Utilizing Episodic Context in Search of Personal Information Archives

Yi Chen
School of Computing
Dublin City University
ychen@computing.dcu.ie

Gareth Jones
School of Computing
Dublin City University
Gareth.jones@computing.dcu.ie

ABSTRACT
An important factor in successful search of personal information archives is the searcher's memory of the contents of the archive.. Recently, increasing attention has been paid to episodic memory or source memory related to finding personal information. Unfortunately, due to the absence of records for episodic context, most existing studies have only been able to explore frequently recalled episodic information without examining the correctness of the recalled details. In this paper, we report our studies which not only further explore the information that people tend to recall when looking for previously encountered items, but also examine the features that are likely to be remembered correctly. This study is based the records of three subjects who have very rare collections of two-year long episodic context data from their daily life experiences. Based on these findings, we have developed a system which aims to enable users to search for a variety of personal information with what they are most likely to correctly recall about it. This system is evaluated with the three subjects and their data collections. The results support our hypothesis that utilizing episodic context brings better user experience and effectiveness for searching person information.

Categories and Subject Descriptors
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human Factors; Design; Measurement.

Keywords
Memory, search, refinding, context

1. INTRODUCTION
Personal information generally refers to all the items or information that belong to an individual. According to Jones and Teevan [9], any information that a person keeps, receives, experienced, or is about the person, is this individual's personal information. Some typical examples of personal information include: created documents, received emails, visited web pages, and personal digital photos. The behaviour of looking for previous encountered information is called *refinding*. It is a predominant finding behaviour in the personal information space, since data owners have previously been exposed to most of their personal information. With the rapid growth of amount and variety of personal information that is stored, particularly in digital forms, it is becoming increasingly difficult for people to manage and locate this information. Desktop search systems were envisaged with the intention to relieve people of the burden of personal information management. These systems intended to enable users to find items without organizing them or remembering where they were stored [4]. Since the majority of personal information has already been experienced by the individual, it is usually expected that users can remember details of the targets that they encountered before, so that they can search for these targets with "remembered" content and attributes of a target, such as filename, path to the file, or keywords relating it.

However, users do not always remember the exact details that the search system accepts. For example, the search system may require users to search with the exact words in the textual content, but the user may only remember a sense of the meaning of the words rather than the exact expressions [10]. In another case, the user may remember that there is a graph in the document, but cannot query the search system using this knowledge since the system does not have a field for visual features in documents. The gap between the user's knowledge (what they can recall) and the attributes or information a search system accepts significantly compromises the power of potential power of search functions. To close the gap, it is essential to understand what users tend to recall, and to develop search systems which enable them to generate queries using data relating to their targets which they are most likely to recall reliably

The remainder of this paper is structured as follows. The next section reviews methods for study of memory features relating to information or electronic objects. This is followed by details of several studies exploring memory of features, reliability of memory of features and exploration in the context of our prototype system. Finally, we end with a summary of the conclusions of these studies.

2. RELATED WORK
Several groups have sought to explore what people remember about the information or electronic objects that they have seen before, e.g. documents (e.g. [1]), photos (e.g. [11]) and video clips (e.g. [8]). There are two broad types of approaches that have been used to explore this question: 1) *log analysis approaches* in which the system captures user activities (such as key strokes and queries entered) during search tasks and the researchers explore the research questions using the logged data (e.g. [5, 12]); 2) *self-report approaches* in which the subjects explicitly report to the researcher regarding what

Irish HCI '2013, June 12-13, 2013, Dundalk, Ireland.

they can recall (e.g. [1, 7]]). The former is usually less intrusive or effort-consuming compared to the latter. Therefore, studies with this method are much easier to extend to larger scale research studies than the self-report method. Yet, this approach can only capture what users "tell" the information system, which is not necessarily equivalent to what users remember. Thus, this method can only record the types of remembered information that search system is able to accept. If the system only has two search fields, e.g., title and author, data from the implicit-logging method can only show the frequency that people search with each of these attributes. However, this does not equal the likelihood that people remember these two attributes. For example, if the author field is usually easier to type and tends to return more relevant results than the title field, then users may prefer to search with "author", and they may sometimes not bother to type any terms into the title field just to save the time and effort of doing this, since they don't see any value to do this for their search results. Moreover, even if users remember other types of attributes, they cannot tell the system and the log cannot capture these attributes.

The self-report approach is more difficult to expand to large scale. Yet, it possesses more flexibility in exploring the types and accuracy of information that people remember. For example, Blanc-Brude and colleagues [1] tested recall memory of document attributes. They not only explored the features that are likely to be remembered, but also where people make mistakes or the parts they remember, e.g. which part of a document's path they tend to recall. In another study, Gonçalves and Jorge [7] let their participants "tell stories" about their documents. With this free narrative form, they collected information regarding users' autobiographical memory related to documents, which involves context from the physical world when they accessed the items previously. In addition, their participants also recalled related "real world" events, e.g. "I went to the library", "I printed the document". Their findings suggested that people usually have episodic memory related to accessing electronic documents, and that other associated real life events are also likely to be recalled. Unfortunately, this study did not evaluate how correctly the information was recalled as no ground truth was recorded.

Most of these short-term self-report experiments lack a real task context, as they usually asked the participants to find or recall detail of pre-selected specific files. In fact, there are at least two other types of task with different types of targets [6]. For example, looking up a specific piece of information (target), which could exist in one or more sources (files, webpages, emails and so on). Indeed, what a person remembers at the time of search depends not only on the target itself, but also on a complex combination of factors such as physical context and the internal state of the person who conducts the refinding task.

An alternative to short-term experiments are in-situ methods, such as experience sampling methods and diary studies. Experience sampling methods usually sample participants' status or experiences at certain intervals controlled by the researchers [3]. For example, the researchers can send them a message every hour or at a random interval to ask them what they experienced during the last hour or what they feel "at the moment". The diary study method usually gives more freedom to the participants, letting them decide when to report. The diary can be taken on a daily basis or at more flexible intervals, e.g. when some target event occurs (event-triggered). While the diary study method possesses more instancy and flexibility, it largely relies on the subjects' prospective memory, that is, their remembering to make a diary entry when a target event occurs. Combining experience sampling methods and a diary study may solve the problem. For example, let the participants make a diary entry when a target event happens, and send alerts to them as reminder as well as a signal to make a diary entry if they remember any target events which happened shortly before the alert, e.g. within the last hour. Of course, one challenge for a diary study is the difficulty of collecting enough data in a short period of time. Due to the involvement of effort and time, researchers usually cannot afford to hire many participants.

The cross-sectional one-off survey method is another alternative. Instead of sampling one subject's re-occurring behaviour over a long period to get a significant amount of data, this method samples the behaviour of a large sample of subjects who conducted certain behavior. Compared to in-situ methods such as experience sampling or diary study, one-off surveys are usually criticized for the low reliability of answers as they largely rely on the participant's memory. For example, if we want to explore people's refinding behaviour during the last week, it is more likely that the participants would miss report the details, than if they report their refinding tasks for the last hour every two hours over a week with the experience sampling method. However, since the cross-sectional approach usually requires much less effort than long term studies and can be done immediately after the participants were informed of it, it is more likely that many more participants can be recruited.

In this paper, we report three studies. In the first study, we extensively explore the features or attributes that tend to be remembered, in particular, the types of episodic context information related to target information, with both in-situ methods and a one-off survey. Of course, similar to the study by Gonvalves [7], we could not validate the recalled content from our subjects in these studies. However, as part of a larger project, we have three subjects who collected a variety of episodic context data for about two years. Therefore, in the second study, we examine the correctness of recall for each type of information from these subjects. Finally, We built up a search system based on the findings from the above two studies, and tested the hypothesis: enabling people to search with episodic context makes personal information search easier.

3. Study 1: Exploring remembered features

In the first study, we tried to investigate as many types of facets people may remember as possible. We adopted a combined diary and experience sampling method, and a cross-sectional one-off survey to explore the likely remembered features or related information of the search target.

3.1 Methodology

3.1.1 Participants

For the diary study, 11 participants were recruited via email and leaflets, including four males and three females, with age ranging from 22 to 39. Five of them were researchers, and the

rest were undergraduates or taught masters students. To encourage users to provide more details, we decided to make the study anonymous. This means that we cannot keep track of an individual participant so as to explore individual differences, thus we assigned each participant an ID generated automatically when they registered on online. All subjects, who agreed to participate, were initially required to complete a questionnaire regarding: gender, age group, usage of computers and information management habits. On completing the questionnaire, they were assigned with an ID, which was made up of six randomly generated numbers.

The one-off survey was posted on a questionnaire hosting website (www.sojump.com), which had more than 2 million subscribers at the time of the study. These subscribers receive onsite messages and emails informing them of new surveys that offer credit payment. The participants for our one-off survey study were recruited via the survey's hosting website, mail list and social networks. To reduce the memory problem of a typical one-off survey (e.g. recalling what happened during the last few months may be not very accurate), we limited the time range to the last hour. That is, we only asked each participate to recall one refinding task which they had carried out within an hour prior to answering the questionnaire, if there was any. In the instructions of the questionnaire (a brief description of the survey), we stated clearly that only those who have just looked for some previously encountered information could participate. We also asked the participants if they had looked for or wanted to find any previously encountered information at the beginning of the questionnaire, in order to filter out unwanted participants, i.e those who had not undertaken a refinding task in the last hour. Therefore, the participants who completed the questionnaire are those who had carried out a "refinding" task within an hour prior to completing the questionnaire. Due to the popularity of the website, paid-surveys can often get hundreds of replies within an hour. This means that if the survey is posted at 11.00 at night, most of the answers were from subjects who are still using their computer at 11.00 or later. Since people may do different types of things at different times of the day, e.g. they may do more activities relating to their work in the daytime, and more casual things or personal interest in the evening, and some people only use computers during certain periods of the day, there could be a bias of the sample. To avoid such bias introduced by time, the survey was posted four times on four different days and at different times of the day: morning (around 9am), afternoon (2pm), evening (7pm), and late evening (10pm). A total of 634 subjected completed the questionnaire, including 258 female and 376 male.

3.1.2 Material

Both studies used online questionnaires with almost the same content. The main questions in the online questions were:

Description of the target, scenario, and purpose.

This part started with an open question asking the subjects to describe the target. We expected to see some types of remembered features from the description, as the descriptions of task and recalled information can hardly be separated. This was followed by a list of options asking the subject to select or add a new type of the target. The types in the list were based on the findings from some pilot studies. Similarly, the purpose of the refinding tasks (why they needed the information), and the scenario of the task (what the subject was doing) were also presented in a similar form, that is, an open question followed by a multiple choice(s) question.

What they remember

This part also began with a free narrative box to collect types of remembered information from free recall, without directing or limiting the types of information to be recalled. However, people cannot always recall everything which they actually remember, if no proper memory cues are provided. Multiple choices questions were presented after this question, asking the participants to select the remembered types of information in the list. This enabled us to statistically explore the frequency that a type of information is recalled. The options were designed based on findings from related works (e.g. [4]) such as, the author, the source (where did you download it from or who sent it to you), the type, and the physical context of the subject (where he or she was, the approximate time, other activities, personal events or public events).

Information about the person

We asked question about the subject's gender, age group (e.g. 26-30), usage of computers, personal information management habits, and frequency of travel. These questions were asked in a separate questionnaire for diary subjects as they were expected to answer the main questionnaire more than once, and the personal information is unlikely to change during a short time interval of a couple of weeks.

For the diary study, a physical diary book was designed to take instant notes of a recent refinding task. Participants were expected to take notes right before or shortly after undertaking any refinding tasks. They could complete the online diary entry at any time afterwards when it is convenient for them. The physical book version questionnaire included a description of the target, the reason and purpose for finding it, a free description of whatever related information the subject remembered about it, and finally a multiple choice question to select the remembered features from a list. This was to make sure that what the user reported to be remembered in the online question was the same as what they recalled at the time of the refinding task. Emails were programmed to be sent to the participants randomly about twice a day. Since the diary study is anonymous, no face-to-face interviews were conducted afterwards, nor were any physical diary books collected from the participants. Only the data from the online questionnaire was downloaded one month later.

3.1.3 Quality control

Due to the sample distribution of the one-off survey, a Chinese version was provided for participants called for via the survey hosting website, as the subscribers to the website are expected to all be Chinese speakers. A significant disadvantage of online surveys is the un-monitored quality of the answers. Therefore, it is necessary to include certain mechanisms to avoid or to filter out answers which were not clear or treated seriously. The combination of both open-ended questions and multiple choice for most of the questions provided one solution. Apart from this, a couple of the questions were presented twice, but in a different

form, to roughly measure the reliability of the answer from each subjects. Minimum answering time was set for some pages in order to increase the likelihood a subject had read the question carefully and thought about the answers that they were giving. The answer to open-ended questions was also considered as an important criterion to assess the quality of the questionnaire responses.

3.2 Results and Discussion

A total of 515 valid tasks were collected in the diary study and cross-sectional survey.

3.2.1 Targets and Tasks

The targets range from documents, files, video clips, music, and photos to information on topics, which involves information from single or multiple sources, such as web pages, emails, chat records. Among the reported tasks, 47% of targets were easy to find, 19% were not found, 10% were difficult to find, and 24% of them had to-be searched for.

We combined the category selectd by the participants and the free description of targets, and found that most of the targets fell into following categories: a specific website (30%), a specific document, email or article (29%), a specific piece of information such as a number (8%), software or applications (10%), folder or directory (6%) , objects or entities such as an image or online shopping object(5%), and finally, the source (6%), that is, where or when an object or information was encountered, e.g. where I heard this tune, "the name of a TV series in which I saw the actress ...". The target types in our study showed a richer variety of targets than those usually studied in personal information management (PIM) literature. The targets are not always known-items or specific encountered information. For example, the targets in exploratory tasks are usually very uncertain. We concluded with the following list of types based on the specificity of the target:

- *Specific piece of information*: these are usually small pieces of information that the user needs to use directly. Examples include a phone number, an email address, a reference of a paper. Some of these attributes were used as a support for planning real life events, e.g. opening hours, exact name of an event.
- *Specific items (known-items)*: such as a specific document, email, application (software) and multimedia objects (e.g. YouTube video, images). These types of items are usually used directly, transferred (given to someone else) or as a source for browsing and finding other information.
- *Specific source*: Examples include a specific website, folder or document. These targets are usually a middle step in an information seeking task, the user usually proceeds to find other information as the task continues. The information may or may not have been encountered previously.
- *Details from specific source*: The "details" can involve both what one has seen and what one has not seen. This type of target is usually for learning purposes. Of course, learning about the information is usually not the ultimate purpose. Such information is acquired to support better decision making or planning, and to apply newly learned knowledge to current work, e.g. learning about a function for current programming task, learning about an agenda to have a better plan for the next couple of days.
- *Topic*: This type of target is usually multiple pieces of information taken from multiple sources, and is what one has seen before but for which one cannot recall all the details. For example, all of "my previous papers on this topic, information of all the recent movies, prices of the flights that I saw the other day.

Most of tasks found here could fit into the categorr of refinding tasks as defined by Elsweiler [6], but there are some exceptions. Based on the categorization by Elsweiler [6], we conclude the following types of refinding tasks reported in the diary study:

- *Look up tasks*: For exact details (e.g. phone number, address, contact names), attributes of an object such as price, date and/or time of an event, source (e.g. name of the song which sounds like this, the name of a book which have an episode like this.), etc.
- *Known item tasks*: The targets are usually specific objects that were encountered previously, such as a file, email, specific article, or software, multimedia object (e.g. images or videos clips), online shopping items. This type does not always stand-alone. It is sometimes one stage of an exploratory task or a look up task.
- *Exploratory tasks (topic learning)*: In these tasks, people usually do not have a clear idea of the exact information that they wish to obtain. The required information could be what the user has viewed before or something that they have not seen previously. Yet, they usually have some idea of potential sources where they encountered the relevant information. This type of task usually happens after a known item finding task, where the subject looks for some specific potential sources (e.g. a folder, a collection, a web site, or a document), and browses for interesting content.
- *Navigational tasks*: The target is usually a website, folder or directory, or group of pages such as a person's home page, blogs. The targets of this type of task are seldom a final stop of a finding process. Any of the above three types of tasks, in particular, an exploratory task is usually followed after reaching the target, by navigating or browsing in it.

3.2.2 Context of the Tasks

Regarding the context of the finding task, that is, what the subject was doing when he or she wanted to find the target, most of the participants were doing things related to the information (77.6%), including: reading or doing related work on a computer (41.5%), or talking about related things (23.7%). Other related context (12.5%) includes working on or seeking for information in the physical world, e.g. viewing physical photo albums. Other types of context were reported as irrelevant to the information they were looking for, including: working on other things on the computer, traveling, or in other casual settings. As for the purpose of the finding tasks, most of them (68.7%) were needed to continue with other work, about 12.1% of the targets were required by other people, and 19.2 % were just for casual reviewing. Some of subjects said that they just wanted to pick up the feeling they had when reading the book or watching the movie a few years ago.

3.2.3 Memory: what do people recall?

We coded the participant's description of the target, their free description of remembered information, and the types of remember attribute/meta data we listed.

Since we could not test the accuracy and reliability of the recalled information in this study, we do not focus on the frequency of recalling each feature. The frequency we report is just an indicator of the approximate rate that people recall such information.

When reporting memory of the target item itself, subjects usually described the perceived summary of content (of the story, article or movie). They also reported remembering part of the content (e.g. part of lyric, script or word). We asked participants who did not find their target to select the types of information that they remember about their target. Among the list of options, the keywords or sentences in textual target are the most well remembered features (36%), the name of websites, titles of articles or subjects of emails were also claimed to be remembered in around one third of reported tasks where such attributes are applicable. Visual features such as layout, background colour or salient visual elements were reported to be a remembered feature for a quarter of the reported tasks for finding web pages, online articles, blogs, and about 5% of the tasks for files on computers. Author of documents or article was selected as a remembered attribute for only 3 tasks. This is very different from what was claimed in [5], which found that people (author) was a well-remembered feature. This is probably due to the difference of subjects the two studies sampled. Participants in their study [5] were mostly researchers who read research papers, the authors' names of which are usually more important and easily visible than the editors' or bloggers' names of the web pages or on one of numerous articles.

Other types of information people tend to remember about electronic items are:

- Summary, gist of the meaning or other details (not exact words) of some content within or surrounding the target such as descriptions and other comments on the page of an online shopping item.
- Function of the web site or application.
- Self-created content in it, e.g., "my comments" on an article.

According to the participants' descriptions, most of them (93%) remember the types of the source where they encountered the target previously, e.g. whether it was from the web, told to them by other people via email or conversation, or created by him or herself. For information received from other people, 54% of subjects claimed to remember the contact name of the sender. As for information previous found on the web, 70% of the subjects who had searched for the target previously claimed to remember part of the queries they used. Some participants also remembered how they found the target previously. For example, "…but I know how to find it… last time I used Google, and the keywords are " …. ", I found it easily" , "When I read about the museum on wiki, I saw the movie, so I tried to find the wiki page again".

For 91% of the diary entries, the participants claimed to remember at least one occasion of interacting with the targets, although for 60% of these they only remember a general context. For example, "I was working on it day and night to beat the deadline". People sometimes also remember why they accessed that item previously, associated events or tasks, or people involved in those events. Another interesting finding is that in many of the diary entries, the subjects claimed to remember how they found the target previously, sometimes even remembering the exact queries they used to find the items. In three of the diary entries, the participants mentioned that they remember particularly well their first encounter with the item, e.g. when it was created, first received, or found. According to the subjects' reports, about 32% of the targets were encountered only once, 43% encountered several times in a separate time period, 29% of them were used or visited frequently in a certain period of time, and 2.1% of them were not sure of the number of previous encounters.

About 53% of the subjects claimed to remember some of the applications that they were using, and 33% of them remembered the name of the websites or documents that they were visiting around that time.

Most of the participants remembered where they were (65%), of course, most of these participants reported that they were in their regular locations (83%). About 41% of the participants claimed to remember the name of the exact address, name of street or estate, 19% of participants remembered locations around that place, and most of these people also claimed to be able to find the place on a map.

Apart from location, approximately 37% of participants reported remembering who was nearby, the weather status (28%), the light status (16%) and their emotional status (15%), most of emotional status values reported were either excited or depressed.

Many of the participants remembered what they were doing during that this period of time (67%), 37% of them remembered what happened in their organization (e.g. school, company), and public events (21%).

In short, episodic context and the source of the information is usually recalled. Similar to [7], we could not test how reliably people recall these features. However, this study collected rich information regarding the types of things people may remember. In the next study, we select some of the frequently recalled information and explore how likely it isthat people can correctly recall various features.

4. Study 2: What do people reliably recall?

With three subjects who had collected context information from their lives over a two years timespan, we conducted another experiment to explore the accuracy of the features people tend to recall.

4.1 Methods

4.1.1 Participants

The participants were three researchers who continuously collected a variety of personal data for 20 months, which we call their "lifelogs". Their data include: details of their computer activities (documents, emails, etc.), their location (captured using GPS devices), name of people nearby (noted by the name of Bluetooth devices people carries with them), weather, light status and so on (more details can be found in [2]).

4.1.2 Procedure

The subjects were required to generate 50 tasks based on free recall. We tried to encourage the subjects generate richer varieties of tasks, and provided suggestions according to the types we concluded in the previous study, such as: some information related to conferences you went, some interesting videos, websites, articles, some papers you worked on. For each of the tasks, the subjects were required to recall a list of information as (shown in table 1, column 1), and TO enter them into a spreadsheet.

The values were extracted and inserted into a search engine to retrieve potentially relevant items. The retrieved results were presented to the subjects to select judge their relevance.

For each of the relevant items, we retrieved its corresponding attributes and metadata, including: *keywords, extension (type of target item), date of visiting, month, season, day of week, part of week (week end, weekday), part of day, time range (e.g. 8am-9pm), people present, location, weather, file path, country, file name, from contact, to contact, device, year. Apart from above listed features, the participants also recalled and reported following types of information, though they could not be validated: emotion, phone calls and SMS around that time, other items accessed at that time, related people and related location.*

4.1.3 Results

We developed an application to compare the extracted value and the value reported by the subjects. This code calculated the hit rate and false query rate for each field. The *hit rate* refers to the percentage of relevant items for each task which matches the recalled attribute or metadata. The *false query* rate describes how many terms or values for a field (attribute or metadata) that the subjects recalled that has no matching items in the relevant results. For example, a subject may recall that the target documents were encountered probably in May, June or/and July. There are two documents, one of them was visited in May, and the other is encountered in August. Therefore, the query for the "month" field only matches one of the two documents, and the hit rate is 50%. This means, the query alone may only retrieve half of the potentially relevant documents. At the same time, only one of the three values ("May") matches at least one of the relevant items, therefore the false query rate = 66.7% (2/3 of the query may bring totally irrelevant results). The higher the hit rate and lower the false query rate, the better the accuracy and reliability. The following formula is used to assess reliability of recalled attributes based on hit rate and false query rate:

$$\textit{Reliability of recall} = \textit{hit rate} \times (1 - \textit{false query rate})$$

Of course, reliably recalled attributes may not necessarily be the most useful search fields for these participants, if these attributes or features were rarely remembered. Therefore, the score of reliability and the frequency of recall are combined to predict potential usefulness of the corresponding search field.

$$\textit{Usefulness} = \textit{hit rate} \times (1 - \textit{false query rate}) \times \textit{frequency of recall}$$

Table 1. Validation of recalled content

Attributes	Average Length	Frequency (%)	Reliability	Usefulness
Keywords	2.96	96	0.40	38*
Extension	1.15	99	0.63*	62*
File path	1.02	8	0.28	2.2
File name	2.25	15	0.14	2.0
From Contact	0.38	9	0.53*	4.8
To Contact	1.63	4	0.47	1.9
Device	1	71	0.15	11
Country	1	81	0.71*	58*
Date Range	2.9 (days)	9	0.14	1.2
Date	2.87	22	0.05	1.0
Month	1.09	83	0.38	31*
Season	1.04	95	0.36	34*
Day of week	1.17	6	0.19	1.2
Part of week	1	54	0.83*	45*
Time Range	3.47	7	0.40	2.8
Light status	1	6	0.19	1.2
People Present	2.3	17	0.09	1.5
Geo-location	1.02	94	0.25	24*
Weather	1.2	9	0.15	1.3
Surrounding Items	1.59	33	NA	NA
Emotion	1	8	NA	NA
Phone Call	1	6	NA	NA
SMS	1	3	NA	NA
Related People	2.67	3	NA	NA
Related Location	1.5	2.1	NA	NA

* NA indicates the recalled content for this field is not examined, usually because the information is not captured.

According to Table 1, the extension (item type), country of the person, name of the contact who sent the email or SMS and part of week, are the most reliably remembered features. Attributes or features which seem to be most useful (frequently and reliably recalled) include: extension, keywords, country, Geo-location, month, season, and part of week.

5. STUDY 3: EVALUATION WITH PROTOTYPE SYSTEM

In the above studies, we found that there are quite a few types of information which people may recall. Yet, not all of them tend to be recalled accurately or correctly. Considering the personal and contextual differences, we decided to include all the likely recalled and supported features (information that can be recorded and processed) in our a prototype system for people to search personal information archives. Of course, to evaluate how helpful the extended search options (context information) are, a long term (e.g. at least several months) collection is needed. This collection should not only include the personal information that people may look for, but also contextual data, which can be used to search. In this study, we recruited the three subjects who participated in the above study, and explored the question with their lifelog data.

5.1 Prototype Search System

The prototype search system was built on a search engine developed specifically for searching information in this specific prototype data collection from the three subjects. It indexed each computer activity with details of the computer item and the context. Fields of the target items include: title of the activity, name of the main file the activity is on, the path of the file or URL of webpage, type of the item, subject and contact names of emails, full text content. As each of the records in the index corresponds to a computer activity rather than a computer item, the context is unique, represented by fields including: location, device, people nearby (represented by Bluetooth names), weather, light status and so on. The search engine allowed the users to retrieve records with partially matched queries. For example, the query" type "pdf" +

title "hello world" could retrieve all items with the title "hello world" regardless the type and all "pdf" files.

In this study, we have two groups:

1) The first group of search options includes the following fields: keywords, title, filename, item type, extension, and "from who/where" (path for files and URL for online items, "received from" and "send to" for emails and text messages).

2) The second group of search fields includes:

a. Date and time: People seldom remember the exact number, which they have not necessarily paid attention to; these numbers include the exact time and exact date. Yet, people could perceive the date and time, e.g. shortly before Christmas as there were many Christmas decorations, in the evening as the light is dark. People may remember or be able to infer the numbers for the approximate month, year, day of week (e.g. Fridays as they always have such meeting on Fridays), or part of the week (e.g. weekend). They may also be told about approximate time based on their own schedules, e.g. after coffee, so that's around 3-4 pm as the person always have coffee at around 3pm. For these reasons, we allowed plenty of flexibility to the query on date and time. In this system, users could search by: year, month(s), part or day(s) of the week, part or hours of the day. Range sliders were available as an alternative to text format for people who remember the range.

b. Physical context: Location, weather, light status and people appeared at the time of any instances of access or interacting with the target items.

To provide better cues to let the users recall any possible memory traces that could be used by the search engine, we tried to make all the search options visible to the users. However, since there are a considerable number of the search options and some of them take more than a height of a single text line, it is difficult to present them all on a single view (without scrolling up and down). Even if the screen is big enough to accommodate all of them, users are unlikely to scan all the options when they are in a hurry or if they are lazy. We group together the options for similar type of the attributes, and represent them with an easily visible label, such as "where were you", "who was there". We expected that these labels could help users to recall corresponding content. By clicking the label, the search options in that group is unfolded and ready for to accept input. In this way, we could save space and reduce the amount of information presented to the users. When several options were unfolded, it became similar to a traditional multi-field search interface again, with many search options listed, and required the users to take time to browse for the desired search option. To avoid this, side labels were used. These labels, representing single or groups of search options, remained in the same place, and were always visible. When clicking a label, the search panel scrolled to the corresponding search option(s).

One problem people usually encounter is the difficulty of recalling the exact information such as the name of a file or exact spelling of a city. For this reason, "automatic completion" is used in some options such as file type, name of locations and people. This system searches for matching names with what the users partially typed.

5.2 Evaluation

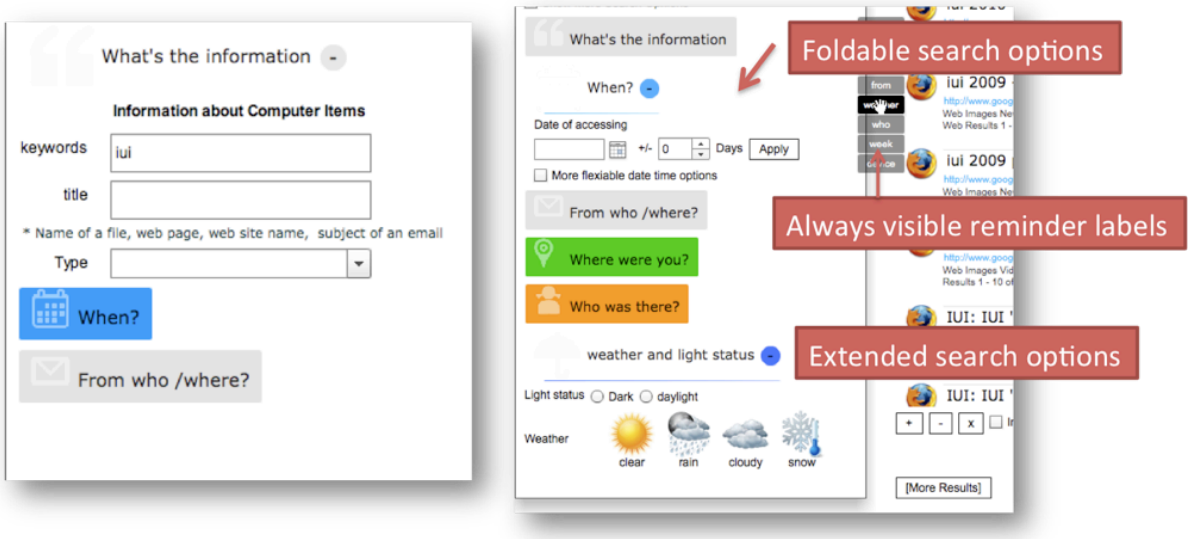

Figure 1. Prototype system: baseline interface (left) and experimental

Evaluation of the system took place a year after the previous experiments to allow plenty of time for forgetting information learnt during the previous experiment. The evaluation of search systems are usually based on three criteria: effectiveness (if the user can find the needed information); efficiency: how easy it is to retrieve and use the information; and satisfaction, which is a subjective evaluation of the user's preference of using the system.

5.2.1 Method

We evaluate the systems with the three subjects. Again, they were asked to generate 30 tasks for: specific items/files, specific pieces of information, topic search. We encouraged them to create 10 tasks for each of them. Ideally, we would like the subjects to use the system in their daily life for real information needs. However, we found that this difficult. According to one of the subjects, the work that she is conducting at the time of generating these tasks is very different from what she was doing a year ago, therefore, she hardly needed to re-find things from that "old" archive. Despite this, we wanted the information needs to be contextualized. We required the subjects to imagine the situation in which they would need the information, so that the tasks are closer to real ones, and

we can also evaluate how easy it is for the users to use the information the system provides.

We created a baseline interface which only included basic search options such as keywords, contacts, path, type, and exact date range. For the first 10 tasks, the participants were forced to use baseline interface in which they cannot search by contextual information. This was to make sure that at least 10 tasks are conducted with baseline interface. The next 10 tasks uses the interface with context options. The subjects were only introduced to the experimental interface after these 10 tasks. A couple of training tasks were conducted with the targets assigned by the researchers. In another 10 tasks, the context option was default hidden (not selected), but could be displayed if the users select them for to use. After each task, the subjects were presented with a short questionnaire to rate on a five point rating scale: the effectiveness (if they found their target), efficiency (how difficult is it, 1=extremely easy, 5=extremely difficult), difficulty of recall, and provide further comments if they want. The difficulty level was compared with that they rated before given this system. That is, the difficulty level of finding the target without this system.

5.2.2 Result

We logged the types of queries they used, and calculated the frequency of using each type of queries for 217 search sessions. A search session refers to a submission update to query. There were about 2.4 search sessions per task. We found that the most often used types of query with the experimental interfaces were still keywords (50%). Apart from this, extension (29%), from contact (16%), date range (24%), day of week (27%), year (23%), month (27%), season (14%), country (18%), region (17%), city (15%) were also used comparatively often. Marginally features used include: path (4.6%), people nearby (3.7%), weather (3.6%), and light status (2.8%). In 30 tasks that the users could select to extend their default baseline interface to the advanced experiment interface, they only did so for 16 of them. The rest of these tasks were found to be comparatively simple (mean=2.2, SD=1.8).

Our findings suggest that users are still used to keywords based search even when advanced search options are provided. That they used some of the extended options indicated a need to include these options in a search system.

Congruent with our hypothesis, we found that searches with extended options are more likely to find desired information than the baseline system ($\chi 2=4.43$, $p < .05$). We calculated the difference between the post tasks rating of difficulty and the rating of difficult rated when generating the tasks. A small but significant advantage was found ($p<.05$). This further supported our hypothesis that providing richer, and in particular, episodic contextual search options can make the search tasks easier.

6. CONCLUSION

In this paper, we went beyond attributes and extensively explored the remembered information of encounter information. We integrated these types of information into our prototype system to allow users search by a variety of information. Due to the difficult of collecting the data, both our validation of recall and the evaluation of the prototypes system are only based on three subjects, who showed considerable difference between each other. The evaluation of the results showed promising results of including rich contextual fields in search systems. With the development of ubiquitous computing, context such as location, weather or people may be captured easily with devices that most people have, e.g. smart phones. The future work may look at how to intergrade such information with current desktop search systems, or to personalize web search engines to provide better support in findings tasks for previously encountered information.

7. REFERENCES

1. Blanc-Brude, T. and Scapin, D. L. What do people recall about their documents?: implications for desktop search tools. In Anonymous *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 2007, 102-111.
2. Byrne D, Kelly L and Jones G.J.F. *Multiple Multimodal Mobile Devices: Lessons Learned from Engineering Lifelog Solutions*. In: Handbook of Research on Mobile Software Engineering: Design, Implementation and Emergent Applications, IGI Publishing, 2012
3. Consolvo, S. and Walker, M. Using the experience sampling method to evaluate ubicomp applications. Pervasive Computing, IEEE, 2, 2 (2003), 24-31.
4. Cutrell, E., Dumais, S. T. and Teevan, J. Searching to eliminate personal information management. Commun ACM, 49, 1 (2006), 58-64.
5. Dumais, S., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R. and Robbins, D. C. Stuff I've seen: a system for personal information retrieval and re-use. In Anonymous *Proceedings of the 26th annual international ACM SIGIR conference.* ACM, 2003, 72-79.
6. Elsweiler, D. and Ruthven, I. Towards task-based personal information management evaluations. In Anonymous *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, 23-30.
7. Gonçalves, D. and Jorge, J. "Tell Me a Story" Issues on the Design of Document Retrieval Systems. Engineering Human Computer Interaction and Interactive Systems, 2005, 117-134.
8. Jaimes, A., Omura, K., Nagamine, T. and Hirata, K. Memory cues for meeting video retrieval. In Anonymous *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences.* (). ACM, , 2004, 74-85.
9. Jones, W. Personal information management. Annual review of information science and technology, 41, 1 (2007), 453-504.
10. Lansdale, M. W. The psychology of personal information management. Appl. Ergon., 19, 1 (3 1988), 55-66. DOI=DOI: 10.1016/0003-6870(88)90199-8.
11. Naaman, M., Harada, S., Wang, Q. Y., Garcia-Molina, H. and Paepcke, A. Context data in geo-referenced digital photo collections. In Anonymous *Proceedings of the 12th annual ACM international conference on Multimedia.* (). ACM, 2004, 196-203.
12. Teevan, J., Adar, E., Jones, R. and Potts, M. A. S. Information re-retrieval: repeat queries in Yahoo's logs. In Anonymous *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, 151-158.