Personal information management through interactive visualizations

Florian Evequoz and Denis Lalanne

Abstract—The PhD thesis 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) cross-media mining to generate indexes and alignments, (2) data clustering to organize personal information and (3) information visualization techniques, that should provide means to easily navigate through the personal information space. In this paper, we present our approach to these challenges using the personal email archive as an entry point into the personal information space. Indeed, we believe that emails are a particularly rich source of metadata for indexing personal information, as well as a representative subset of the whole personal information space. We finally present the work done during the first year of this PhD thesis and the roadmap for the future.

Index Terms—Personal information management (PIM), email, visualization.

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. 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. The very hierarchy of file storage system is the main reason for this failure. Indeed, as Bush pointed out in [2], our mind works by association rather than by following the rules of a static hierarchy. 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 of visualizing the personal information space interactively. The accent will be put on browsing capabilities, taking advantage of the structure of the personal information and the links existing between different pieces of information. The personal email archive will be used 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 brief state-of-the-art of personal information management (PIM), with a particular interest for existing email management and visualization systems. Section 3 is devoted to a more concrete presentation of the work envisioned for our particular research project. Sections 4 and 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, described in particular in [10]. As a complete state of the art is not the point of our discussion, we simply want to recall that previous research mainly focused on the data management perspective of PIM, trying to apply semantic modelling to PI. Some other works focused only on specific parts of PI, like emails or agendas, or offer a port of web search engines to the desktop. Our works differs from the cited ones because it focuses on finding similarities between different pieces of data and taking advantage of the links inferred from them to browse the whole PI with the help of information visualization techniques.

3 Project overview

The main goal of this PhD thesis is to investigate how interactive information visualization techniques combined with cross-media mining can help face the challenges of PIM. More specifically, the purpose of the project is to provide solutions and techniques to :

Create a PI network Cross-media information mining techniques shall be used to generate indexing metadata. On top of this metadata, thematic, temporal or social-network based links shall be created, connecting information of homogeneous or heterogeneous types (documents, pictures, music . . .), and thus building a PI network.

Organize PI The system shall provide means to organize PI in flexible ways, combining automatic clustering techniques based on indexes along with user-assisted clustering. This method will give the user a feeling that he controls to some extent the archiving and organization of his PI.

Navigate through PI The use of visualization suits well for navigating into a PI space. Therefore, synchronized views of personal data should offer different levels of details, provide details on demand and filtering mechanisms, and introduce browsing information as an alternative or a complement to more traditional search engines.

Support PIM Thanks to the novel access to personal information it provides, the system shall help the user elaborate new strategies to manage its PI and avoid being overloaded.

To reach these goals, we follow a user-centered approach, gathering user requirements at the beginning of the project and conducting user-satisfaction evaluation once a working system is available.

- *Florian Evequoz is a first-year PhD student at University of Fribourg, Switzerland. Email : florian.evequoz@unifr.ch.*
- *Denis Lalanne is a senior research assistant at University of Fribourg, Switzerland. Email : denis.lalanne@unifr.ch.*

Manuscript received 31 March 2007; accepted 1 August 2007; posted online 27 October 2007.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

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 [9]. 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 recipients 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. 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 [7] 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. Our system will then be an email-centric personal information manager. More specifically, the following steps need to be performed:

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

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 to the user several visualization techniques 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 [5]). We believe that the combination and synchronization of several visualization techniques applied to different dimensions will help the user browse instinctively through his PI. Therefore, a goal of this thesis 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.

As an extension, we also plan to provide access to professionnal data through the PI structure, thus offering an ego-centric view of professional data, for instance meeting recordings [8].

4 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 [4] and (c) the AMI meeting corpus [3], 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. Moreover, using the similarity based on the co-occurrences of words in the subjects and contents of emails, a 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 [6]. 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. 1). 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.

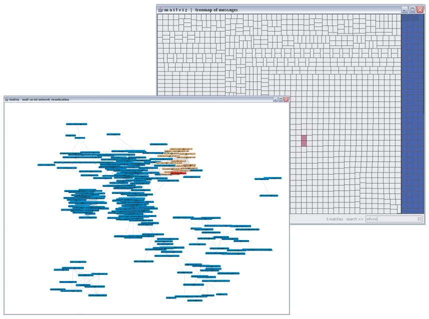
mailviz | treemap of messages
Mailviz : mail social network visualization
6 matches search >> infovis

Fig. 1. The implemented views of a personal email archive. The left window presents the social-network graph, where nodes map to email addresses. A subgroup of connected addresses is highlighted. The right window shows a treemap of emails, clustered by textual similarities. Results of a search with the keyword 'infovis' are highlighted in purple.

5 Future works and applications

In the short-term future, we will investigate further data mining and clustering methods on email data, in particular including the temporal dimension. However, the main focus will be on the visualization aspects, following the directions presented in section 3, namely introducing synchronized views of different dimensions, overview capability, details on demand and filtering mechanisms. The next step of the project in the long-term will be the alignement of the email structure with the remaining personal information. Finally, a user-satisfaction evaluation will validate our approach. The deliverables of this PhD thesis will serve two different research projects. In the first place, in the scope of Hasler Stiftung's Memodules, this thesis will provide an automatically extracted PI structure [1] on top of which the user will be able to connect tangible shortcuts to his personal information. In the second place, a goal of NCCR IM2.HMI project is to develop methods for accessing recorded meetings data. In this context, our thesis will facilitate personalized browsing of huge amounts of recorded data, thanks to similarity links that can be drawn between personal and professional information, thus opening the door to ego-centric professional data browsing [8].

References

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

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

[3] J. Carletta et al. 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.

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

[5] S. Havre et al. 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.

[6] J. Heer et al. 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.

[7] D. Lalanne et al. Using static documents as structured and thematic interfaces to multimedia meeting archives. In *MLMI*, pages 87–100, 2004.

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

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

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