, E., Gelernter, D., 1996. Lifestreams: a storage model for personal data. *SIGMOD Record (ACM Special Interest Group on Management of Data)* 25 (1), 80–86.
- Golder, S., Huberman, B. A., 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science* 32 (2), 198–208.
- Gonçalves, D., Jorge, J. A., 2004. Describing documents: what can users tell us? In: *IUI '04: Proceedings of the 9th international conference on Intelligent user interfaces*. ACM, New York, NY, USA, pp. 247–249.
- Gwizdka, J., 2002. Reinventing the inbox: supporting the management of pending tasks in email. In: *CHI '02: CHI '02 extended abstracts on Human factors in computing systems*. ACM, New York, NY, USA, pp. 550–551.

- Harvey, M., Elsweller, D., Baillie, M., Ruthven, I., 2009. Folksonomic tag clouds as an aid to content indexing. In: in preparation for ICWSM 2009 - International AAAI Conference on Weblogs and Social Media.
- Herrmann, D. J., 1982. Know thy memory: The use of questionnaires to assess and study memory. *Psychological Bulletin* 92 (2), 434–452.
- Jones, W., Bruce, H., Foxley, A., Munat, C., 2005. The universal labeler: Plan the project and let your information follow. In: Grove, Andrew (Eds.), *Proceedings 68th Annual Meeting of the American Society for Information Science and Technology (ASIST) 42*, Charlotte (US).
- Jones, W., Teevan, J. (Eds.), 2007. *Personal Information Management*. Seattle: University of Washington Press.
- Jones, W. P., Bruce, H., Dumais, S. T., 2001. Keeping found things found on the web. In: *Proceedings of ACM's CIKM'01, Tenth International Conference on Information and Knowledge Management*. pp. 119–126.
- Jones, W. P., Dumais, S. T., 1986. The spatial metaphor for user interfaces: experimental tests of reference by location versus name. *ACM Trans. Inf. Syst.* 4 (1), 42–63.
- Kalnikaité, V., Whittaker, S., 2007. Software or wetware?: discovering when and why people use digital prosthetic memory. In: *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, New York, NY, USA, pp. 71–80.
- Kaptelinin, V., 2003. Umea: translating interaction histories into project contexts. In: *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press, New York, NY, USA, pp. 353–360.
- Kelly, D., Teevan, J., 2007. *Personal Information Management*. Seattle: University of Washington Press., Ch. Understanding what works: Evaluating personal information management tools, pp. 190–204.
- Kelly, L., Chen, Y., Fuller, M., Jones, G., October 2008. A study of remembered context for information access from personal digital archives. In: *Proceedings of the Second International Symposium on Information Interaction in Context*.
- Lansdale, M., 1988. The psychology of personal information management. *Appl Ergon* 19 (1), 55–66.
- Lansdale, M. W., Simpson, M., 1990. A comparison of words and icons as external memory aids in an information retrieval task. *Behaviour & Information Technology* 9, 11–131.
- Mackay, W. E., 1988. More than just a communication system: diversity in the use of electronic mail. In: *CSCW '88: Proceedings of the 1988 ACM conference on Computer-supported cooperative work*. ACM, New York, NY, USA, pp. 344–353.
- McCullagh, P., Nelder, J. A., 1989. *Generalized Linear Models*, 2nd Edition. Chapman & Hall/CRC.
- Ringel, M., Cutrell, E., Dumais, S., Horvitz, E., 2003. Milestones in time: The value of landmarks in retrieving information from personal stores. In: *Proc. INTERACT 2003*. pp. 184–191.

- Rivadeneira, A. W., Gruen, D. M., Muller, M. J., Millen, D. R., 2007. Getting our head in the clouds: toward evaluation studies of tagclouds. In: CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, New York, NY, USA, pp. 995–998.
- Robertson, G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D., van Dantzich, M., 1998. Data mountain: using spatial memory for document management. In: UIST '98: Proceedings of the 11th annual ACM symposium on User interface software and technology. ACM Press, New York, NY, USA, pp. 153–162.
- Rocchio, J., 1971. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Englewood Cliffs, NJ, Ch. Relevance feedback in information retrieval., p. 313323.
- Schraefel, M. C., Smith, D. A., Owens, A., Russell, A., Harris, C., Wilson, M. L., 2005. The evolving mspace platform: leveraging the semantic web on the trail of the memex. In: Proceedings of Hypertext, 2005, Salzburg.
- Sellen, A. J., Harper, R. H. R., 2003. The Myth of the Paperless Office. MIT Press, Cambridge, MA, USA.
- Weiland, M., Dachsel, R., 2008. Facet folders: flexible filter hierarchies with faceted metadata. In: CHI '08: CHI '08 extended abstracts on Human factors in computing systems. ACM, New York, NY, USA, pp. 3735–3740.
- Whittaker, S., Bellotti, V., Gwizdka, J., 2007. Personal Information Management. Seattle: University of Washington Press., Ch. Everything through Email, pp. 167–189.
- Whittaker, S., Jones, Q., Terveen, L., 2002. Contact management: Identifying contacts to support long term communication. In: Proceedings of CSCW 2002 Conference on Computer Supported Cooperative Work. New York: ACM Press., pp. 216–225.
- Whittaker, S., Sidner, C., 1996. Email overload: exploring personal information management of email. In: Tauber, M. J. (Ed.), CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press, New York, NY, USA, pp. 276–283.
- Wilson, M. L., André, P., mc schraefel, 2008. Backward highlighting: enhancing faceted search. In: UIST '08: Proceedings of the 21st annual ACM symposium on User interface software and technology. ACM, New York, NY, USA, pp. 235–238.
- Yee, K., Swearingen, K., Li, K., Hearst, M., 2003. Faceted metadata for image search and browsing. In: Cockton, G., Korhonen, P. (Eds.), CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press, New York, NY, USA, pp. 401–408.