Indexing and visualizing digital memories through personal email archive

Florian Evequoz
Departement of Computer Science
Bd de Perolles 90
CH-1700 Fribourg, Switzerland
florian.evequoz@unifr.ch

Denis Lalanne
Departement of Computer Science
Bd de Perolles 90
CH-1700 Fribourg, Switzerland
denis.lalanne@unifr.ch

ABSTRACT

The research work presented in this paper tackles the personal information overload phenomenon. Our purpose is to offer a set of interactive visualization techniques allowing one's personal digital memory to be organized and overviewed. This system will provide easy browsing, guiding towards the wanted piece of information and allow a free exploration of the personal information space. Three main axis of research are involved: (1) emails clustering to organize personal information, (2) cross-media alignments to connect clustered emails with personal information, and (3) information visualization techniques, to provide interactive means to navigate through personal information. As an alternative to the presented state-of-the-art works, our approach to these challenges takes the personal email archive as entry point into the personal information space. We finally present the work done during the first year of this research and the roadmap for the future.

Keywords

Personal information management, information visualization, emails, memory associations

1. INTRODUCTION

With all types of media becoming digital and storage devices regularly increasing in capacity, we tend to accumulate a growing amount of information that becomes "personal" as soon as we decide to keep it. However, this personal information (PI) grows very fast, challenging our natural wish for order and the capacity of our memory. Thus, e-mails, pictures, videos, music, personal documents and every other pieces of information creating our individual digital memory are often stored anarchically and become hard to retrieve or correlate. Even when we try to keep it organized, the rigid hierarchy of filesystems or mail archives forces us to take decisions on classification schemes that may not be relevant in the long-term, eventually leading to frustration if we fail to retrieve a piece of information that we know for sure is in our collection. We may then feel overwhelmed and have the feeling of losing control over our personal digital memory.

The very hierarchy of file storage system is the main reason for this failure. Indeed, as Bush pointed out in [3], our mind works by association rather than by following the rules of a static hierarchy. An email from a friend may remind us of the holidays we spent together, the places we saw and of which we took the pictures and the music we used to listen to at the time. Gathering this information at once would require painful search in crowded data repositories. However with current systems, no explicit links connect heterogeneous pieces of information about the same topic, related to the same people or having another characteristic in common. This implies that we often have to perform repetitive searching tasks using several different applications in order to gather the obviously correlated information we look for. All in all, the inherent or instinctive structure of our personal information is hidden, and the current searching mechanisms do not put it into evidence. It remains obscure also because we cannot get an overall view of it.

The purpose of the research work presented in this paper is to investigate means for restructuring and visualizing the personal information space interactively. The focus will be put on browsing capabilities, taking advantage of the structure of the personal information and the various links existing between different pieces of information, e.g. thematic, temporal or social links. We will use the personal email archive as the main source for generating a personal information structure. Other personal information (texts, pictures, music) will then be aligned with this structure and integrated together into an interactive visualization tool.

In the following sections, we present a state-of-the-art of personal information management (PIM) and email management and visualization systems. Section 3 is devoted to a more concrete presentation of the work envisioned for our research project. Section 4 present the use cases and applications of the project. Finally, section 5 present the work achieved during the first year of the project and the future plans.

2. STATE OF THE ART

PIM research has been receiving a growing interest in the recent years, leading to the developement of several tools and methodologies. We present here some innovative works that motivated our approach. *Stuff I've Seen* [8] is a search

Supporting Human Memory with Interactve Systems, workshop at the HCI 2007 (British HCI conference 2007), September 4th, 2007, Lancaster, UK

system providing a unified view over all types of PI (files, emails, web pages, etc.) and enabling both queries and filters based on available metadata like date, author or type of document. The search results are presented as a textual list that can be (automatically) ranked by relevance or sorted chronologically. The study conducted at the end of the project notes the importance of people and time for retrieving information. MyLifeBits [9] is a database of resources and links. Links represent either user-created collections of resources or so-called transclusion, that happen when a resource cites or uses another one (e.g. a picture included in a Powerpoint presentation). This work also recalls the need for annotation on non-text media. Results of queries in MyLifeBits can be viewed using traditional detailed or thumbnail view, or using more original and flexible time-based visualizations. FacetMap [18], built on top of MyLifeBits data store, offers a query-refinement mechanism based on facets and therefore allows to browse the data instead of searching. Other works, described in particular in [21] focus on the semantic aspect of personal information. A most interesting user-study of personal information management strategies has been conducted in 2004 by Boardman [2]. Among other conclusions, Boardman notes that users generally prefer to browse than to search their PI and that the email archive has a potential for being integrated with the files, as similarities are strong between files and filed emails. We want to explore the direction he suggests and try to structure the PI around the email archive's own structure. Finally, an interesting research work dating back to 1994 presented in [15] suggests to exploit the human episodic memory, i.e. the ability we have to associate things to a context, in particular places and people, in our memory. [17] also tries to replace information in context, specifically in their temporal context. Following this latter idea, our work tries to recover the natural context of information, focusing on finding similarities between different pieces of data, exploiting the social temporal and thematic dimensions.

In the particular domain of email management, recent works introduce visualization techniques to tackle the issues raised by email overload. Some works use visualizations to help handle the current inbox and keep a synthetic view of tasks. Thread Arcs [12] for example, presents a novel visual approach that helps to understand threads of messages. The works of Dredze [7] or Cselle [6] use machine-learning methods to classify emails into activities, helping to keep trace of current (but also possibly past) tasks. The visualization of activities relies on color schemes. Nevertheless, few works really address the problem of managing and exploring an individual's whole email archive. In [19] the authors propose to visualize the "conversational history" between the mailbox owner and a chosen contact during a certain period of time. While this is an invaluable tool for a psychological self-analysis, it does not provide an overall view of the email archive. Similarily [16] explores relationships through past emails, but consists in an analytical tool rather than an exploration tool. Our system will try to provide a synthetic view of the whole personal email archive that can then be extended to include other types of connected PI.

3. PROJECT OVERVIEW

The PIM works presented in the previous section consider all types of PI equally important with respect to indexing. We chose one as being prominent. As suggested by [2] the email archive has the potential of being taken as a core around which other PI gets connected. Therefore, in order to generate metadata on PI, we use the personal email archive as main source of metadata. Email is indeed a rich subset of PI [20]. A single email inherently connects together people, topics and time. Therefore, a whole personal email archive contains invaluable thematic, temporal and social metadata that would be hard to obtain with other types of PI: people knowing each other usually appear together as recipents of a message, some topics are related to particular groups of contacts, topics and relations are closely related to time periods, etc. Our purpose is to gather this metadata and retrieve clusters pertaining to the social, thematic and temporal dimensions. These three dimensions are preferred by users looking for documents, according to the study in [17]. In the next step of analysis, the remaining PI will be aligned with the dimensional structures extracted from emails, following the approach successfully used in [13] to align multimedia data with textual documents.

Once the metadata is available, relevant visualization techniques will be applied in order to allow browsing the whole PI. As we will not use data-driven classification of PI which requires a training set of already classified data, but statistical analysis, the role of the personal information visualization will be particularly emphasized. We plan to present the user with several visualizations simultaneously, in order to enable visual query refinement using known interaction methods such as on-demand filtering, link & brush, etc. Each of the visualizations proposed will focus on one of the aforementioned dimensions, or a combination of two of them (e.g. variation of themes over time, using a technique similar to ThemeRiver [10]). We believe that the combination and synchronization of several visualization types applied to different dimensions will help the user browse instinctively through her/his PI. Therefore, a goal of this research work will be to confirm or invalidate our hypothesis that a good use of visualization can be efficient for handling PIM, without using any semantic modelling. More specifically, the following steps need to be performed by our email-centric PIM system in order to fulfill its goals:

1. Features or metadata extraction from emails
2. Clustering according to social, thematic and temporal dimensions, based on similarity computations
3. Alignment of the structure extracted from emails with the remaining PI
4. Visualization of the PI, based on the structure and taking advantage of similarity links

4. USE CASES AND APPLICATIONS

The primary use case regards the outcome of this project as a memory association facilitator. In the context of Hasler Stiftung's Memodules [1] project, tangible objects serve as shortcuts to digital information. A user can tag a "real" physical holidays souvenir, associate it to digital pictures, sounds, etc. and use it later to retrieve the images and other data of her/his vacation. One of the main issues raised by this approach is the actual association between digital information and the physical object, in particular the work

needed to gather all the pertaining digital information that one want to associate to a specific object. This is where our system comes into play. Indeed, the views of PI it provides, exhibiting a meaningful PI structure, may be used to facilitate the association task. If the emails, documents, pictures and other information about a holiday are gathered and presented together, associating the whole episode to a physical object becomes easier. In addition, physical objects already associated to digital information may serve as query parameters to retrieve correlated pieces of PI. A combination of two physical objects may be used to retrieve the PI pertaining to both objects' associations.

The secondary use case handles professional data access. NCCR IM2 project deals with multimodal information management, and in particular meeting data management (documents, emails, audio/video recording of meetings, slideshows, etc.). Our project comes within the scope of IM2.HMI, of which a specific goal is to develop methods for accessing recorded meetings data. The benefit of our project in this context is to be found in the assistance it can provide for browsing huge amounts of professional data. Indeed, the extracted PI structure may serve as a filter to help navigate through such professional data, and to assist in finding information thanks to similarity links that can be drawn between personal and professional information [14]. Fig. 1 summarizes our approach and the link to both applications.

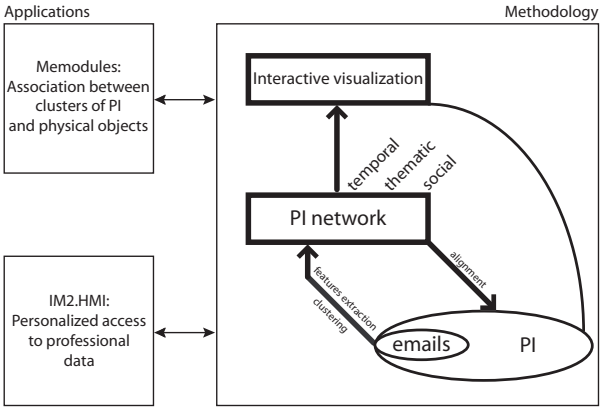

Figure 1: A PI network based on social, temporal and thematic dimensions is extracted from emails and aligned with the remaining PI. Interactive visualization techniques take benefit of this network to help browsing personal information. Two chosen applications are shown on the left.

5. ACHIEVED WORK

The first research efforts were centered on emails. In a first phase, we collected personal data. It consists of (a) a personal email archive containing around 6000 emails and 3500 addresses, (b) the Enron public email archive [5] and (c) the AMI meeting corpus [4], that contains textual, audio, video and email data. A user requirements questionnaire was set up, focusing on the relationship between personal and professional information and the preferred way of accessing them. We then developed a tool for extracting email data from IMAP servers and local archives into a database and perform statistical analysis on textual content and addresses to gather relevant features. More specifically, textual similarities were computed using a cosine similarity measure on tf.idf values. With this method, the similarity between two messages is proportional to the number of words they have in common, and depends on the discriminancy of each word (a rare word, like "Cappadoce" has more weight than a very common word like "meeting"). Likewise, the similarity between two contact's addresses is proportional to the number of times they appear together in the headers of emails. Using the similarity based on the co-occurrences of words in the subjects and contents of emails, a simple hierarchical clustering was performed, which aims at finding a thematic organisation of emails. As well, a social network was built based on similarity measures between email addresses. For the information visualization part, simple views of the email archive were implemented with the help of the prefuse toolkit [11]. The result of thematical clustering was fed into a treemap-visualization, while the result of the social analysis is visualized as a social network graph (see Fig. 2) using a spring layout algorithm to separate main clusters from one another. Even if the two visualizations still need refinement and have not been synchronized yet, limiting the possibilities of visual querying, they show interesting trends that would not have been highlighted by traditional mail clients.

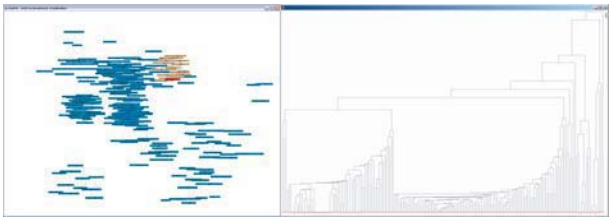

Figure 2: The implemented views of a personal email archive. The left window presents the social-network graph and the right one shows the thematic hierarchical clustering of emails.

6. CONCLUSION AND FUTURE WORKS

In this paper we have presented our approach of personal information management using information visualization. Taking into account the previous work in the field, we compute similarities between differences pieces of data to build a network of PI, in an attempt to gather memory episodes the same way our human memory does. The personal email archive is used as the main source for metadata on PI, allowing to generate thematic, social and temporal links between pieces of information. Our plan is to align the structure extracted from emails with the remaining PI and use this structure as an entry point into the PI. Achieved work covered the phases of data collection and metadata extraction using statistical analysis and clustering methods. We also developed visualizations of the email archive based on the social and thematic dimensions. In the near future, we plan to synchronize the different email archive views to enable more advanced visualization techniques such as link & brush, details-on-demand, filtering, etc. Indeed, we believe that these visualization techniques build upon the similarity

links extracted from the data can enhance the PI browsing experience. Further, the alignment of email archives with the rest of the PI will be considered, and various email-centric PI browsers will be implemented through our two use-cases: ego-centric meeting browsing and tangible access to PI. In a final phase, user-evaluations will be conducted, comparing our set of tool with standard mail clients and desktop managers.

7. REFERENCES

[1] Memodules homepage. `http://www.memodules.ch/`.

[2] R. Boardman and M. A. Sasse. "stuff goes into the computer and doesn't come out": a cross-tool study of personal information management. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 583–590, New York, NY, USA, 2004. ACM Press.

[3] V. Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945.

[4] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. A. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus: A pre-announcement. In S. Renals and S. Bengio, editors, *MLMI'05: Proceedings of the Workshop on Machine Learning for Multimodal Interaction*, number 3869 in LNCS, pages 28–39. Springer-Verlag, 2005.

[5] W. Cohen. Enron email dataset. Retrieved May 5, 2005, from `http://www.cs.cmu.edu/~enron/`, 2005.

[6] G. Cselle, K. Albrecht, and R. Wattenhofer. Buzztrack: topic detection and tracking in email. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, pages 190–197, New York, NY, USA, 2007. ACM Press.

[7] M. Dredze, T. Lau, and N. Kushmerick. Automatically classifying emails into activities. In *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*, pages 70–77, New York, NY, USA, 2006. ACM Press.

[8] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 72–79, New York, NY, USA, 2003. ACM Press.

[9] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. Mylifebits: fulfilling the memex vision. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 235–238, New York, NY, USA, 2002. ACM Press.

[10] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *INFOVIS '00: Proceedings of the IEEE Symposium on Information Vizualization 2000*, page 115, Washington, DC, USA, 2000. IEEE Computer Society.

[11] J. Heer, S. K. Card, and J. A. Landay. prefuse: a toolkit for interactive information visualization. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430, New York, NY, USA, 2005. ACM Press.

[12] B. Kerr. Thread arcs: an email thread visualization. In *IEEE Symposium on Information Visualization (INFOVIS)*, pages 211–218, 2003.

[13] D. Lalanne, R. Ingold, D. von Rotz, A. Behera, D. Mekhaldi, and A. Popescu-Belis. Using static documents as structured and thematic interfaces to multimedia meeting archives. In *MLMI'04: Proceedings of the Workshop on Machine Learning for Multimodal Interaction*, LNCS, pages 87–100. Springer-Verlag, 2004.

[14] D. Lalanne, M. Rigamonti, F. Evequoz, B. Dumas, and R. Ingold. An ego-centric and tangible approach to meeting indexing and browsing. In *Machine Learning for Multimodal Interaction (MLMI' 07)*, 2007. Accepted for publication.

[15] M. Lamming and M. Flynn. Forget-me-not: intimate computing in support of human memory. In *Proceedings FRIEND21 Symposium on Next Generation Human Interfaces*, 1994.

[16] A. Perer, B. Shneiderman, and D. W. Oard. Using rhythms of relationships to understand e-mail archives. *J. Am. Soc. Inf. Sci. Technol.*, 57(14):1936–1948, 2006.

[17] M. Ringel, E. Cutrell, S. T. Dumais, and E. Horvitz. Milestones in time: The value of landmarks in retrieving information from personal stores. In M. Rauterberg, M. Menozzi, and J. Wesson, editors, *INTERACT*, pages 184 – 191, Zurich (Switzerland), 1 -5 September 2003. IOS Press.

[18] G. Smith, M. Czerwinski, B. Meyers, D. Robbins, G. Robertson, and D. S. Tan. Facetmap: A scalable search and browse visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):797–804, 2006.

[19] F. B. Viegas, S. Golder, and J. Donath. Visualizing email content: portraying relationships from conversational histories. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988, New York, NY, USA, 2006. ACM Press.

[20] S. Whittaker, V. Bellotti, and J. Gwizdka. Email in personal information management. *Commun. ACM*, 49(1):68–73, 2006.

[21] H. Xiao and I. F. Cruz. A multi-ontology approach for personal information management. In *ISWC 2005: Proceedings of the Semantic Desktop Workshop*, 2005.