Applying Contextual Memory Cues for Retrieval from Personal Information Archives *

Marguerite Fuller, Liadh Kelly and Gareth J.F. Jones
Centre for Digital Video Processing &
School of Computing
Dublin City University, Dublin 9, Ireland
{mfuller,lkelly,gjones}@computing.dcu.ie

ABSTRACT

Advances in digital technologies for information capture combined with massive increases in the capacity of digital storage media mean that it is now possible to capture and store one's entire life experiences in a Human Digital Memory (HDM). Information can be captured from a myriad of personal information devices including desktop computers, PDAs, digital cameras, video and audio recorders, and various sensors, including GPS, Bluetooth, and biometric devices. These diverse collections of personal information are potentially very valuable, but will only be so if significant information can be reliably retrieved from them. HDMs differ from traditional document collections for which existing search technologies have been developed since users may have poor recollection of contents or even the existence of stored items. Additionally HDM data is highly heterogeneous and unstructured, making it difficult to form search queries. We believe that a Personal Information Management (PIM) system which exploits the context of information capture, and potentially of earlier refinding, can be valuable in effective retrieval from an HDM. We report an investigation into how individuals perform searches of their personal information, and use the outcome of this study to develop an information retrieval (IR) framework for HDM search incorporating the context of document capture. We then describe the creation of a pilot HDM test collection, and initial experiments in retrieval from this collection. Results from these experiments indicate that use of context data can be significantly beneficial to increasing the efficient retrieval of partially recalled items from an HDM.

General Terms

Personal Information Management (PIM), Tools and Techniques in Support of PIM, Human Digital Memories (HDM), Information Retrieval, Context data, Context-based retrieval

INTRODUCTION

Vannevar Bush could never have envisaged the impact his 1945 article 'As We May Think' [10] would have on computing science and society. His article is largely credited with predicting many recent innovations in computing including the World Wide Web. However, Bush presented a vision encompassing far more than the idea of linking pages of information. He foresaw a world where all information associated with someone's life could be stored and, importantly, retrieved at a later stage. Advances in a myriad of digital technologies mean that information relating to an individual's life can now be captured from an ever expanding range of devices including desktop computers, PDAs, digital cameras, video and audio recorders, and various sensors, including GPS, Bluetooth, and biometric devices. These devices can be used to record details of documents authored or read by a person, their communications with other people and machines, as well as the images they see, the sounds they hear and even the mood they experience at particular points in time. Over time a person's digital memory increases in size thus growing in scope and complexity. When captured over an extended period these vast digital archives of one's personal life experiences are more formally referred to as *Human Digital Memories (HDMs)*. Technologies relating to capture and management of HDMs, referred to as Personal Information Management (PIM) tools, form a rapidly growing research area [24].

In recent years, progress in digital storage capacity has been staggering to such a degree that they should be able to satisfy a person's lifetime recording needs. It is estimated that 100 email messages a day (5KB each), 100 web pages a day (50KB each), 5 scanned pages a day (100KB each), 1 book every 10 days (1MB each), 10 photos per day (400KB jpeg each), 8 hours per day of sound - e.g. telephone, voice annotations, and meeting recordings (8KB/s), and 1 new music CD every 10 days (45 min each at 128KB/s) would take five years to fill up an 80GB drive [7]. Utilization and management of the stored data however, is still a significant challenge. HDMs can potentially make it possible for individuals to vividly relive an event with sound and images, enhancing personal reflection. It is probable that they could also enable a person to retrieve important information they once came across but have since forgotten. The human memory is maddeningly elusive; we stumble upon its limitation everyday, when we forget the title of a document we were working on last week etc. Forgetfulness is a significant contributor to the problems surrounding the utilization and management of digital memories, as a user often does not

*

recall sufficient details about an item to retrieve it by traditional IR methods or in some cases they totally forget about the existence of the item. We believe that searching an HDM for a specific event, a person will often be able to reliably retrieve the correct result when key details about the event are remembered - document type, its name, or keywords contained in it for example. However, the quality and quantity of key details a person remembers about a specific event diminishes over time, thus making searching for a specific event in an HDM spanning many years a challenge. However some of the contextual details relating to item capture or subsequent access will often still be recalled, for example who they were with, where they were, what was happening at the time of the event etc. As an example consider the following scenario. *I clearly remember the rain distracting me from my work in the last two months as it pelted against my office window. At the time I was chatting to a friend, Jane, over instant messaging (IM), while simultaneously working on a piece of code.* Associating this contextual information with items would potentially allow for new types of retrieval based on an individual's memory of the item.

An HDM is typically a combination of many types of media, including audio, video, images, and many types of texts, and retrieving information from these archives presents many challenges. There is the potential for a large percentage of noisy data in these archives (such as items that the user would not wish to be captured, or items they will never want to see again); many items in the archive may be very similar, repeatedly covering the same topic; a user may forget about captured items or not even be aware that a particular piece of data was captured, and is therefore available for retrieval; the user may not be able to describe clearly what they are looking for; items may not have formal textual descriptions, meaning that they cannot be retrieved using standard text or meta-tag based retrieval methods; and items may not be joined by inter-item links, meaning link structure cannot be utilized in the retrieval process. It is this unique combination of attributes of HDMs that motivate our research into creation of retrieval techniques specifically for the personal archive domain. This domain is fundamentally distinct from traditional IR domains due to both the above combination of factors, and the fact that items in HDMs are personal to the individual and the individual has personal memories about items related to such things as item creation and subsequent access. We believe memory cues relating to the context of information capture, and potentially the context of earlier re-finding of items, can play a strong role in IR for the HDM domain, within our work we are exploring methods to enable context data to be used for search within HDMs. We are particularly interested in how context can be used to link and annotate items to increase retrieval performance from HDMs.

This paper is structured as follows: the next section reviews studies in memory and context, and summarises existing work in systems for searching HDMs, the following section provides details and results from a perliminary study we carried out to test our hypothesis that context data can improve retrieval from HDMs, and the results and conclusions suggest directions for further research and experimentation.

BACKGROUND - MEMORY AND CONTEXT

Large amounts of data are captured in HDMs, with the result that searching through an HDM to find individual items is difficult particularly when we often only remember a few keywords relating to them. Remembering the context of our experience with items can frequently be easier than remembering these crucial keywords [4].

In a diary study conducted by Teevan et al [34], a participant attempted to locate a document that they knew existed in their file system. Although, they knew exactly what document they were looking for, they could not describe its contents or its location in advance. However, the participant was able to locate the document successfully through a series of small steps, using the local context at each stage of their search to inform them of the next step to take. Teevan et al suggest that the files surrounding the particular file one searches for are important when determining whether a certain file is the correct one. This observation raises the question, how to capture this contextual information and present it in such a manner that it will act as a memory cue to identify the document the participant seeks.

The most common memory problems experienced by people are described in [33]. Of particular interest here, three 'sins' of forgetting are put forward:

- Transience - The gradual loss of memory. Memory is lost, diminishes over time. Example: Being unable to accurately recollect the details of your 21st birthday celebrations.

- Absent-mindedness - The condition of being so lost in solitary thought as to be unaware of one's surroundings. This becomes a sin when one is unable to direct attention to the things that will be needed to be remembered later, or if one encoded relevant things on a level that is too shallow for long-term retention. Example: Daydreaming in lectures, not paying attention to where one puts one's keys.

- Blocking - A failure to retrieve or access deeply encoded information; one's memories are temporarily unavailable. Example: Being unable to recall the name of a former college lecturer because one can only think of another lecturer.

And three 'sins' of distortion are:

- Misattribution - A failure to source-memory; being able to remember the content, but forgetting the actual source of the information and attributing it to

some other source. Example: Reading an academic paper and remembering the gist of the content, but attributing the source to a similar paper read at the same time.

- Suggestibility - This refers to the possibility to 'remember' something while the only basis for this memory is that it was suggested to one by someone else. Example: Being asked by someone to remember something vividly that never happened and 'remembering' the event. This happens in courtrooms every day.

- Bias - Misremembering due to the influence of current knowledge, emotions, beliefs. Usually the recall is selective or distorted. Example: Remembering bad actions of a person that one immensely dislikes better than the good actions of this person.

The above shortcomings provide greater awareness of the human memory pitfalls especially the sins of forgetting. In order to capture as much information as possible that people remember about items and to minimize (or compensate) for forgetting, PIM systems which stimulate remembering and allow the user to follow memory cues are required. These systems exploit humans' recognition and recall processes [29]. Both suitable user interfaces, which enable the user to follow memory cues, and backend context-based search algorithms are required. In [27] we explore the user interface element of this challenge in some detail, and present a sample PIM interface which allows users to search through their HDM using memory cues. In [18] an interesting PIM interface for email access is presented, here the user can locate emails based on memory cues, for example emails are represented as thumbnails which show a picture of the author of the email. A number of PIM interfaces enabling users to more easily locate files are presented in [13]. The remainder of this paper explores suitable IR techniques for PIM systems. To this end the remainder of this section looks at existing studies which have investigated the elements of files that people tend to remember, and the current use of context data in PIM retrieval systems.

Remembered Context

In recent years researchers have begun to explore the possibility of using various types of context data to aid retrieval. Standard forms of context data such as time, date, number of accesses, etc have proved beneficial in retrieval from various collections (e.g. [19], [31], [32]). Context data allows us to harness this crucial component of memory for recollection of items we wish to re-retrieve from a HDM. Elsweiler et al [17] put forward the hypothesis that memory lapses make it difficult for people to find items in their personal archives using traditional IR systems, because they require them to remember enough information about the item they are looking for to perform a content-based query. They carried out an extensive user study to help them understand how people recover from memory lapses, from which they found evidence to support the use of context data in HDM retrieval.

Research has investigated how individuals manage their personal information; the steps they execute when searching for desired documents, the attributes they actually recall about their documents and the characteristics of their recall. Over a decade ago, Barreau and Nardi [6] summarized two studies of the way individuals organize and find files on their computers. The first study investigated information organization practices among users of DOS, Windows and OS/2, while the second examined the finding and filing practices of Macintosh users. Both studies concluded that individuals preferred location-based search over text-based search for finding files. Location-based searching involves the individuals taking a guess at the directory/folder where they think the file might be located, going to that location and browsing the list of files in the location until they find the file they are looking for. The process is iterated as needed. Barreau and Nardi reported that users in their study preferred browsing a list of files rather than trying to remember the exact file names. They hypothesized that users prefer location-based filing because it more actively engages the mind and bestows a greater sense of control. They noted that individuals preferred filing by location because it aided in helping them find what they need as well as serving a crucial reminding function.

In another study [8] further evidence supporting the use of context data in retrieval is presented. In this study they found the following to be attributes that people use to relocate files:

- Location: name and path direction
- Type of Format: Word, PDF, PowerPoint etc
- File name
- Title: title within the document
- Size: in terms of pages, numbers of lines
- Time: time of last usage of the document
- Keywords: meaningful words within the document
- Visual elements: existence of graphics, tables, colours within the document
- Associated event: significant events that occurred in association with the last usage of the document e.g. emails, telephone calls
- Links: documents that are related to the target document e.g. previous versions
- Actions: operations performed on the document by the user e.g. printing, copying, paste, cut, insert.

This list of context data can be extended by incorporating findings from research undertaken in [20][34] which

found (among other things) local time, daylight status, weather, and location of individual to be good sources of contextual information. This provides a balance to the above list of context data, as before, it mainly concentrated on context data regarding documents themselves. The identified contexts now stand at: location of individual; source of file; weather; type; time of day; format; month; keywords; year; size in pages; location of file; status draft; name of file; topic; title of file; actions preformed on the document backup, printed etc; Period of day morning, afternoon, evening.

The above list differs from our earlier list of attributes as it now incorporates contexts outside the computer i.e. the weather and location of an individual, and also properties mentioned in Presto [14].

Existing Systems

Others have previously looked at developing systems to allow individuals to search through their personal collections. In this section we give examples of some of these systems.

As part of the *MyLifeBits* project [7][23] Gorden Bell has dedicated much time to digital capture of all of his personal data. This project has also developed interfaces to help people browse and search through their HDMs. In similar work the *Haystack* project [16] investigates methods to allow users to organize all their personal information in whatever way makes most sense to them. This system removes the barriers that normally exist between different types of items, such as email and photos for example. The *LifeStreams*[21] system offers an alternative approach to existing personal file organizational hierarchies by using a time-ordered stream as a storage model and stream filters to organize locate, summarize and monitor incoming information. These streams exist for the past, present and future. In *Life Log* [4] a person's life is continuously recorded. Information captured includes video, audio, acceleration sensors, gyro, Global Positioning System (GPS), annotations, documents, web pages and emails. This project also examines the use of context and content data for retrieval.

Presto [14] is a prototype system enabling users to organize their documents entirely in terms of document properties. Properties are name/value pairs e.g. author = bloggs. Properties emphasized in the paper included: author, topic, type, status, format, backup and conference. The properties provided a uniform mechanism for managing, coding, searching, retrieving and interacting with documents. The prototype aimed to offer an alternative to the rigid hierarchical schemes that most information management systems rely on to organize documents. A benefit of this approach is that properties were directly associated with documents, rather than with document storage locations. If documents are moved from one place to another, they still retain their properties. A major feature of Presto, was that end users did not have to find a unique name for a document or a place to store it where it was likely to be found again. Instead the application automatically tagged data items with relevant contextual information, any of which could be used to retrieve documents again. 'Mail.From = dourish' and 'read within 1 month' was provided as an example query used to retrieve information. Presto illustrates that the context data surrounding a document can be used in its retrieval.

Stuff I've Seen [15] is an IR system that facilitates information re-use. The system provides a unified index of information that a person has seen such as email, web page, documents, appointments. Since one has seen the information it enables rich contextual cues to be built into the search interface. It was found that filters such as date and type were frequently used to hone in on relevant items. Also, date and people names in particular provided rich contextual cues for retrieval.

In [28] the usefulness of different types of context data for the personal photo retrieval task is examined. In particular they looked at the usefulness of the following types of context data for a photo browser interface: location, time of day the photo was shot, the light status, weather status and temperature, and events in time. Results from their survey showed that context involving location and people were rated highly in both usefulness for retrieval and also being well remembered. Also, location and time-derived context such as weather, local time, daylight status, and season proved useful. They noted that local time and daylight status seemed to be stronger cues than date/time.

PILOT USER STUDY

In this section we describe a perliminary experimental investigation using context with content for searching an HDM. The results of this study suggest that context data is beneficial in HDM retrieval. We first motivate and describe the experiment and test collection used for this study, and then give details of the user study and results obtained.

Experimental Methodology - Motivation

Creating a New HDM

A small experimental HDM was created for this experiment. This was necessary since we didn't have access to a suitable HDM archive, but also for other important reasons. Since the memories in an HDM belong to the HDM owner, the owner of the HDM should be the subject of retrieval experiments. Also, as HDMs are personal to an individual they may contain sensitive information about a person, thus making it inappropriate to use someone elses HDM for this experiment.

Choice of Recorded Event Types

Each action performed on the individual's computer was recorded as an event, e.g. viewing a pdf file or engaging in an IM chat. For each event we recorded the contents of the event along with the following types of

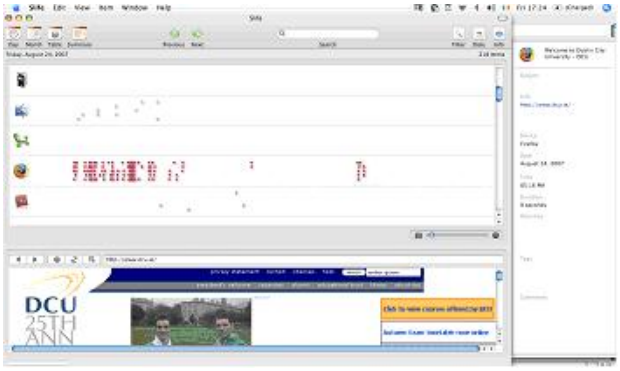

Figure 1. Slife in Action

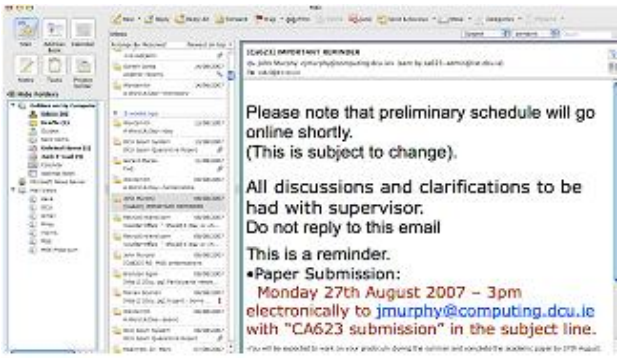

Figure 2. Overall view of Inbox

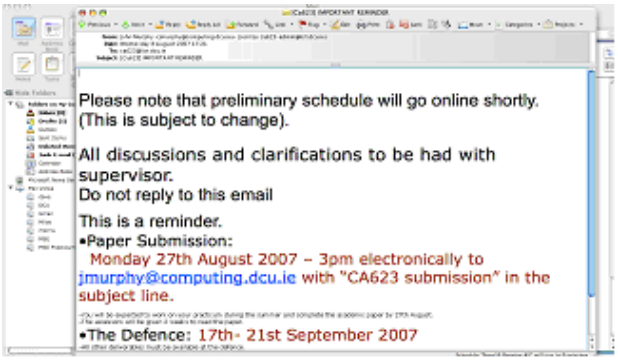

Figure 3. Email encapsulated its own window.

context data: date, month, year, hour, minute, weekday (e.g. Mon, Tues) of the event; file title; file source (e.g. Word, Firefox); file type (e.g. document, chat, web); event duration; surrounding events types; surrounding events sources; surrounding events content; location of file access (e.g. kitchen, lab); weather conditions at time of file access (e.g. showers, cloudy); season at time of file access (e.g. Spring, Winter).

Additional sensors can be used to gather further data about users' context. For example, BodyMedia devices [2] can be used to record biometric response, a Microsoft SenseCam [7] can be used to take images of a person's surroundings, or mobile devices equipped with Bluetooth can record who was in a persons company [11]. However, due to time constraints it was not possible to include context data relating to the individual's biometric response, information obtained from a SenseCam or information obtained from Bluetooth devices in this collection.

Recording of Computer-User Interaction

The Slife system [3] was used to capture the activities on the individual's Mac personal computer. Slife is similar to the MyLifeBits logging software [22] for the Windows OS but runs under the Mac OS X. It both records and can graphically display user computer activity/events. An event is generated each time a window is brought to the foreground. For each event Slife records: time and date accessed, the duration the window was open for, events linked to (if applicable), type of application e.g. web, chat, document, source e.g. Microsoft Word, window title, and the textual content inside the window. Slife can monitor interactions from more than 15 applications, including web browsers, email clients, text editors, instant messaging clients and media players. It is also possible to write customized scripts for Slife, using AppleScript [1], that fetch data e.g. scripts that collect data for a specific event or day, range of dates etc. Captured activities can be browsed using a graphical interface supplied as part of Slife. Figure 1 shows the browsing interface.

Choice of Retrieval Search System

Having collected our data, the Lucene [5] search system was then used for retrieval of items from it. Apache Lucene is a mature open source text search engine library implemented in Java[1]. Lucene provides a simple yet powerful core API that allows indexing and searching capabilities to be added to applications. Search features include ranked searching, where best results are returned first, powerful query types, phrase queries, wildcard queries, proximity queries, range queries, fielded searching e.g. title, author, content, sorting by field, date-range searching and the ability to simultaneous update and search. Lucene's logical architecture allows for its API to be agnostic of file format. Text from PDF and HTML files, Microsoft Word documents, as well as many others can be indexed, so long as their textual information can be extracted. The core concepts behind its architecture are: index, field, and term. An index contains a sequence of documents. A document is a sequence of fields. A field is a named sequence of terms and a term is a string. The index stores statistics about terms in order to make term-based search more efficient. Lucene's index falls into the family of indexes known as inverted files. Indexing facilitates rapid searching as it converts original data into a highly efficient cross-reference lookup.

[1] http://lucene.apache.org/

Data Collection

One subject's computer interaction was recorded over a period of six weeks using Slife [3]. The basic Slife was augmented by writing scripts to support 'Finder', 'Terminal' and 'Script Editor' applications - the 'Finder' script captured when a particular folder was opened; the 'Terminal' script captured the details typed on the Terminal command line; the 'Script Editor' script captured the contents of scripts written. The following additional context data was also created for events:

- Using time and date, information functions were written to determine, the hour, minute, second, season and period of the day in which the event took place e.g. morning, afternoon, evening, night.
- The file name and location of documents viewed with the 'preview' application.
- Geographic location, e.g. office, lab, etc was captured using the Experience-Sampling Method (ESM) [12]. More specifically the user was prompted to enter their current location on computer startup and at varying intervals during computer use. While this method requires a minor burden to be placed on the user, the technique sufficed for this small scale single user study. On a larger scale study GPS devices could be used to automatically determine a person's outdoor location [31]. And Bluetooth stations located at appropriate positions indoors along with Bluetooth enabled devices e.g. mobile phone, could be used to automatically determine a person's indoor location at a particular point in time [4][7]. The test cases were generated in a fashion similar to the Day Reconstruction Method (DRM) introduced in [26].
- Weather data. Weather history was obtained from [25], and parsed to extract the weather history for each hour of a day.
- Linking of related events. This process is described in detail in the next section.

We created a script which generated XML files containing details of events and associated context. The XML files were parsed to extract event information and create event objects. Slife, captures the links to WebPages that were viewed, it does not however, store the contents of these pages. During the instantiation of Event Objects it was verified to see if the event was of type Web, and if so the content of the viewed Webpage was downloaded and then added to the Event Object with the HTML tags removed. Events were also checked during instantiation to see if they were of type PDF. If they were, the textual content was extracted from the PDF file and added to the Event Object.

Issue with Slife

The logging nature of Slife presented a few issues. While working, users may often rapidly switch between windows; changing which one appears in the foreground. Each time a window is brought to the foreground Slife logs it as a new event, even though it usually would not be considered so by the user. This problem is illustrated in Figure 4, where a lot of time is spent working on a Microsoft Word document. A user would generally class this as one activity. The screenshot gives the impression of one continuous event, however on closer inspection, this is actually a string of closely timed events, as shown in Figure 5. It is apparent in this blownup image from a portion of Figure 4 that each switch of window displayed in the foreground has actually been classed as a distinct event, when in fact only one event occurred. This form of logging presented a serious issue in our experimentation, as searching for this particular document would result in the retrieval of possibly 100s of copies of the same document. To overcome this problem we checked the filename of the event, if the filename already existed in the collection the previous event was removed and replaced with the new event. One problem with this solution is the existence of situations where it is useful to have different versions of the same document. Such as when roll back changes need to be made. We plan to support this facility in a revised version of the application.

Figure 4. Working on the same document.

Figure 5. Blown up section of Figure 4.

Event Linking and Indexing

Linking has great potential to assist in context-based search. We have explored this idea concepually in more detail in previous work [27]. In this paper we make some initial steps to investigate and implement the usefulness of linking items to incorporate context in the search process. For this experiment we looked at linking, or relating items, based on time proximity, thus allowing searches of the form *'when I was working on the document I'm looking for I received an email from Jill'*. This approach to linking events, for a given day,

consisted of the following:

1. The events of the day were sliced into different sections. These sections comprised bursts of activity that occurred around the same point in time. The captured data was analyzed to determine if the time between events exceeded 30 minutes, if this was true, the events grouped into different slices.

2. Three attributes were added to the Event Object, an attribute for the type of events in the slices (e.g. surrounding events types = ['chat', 'document']), another attribute for the source of events (e.g. surrounding events sources = ['Excel', 'Word']), and a final attribute that contained the content of these events.

3. Following event slicing/partitioning, events were annotated with the weather and location context data. To do this all events in a given partition were annotated with the weather and location data occurring at the median time of the partition.

Table 1 provides a summary of the complete set of context data associated with each event. This context data forms part of the index of an event. In Lucene this is referred to as the fields of a document. The remainder of an event index is derived from the content of the event, using the StandardAnalyzer built into Lucene. The StandardAnalyzer tokenizes fields based on a sophisticated grammar that recognizes email address, acronyms, alphanumeric and more; converts capitals to lowercases, and removes stop words.

General Information	
Event ID	Event content
Context Data	
Day	Month
Title	Year
Source e.g. Word, Firefox	Hour
Type e.g. document, chat, Web	Minute
Location e.g. college, kitchen	Weekday e.g. Mon, Tues
Weather e.g. showers, cloudy	Surrounding Events Types
Season e.g. summer, winter	Surrounding Events Sources
Duration	Surrounding Events Content

Table 1. Complete set of Content & Context

The investigation into the usefulness of context data in the retrieval process began upon completion of the indexing phase. This investigation involved the generation of a set of test case scenarios and a set of queries; the generated queries were then applied to the test case data and results for each query collected. The queries were evaluated using a known-item search scenario: given a known remembered item, can it be successfully retrieved from the collection.

Test Case Generation

Comparing the relative effectiveness of different retrieval approaches through laboratory experiments is a well-established tradition in IR. Experimenters use several rules-of-thumb to evaluate IR systems, such as the number of queries needed for a good experiment. Buckley and Voorhees [9] validate that at least 25 queries, although 50 is better, are needed for a good experiment. Using too few queries produces unreliable averaged results, since they may be overly influenced by results for individual queries.

Thus a set of 30 test case scenarios were created from the participant's memory without looking at the data set. These test case scenarios were drawn from events that covered the time span of the data collection. The collected data ranged from the middle of July 2007 to the end of August 2007. To generate these test cases the test participant identified the key events that occurred during the six week data collection period, these included: a friend's birthday, a Christening, several meetings in DCU, and dinner in a restaurant. Following this, the participant was required to recall the activities preformed on their computer around or close to these key events. The following is an example of a typical recall: *'I was in work on a Friday in July, I remember thinking I couldn't find my umbrella that morning, and I hoped it wouldn't rain. I decided to meet friends in the Golden Lion restaurant for dinner after work. It was fairly late when I went to get the bus home, however, it wasn't the last bus. On the way to the bus stop, I remember thinking how lucky I was that it wasn't raining, as I didn't have my umbrella with me. On returning home, I logged onto the computer in my office. My friend, Sarah, was online, I remember telling her about the tasty curry I had for dinner'.* Each test case then consisted of a remembered scenario and computer file (email, word document, IM, web page, pdf) acessed around or close to the scenario.

Query Types

After establishing the 30 test case scenarios, the data from the scenarios was converted into queries. Different types of queries were constructed to assess the usefulness of the various types of context data on their own and also in unison. Eight types of queries relating to the document were constructed.

Query One: Content only

Query Two: Context only, this incorporated, the following fields: title, source, type, location, weather, year, month, day, hour, minute, weekday, period, surrounding event types, surrounding event source, surrounding event content, duration, season.

Query Three: Combination of content and context.

Query Four: Combination of content and time, i.e. hour, minute, weekday, day, month, year, season and period of the day.

Query Five: Combination of content and weather.

Query Six: Combination of weather and location.

Query Seven: Combination of content and type and source.

Query Eight: Combination of content and surrounding type, content and source.

Query type one represents the current standard approach for retrieval using search engines. Results generated from this content only query were used as the benchmark.

Querying

The goal of known-item querying is to retrieve the correct file for a given remembered scenario. The 8 sets of 30 were entered into Lucene, via the command line. Upon entering a query, the query was processed using the same analyzer used for indexing to convert the queries into terms to facilitate searching. After the query processing Lucene returned a set of documents matching the query. Examples of queries include:

Query One: content: (+ curry + golden + lion)

Query Two: type: chat source: adium year: 2007 month: July period: night weekday: Friday season: summer location: office weather: cloudy surType: web surSource: firefox

Query Three: content: (+ curry + golden + lion) type: chat source: adium year: 2007 month: July period: night weekday: Friday season: summer location: office weather: cloudy surType: web surSource: firefox

Query Four: content: (+ curry + golden + lion) year: 2007 month: July period: night weekday: Friday season: summer

A typical querying scenario would be where the user searched for an email they recalled writing on a 'Thursday' with the word 'Friday' in the content of the mail. This search would be matched by events with the following index terms: content index term: 'Friday', day index term: 'Thursday', source index term: 'email application'.

Results

For each query, the rank of the target document in the list returned by Lucene was noted. The Mean Rank and the Mean Reciprocal Rank were then calculated for each query set. The Mean Rank as its name suggests, is the average rank of the target documents. The reciprocal rank of a query response is the multiplicative inverse of the rank of the correct answer. The mean reciprocal rank is recommended as a measure of evaluation for known item search as it is not severely influenced by target documents retrieved at low ranks. Figure 2 shows the results obtained. The combination of content and context performed very well, for nearly two thirds of the queries the target document appeared top in the list of documents returned. This combination performed considerably better than content only with an increase of 0.26 in the results. Content and time, and content and location also performed well. Context only, content and weather combinations preformed under the benchmark of content only. Possible reasons for this are: for context only, the circumstances surrounding when a document was viewed was not unique to that document and applied to all the documents viewed at the same time as it; for content and weather, the time span of the collection only spanned the summer season, and the weather experienced was consistently cloud and rain, thus not showing much variation. The content and surrounding topic, content and source queries performed slightly under the benchmark, this suggested that a new annotation technique is needed for capturing the surrounding events. It is probable that too many documents had the same surrounding events; perhaps reflecting the participant developing a certain routine when logging into and using their computer. The routine typically involves first looking at email, then saying hello to friends on IM, and finally checking out websites of interest. These events tend to happen multiple times each day and in the same order.

Query Type	Mean Rank	Mean Reciprocal Rank
Content only	117.33	0.44
Context only	106.97	0.33
Content & Context	2.433	0.70
Content & Time	22	0.55
Content & Weather	180.46	0.33
Content & Location	12.06	0.54
Content & Type and Source	173.26	0.46
Content & Surrounding Type, Content and Source	301.17	0.42

Table 2. Experimental Results

The combination of content & context performed very well, for nearly two thirds of the queries the target document appeared top in the list of documents returned.

Considerations

The above results are promising, however, a few factors need to be taken into account when considering their authority and extensibility to larger data sets.

- The collection was limited in size; it only spanned six weeks, HDMs have the potential to span many

years, a lifetime. Searching involved a few thousands documents compared to possible millions in a lifetime HDM.

- The collection contained recent events; the contexts of these events, such as location, weather and time were relatively fresh in the participant's memory.
- The results are specific to one person. A wider set of participants would be needed to gauge the full potential of this IR technique.
- Lucene sorts documents returned by score and document number. Therefore documents which have the same score are sorted by document number. Since the document numbers increase as documents are added to the index, older documents have higher ranking than newer ones. The side effects of this need to be investigated.

Result Conclusions

Despite the above mentioned shortcomings of the perliminary user study, the results obtained are promising. They show that remembered context combined with content only retrieval improves results in the HDM domain. Further investigation using more participants, over a longer time frame will test this hypothesis further.

In future experiments we also need to investigate how easy such context cues would be for a user to recall and employ in 'real terms'. In our test environment the experiment subject had a considerable amount of time to think about events and remembered context. However in real world operation a user needs to obtain required information as quickly and effortlessly as possible from the system. Experimentation needs to verify what and how much context data a person can easily recall a significant time after an event has occurred. We hypothesize that memory cues provided by a system may aid and trigger this remembering process. The challenge then becomes to develop an interface that provides such cues. This may consist of suggesting prompts such as 'was it raining when you previously saw this item?'; Or possibly simply providing fields for the available types of context data may prove a sufficient memory cue. Some work has already been carried out in this area with the MediAssist system [30]. In [27] we provide one possible solution to this challenge in an interface that allows a user to browse through their HDM using memory cues.

CONCLUSIONS AND FUTURE WORK

Research into HDMs is growing at a rapid rate. In this paper we have presented a proposed solution to retrieval from these often vast personal archives, using context to harness what people remember about personal information. Our pilot user experiment supports the notion of using a combination of context and content data for HDM retrieval. A larger user study will provide more concrete results. Our pilot also illustrated that time and location serve as valuable contexts, this observation is consistent with findings in the MediAssist paper [30]. MediAssist enables organization and searching of personal digital photo collections based on contextual information, content-based analysis and semi-automatic annotations. The results of their evaluation clearly showed the strength of indexing by time and location.

In the future it would also be interesting to investigate the use of other richer sources of context data, such as social and biometric context. Additionally we would like to further investigate the time slicing technique used for linking items and explore other methods of linking items within HDMs. Other issues which could be explored include:

- The expansion of the query set and the construction of different types of queries, such as range queries, multi-field queries, etc.
- The incorporation of boosting for documents, fields, and queries. The duration one spends doing an activity should factor into how important that document is. If one spends a duration of 0.1 seconds looking at a website for example, the likelihood of that website being important is slim.
- Linking of queries to documents. If a document were selected for a specific query, that query could be linked to the document as additional annotating metadata, therefore improving future searches for the document if a similar query is entered.
- The benefits of keeping a digital diary of events that happen in one's life (e.g. meeting with funding body, friend's wedding), and using the information stored in this diary to link to items in the HDM. This would enable the retrieval of HDM items based on a specific event, e.g. all items that occurred within one day of the due date of this paper.

Context data can be used to harness the way people remember items in their HDMs. If this context information is exploited correctly, we believe it will be possible to create a system that retrieves items based on both an individual users unique information needs and on what they remember about items. We envisage a system which responds and adapts to the user; a system which works and evolves with the user's needs and the way they remember information. HDMs are increasingly becoming part of our present. In the near future we believe it will be hard for people to imagine a world where HDMs did not exist. We are embracing this future by developing retrieval techniques that address the unique retrieval requirements of HDMs.

ACKNOWLEDGMENTS

This work is funded by a grant under the Science Foundation Ireland Research Frontiers Programme 2006.

REFERENCES

1. Apple -Mac OS X-ApplesScript, Copyright ©2007 Apple Inc. viewed July 2007, http://www.apple.com/macosx/features/applescript/.

2. BodyMedia - http://www.bodymedia.com.

3. Slife Labs - http://www.slifelabs.com.

4. K. Aizawa, D. Tancharoen, S. Kawasaki, and T. Yamasaki. Efficient Retrieval of Life Log Based on Context and Content. In *The 1st ACM Workshop on Capture, Archival and Retrieval of Personal Experiences (CARPE 2004)*, pages 22–31. New York, New York, USA, October 2004.

5. Apache Lucene. Overview, Copyright ©2006 The Apache Software Foundation. http://lucene.apache.org/java/docs/.

6. D. Barreau and B. Nardi. Finding and reminding: file organization from the desktop. *SIGCHI Bull*, 27(3):39–43, 1995.

7. G. Bell. Challenges in Using Lifetime Personal Information Stores based on MyLifeBits. Alpbach Forum, 26 August 2004.

8. T. Blanc-Brude and D. Scapi. What do people recall about their Documents? Implications for Desktop Search Tools. In *IUI'07*. Honolulu, Hawaii, USA, January 2007.

9. C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd ACM Annual International Conference on Research and Development in Information Retrieval*, 2000.

10. V. Bush. As We May Think. *The Atlantic Monthly*, pages 101–108, 1945.

11. D. Byrne, B. Lavelle, A. Doherty, G. Jones, and A. Smeaton. Using Bluetooth and GPS Metadata to Measure Event Similarity in SenseCam Images. In *IMAI'07 - 5th International Conference on Intelligent Multimedia and Ambient Intelligence*, pages 18–24. Salt Lake City, Utah, USA, July 2007.

12. M. Csikszentmihalyi and R. Larson. Validity and reliability of the Experience-Sampling Method. *J. Nerv. Ment. Dis.*, 175(9):526–536, September 1987.

13. E. Cutrell. Search User Interfaces for PIM. In *Personal Information Management, SIGIR 2006 Workshop*, pages 32–35. Seattle, Washington, 2006.

14. P. Dourish, W. Edwards, A. LaMarca, and M. Sailsbury. Using Properites for Uniform Interaction in the presto Document System. In *UIST'99*. Asheville, NC, 1999.

15. S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins. Stuff I've seen: a system for personal information retrieval and re-use. In *SIGIR '03: The 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 72–79, New York, NY, USA, July 2003. Toronto, Canada, ACM Press.

16. E. Adar and D. Kargar and L. A. Stein. Haystack: per-user information environments. In *The eighth international conference on Information and knowledge management (CIKM 1999)*, pages 413–422, 1999.

17. D. Elsweiler, I. Ruthvan, and C. Jones. Dealing with Fragmented Recollection of Context in Information Management. In *Context-Based Information Retrieval (CIR-05) Workshop in Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05)*, 2005.

18. D. Elsweiler, I. Ruthvan, and L. Ma. Considering Human Memory in PIM. In *Personal Information Management, SIGIR 2006 Workshop*, pages 49–52. Seattle, Washington, 2006.

19. D. Elsweiler and I. Ruthven. Towards task-based personal information management evaluations. In *Proceedings of the 30th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '07)*, 2007.

20. D. Elsweiler, I. Ruthven, and C. Jones. Towards Memory Supporting Personal Information Management Tools. *Journal of the American Society for Information Science and Technology*, 2007.

21. E. Freeman and D. Gelernter. Lifestreams: A Storage Model for Personal Data. *SIGMOD Record*, 25(1):80–86, March 1996.

22. J. Gemmell, G. Bell, and R. Lueder. MyLifeBits: A Personal Database for Everything. *Communications of the ACM. Personal Information Management*, 49(1):88–95, January 2006.

23. J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. MyLifeBits: Fulfilling the Memex Vision. In *ACM Multimedia '02*. Juan Les Pins, France, December 2002.

24. J. Gemnell and H. Sundaram, editors. *CARPE 2004 The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, 2004.

25. Ireland Weather. Hourly weather history for DUBLIN AIRPORT, Ireland. http://www.freemeteo.com.

26. D. Kahneman, A. Krueger, D. Schkade, N. Schwarz, and A. Stone. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *SCIENCE*, 306:1776–1780, December 2004.

27. L. Kelly and G.J.F. Jones. Venturing into the Labyrinth: the Information Retrieval Challenge of Human Digital Memories. In *Proceedings of the Workshop on Supporting Human Memory with Interactive Systems, at HCI 2007: The 21st British HCI Group Annual Conference*, September 2007.

28. M. Naaman and S. Harada and Q. Wang and H. Garcia-Molina and A. Paepcke. Context Data in Geo-Referenced Digital Photo Collections, 2004.

29. D. Norman. *The Design of Everyday Things*. The MIT Press, London, England, 1998.

30. N. O'Hare, C. Gurrin, G. J. F. Jones, H. Lee, N. E. O'Connor, and A. F. Smeaton. Using text search for personal photo collection with the mediassist system. In *Proceedings of ACM-SAC'07*, Soeul, Korea, 2007.

31. N. O'Hare, H. Lee, S. Cooray, C. Gurrin, G. Jones, J. Malobabic, N. O'Connor, A. Smeaton, and B. Uscilowski. MediAssist: Using Content-Based Analysis and Context to Manage Personal Photo Collections. In *CIVR2006 - 5th International Conference on Image and Video Retrieval*, pages 529–532. Springer-Verlag, July 2006.

32. M. Ringel, E. Cutrell, S. Dumais, and E. Horvitz. Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores. In *Interact 2003*, 2003.

33. D. Schacter. The Seven Sins of Memory: Insights from Psychology and Cognitive Neuroscience. *American Psychologist*, 54(3):182–203, 1999.

34. J. Teevan, C. Alvarado, M. Ackerman, and D. Karger. The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Directed Search. In *CHI 2004: Conference companion on Human factors in computing systems*, pages 415–422. Vienna, Austria, April 2004.