SEARCHING HETEROGENEOUS HUMAN DIGITAL MEMORY ARCHIVES

Liadh Kelly

Centre for Digital Video Processing, School of Computing, Dublin City University, Dublin 9, Ireland

ABSTRACT

Advances in digital storage technologies mean that vast digital archives of ones personal life experiences can now be generated. These personal archives (Human Digital Memories (HDMs)) can contain many types of data in various media created or accessed by the individual. These archives are of little benefit if an individual cannot locate and retrieve significant items from them. Existing search techniques are not sufficient for retrieval of items from these new unstructured spaces. This research proposes to develop effective HDM search by the integration of rich sources of context data, such as biometric information, with items in the HDM as a method to aid effective retrieval in this new domain.

1. INTRODUCTION

Vannevar Bushs 1945 seminal article As We May Think [1] provided a vision for a world where all a persons personal information could be stored and importantly retrieved at a later stage. With advances in modern technology Bushs ideas are now being realized. Vast digital archives of ones personal life experiences can now be generated. Items read, written, and downloaded; footage from life experiences; photographs taken; videos seen; music heard; details of places visited etc, can all be captured using digital devices.

While a persons entire life experiences can now reasonably be stored on computer, little attention has been given to how the individual (or a descendent of the individual) might locate important relevant items in this new type of vast archive. This space is unique in many ways and provides a number of challenges for retrieval in that: items will often not have formal textual descriptions; many items will be very similar, repeatedly covering common features of the users life; items will often not be joined by inter-document links; the archive will contain much non-useful data; the user may be unable to describe clearly what they are looking for; and the user may not even be aware that the data was captured and is available. We suggest that rich sources of context data can be used to help overcome these problems and to provide people with effective ways to retrieve from their HDMs.

The remainder of this paper discusses the integration of existing technologies into the system, automated analysis and organisation of the data to assist search, annotation of HDM items with context data and how this information can be used for future retrieval.

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

2. INFORMATION COLLECTION

Standard forms of context data, such as date and location, are already in use in Personal Retrieval Systems (PRSs), examples include *MyLifeBits* [2] and *MediAssist* [3]. We believe that there are many other rich sources of context which can be used to improve retrieval in this personal archive domain.

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 how the person was feeling when the photo was taken.

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 [3].

Bluetooth tracking devices allow for the detection of other Bluetooth devices in the nearby vicinity - in todays society many people have Bluetooth technology activated on their mobile phones. This enables us to maintain a record of who was present when creating or accessing items from an HDM [4]. 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, for example Heart Rate Monitor and BodyMedia SenseWear armband, provide information on a subjects physiological state, which can be used as an indication of different types of arousal, such as excitement or boredom, which may correlate with more significant events in the individuals life.

There are a number of challenges associated with HDM generation, for example existing technologies do not integrate all of the previously mentioned sensors in one device. There are also a number of privacy issues raised by the capturing of personal data.

3. INFORMATION RETRIEVAL

The iCLIPS project at DCU is dedicated to finding new retrieval methods for the HDM space. As part of this we will also explore methods of linking items in an HDM based on a users past interaction with these items, and also based on associations between items. We are interested in investigating if extensions to PageRank [5] 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 users 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.

Some data items will have no obvious connections or may be highly redundant, as well as being in forms that are difficult to index. For example, it is difficult to reliably search for individual digital photos based on context analysis. In iCLIPS we aim to address this problem by building on existing work in the CDVP at DCU. The MediAssist project [3] provides an interface for efficient searching of digital photos using context and content information. Photos are segmented into events based on time proximity and thumbnails of five representative photos are displayed for each event. Among other things this system suggests names for unknown faces by matching body patches of known faces to those of unknow faces.This technology would allow us to automatically detect all photos containing a searched for person, for example. An emerging area of lifelogging relevant to HDMs is proactive capture of images using devices such as the Microsoft Research SenseCam. The SenseCam is a small device worn around the users neck which captures 3,000 images on an average day both automatically and based on changes in sensor data captured by the device. These images due to both sheer volume and unstructured nature are impractical for a user to search through. [6] segments these images into events using SenseCam sensor data. Key images, similar to keyframes in video search, are then extracted from each event. These key images can be used as a summary of a day or as an entry point into more images of a particular event. We could also link these key images with other related forms of data, such as emails, web pages etc.

4. FURTHER RESEARCH DIRECTIONS

Coupled with integration of existing related technologies we will also examine 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.

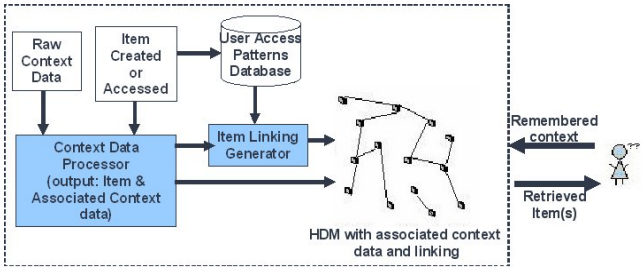

Figure 1. Context-based retrieval from a HDM using memory cues

In addition to existing sources of personal context information we believe there are other rich sources waiting to be discovered. If existing and new forms of context information are exploited correctly, 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.

5. REFERENCES

[1] V. Bush, "As We May Think," *The Atlantic Monthly*, pp. 101–108, 1945.

[2] J. Gemmell, G. Bell, and R. Lueder, "MyLifeBits: A Personal Database for Everything," *Communications of the ACM. Personal Information Management*, vol. 49, no. 1, pp. 88–95, January 2006.

[3] N. O'Hare, H. Lee, S. Cooray, C. Gurrin, G.J.F. Jones, J. Malobabic, N.E. O'Connor, A.F. Smeaton, and B. Uscilowski, "MediAssist: Using Content-Based Analysis and Context to Manage Personal Photo Collections," in *CIVR2006*, 2006, pp. 529–532.

[4] D. Byrne, B. Lavelle, A. Doherty, G.J.F. Jones, and A.F. Smeaton, "Using Bluetooth and GPS Metadata to Measure Event Similarity in SenseCam Images," in *IMAI'07*, 2007, pp. 18–24.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Tech. Rep., Computer Science Department, Stanford University, January 1998.

[6] A. Doherty, A.F. Smeaton, K. Lee, and D.P.W. Ellis, "Multimodal segmentation of lifelog data," in *RIAO 2007 - Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*, 2007.