Personal Information Manager to Capture and Re-Access What We See on Computers

Zaher Hinbarji
Dublin City University
zaher.hinbarji@dcu.ie

Moohamad Hinbarji
Dublin City University
moohamad.hinbarji@dcu.ie

Rami Albatal
Heystaks Technologies Ltd.
rami.albatal@heystaks.com

Cathal Gurrin
Dublin City University
cathal.gurrin@dcu.ie

ABSTRACT

Nowadays we live in a world where many of us engage with computers more than humans as a result of spending a major part of our life in front of a range of computing devices. Consequently, it's becoming important to shed more light on our interactions with computing devices, which we see as a special domain of lifelogging (information-lifelogging), where capturing and archiving what we see on our computer screens can be utilised for several useful applications such as user profiling, personalization and memory support. In this work, we present a tool that allows us to passively capture the digital content we see on our screens for later re-access. It can be considered as a type of digital memory that stores user's computer usage to recall a user's information creation and access activities. This has potential to assist users to better achieve their daily tasks by having access to a digital backup where their previous content and experience can be recalled as required.

CCS Concepts

•Information systems → Digital libraries and archives; •Human-centered computing → Human computer interaction (HCI);

Keywords

Personal information management, Lifelogging, Personal digital archive, User behavior modeling

1. INTRODUCTION

Personal information management is the process whereby individuals capture, store, organise and retrieve their personal information (paper or digital) to assist in normal daily life tasks [8], such as recalling past events, supporting decision making, enhancing performance in a job, or even for supporting memory. Lifelogging, is a unique form of personal information management where it is defined as "a phenomenon whereby individuals can digitally record their own daily lives in varying amounts of detail and for a variety of purposes" [3], and as such, it is concerned with the management of potentially large and detailed personal data archives.

As most of the related work on lifelogging is focused on gathering and analysing visual and physical activity data [4], we focus here on logging and managing the digital content the user engages with while working on computers; we call this information-lifelogging. Nowadays, computers have secured an important role in both of our professional and personal lives. This means that the logs of our interactions with information (via devices) can be seen as a rich stream of evidence to mine and analyse for insights about our life and to support us to better complete everyday tasks. We believe that many assistive applications can be built on the top of information-lifelogging, for example more focused forms of personalization, detailed user profiling and personal information management systems.

An important challenge that faces all lifelogging technologies is having a proper information system in place that can be used to filter and re-access the previously logged information efficiently [4]. It is our conjecture that a tool for capturing content from computer interactions and archives it for later access could support new types of memory recall applications. Such a system could be seen as a digital memory that stores our computer usage, which would assist in finding answers to questions such as: when was the last time I emailed a particular person and what did we say? what did I last read (or write) about a certain topic? or to bring together all communications with a person or regarding a certain topic. Such a system could also allow us to filter the data across various facets such as date/time or application(s) used. While many individuals employ a backup system to protect our documents and files on computing devices, we never had before a similar solution for backing up what we see and do on computers.

In this work we present information-lifelogging via a new type of personal information manager, developed in order to passively capture, organise and retrieve the different digital content users interact with while using computers.

LTA'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2983576.2983580

2. BACKGROUND

In 1945 Vannevar Bush described a hypothetical mechanised device, Memex, which can be used by an individual to store all of his/her communications, records, books and documents and which "may be consulted with exceeding speed and flexibility". It is described as being an "enlarged intimate supplement to one's memory" [9]. This was a pioneering description of what we feel is still needed for personal information systems.

In terms of managing personal information, many works have focused on single content source, in [7], for example, an observational study about how people organise web information for later reuse has been conducted. People in this study utilised several methods to manage their web content such as: printing web pages, saving them locally or keeping a list of URLs along with the typical bookmarking and history lists web browsers provide. Researchers in [12] have reported that 58% of the pages the subjects visit were to pages that they have been visited previously. Approaches to enhance web pages re-access have been proposed such as: indexing web pages text in [10] and using a richer graphical representation in [1]. Authors in [13] have focused on using email clients as personal information managers since people used to utilised them for contact management, documents archiving and task management as well as for communication.

However, there are some works that deal with multiple content sources for re-access. Haystack [6], for example, allows users to organise their emails, appointments, documents and tasks in a single interface. It connects items using a semi-structured data model based on both available meta data and user annotations. Another example is MyLifeBits [2] which is novel software that aims to fulfill the Memex vision that we previously talked about. Its main goal is to store all the individual's personal information including sounds and videos with support for annotation, visualization and search. It differs from our work as it does not include data capturing but rather it focuses on storing and organising methods.

One can find several commercial tools for managing specific personal information, but using different applications can lead to the problem of *information fragmentation* as these different solutions have their own separate methods for archiving and exploring the data. Our approach, however, provides a single access point to all different captured contents that stored in a single database. Unlike other systems that are only concerned about managing the data regardless of its source, our proposal here covers the whole process, starting from data capturing to the retrieval phase.

3. SYSTEM OVERVIEW

3.1 Architecture

To collect all types of digital content the user engages with, we built client-side software that works passively in the background of the user's computer and captures the screen when the user is interacting with the device. A screenshot of the current active window is taken periodically and whenever the window is switched. Considering only the active window instead of the full screen is to make sure we get only what the user is currently focusing on. That would reduce data redundancy and enhance retrieval later on. Each screenshot image is also associated with both the app name and the window title. This data is uploaded automatically to our cloud indexing service. Users can always control their privacy preferences through their online dashboard as described later in section 3.3.

Next, we extract textual content from screenshot images by applying off-the-shelf optical character recognition (OCR) tool. That allows screenshots to be indexed based on their textual content. The extracted text is associated with its corresponding screenshot. Typical OCR software works well for this task as most of the text in the screenshots is already computerised (computer-generated) and the contrast between text and background is typically good. Our design allows the user to link multiple devices, if available, under the same account in order to get a single entry to all of his/her data with the ability to distinguish between devices based on their given names. Figure 1 shows the system architecture described above.

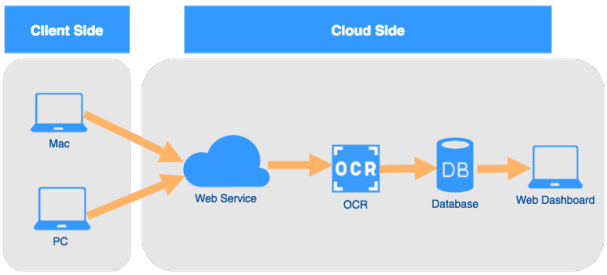

Figure 1: System Architecture

3.2 Search and Filtering

One of the main added values of our tool is to provide the user with an easy and convenient facility to re-access any previously observed (or created) content. To achieve this, our software allows the user to search and retrieve the text extracted from screenshots. Here, the screenshot with its corresponding text are considered equivalent to a document in a typical retrieval system. Microsoft full-text indexing technology was used for retrieval, which utilises a standard information retrieval algorithm. All the retrieved documents (screenshots) are presented along with their timestamp, window's title and application name as shown in Figure 2.

However even with modest computer usage, the volume of data for indexing would become quite large and we felt that there was a need to include faceted-search functionality. Without implementing a faceted-search interface, the huge amount of data for a user to browse in response to typical search queries would be a major challenge, as is the case with other forms of lifelog data [4].

Our tool supports the user to focus a query via a number of facets (filters), such as application category (e.g., communication, browser, utility,..), date range, or even by a specific application name (e.g., Skype, Chrome,..). We think application categories are an information-lifelogging form of user context, which will be meaningful to the user when retrieving information, and essential in reducing the size of the ranked list.

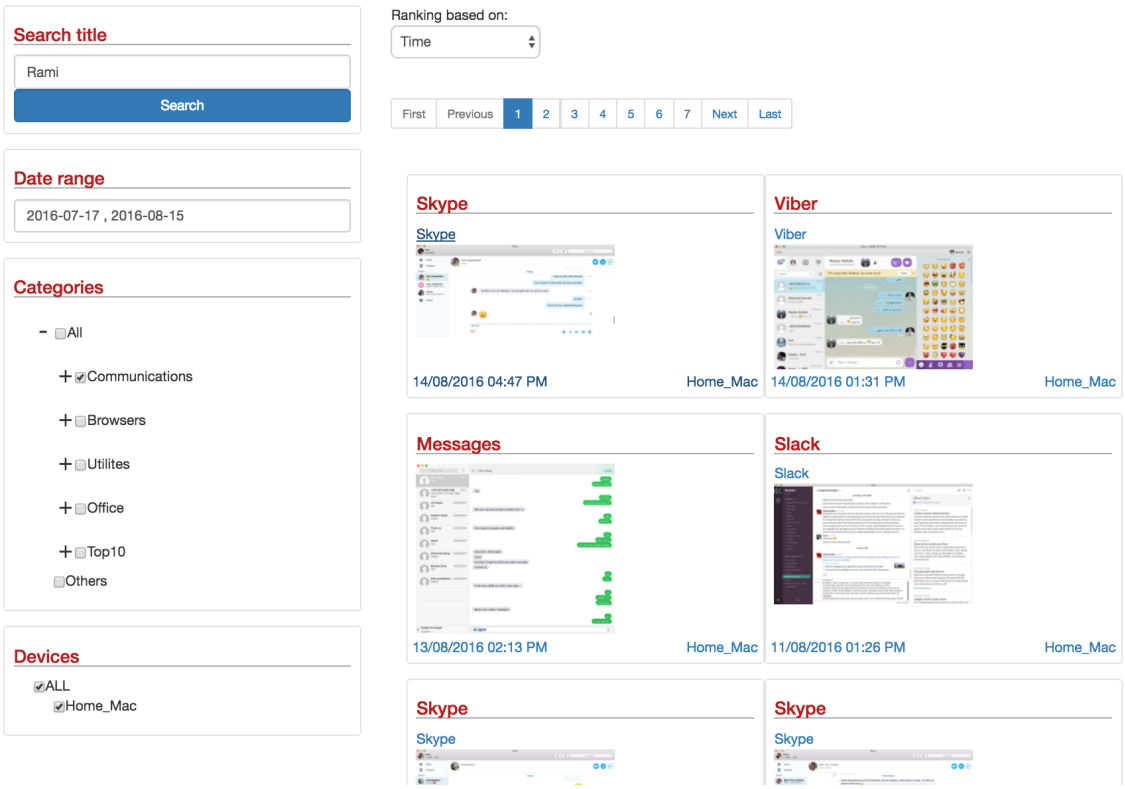

Figure 2: User interface showing a query for all conversations with a named person

Figure 3 illustrates the type of application categories filtration interface. The user can also control the date range (e.g., today, yesterday,..) of the query to limit the search space based on their recollection of when the information event took place. Finally, for easier navigation, results are presented in chronological order (be default) from recent to oldest by default although the user has the option to use text-similarity as the ranking option (as shown in Figure 2).

3.3 Privacy

Privacy consideration is an important challenge for personal information software in general. Given that all the contents the user sees on the screen is captured and uploaded to the cloud, such systems should consider privacy as a central tenet of the system. Our proposed solution for this involves both client-side and cloud-side arrangements. On the client side, the capturing tool maintains a black-list of private applications and another black-list of window titles keywords. These lists are used to prevent unwanted content to be captured in the first place.

All data captured locally is also automatically synced to the cloud (if the user agrees) and the cloud data is manageable by the user via the online dashboard, which facilitates review and deletion of previously uploaded data. The cloud service also implements black-list and white-list user-controlled filtering. Furthermore, through the same search interface users can manually delete the unwanted returned results should they so wish. Finally, users accounts are password protected for a secured access and our data capture client-side software allows pausing and restarting of the capture process, as the user requires.

3.4 Technical Specification

The current version of our tool works on Mac OSX 10.7 or later. An equivalent windows version is currently under development. After installing our client app (LoggerMan [5]), the app icon will show in the upper menu; the app icon shows the current status of the app. The default status is running, however, the user has always the option to pause data capturing whenever he/she wants. An idle status of the computer is also taken into consideration to avoid capturing the exact same screenshot twice if nothing has changed on the screen since the last capture. Data uploading is performed efficiently where the app always observes cloud availability. If the cloud is not accessible for any reason (e.g. no internet connection), the data will be cached locally on the machine until it can be uploaded successfully at a later time.

As mentioned earlier, the Microsoft full-text indexing has been utilised in this research. It provides standard term-frequency-based search techniques for textual data, called Okapi BM25 [11]. Word tokenization, stop words removal, stemming and query expansion are all part of the retrieval process, which results in a higher standard of retreval and also makes the system tolerant of spelling errors.

User	Weekday	Weekend	Weekday Apps	Weekend Apps
1	1103.28	1017.06	Chrome, Safari, Xcode, Preview, Word	Chrome, Xcode, Simulator, Skype, Viber
2	1015.72	796.19	Chrome, Eclipse, Slack, Terminal, gedit	Chrome, Eclipse, Word, Skype, Android Transfer
3	713.75	393.5	Chrome, Mail, Word, Keynote, Postbox	Chrome, Mail, Skype, Excel, Preview

Table 1: Weekdays VS. Weekends Computer Usage Comparison

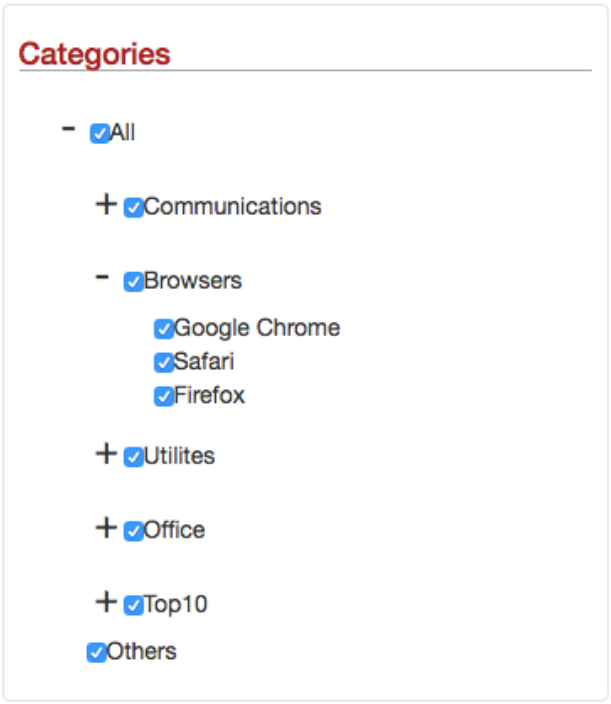

Figure 3: Result filtration based on the application category

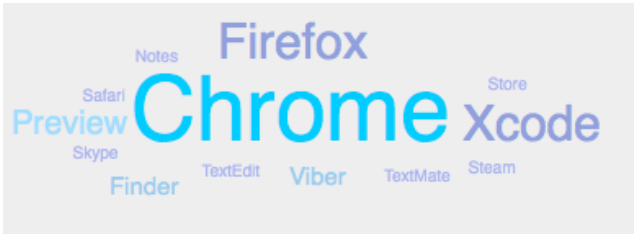

Figure 4: Most frequent used applications cloud

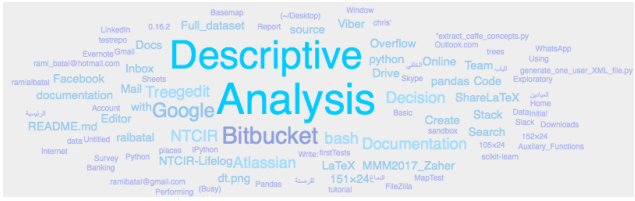

Figure 5: Most frequent words found in windows titles

4. EVALUATION

Our software for information-lifelogging has been designed for long-term use, so as part of our research, three users have been using our tool for the last 6 months on their work and home computers, generating an average of 150,000 screenshots each. Our subjects display typical and expected patterns of use, where they use the computer for different durations during weekdays comparing to the weekends. Table 1 shows the average number of screenshots generated by each user during weekdays versus weekends along with the top 5 used applications. We can see for example that user 1 uses the computer during weekends almost as much as during weekdays but for different purposes based on the used applications. Also, Skype, for instance, seems to occupy a major part of both of the three users' weekends computer usage. Even this simple data illustrates the potential to reveal useful insights about the user and their lifestyle.

As previously mentioned, our goal in this research is to provide users with the ability to recall and re-access their previous computer usage and the content they engage with. To see how effective our tool is at such information-logging, we carried out an initial small user experiment where we asked three users (who are already long-term users of loggerman[1]) to complete five different known-item search tasks using our tool, where they had a maximum of one minute to complete each task. Table 2 shows the list of questions and the number of seconds each user took to do the task ('DNF' in the table means the user Did Not Finish, or failed to find the information in less than a minute). The experiment ended up with 80% of the tasks having been completed successfully during the time limit. According to a qualitative evaluation carried out post-experiment, the current version of the system is very useful when a user is seeking to recall specific information that they can generate a query for (i.e. remember some keywords related to it). However, for some queries, users noted that the tool returned too many results that contain the required keywords but not in the intended area (e.g. the name of another person appears in the side menu of a chat application while the user is chatting with someone else). That's because currently the system deals with the text in the screenshots regardless of their position in the image i.e., there's no difference between a keyword appears in the body of an article and a keyword appears in the menu bar, title, or even in an advertisement banner in the page. To give different importance for texts depending on their functional and locational context, image processing and computer vision techniques would need to be developed to do further analysis of the image content and associate a value weight to the different types of textual content within the screenshot.

[1]http://loggerman.org/

Task	User 1	User 2	User 3
Find the most recent email from your immediate line manager.	21	19	DNF
Find the most recent bank balance from your online banking.	25	56	51
What was the last communication you had with a named family member.	20	DNF	41
Find the first mention on BREXIT in the online news.	14	12	11
What was the price of the last item you purchased online.	18	DNF	14

Table 2: Experiment's Questions and Results

5. REPORTING AND FUTURE WORK

In order to give users insight about their computer usage and the digital content they observe, different types of reports and statistics are already under development as part of our future work with the information-lifelogging tool. For example, Figure 4 and Figure 5 both show tag clouds that represent relatively the most frequent used applications and the most frequent words found in windows titles for a user during a specific date range. Several future improvements can be done including different interfaces and retrieval algorithms that emphasize on the personal nature of the data and the familiarity of the user with their content. In addition, text analysis techniques are already planned to be part of the system such as entity extraction (people, places and contact information), topic identification and summarization techniques. This is to conform to our wider research plans of automatically building enhanced user profiles based on human-computer interaction data, which has potential to extract new insights about the user that was not possible before. Finally, we are extending the data gathering process to include windows applications, mobile device applications and also directly accessing the content of web pages that the user has read.

6. CONCLUSIONS

In this paper we have presented a information-lifelogging tool to capture, store and re-access the different digital content the user engages with while working on a computer. It provides the user with a single access and store point to all of their captured content even if they have originated on different devices. Unlike other available systems that provides general personal data management functionalists, we offer here a complete solution that covers data capturing as well as data managing and re-access in one integrated solution that takes into consideration privacy concerns. We have shown by means of a small experiment that the information-lifelogging tool can be effective and we have discussed several future plans that include developing and customising current information retrieval algorithms to work better on personal data, as well as applying text analysis techniques to further enhance accessibility and to provide statistics and insights about the user's preferences, interests or other personal behaviour patterns.

7. ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289.

8. REFERENCES

[1] A. Cockburn and S. Greenberg. Issues of page representation and organisation in web browser's revisitation tools. *Australasian Journal of Information Systems*, 7(2), 2000.

[2] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. Mylifebits: Fulfilling the memex vision. In *Proceedings of the Tenth ACM International Conference on Multimedia*, pages 235–238, New York, NY, USA, 2002. ACM.

[3] C. Gurrin, R. Albatal, H. Joho, and K. Ishii. A privacy by design approach to lifelogging. In *O Hara, K. and Nguyen, C. and Haynes, P., (eds.) Digital Enlightenment Yearbook 2014*, pages 49–73. IOS Press, The Netherlands, 2014.

[4] C. Gurrin, A. F. Smeaton, and A. R. Doherty. LifeLogging: Personal Big Data. *Foundations and Trends in Information Retrieval*, 8(1):1–125, 2014.

[5] Z. Hinbarji, R. Albatal, N. E. O'Connor, and C. Gurrin. Loggerman, a comprehensive logging and visualisation tool to capture computer usage. In *22st International Conference on MultiMedia Modelling (MMM 2016)*, pages 342–347, 2016.

[6] D. Huynh, D. R. Karger, D. Quan, and V. Sinha. Haystack: A platform for creating, organizing and visualizing semistructured information. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 323–323, New York, NY, USA, 2003. ACM.

[7] S. Jones, William; Dumais and H. Bruce. Once found, what next? a study of âĂŸkeepingâĂŹ behaviors in the personal use of web information. In *Proceedings of the ASIST Annual Meeting*, pages 391–402, 2002.

[8] W. Jones. *Keeping Found Things Found: The Study and Practice of Personal Information Management: The Study and Practice of Personal Information Management*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[9] L. Manovich. *As We May Think, The New Media Reader*. The MIT Press, 2003.

[10] B. McKenzie and A. Cockburn. An empirical analysis of web page revisitation. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, HICSS '01, pages 5019–, Washington, DC, USA, 2001. IEEE Computer Society.

[11] S. E. Robertson and K. S. Jones. Simple, proven approaches to text retrieval. Technical report, 1997.

[12] L. TAUSCHER and S. GREENBERG. How people revisit web pages. *International Journal of Human-Computer Studies*, pages 97–137, 1997.

[13] S. Whittaker, V. Bellotti, and J. Gwizdka. Email in personal information management. *Commun. ACM*, 49:68–73, 2006.