Towards Task-based Personal Information Management Evaluations

David Elsweiler
Department Computer and Information Sciences, University of Strathclyde
dce@cis.strath.ac.uk

Ian Ruthven
Department Computer and Information Sciences, University of Strathclyde
ir@cis.strath.ac.uk

ABSTRACT

Personal Information Management (PIM) is a rapidly growing area of research concerned with how people store, manage and re-find information. A feature of PIM research is that many systems have been designed to assist users manage and re-find information, but very few have been evaluated. This has been noted by several scholars and explained by the difficulties involved in performing PIM evaluations. The difficulties include that people re-find information from within unique personal collections; researchers know little about the tasks that cause people to re-find information; and numerous privacy issues concerning personal information. In this paper we aim to facilitate PIM evaluations by addressing each of these difficulties. In the first part, we present a diary study of information re-finding tasks. The study examines the kind of tasks that require users to re-find information and produces a taxonomy of re-finding tasks for email messages and web pages. In the second part, we propose a task-based evaluation methodology based on our findings and examine the feasibility of the approach using two different methods of task creation.

Categories and Subject Descriptors

H3.3 [**Information Search and Retrieval**]:

General Terms

Measurement,Management,Experimentation, Human Factors

Keywords

Personal Information Management, User Evaluation

1. INTRODUCTION

Personal Information Management (PIM) is a rapidly growing area of research concerned with how people store, manage and re-find information. PIM systems - the methods and procedures by which people handle, categorize, and retrieve information on a day-to-day basis [18] - are becoming increasingly popular. However the evaluation of these PIM systems is problematic. One of the main difficulties is caused by the personal nature of PIM. People collect information as a natural consequence of completing other tasks. This means that the collections people generate are unique to them alone and the information within a collection is intrinsically linked with the owner's personal experiences. As personal collections are unique, we cannot create evaluation tasks that are applicable to all participants in an evaluation. Secondly, personal collections may contain information that the participants are uncomfortable sharing within an evaluation. The precise nature of this information - what information individuals would prefer to keep private - varies across individuals making it difficult to base search tasks on the contents of individual collections. Therefore, experimenters face a number of challenges in order to conduct realistic but controlled PIM evaluations.

A particular feature of PIM research is that many systems have been designed to assist users with managing and re-finding their information, but very few have been evaluated; a situation noted by several scholars [1, 6, 7]. Recently, however, researchers have started to focus on ways to address the problem of PIM evaluation. For example, Kelly [16] proposes that numerous methodologies must be taken to examine and understand the many issues involved in PIM, although, she makes explicit reference to the need for laboratory based PIM studies and a common set of shared tasks to make this possible. Capra [6] also identifies the need for controlled PIM lab evaluations to complement other evaluation techniques, placing specific emphasis on the need to understand PIM behaviour at the **task** level.

In this paper, we attempt to address the difficulties involved to faciliate controlled laboratory PIM evaluations. In the first part of this paper we present a diary study of information re-finding tasks. The study examines the kind of tasks that require users to re-find information and produces a taxonomy of re-finding tasks for email messages and web pages. We also look at the features of the tasks that make re-finding difficult. In the second part, we propose a task-based evaluation methodology based on our findings and examine the feasibility of the approach using different methods of task creation. Thus, this paper offers two contributions to the field: an increased understanding of PIM behaviour at the task level and an evaluation method that will facilitate further investigations.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

2. RELATED WORK

A variety of approaches are available to study PIM. **Naturalistic approaches** study participants performing naturally, completing their own tasks as they occur, within familiar environments. These approaches allow researchers to overcome many of the difficulties caused by the personal nature of PIM. As the tasks performed are "real" and not simulated, the participants can utilise their own experiences, previous knowledge and information collections to complete the tasks. A benefit of the approach is that data can be captured continuously over extended time periods and measurements can be taken at fixed points in time within these [15]. Naturalistic approaches can be applied by conducting fieldwork [17, 8], ethnographic methods as suggested by [15] or via log file analysis [9, 7]. Both ethnographic and fieldwork methods require the presence of an experimenter to assess how PIM is performed, which raises a number of issues. Firstly, evaluation in this way is expensive; taking long time periods to study small numbers of participants and these small samples may not be representative of the behaviour of larger populations. Secondly, because participants cannot be continually observed, experimenters must choose when to observe and this may affect the findings. An alternative strategy to conducting naturalistic evaluations is to utilise log file analysis. This approach makes use of logging software that captures a broad sampling of user activities in the context of natural use of a system. In [9] a novel PIM search tool was deployed to 234 users and the log data provided detailed information about the nature of user queries, interactions with the query interface and about properties of the items retrieved. Log file analysis is a powerful methodology as it allows the capture of a large quantity of detailed information about how users behave with the system without the expense and distracting influence of an observer. Nevertheless, there are limitations to this strategy. Firstly, to attain useful results, the deployed prototype must be something that people would use i.e. it has to be a fully functional piece of software that offers improvement on the systems ordinarily available to participants. Developing a research prototype to this standard is beyond the resources of many researchers. Further, caution must be taken when analysing logs, as the captured data shows nothing about the goals and intentions that the user had at the time. It is, therefore, difficult to make any concrete statements about the reasons for the behaviour depicted in the logs. This reveals a need to complement naturalistic studies with controlled experiments where the experimenter can relate the behaviour of study participants to goals associated with known search tasks.

Laboratory-based studies simulate users' real world environment in the controlled setting of the laboratory, offering the ability to study issues that are tightly defined and narrow in scope. One difficulty in performing this kind of evaluation is sourcing collections to evaluate. Kelly [16] proposes the introduction of a shared test collection that would provide sharable, reusable data sets, tasks and metrics for those interested in conducting PIM research. This may be useful for testing algorithms in a way similar to TREC in mainstream IR [13]. However, a shared collection would be unsuitable for user studies because it would not be possible to incorporate the personal aspects of PIM while using a common, unfamiliar collection. One alternative approach is to ask users to provide their own information collections to simulate familiar environments within the lab. This approach has been applied to study the re-finding of personal photographs [11], email messages [20], and web-bookmarks [21]. The usefulness of this approach depends on how easy it is to transfer the collection or gain remote access. Another solution is to use the entire web as a collection when studying web page re-finding [4]. This may be appropriate for studying web page re-finding because previous studies have shown that people often use web search engines for this purpose [5].

A second difficulty in performing PIM laboratory studies is creating tasks for participants to perform that can be solved by searching a shared or personal collection. Tasks relate to the activity that results in a need for information [14] and are acknowledged to be important in determining user behaviour [26]. A large body of work has been carried out to understand the nature of tasks and how the type of task influences user information seeking behaviour. For example, tasks have been categorised in terms of increasing complexity [3] and task complexity has been suggested to affect how searchers perceive their information needs [25] and how they try to find information [3]. Other previous work has provided methodologies that allow the simulation of tasks when studying information seeking behaviour [2]. However, little is known about the kinds of tasks that cause people to search their personal stores or re-find information that they have seen before. Consequently, it is difficult to devise simulated work task situations for PIM. The exception is the study of personal photograph management, where Rodden's work on categorising personal photograph search tasks has facilitated the creation of simulated work task situations [22]. There have been other suggestions as to how to classify PIM tasks. For example, [5] asked participants to classify tasks based on how frequently they perform the task type in their daily life and how familiar they were with the location of the sought after information and several scholars have classified information objects by the frequency of their use e.g. [24]. While these are interesting properties that may affect how a task will be performed, they do not give experimenters enough scope to devise tasks.

Personal collections are one reason why task creation is so difficult. Rodden's photo task taxonomy provides a solution here because it allows tasks, tailored to private collections to be categorised. Systems can then be compared across task types for different users [11]. Unfortunately, no equivalent taxonomy exists for other types of information object. Further, other types of object are more sensitive to privacy than photographs; it is unlikely that participants would be as content to allow researchers to browse their email collections to create tasks as they were with photographs in [11]. This presents a serious problem - how can researchers devise tasks that correspond to private collections without an understanding of the kinds of tasks people perform or jeopardising the privacy of study participants? A few methods have been proposed. For example, [20] studied email search by asking participants to re-find emails that had been sent to every member in a department; allowing the same tasks to be used for all of the study participants. This approach ensured that privacy issues were avoided and participants could use things that they remember to complete tasks. Nevertheless, the systems were only tested using one type of task - participants were asked to find single emails, each of which shared common properties. In section 4 we show that

people perform a wider range of email re-finding tasks than this. In [4], generic search tasks were artificially created by running evaluations over two sessions. In the first session, participants were asked to complete work tasks that involved finding some unknown information. In the second session, participants completed the same tasks again, which naturally involved some re-finding behaviour. The limitations of this technique are that it does not allow participants to exploit any personal connections with the information because the information they are looking for may not correspond to any other aspect of their lives. Further, if time is utilised by a system or interface being tested the approach is unsuitable because all of the objects found in the first session will have been accessed within the same time period.

Our review of evaluation approaches motivates a requirement for controlled laboratory experiments that allow tightly defined aspects of systems or interfaces to be tested. Unfortunately, it has also been shown that there are difficulties involved in performing this type of evaluation - it is difficult to source collections and to devise tasks that correspond to private collections, while at the same time protect the privacy of the study participants.

In the following section we present a diary study of re-finding tasks for email and web pages. The outcome is a classification of tasks similar to that devised by Rodden for personal photographs [22]. In section 5 we build on this work by examining methods for creating tasks that do not compromise the privacy of participants and discuss how our work can facilitate task-based PIM user evaluations. We show that by collecting tasks using electronic diaries, not only can we learn about the tasks that cause people to re-find personal information, but we can learn about the contents of private collections without compromising the privacy of the participants. This knowledge can then be used to construct tasks for use in PIM evaluations.

3. METHOD

Diary Studies are a naturalistic technique, offering the ability to capture factual data, in a natural setting, without the distracting influence of an observer. Limitations of the technique include difficulties in maintaining participant dedication levels and convincing participants that seemingly mundane information is useful and should be reported [19]. [12] suggest that the effects of the negatives can be limited, however, with careful design and good implementation. In our diary study, we followed the suggestions in [12] to achieve the best possible data. To this end, we restricted the recorded tasks to web and email re-finding. By asking users to record fewer tasks it was anticipated that participant apathy would be reduced and dedication levels maintained. The participants were provided with a personalised web form in which they could record details about their information needs and the contexts in which these needs developed. Web forms were deployed rather than paper-based diaries because to re-find web and email information the user would be at a computer with an Internet connection and there would be no need to search for a paper-based diary and pen.

The diary form solicited the following information: whether the information need related to re-finding a web page or an email message and a description of the task they are performing. This description was to contain both the information that the participant wished to find and the reason that they needed the information. To help with this, the form gave three example task descriptions, which were also explained verbally to each participant during an introductory session. The experimenter ensured that the participants understood that the tasks to be recorded were not limited to the types shown in the examples. The examples were supplied purely to get participants thinking about the kinds of things they could record and to show the level of and type of details expected. The form also asked participants to rate each task in terms of difficulty (on a scale from 1-5, where 1 was very easy and 5 was very hard). Finally, they were asked when was the last time they looked at the sought after information. Again, they were able to choose from 5 options (less than a day ago, less than a week ago, less than a month ago, less than a year ago, more than a year ago). Time information was used to examine the frequency with which the participants re-found old and new information, and when combined with difficulty ratings created a picture of whether or not the time period between accessing and re-accessing impacted on how difficult the participants perceived tasks to be.

36 participants, recruited by mass advertisement through departmental communication channels, research group meetings and undergraduate lectures, were asked to digitally record details of their information re-finding tasks over a period of approximately 3 weeks. The final population consisted of 4 academic staff members, 8 research staff members, 6 research students and 18 undergraduate students. The ages of participants ranged from 19-59. As both personal and work tasks were recorded, the results collected cover a broad range of re-finding tasks.

4. RESULTS

Several analyses were performed on the captured data. The following sections present the findings. Firstly, we examine the kinds of re-finding tasks that were performed both when searching on email and on the web. Next, we consider the distribution of tasks - which kinds of tasks were performed most often by participants. Lastly, we explore the kinds of re-finding tasks that participants perceived as difficult.

4.1 Nature of Web and Email Re-finding Tasks

During the study 412 tasks were recorded. 150 (36.41%) of these tasks were email based, 262 (63.59%) were web-based. As with most diary studies, the number of tasks recorded varied extensively between particpants. The median number of tasks per participant was 8 (interquartile range (IQR)=9.5). More web tasks (median=5,IQR=7.5) were recorded than email tasks (median=3, IQR=3). This means that on average each participant recorded approximately one task every two days.

From the descriptions supplied by the participants, we found similar features in the recorded tasks for both email and web re-finding. Based on this observation a joint classification scheme was devised, encompassing both email and web tasks. The tasks were classified as one of three types: lookup tasks, item tasks and multi-item tasks. Lookup tasks involve searching for specific information from within a resource, for example an email or a web page, where the resource may or **may not** be known. Some recorded examples of lookup tasks were:

- LU1: "Looking for the course code for a class - it's used in a

script that is run to set up a practical. I'd previously obtained this about 3 weeks ago from our website."

- LU2: "I am trying to determine the date by which I step down as an External Examiner. This is in an email somewhere"
- LU3: "Looking for description of log format from system R developed for student project. I think he sent me in it an email"

Item tasks involve looking for a particular email or web page, perhaps to pass on to someone else or when the entire contents are needed to complete the task. Some recorded examples of item tasks were:

- I1: "Looking for SIGIR 2002 paper to give to another student"
- I2: "Find the receipt of an online airline purchase required to claim expenses"
- I3: "I need the peer evaluation forms for the MIA class E sent me them by email"

To clarify, lookup tasks differ from item tasks in two ways - in the quantity of information required and in what the user knows about what they are looking for. Lookup tasks involve a need for a **small piece** of information e.g. a phone number or an ingredient, and the user may or **may not** know exactly the resource that contains this information. In item tasks the user knows exactly the resource they are looking for and needs the entire contents of that resource.

Multi-item tasks were tasks that required information that was contained within numerous web pages or email messages. Often these tasks required the user to process or collate the information in order to solve the task. Some recorded examples were:

- MI1: "Looking for obituaries and other material on the novelist John Fowles, who died at the weekend. Accessed the online Guradian and IMES
- MI2: "Trying to find details on Piccolo graphics framework. Remind myself of what it is and what it does. Looking to build a GUI within Eclipse"
- MI3: "I am trying to file my emails regarding IPM and I am looking for any emails from or about this journal"

There were a number of tasks that were difficult to classify. For example, consider the following recorded task:

- LU4: "re-find AS's paper on graded relevance assessments because I want to see how she presented her results for a paper I am writing"

This task actually consists of two sub-tasks: 1 item task(re-find the paper) and 1 lookup task (look for specific information within the paper). It was decided to treat this as a lookup task because the user's ultimate goal was to access and use the information within the resource.

There were a number of examples of combined tasks, mainly of the form item then lookup, but there were also examples of item then multi-item. For example:

- MI4: "re-find Kelkoo website so that I can re-check the prices of hair-straighteners for my girlfriend"

A second source of ambiguity came from tasks such as finding an email containing a URL as a means of re-accessing a web page. It was also decided to categorise these as lookup tasks because in all cases these were logged by participants as email searches and, within this context, what they were looking for was information within an email.

Another problem was that some of the logs lacked the detail required to perform a categorisation e.g.

- U1: "searching for how to retrieve user's selection from a message box. Decided to use some other means"

Such tasks were labelled as U for "unclassifiable". To verify the consistency of the taxonomy, the tasks were re-categorised by the same researcher after a delay of two weeks. The agreement between the results of the two analyses was largely consistent (96.8%). Further, we asked a researcher with no knowledge of the project or the field to classify a sample of 50 tasks. The second researcher achieved a 90% agreement. We feel that this high agreement on a large number of tasks by more than one researcher provides evidence for the reliability of the classification scheme.

The distribution of task types is shown in table 1. Overall, lookup and item tasks were the most common, with multi-item tasks only representing 8.98% of those recorded. The distribution of the task types was different for web and email re-finding. The majority of email tasks (60%) involved looking for information within an email (lookup), in contrast to web tasks where the majority of tasks (52.67%) involved looking for a single web page (item). Another distinction was the number of recorded multi-item tasks for web and email. Multi-item tasks were very rare for email re-finding (only 2.67% of email tasks involved searching for multiple resources), but comparatively common for web re-finding (12.6%).

	Lookup	Item	Multi-item	Unclass.
Email	90(60%)	52(34.67%)	4(2.67%)	4(2.67%)
Web	87(33.21%)	138(52.67%)	33(12.60%)	4(1.53%)
All	177(42.96%)	190(46.12%)	37(8.98%)	8(1.94%)

Table 1: The distribution of task types

In addition to the three-way classification described above, the recorded tasks were classified with respect to the temperature metaphor proposed by [24], which classifies information as one of three temperatures: hot, warm and cold. We classified the tasks using the form data. Information that had been seen less than a day or less than a week before the task were defined as hot, information that had been seen less than a month before the task as warm, and information that had been seen less than a year or more than a year before the task as cold. Unfortunately, a technical difficulty with the form only allowed 335(81.3%) of the tasks to be classified. The remainder were defined as U for "unclassifiable". A cross-tabulation of task types and temperatures is shown in table 2.

	Hot	Warm	Cold	Unclass.
Email	50(33.33%)	36(24.00%)	37(24.67%)	27(18%)
Web	112(42.75%)	60(22.90%)	40(15.27%)	50(19.08%)
All	162(39.32%)	96(23.30%)	77(18.69%)	77(18.69%)

Table 2: The distribution of temperatures

Most of the tasks that caused people to re-find web pages (42.75%) and email messages (33.33%) involved searching for information that has been accessed in the last week. However there were also a number of re-finding tasks that involved searching for older information: 23.30% of the tasks recorded (24.00% for email and 22.90% for web) involved searching for information accessed in the last month and 18.69% of the tasks recorded (24.67% for email and 15.27% for web) were looking for even older information. This is important with respect to evaluation because there is psycho-

logical evidence suggesting that people remember less over time e.g. [23]. This means that users may find searching for older information more difficult or perhaps alter their seeking strategy when looking for hot, warm or cold information.

4.2 What tasks are difficult?

We looked for patterns in the recorded data to determine if certain tasks were perceived as more difficult than others. For example, we examined whether the media type affected how difficult the participants perceived the task to be. There was no evidence that participants found either email (median=2 IQR=2) or web (median=2 IQR=2) tasks more difficult. We also investigated whether the type of task or the length of time between accessing and re-accessing made a task more difficult. Figure 1 shows this information graphically.

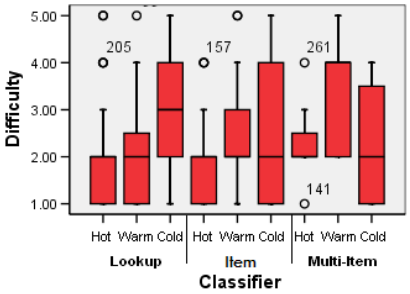

Figure 1: Difficulty ratings for task types

From figure 1, it does not appear that any particular task type was perceived as difficult with respect to the others, although there is a suggestion that lookup tasks were perceived more difficult when looking for cold information than hot and item tasks were perceived more difficult for warm information than hot. To assess the relationship between information temperature and the perceived difficulty, we used Mood's median tests to determine whether the rank of difficulty scores was in agreement for the information temperatures being compared ($p<0.05$). For the look-up task data, there was evidence that hot tasks were perceived easier than cold ($p=0.0001$) and that warm tasks were perceived easier than cold tasks($p=0.0041$), but there was no evidence to distinguish between the difficulty ratings of hot and warm tasks($p=0.593$). For the item task data, there was evidence that hot and cold tasks were rated differently ($p=0.024$), but no evidence to distinguish between hot and warm tasks($p=0.05$) or warm and cold tasks($p=0.272$).

These tests confirm that the length of time between accessing and re-accessing the sought after information indeed influenced how difficult participants perceived the task to be. Nevertheless, the large number of tasks of all types and temperatures rated by participants as easy i.e. < 3, suggests that there are other factors that influence how difficult a task is perceived to be. To learn about these factors would require the kind of user evaluations proposed by [16, 6] - the kind of evaluations facilitated by our work.

4.3 Summary

In the first part of this paper, we described a diary study of web and email re-finding tasks. We examined the types of task that caused the participants to search their personal stores and found three main categories of task: tasks where the user requires specific information from within a single resource, tasks where a single resource is required, and tasks that require information to be recovered from multiple resources. It was discovered that look-up and item tasks were recorded with greater frequency than multi-item tasks. Although no evidence was found that web or email tasks were more difficult, there was some evidence showing that the time between accessing and re-accessing affected how difficult the participants perceived tasks to be. These findings have implications for evaluating PIM behaviour at the task level. The remainder of this paper concentrates on this, discussing what the findings mean with respect to performing task-based PIM user evaluations.

5. TASK-BASED PIM EVALUATIONS

The findings described in section 4 are useful with respect to evaluation because they provide experimenters with enough knowledge to conduct controlled user evaluations in lab conditions. Greco-Latin square experimental designs can be constructed where participants are assigned n tasks of the three types described above to perform on their own collections using x systems. This would allow the performance of the systems or the behaviour of the participants using different systems to be analysed with respect to the type of task being performed (look-up, item, or multi-item). In the following sections we evaluate the feasibility of this approach when employing different methods of task creation.

5.1 Using Real Tasks

One method of creating realistic re-finding tasks without compromising the privacy of participants is to use real tasks. Diary-studies, similar to that described above, would allow experimenters to capture a pool of tasks for participants to complete by searching on their own collections. This is extremely advantageous because it would allow experimenters to evaluate the behaviour of **real** users, completing **real** search tasks on **real** collections while in a **controlled** environment. There is also the additional benefit that the task descriptions would not make any assumptions about what the user would remember in a real life situation because they would only include the information that had been recorded i.e. the information that was available when the user originally performed the task. Nevertheless, to gain these benefits we must, firstly, confirm that the task descriptions recorded are of sufficient quality to enable the task to be re-performed at a later date. Secondly, we must ensure that a diary-study would provide experimenters with enough tasks to construct a balanced experimental design that would satisfy their data needs.

To examine the quality of recorded tasks, 6 weeks after the diary study had completed, we asked 6 of our participants, selected randomly from the pool of those who recorded enough tasks, to re-perform 5 of their own tasks. The tasks were selected randomly from the pool of those available. The issued tasks consisted of 10 email and 20 web tasks, 9 of which were lookup tasks, 12 were item tasks, and 8 were multi-item tasks. The issued tasks represented a broad-sampling of the complete set of recorded tasks. They also included tasks with vague descriptions e.g.

- LU5:"Find a software key for an application I required to re-install".
- LU6:"Trying to find a quote to use in a paper. Cannot remember the person or the exact quote"

The usefulness of such tasks would rely on the memories of participants i.e. would the recorder of task LU5 remember which application he referred to and would the recorder of LU6 remember enough about the context in which the task took place to re-perform the task?

Presented with the tasks exactly as they recorded them, the participants were asked to re-perform each task with any system of their choice. Of the 30 tasks issued, 26 (86.67%) were completed without problems, 2 (6.67%) of the tasks were not completed because the description recorded was insufficent to recreate the task, and 2 tasks (6.67%) were not completed because the task was too difficult or the required web page no longer existed. Experimenters are likely to be interested in the final group of tasks because it is important to discover what makes a task difficult and how user behaviour changes in these circumstances. Therefore, from the 30 tasks tested, only 2 tasks were not of sufficient quality to be used in an evaluation situation. Further, there did not seem to be any issue of the type, temperature or difficulty ratings affecting the quality of the task descriptions. These findings suggest that the participants who recorded most tasks in the diary study also recorded tasks with sufficient quality. However, did the diary study generate enough tasks to satisfy the needs of experimenters?

Participant	Tasks	Lookup	Item	Multi-item	Unclass.
10	26	16	8	2	0
43	9	4	5	0	0
26	9	5	4	0	0
8	9	8	1	0	0
40	8	5	3	0	0
18	7	3	4	0	0
4	6	5	1	0	0
7	6	5	0	1	0
12	5	4	0	0	1
22	5	4	1	0	0
36	5	0	5	0	0
46	5	2	2	0	1
3	5	3	2	0	0

Table 3: The quantities of recorded email tasks

Participant	Tasks	Lookup	Item	Multi-item	Unclass.
26	32	7	20	5	0
32	31	11	18	2	0
10	19	0	10	7	2
33	18	5	13	0	0
5	15	0	7	2	4
8	11	0	6	5	0
22	10	0	3	5	2
28	10	1	7	2	0
37	10	1	9	0	0
35	9	7	2	0	0
6	9	0	1	8	0
40	7	1	5	1	0
9	7	0	0	5	2
12	7	1	0	3	2
42	6	0	4	2	0
29	6	0	3	3	0
15	5	0	2	1	2
4	5	0	4	1	0
43	5	2	3	0	0
18	5	0	0	3	2

Table 4: The quantities of recorded web tasks

Naturally the exact number of tasks required to perform a user evaluation will depend on the goals of the evaluation, the number of users and the number of systems to be tested etc. However, for illustrative purposes we chose 5 tasks as a cut-off point for our data. From tables 3 and 4, which show the quantities of email and web tasks recorded for each participant, we can see that of the 36 participants, only 13 (36.1%) recorded 5 or more email tasks and 20 (55.6%) recorded 5 or more web tasks. This means that many of the recruited participants could not actually participate in the final evaluation. This is a major limitation of using recorded tasks in evaluations because participant recruitment for user tests is challenging and it may not be possible to recruit enough participants if experimenters lose between half and two-thirds of their populations.

Further, there was some imbalance in the numbers of recorded tasks of different types. Some participants recorded several lookup tasks but very few item tasks and others recorded several item tasks but few lookup tasks. There was also a specific lack of multi-item email tasks. This situation makes it very difficult for experimenters to prepare balanced experimental designs. Therefore, even though our first test suggests that the quality of recorded tasks was sufficient for the participants to re-perform the tasks at a later stage, the number of tasks recorded was probably too low to make this a viable option for experimental task creation. However, it may be possible to increase the number of tasks recorded by frequently reminding participants or by making personal visits etc.

5.2 Using Simulated Tasks Based on Real Tasks

Another benefit of diary-studies is that they provide information about the contents and uses of private collections without invading participants' privacy. This section explores the possibility of using a combination of the knowledge gained from diary studies and other attributes known about participants to artificially create re-finding tasks corresponding to the taxonomy defined in section 4.1. We explain the techniques used and demonstrate the feasibility of creating simulated tasks within the context of a user evaluation investigating email re-finding behaviour. Space limitations prevent us from reporting our findings; instead we concentrate on the methods of task creation.

As preparation for the evaluation, we performed a second diary-study, where 34 *new* participants, consisting of 16 post-graduate students and 18 under-graduate students, recorded 150 email tasks over a period of approximately 3 weeks. The collected data revealed several patterns that helped with the creation of artificial tasks. For example, students in both groups recorded tasks relating to classes that they were taking at the time and often different participants recorded tasks that involved searching for the same information. This was useful because it provided us with a clue that even though some of the participants did not record a particular task, it was possible that the task may still be applicable to their collections. Other patterns revealed included that students within the same group often searched for emails containing announcements from the same source. For example, several undergraduate students recorded tasks that included re-finding information relating to job vacancies. There were also tasks that were recorded by participants in both groups. For example, searching for an email that would re-confirm the pin code required to access the computer labs.

To supplement our knowledge of the participants' email collections, we asked 2 participants from each group to provide email tours. These consisted of short 5-10 minute sessions, where participants were asked to explain why they use email, who sends them email, and their organisational strategies. This approach has been used successfully in the past as a non-intrusive means to learn about how people store and maintain their personal information [17]. Originally, we had planned to ask more participants to provide tours, but we found 2 tours per group was sufficient for

our needs. Again, patterns emerged that helped with task creation. We found content overlap within and between groups that confirmed many of our observations from the diary study data. For example, the students who gave tours revealed that they received emails from lecturers for particular class assignments, receipts for completed assignments, and various announcements from systems support and about job vacancies. Importantly, the participants were also able to confirm which other students had received the same information. This confirmed that many of tasks recorded during the diary study were applicable, not only to the recorder, but to every participant in 1 or both groups.

Based on this initial investigatory work, a set of 15 tasks (5 of each type in our taxonomy) was created for each group of participants. We also created a set of tasks for a third group of participants that consisted of research and academic staff members, based on our knowledge of the emails our colleagues receive. Where possible we used the information recorded in the diary study descriptions to provide a context for the task i.e. a work task or motivation that would require the task to be performed. When the diary study data did not provide sufficient context information to supply the participants with a robust description of the information need, we created simulated work task situations according to the guidelines of [2]. A further advantage of using simulated tasks in this way, rather than real-tasks, is that some of the users will not have performed the task in the recent past and this allows the examination of tasks that look for information of different temperatures. If only real-tasks had been used all of the participants would have performed the tasks during the period of the diary study.

The created tasks were used in a final evaluation, where we examined the email re-finding behaviour of users with three different email systems. 21 users (7 in each group) performed 9 tasks each (1 task of each type on each system) using their own personal collections in a Greco-Latin square experimental design. Performing a PIM evaluation in this way allowed the examination of re-finding behaviour in a way not possible before - we were able to observe the email re-finding strategies employed by **real** users, performing **realistic** tasks, on their **own** collections in a **controlled** environment. The study revealed that the participants remembered different attributes of emails, demostrated different finding behaviour, and exhibited different levels of performance when asked to complete tasks of the different types in the taxonomy. The key to both the task creation and the analysis of the results was our taxonomy, which provided the template to create tasks and also a means to compare the behaviour and performance of different users (and systems) performing different tasks of the same type. Some of the findings of the evaluation will be published in [10].

Summarising the approach, to conduct a user experiment using our methodology, researchers would be required to perform the following steps: 1)Conduct a diary study as above [1]. 2)Analyse the recorded tasks looking for overlap between the participants. 3)Supplement the gained knowledge about the contents of participants' collections by asking a selection of the participants to provide a tour of their collection. 4)Use the knowledge gained to devise tasks of the three different types defined within the taxonomy. More detailed information on how to use the research described in this paper to perform task-based PIM evaluations can be found at our website (see footnote 1).

[1]Information about this and the diary forms required can be found at http://www.cis.strath.ac.uk/d̃ce/PIMevaluations

6. CONCLUSIONS

This paper has focused on overcoming the difficulties involved in performing PIM evaluations. The personal nature of PIM means that it is difficult to construct balanced experiments because participants each have their own unique collections that are self-generated by completing other tasks. We suggested that to incorporate the personal aspects of PIM in evaluations, the performance of systems or users should be examined when users complete tasks on their own collections. This approach itself has problems because task creation for personal collections is difficult: researchers don't know much about the kinds of re-finding tasks people perform and they don't know what information is within individual personal collections. In this paper we described ways of overcoming these challenges to facilitate task based PIM user evaluations.

In the first part of the paper we performed a diary study that examined the tasks that caused people to re-find email messages and web pages. The collected data included a wide range of both work and non-work related tasks, and based on the data we created a taxonomy of web and email re-finding tasks. We discovered that people perform three main types of re-finding task: tasks that require specific information from within a single resource, tasks that require a single complete resource, and tasks that require information to be recovered from multiple resources. In the second part of the paper, we discussed the significance of the taxonomy with respect to PIM evaluation. We demonstrated that balanced experiments could be conducted comparing system or user performance on the task categories within the taxonomy. We also suggested two methods of creating tasks that can be completed on personal collections. These methods do not compromise the privacy of study participants. We examined the techniques suggested, firstly by simulating an experimental situation - participants were asked to re-perform their own tasks as they recorded them, and secondly, in the context of a full evaluation. Performing evaluations in this way will allow systems that have been proposed to improve users' ability to manage and re-find their information to be tested, so that we can learn about the needs and desires of users. Thus, this paper has offered two contributions to the field: an increased understanding of PIM behaviour at the task level and an evaluation method that will facilitate further investigations.

7. ACKNOWLEDGMENTS

We would like to thank Dr Mark Baillie for his insightful comments and help analysing the data.

8. REFERENCES

[1] R. Boardman, *Improving tool support for personal information management*, Ph.D. thesis, Imperial College London, 2004.

[2] P. Borlund, *The iir evaluation model: A framework for evaluation of interactive information retrieval systems*, Information Research **8** (2003), no. 3, paper no. 152.

[3] K. Byström and K. Järvelin, *Task complexity affects information seeking and use*, Information Processing and Management **31** (1995), no. 2, 191–213.

[4] R. G. Capra and M. A. Perez-Quinones, *Re-finding found things: An exploratory study of how users re-find information*, Tech. report, Virginia Tech, 2003.

[5] R. G. Capra and M. A. Perez-Quinones, *Using web search engines to find and refind information*, Computer **38** (2005), no. 10, 36–42.

[6] R. G. Capra and M. A. Perez-Quinones, *Factors and evaluation of refinding behaviors.*, SIGIR 2006 Workshop on Personal Information Management, August 10-11, 2006, Seattle, Washington, 2006.

[7] E. Cutrell, D.Robbins, S.Dumais, and R.Sarin, *Fast, flexible filtering with phlat*, Proc. SIGCHI '06 (New York, NY, USA), ACM Press, 2006, pp. 261–270.

[8] M. Czerwinski, E. Horvitz, and S. Wilhite, *A diary study of task switching and interruptions*, Proc. SIGCHI '04, 2004, pp. 175–182.

[9] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D.C. Robbins, *Stuff i've seen: a system for personal information retrieval and re-use*, Proc. SIGIR '03:, 2003, pp. 72–79.

[10] D. Elsweiler and I. Ruthven, *Memory and email re-finding*, In preparation for ACM TOIS CFP special issue on Keeping, Re-finding, and Sharing Personal Information (2007).

[11] D. Elsweiler, I. Ruthven, and C. Jones, *Dealing with fragmented recollection of context in information management*, Context-Based Information Retrieval (CIR-05) Workshop in CONTEXT-05, 2005.

[12] D. Elsweiler, I. Ruthven, and C. Jones, *Towards memory supporting personal information management tools*, (to appear in) Journal of the American Society for Information Science and Technology (2007).

[13] D. Harman, *What we have learned, and not learned, from trec*, Proc. ECIR 2000, 2000.

[14] P. Ingwersen, *Information retrieval interaction*, Taylor Graham, 1992.

[15] D. Kelly, B. Bederson, M. Czerwinski, J. Gemmell, W. Pratt, and M. Skeels (eds.), *Pim workshop report: Measurement and design*, 2005.

[16] D. Kelly and J. Teevan, *(to appear in) personal information management*, ch. Understanding what works: Evaluating personal information management tools, Seattle: University of Washington Press., 2007.

[17] B. H. Kwasnik, *How a personal document's intended use or purpose affects its classification in an office*, SIGIR'89 **23** (1989), no. SI, 207–210.

[18] M.W. Lansdale, *The psychology of personal information management.*, Appl Ergon **19** (1988), no. 1, 55–66.

[19] L. Palen and M. Salzman, *Voice-mail diary studies for naturalistic data capture under mobile conditions*, CSCW '02: Proceedings of the 2002 ACM conference on Computer supported cooperative work, 2002.

[20] M. Ringel, E. Cutrell, S. Dumais, and E. Horvitz, *Milestones in time: The value of landmarks in retrieving information from personal stores.*, Proc. INTERACT 2003, 2003.

[21] G. Robertson, M. Czerwinski, K. Larson, D. C. Robbins, D. Thiel, and M. van Dantzich, *Data mountain: using spatial memory for document management*, Proc. UIST '98:, 1998.

[22] K. Rodden, *How do people organise their photographs*, BCS IRSG 21st Annual Colloquium on Information Retrieval Research,Glasgow, Scotland, 1999.

[23] D.C. Rubin and A.E. Wenzel, *One hundred years of forgetting: A quantitative description of retention*, Psychological Bulletin **103** (1996), 734–760.

[24] A. J. Sellen and R. H. R. Harper, *The myth of the paperless office*, MIT Press, Cambridge, MA, USA, 2003.

[25] P. Vakkari, *Task complexity, problem structure and information actions: Integrating studies in on information seeking and retrieval.*, Information Processing and Management **35** (1999), 819–837.

[26] P. Vakkari, *A theory of task-based information retrieval*, Journal of Documentation **57** (2001), no. 1, 44–60.