The Information Retrieval Challenge of Human Digital Memories

Liadh Kelly
Centre for Digital Video Processing, School of Computing, Dublin City University, Dublin 9, Ireland
lkelly@computing.dcu.ie

Abstract

Today people are storing increasing amounts of personal information in digital format. While storage of such information is becoming straight forward, retrieval from the vast personal archives that this is creating poses significant challenges. Existing retrieval techniques are good at retrieving from non-personal spaces, such as the World Wide Web. However they are not sufficient for retrieval of items from these new unstructured spaces which contain items that are personal to the individual, and of which the user has personal memories and with which has had previous interaction. We believe that there are new and exciting possibilities for retrieval from personal archives. Memory cues act as triggers for individuals in the remembering process, a better understanding of memory cues will enable us to design new and effective retrieval algorithms and systems for personal archives. Context data, such as time and location, is already proving to play a key part in this special retrieval domain, for example for searching personal photo archives, we believe there are many other rich sources of context that can be exploited for retrieval from personal archives.

Keywords: Personal Information Management, Human Digital Memories, Context data, Context-based retrieval

1. INTRODUCTION

Vannevar Bush could never have envisaged the impact his 1945 article *'As We May Think'* [1], in which he presented his *Memex* vision, would have on modern science. This article is largely credited with proposing ideas that would lead to the development of the World Wide Web. However Bush presented far more than the idea of linking pages of information. He provided a vision for a world where all a person's personal information could be stored and importantly retrieved at a later stage. With advances in modern technology Bush's ideas are now coming to be realized.

Recent years have seen individuals storing increasing amounts of personal information in digital format. We have now reached the point where all of a person's personal life experiences can be stored digitally – everything from items read, written, or downloaded; to footage from life experiences, e.g. photographs taken, videos seen, music heard, details of places visited, details of people met, etc. While much attention has been given to the generation of these vast personal archives (Human Digital Memories (HDM)), e.g. [2][3], less attention has been given to retrieval of information from them.

HDMs are fundamentally different from traditional content archives for which existing retrieval techniques have been developed, in that: an HDM is typically a combination of many types of media, audio, video, images, and many texts of textual content; there is the potential for a large percentage of noisy data in these archives; many items in the archive may be very similar, repeatedly covering the same topic; a user may not 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 could not be utilized in the retrieval process. It is this unique combination of attributes of HDMs that motivate this 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. These factors combined lead to the requirement of new retrieval techniques specific to this domain.

We make the following definitions, for use throughout this paper:

- Item/file – throughout this paper these terms are used interchangeably where each refers to any type of file stored on the persons computer (e.g. document, email, photograph, audio file, motion picture etc).
- Human Digital Memory (HDM) – is a collection of all digital items stored by the individual.

This paper is structured as follows: Section 2 describes existing work related to this domain. Section 3 describes the scope of this project, some questions we have and proposed solutions. Finally, Section 4 gives conclusions from the project so far, and highlights some issues for open discussion.

2. RELATED WORK

In recent years, Microsoft's Gordon Bell has invested much time in the storing part of Bush's vision by the digital capture of all of his personal data, as part of the *MyLifeBits* project [2][3]. He has captured everything from letters, books, CDs, to items viewed on computer, phone conversations etc. Beyond the capture of personal data, researchers have begun looking at how people might retrieve from these vast personal archives, the remainder of this section examines this further.

2.1 Context in personal file retrieval

MyLifeBits [2][3] and *Stuff I've Seen* [4][5] are systems created at Microsoft to allow retrieval of personal files. These systems associate certain types of context data with items in the HDM in order to allow retrieval based on memory. For example, if you remember the date an item was created you could retrieve based on this. *MyLifeBits* in particular uses standard context data such as location, people or date information. This context data, in addition to being used for retrieval, is used to link items, for example two photographs taken in the same location could be associated with each other. *Stuff I've Seen*, again uses standard forms of context data, such as date, author etc. An extension of *Stuff I've Seen*, *Phlat* [6], uses an extra form of context data, namely tagging. Tagging allows users to add tags (keywords) to items at their discretion, these tags can then be used in future retrieval. *Phlat* by allowing users to add their own tags to items enables users to retrieve based on more memory cues that may be useful to them, this approach though has its limitations in that the burden for tagging is placed on the user, more efficient would be a system that did not place such a burden on the user. In other work, *MediAssist* [7] associates context data such as time, location, people and weather conditions with photos to allow people retrieve based on their memory of photos. While these systems have made the first steps towards the use of context data to allow people retrieve from their personal archives, based on what they remember about items, they are quite limited in that they only capture a subset of the many ways people remember items. We believe there are many other forms of context data that could be captured and used in this domain, this is discussed further in the next section.

As mentioned above *MyLifeBits* uses common item attributes to link items together, thus allowing for the association of items with each other. Another system, *Connections* [8][9], also links files together. Here though the linking is based on the patterns of user file access, from which a relation graph is formed using successor models. This allows for the ranking of results of a user text-based query using algorithms such as *PageRank* [10]. This method stems from the success of *PageRank* in the *Google* search engine, which uses the webs link structure to determine important web pages related to a user's query. The results in [8] and [9] seem quite promising and these techniques appear to be a good avenue for further investigation for retrieval from HDMs.

2.2 Interfaces for personal retrieval systems

Interfaces in this domain are also beginning to move away from standard desktops and filing systems. *MyLifeBits* uses simple interfaces based on timelines and standard text based searching. *Stuff I've Seen* has an interface that takes advantage of the fact that cues such as author, time, thumbnails or previews of the item can help trigger a user's memory. *Haystack* [11] allows users to organize all their personal information in whatever way makes most sense to them. This system, similar to the *MyLifeBits*, *Stuff I've Seen* and *Phlats* systems, removes the barriers that normally exist between different types of items, such as email and photos for example. *Lifestreams* [12], replaces the standard desktop, with an interface that arranges items in time-order. Using this interface a person can filter and order items among other things. The system in [13] adds components to existing application interfaces to attempt to improve on existing filing systems. It automatically groups items into tasks and folders by determining what task a user is engage in. This is achieved by recording both personal items viewed and the context in which they are used. It also uses analysis of the user's current task to predict the next folder that the user may wish to access information from. Similar to the current limited context problem, we believe these interfaces only begin to address the interface requirements of the HDM domain, where there is a strong need for more intuitive interfaces that allow the user to follow more memory cues.

The key idea common to all current efforts in the personal information management domain, is to create systems that make it easier for users to (re-)access personal information. Commercial systems which help people search

through their personal files, while not as sophisticated as some of the systems already mentioned, are also showing signs of beginning to address the personal file retrieval problem. Good examples here are: *Google Desktop* [14] which indexes all items a person views saves on their computer. Users can then perform text based queries to retrieve items at a later stage. *Microsoft's Windows Desktop Search (WDS)* [15] which also indexes files on a person's computer and allows them easily query the index from the Windows taskbar. *i-sho* [16] organizes all a person's (or group of people's) personal data in a horizontal time-line interface. More precisely the interface is like a digital diary, grouping items by the timestamp on them. A separate timeline (layer in the application) can be created for different categories. For example, users can then choose to perform a search or view all photos taken in a given day, month or year. These systems while not overly sophisticated highlight the strong move towards, and need for, improved retrieval systems in the personal archive domain.

2.3 The importance of memory in HDMs

People's memory, or lack thereof, of items in their HDMs is acknowledged as being crucial for effective retrieval and the design of effective HDM interfaces. Key benefits from a greater understanding of how people remember, or forget, items would then be, a greater understanding of what context data and interfaces might prove beneficial in aiding people recover, rediscover, or even discover items in their personal data archives. Elsweiler et al. [17][18] have carried out an extensive user study to help them understand how people recover from memory lapses. They put forward the hypothesis that memory lapses make it difficult for people to find items in their personal archives, because traditional information retrieval (IR) systems require people to remember enough information about the item they are looking for to perform a query. From this study they found that:

- People implicitly make others aware of things they have forgotten by talking about past events.
- Around 40% of action-slips were caused by the user trying to multitask.
- Poor encoding in memory of information required in the future is one of the main causes of memory lapses.

They also found that people use different mechanisms, such as the use of bookmarks, to overcome memory weaknesses. Preventative measures such as leaving objects in specific places as reminders of tasks to be performed are also exploited. It was found that individuals use many different techniques to recover from memory lapses. Findings from the above study can be taken into consideration when designing an efficient HDM retrieval system. Indeed following on from this user study, Elsweiler et al. created a photo browser which exploits people's remembering mechanisms. More specifically the interface created always displays a user's entire photo collection, photos matching a user's query are enlarged and all other photos are shrunk in size. Additionally, when a user hovers over a photo the photo is enlarged and information related to the photo is displayed. They also provide a number of filtering options which they classify as: visual filtering, semantic filtering by free-text, semantic filtering by groups, temporal filtering via date line, spatial filtering by screen location and smart filtering.

In other memory related research, Jaimes et al. [19] use memory cues for a meeting video retrieval system. From a user study, they found what types of items people easily remember and easily forget about meetings. In other words they found what items might act as memory cues in the meeting video retrieval domain. *Location of the meeting room, table layout in the room, seat positions and main speaker names* are among the items remembered. While items like date, time, dress and posture were hard to recall. They used this information in the design of an intuitive interface that uses information which will act as memory cues, for the retrieval of the desired meeting video.

These two examples highlight the possibly benefits from creating interfaces which exploit the ways users remember information. However systems such as these are only the beginning, there is need for the development of interfaces that exploit memory cues across entire HDMs, not just single domains such as photos.

As outlined in this section, the first steps in IR for personal collections have been made. My research seeks to progress this area to find new retrieval strategies and methods for the HDM space by focusing on its unique attributes.

3. PROJECT SCOPE

This project is motivated by the attributes of the personal archive space, as highlighted in Section 1. Ways of addressing and potentially exploiting this combination of features need to be found if we are to successfully provide applications that enable people to browse and search their personal archives.

Conventional IR techniques alone are not sufficient for retrieval in HDMs. Take a sample scenario where a person is looking for a particular, recently viewed, photo from her HDM archive. All she now remembers is that the sun was

glaring in the window when she last saw the photo and that she was talking on the phone to her friend, Jack, at the time. Conventional IR techniques would not be capable of retrieving the correct photo based on these criteria. New approaches to IR using context are required, e.g. a system that could retrieve a photo based on the weather when the photo was last viewed and who else was present. This requires capture and association of potentially useful context data.

3.1 Context

Data relating to context information associated with content creation or access can be obtained in a number of ways, such as through the use of timestamps, GPS technology, Bluetooth, and biometric sensors.

- Timestamps can be useful when a person recalls the time, day, month or year an item was created or accessed.
- GPS technology allows users retrieve an item based on location of item creation or previous access.
- Combining time and location information can be used to determine such things as the light status and weather conditions at the time of item access or creation [7].
- Bluetooth tracking devices allow for the detection of other Bluetooth devices in the nearby vicinity – in today's society many people have Bluetooth technology activated on their mobile phones. This enables us maintain a record of who was present when our subject was creating or accessing items from their HDM [20]. This information may prove useful in subsequent search, for example the user may be able to recall who was present when they were working on or viewing a particular item.
- Biometric sensors provide information on a subject's physiological state. In the CDVP at DCU we are exploring the user of two Biometric sensor devices – a *Polar Heart Rate Monitor* and a *BodyMedia SenseWear Armband*. The *Polar Heart Rate Monitor* is a sensor that the person wears on their chest. This sensor captures the person's heart rate, giving an indication of the subject's current stress level. This information might be used to assist a user in retrieving items they recall creating at a stressful time in their life for example. The BodyMedia SenseWear Armband is worn around the upper-arm. It captures physiological data, namely: Galvanic Skin Response (GSR), which is a measure of skin conductivity which is affected by sweat from physical activity and emotional stimuli; Heat Flux, which is a measure of the heat being dissipated by the body; Skin Temperature, which is a measure of the body's core temperature; and Accelerometer, which is a measure of movement or lack there of. These measures can be used as an indication of different types of arousal, such as excitement or boredom, which may correlate with more significant events in their life.

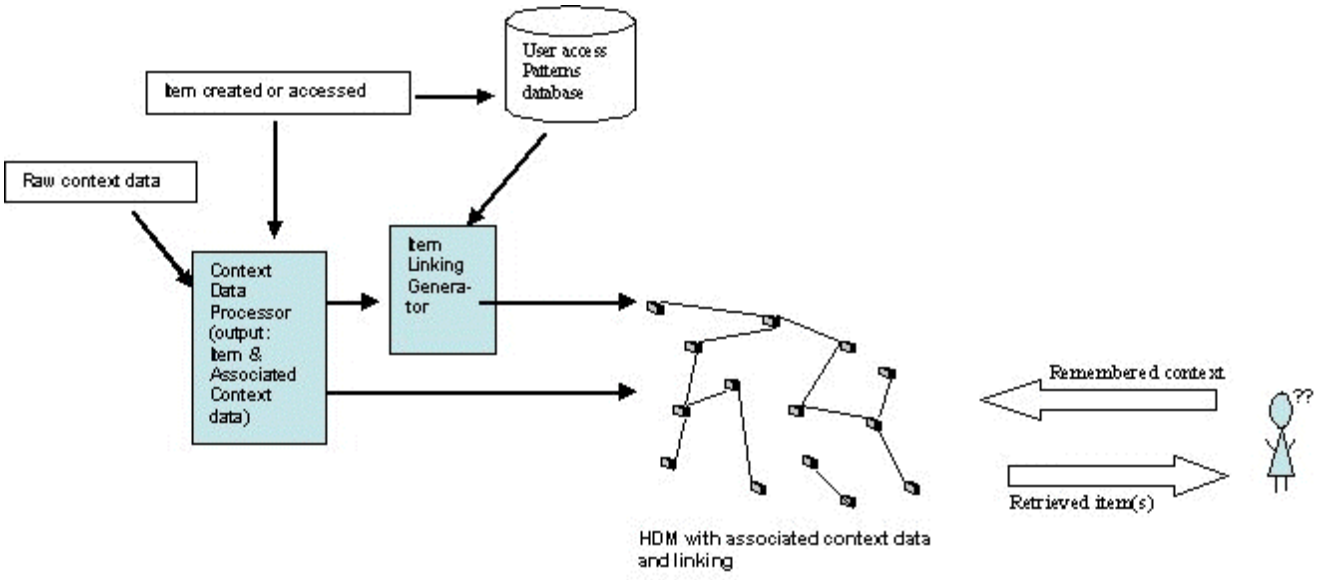

FIGURE 1. CONTEXT-BASED RETRIEVAL FROM A HDM USING MEMORY CUES.

In our research we are seeking to capture as many context sources as possible, to allow us to explore how to exploit context information sources in search. With this in mind we intend linking items within the HDM to further enhance the context information. We will explore methods of linking items based on a user's past interaction with items in their HDM, and also based on associations between items. We are interested in investigating if extensions to *PageRank* type algorithms can help locate interesting items based on users' memories of required items. Figure 1 demonstrates how we expect context information and linking might tie together to aid retrieval in a HDM. More

specifically, it shows how a user's HDM can be transformed into a linked graph using user access patterns and context information associated with items. The user can then query this linked structure using recalled context information.

As part of this investigation we will be examining which forms of context data, or combinations of them, prove more useful for search. Or whether certain types of items are remembered in different ways – do people remember stress level information more with document creation than photo capture for example. Additionally, we would like to investigate if these context cues are consistent across individuals.

This leads us to our first two research questions: How do people remember/recall events from their lives? And what triggers act as memory cues in the remembering process? In other words what context information might prove useful in helping people locate items in their HDMs. We would like to investigate how people remember/recall events from their lives. Following from this we can investigate what context information to use and how best to capture and exploit it. We expect that no single form of context data will provide a solution to the HDM retrieval problem, rather a combination of existing and new forms of context data will prove most effective. For example it seems people are more likely to remember events in which they were particularly aroused, the use of sensors to capture such arousal might prove very effective for future retrieval.

We believe that there are rich sources of personal context that can be exploited to trigger a person's memory and to aid retrieval. By capturing as much context information as possible we can explore memory cues and how people both remember and re-find information.

3.2 User Interface

Of course, people must use some form of an interface to interact with and retrieve information from their HDMs. A correctly designed interface is crucial to the success of an HDM retrieval application. An interface should be intuitive, simple to use, and afford the user all the functionality they require. An interface appropriate for interaction with a HDM is required. This interface should allow the user to easily search and browse through their HDM. This poses obvious problems, due to the possible vastness of the HDM. Various methods, such as fish-eye views, linking and zooming spaces have been successfully exploited in other domains to allow users move through information spaces. We will investigate the appropriateness of such existing techniques for the HDM domain, and also explore the possibility of new techniques which might prove more appropriate for this domain.

This leads us to our third research question: What type of tools can be constructed to meet peoples' needs in the HDM domain? In answering this question, we will consider the role context plays in the retrieval process and the ways in which people remember. *Stuff I've Seen* [4][5], as discussed in Section 2.2, integrates some features which may trigger the user's memory in its interface design. We anticipate that people's remembering processes can be used as a basis for the design of an effective interface to aid IR in HDMs.

4. CONCLUSIONS

This project is dedicated to the development of technologies that provide an intuitive means of retrieval from personal data archives. We intend these to be supported by algorithms which harness the types of information people use in the remembering process.

We postulate that, in addition to existing sources of personal context information there are other rich sources waiting to be discovered. If this context information is exploited correctly, it will be possible to create a system that retrieves items based on both an individual user's 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. Human Digital Memories are increasingly becoming part of our present. In the near future it will be hard for people to imagine a world where HDMs did not exist.

4.1 Open Issues

HDMs are personal to an individual, which means standard data collections cannot be used for experimentation purposes. Ideally a user's personal collection of files, and their interaction with this collection, would be compiled over an extended period of time and then the user would be the subject of experiments on their personal archive. In practical terms this will probably not be feasible for many large scale experiments. However, test collections of personal archives must be developed for the HDM domain. When using such collections a number of questions arise: How would we perform experiments on this test-bed without the archives' owners? Would it be possible for the owners to perform some ground-truthing on their archives which could be used for future experiments? What format would this ground-truthing take? The issue of privacy also arises when dealing with individuals' personal

archives – how much information will people be willing to give?, Will the information provided in archives for experiments be sufficiently similar to a 'real personal archive' and their information needs? And how can we ensure privacy of individuals' data?

ACKNOWLEDGEMENTS

This PhD work is supervised by Dr. Gareth J.F. Jones and funded by grant CMS023 under the Science Foundation Ireland Research Frontiers Programme 2006.

REFERENCES

[1] Bush V. (1945) As We May Think. *In The Atlantic Monthly*, 101-108.
[2] Gemmell J, Bell G, Lueder R, Drucker S and Wong C. (2002) MyLifeBits: Fulfilling the Memex Vision. *In ACM Multimedia '02*. Juan Les Pins, France.
[3] Gemmell J, Bell G, and Lueder R. (2006) MyLifeBits: A Personal Database for Everything. *Communications of the ACM. Personal Information Management*, 49(1):88–95.
[4] Dumais S, Cutrell E, Cadiz J, Jancke G, Sarin R and Robbins D.C. (2003) 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 information retrieval*, pp.~72–79, New York, NY, USA. Toronto, Canada, ACM Press.
[5] Cutrell E, Dumais S.T. and Teevan J. (2006) Searching to Eliminate Personal Information Management. *Communications of the ACM. Personal Information Management*, 49(1):58–64.
[6] Cutrell E, Robbins D.C, Dumais S.T, and Sarin R. (2006) Fast, Flexible Filtering with Phlat - Personal Search and Organization Made Easy. *In CHI 2006: Conference companion on Human factors in computing systems*, pp.~261–270, New York, NY, USA. Montreal, Quebec, Canada, ACM Press.
[7] O'Hare N, Lee H, Cooray S, Gurrin C, Jones G, Malobabic J, O'Connor N, Smeaton A.F and Uscilowski B. (2006) MediAssist: Using Content-Based Analysis and Context to Manage Personal Photo Collections. *CIVR2006 - 5th International Conference on Image and Video Retrieval. Springer Lecture Notes in Computer Science Vol. 4071*, Tempe, AZ, 13-15 July.
[8] Soules C.A.N. (2006) Using context to assist in personal file retrieval. *PhD thesis*, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.
[9] Soules C.A.N and Ganger G.R. (2005) Connections: Using Context to Enhance File Search. *In 20th ACM Symposium on Operating Systems Principles (SOSP'05)*, pp.~119–132. Brighton, United Kingdom.
[10] Page L, Brin S, Motwani R and Winograd T. (1998) The PageRank Citation Ranking:
Bringing Order to the Web. *Technical report.*
[11] Karger D.R, Bakshi K, Huynh D, Quan D and Sinha V. (2003) Haystack: A Customizable General-Purpose Information Management Tool for End Users of Semistructured Data. *In Proceedings of the 2003 CIDR Conference*.
[12] Freeman E and Gelernter D. (1996) Lifestreams: A Storage Model for Personal Data. *ACM SIGMOD Bulletin*.
[13] Stumpf S, Bao X, Dragunov A, Dietterich T.G, Herlocker J, Johnsrude K, Li L and Shen J. (2005) Predicting User Tasks: I Know What You're Doing! *In 20th National Conference on Artificial Intelligence (AAAI-05), Workshop on Human Comprehensible Machine Learning.* Pittsburgh, PA, USA.
[14] Google Desktop. http://desktop.google.com/
[15] Microsofts Windows desktop search. http://www.microsoft.com/windows/desktopsearch/default.mspx
[16] i-sho. http://www.i-sho.com/
[17] Elsweiler D, Ruthven I and Jones C. (2005) 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)*.
[18] Elsweiler D, Ruthven I and Jones C. (2007) Towards Memory Supporting Personal Information Management Tools. *Journal of the American Society for Information Science and Technology*.
[19] Jaimes A, Omura K, Nagamine T and Hirata K. (2004) Memory Cues for Meeting Video Retrieval. *In CARPE'04*, pp.~74–85. New York, New York, USA.
[20] Byrne D, Lavelle B, Doherty A, Jones G and Smeaton A.F. (2007) Using Bluetooth and GPS Metadata to Measure Event Similarity in SenseCam Images. *IMAI'07 - 5th International Conference on Intelligent Multimedia and Ambient Intelligence*, Salt Lake City, Utah, pp. ~18-24.