Email as Electronic Memory: A Spatial Exploration Interface

Florian Müller, Martin Guggisberg, Helmar Burkhart
Computer Science Department
University of Basel
Basel, Switzerland
{*florian.mueller, martin.guggisberg, helmar.burkhart*}*@unibas.ch*

***Abstract*—Recently, electronic memory (e-memory) applications have come into research focus. Using personal data ranging from email to actual life logs, they are to provide us with an interface that facilitates functions such as retrieving, reminiscing, and reflecting information from our past – functions that we know well from our biological memory. We present an exemplary e-memory application based on personal email archives that supports reflection and reminiscence by providing a spatial layout of email communication data. The spatial layout is derived using a physical force relaxation simulation. In order to emphasize various properties of the communication network, the communication is represented as a weighted, directed graph. This allows analysis in terms of various metrics.**

***Keywords*-Multimedia, Information retrieval, Data visualization, Electronic memory, Lifelogging**

I. Introduction

In his Memex vision, Vannevar Bush imagined an information system that would allow the effective storage, editing and retrieval of all information encountered throughout a lifetime (see [1]). Today, his vision is often cited by proponents of lifelogging. In lifelogging, data arising from everyday activities both in virtual and real spaces is persisted and used for further processing. The recording of visual information has been a dominant component of lifelogging since its early application by Steve Mann, who described his first capture activities as *personal imaging*. However, he also noted that the result of having a lifelog is the ubiquitous availability of a *personal information domain* (see [2]). In this perspective, lifelogging provides the basis for an *electronic memory* (e-memory). Specific applications of e-memory cover memory deficit compensation (recall names and faces, retrieve lost objects), memory-related medical conditions (amnesia, dementia) and applications for reminiscence and self-reflection, which could be called explorative e-memory applications.

While lifelogging is still often an explicit activity carried out only by a few enthusiasts such as Gordon Bell of Microsoft (see [3], [4]), every-day use of computer technology without special focus on lifelogging already provides users with a vast corpus of data that could serve as the basis of a considerable e-memory. Note that a mere collection of data cannot be considered an e-memory. It is only when these data are made accessible and comprehensible (similar to our biological memories) that we can speak of e-memory. In this work, we show how the use of visualization techniques can aid users in exploring their digital corpora with the example of email communications. The visualization of the activities in a user's communication network is based on a spatial layout derived from a physical force relaxation simulation. Contacts are represented as particles, and interactions (emails) are represented as springs between particles. In order to obtain interesting views and insights of the user's communication, these communications are represented as a directed, weighted graph on which various computations can take effect.

This paper is organized as follows. In the remainder of this section, we introduce the notion of electronic memory and describe the status quo of electronic mail interfaces. In the second section, we detail the architecture of our electronic memory system and the data used. Sections three and four provide a detailed account of the mechanisms on which our system is based. In the final section, we conclude our results.

A. Electronic Memory

Specific applications of e-memory cover memory deficit compensation (recall names and faces, retrieve lost objects), memory-related medical conditions (amnesia, dementia) and applications for reminiscence and self-reflection, which could be called explorative e-memory applications. A classification of electronic memory applications was recently proposed by Sellen (see [5]). She defines five classes of electronic memory applications, which she labels as the five "R"-requirements. These are:

- Recollecting (re-experiencing past memories for the purpose of locating specific information items)
- Reminiscing (re-experiencing past memories for emotional reasons)
- Retrieving (retrieving some specific information, without re-experiencing)
- Reflecting (analyzing behavior over time and deducing conclusions from it)
- Remembering intentions (prospective memory, i.e. remembering to execute a decision taken in the past)

Our work focuses on the two aspects of retrieving and especially reflecting. One of the major challenges for reflecting is the reduction of complexity in the available data.

ISBN: 978-1-61208-129-8

It will be shown that the aggregation of spatial and grouping information, combined with several specific communication metrics, are valuable tools in achieving such a reduction and can help in making sense of large amounts of data through the use of a visual interface.

B. Interfaces for Email

Since we have chosen to use personal email communication as the basis for our e-memory application, a short review of state-of-the-art email interfaces is in order. Most current email interfaces are relatively unsuitable for providing e-memory interfaces to personal communication. They are mostly list-based, i.e. the messages are displayed in a chronological (or otherwise sorted) list, and details about the currently selected message are displayed in a separate view component. While search functionalities enable users to retrieve messages that match specific criteria, the list-based view is not suitable for reminiscing or recollecting: going through the list item by item is time-consuming, and the view does not aggregate the information contained in multiple messages – only one at a time is displayed in detail.

Several interfaces have been proposed to support higher-level views of email communication. Frau and others (see [6]) have proposed a dynamic email interface ('Mailview') which displays plots of email communication over time. The interface focuses on visually aggregating existing message attributes such as its size or the folder it is stored in. Viégas has developed several visualizations for email, including the PostHistory interface (see [7], [8]). It presents users with an overview of their email communications through the use of a timeline overview in the form of a calendar and a visualization of the contact network of the user. Depending on time windows and contact selection, the communication with one or several persons is displayed over time in the calendar view, and relevant social network context is displayed in the contact view.

Such proposed visualizations are more suitable for e-memory applications such as reminiscing and recollecting than list-based interfaces. They aggregate the information contained in multiple messages and present them in a single, at-a-glance view. The fundamental advantage of such visualizations is that they abstract from individual messages and display information on the time-aggregated structure and context of email. In this contribution, we would like to provide a new basis for email visualization for e-memory applications, namely a spatial layout for email which, intrinsically, does not have a representation in this domain. As we will show, deriving a spatial layout makes large email corpora comprehensible by providing an overall-view of email communication. Examples of such approaches can be found in social network visualization (see [9], [10]).

II. Architecture and Dataset

The overall architecture of the visualization system is depicted in Figure 1. In a first step, the email communication from several mailboxes is extracted via the IMAP protocol and preprocessed. In the preprocessing, various irregularities resulting from non-standard-conform email clients are eliminated, such as incorrect representation of dates and different character encodings. Once the email data has been regularized, it is stored in a relational database, using one table for the sent messages, and another table for a per-message and per-destination listing of the recipients.

In the next step, which is depicted in the top part of Figure 1, a spatial layout for all communication participants, including grouping information, is derived (*grouped contact map*). This is based on a physical force relaxation simulation, for which details are described in section III. The grouped contact map is the first part of the input for the actual visualization. The second part is generated through a graph processor (depicted in the bottom part of the figure). The graph processor uses the open source JUNG Graph API (see [11]) to represent the entire email communication as a directed, weighted graph. Based on this graph representation, several metrics can be applied and later used in the visualization. Details are described in section IV. The output of the graph processor is a *graph-metric map*, which contains the weights of the nodes and edges of the communication graph according to the applied metric.

The grouped contact map as well as the graph-metric map are used by the visualization unit to generate a graphical representation of the email communication. The visualization unit is written in Processing (see [12]), a Java-based language specifically engineered to support visualizations. It obtains the spatial layout from the contact map, and draws nodes and edges according to the graph-metric map. While the grouped contact map is loaded at start-up, the graph-metric map can be generated on demand in order to switch between various metrics.

Explored Dataset Statistics	
Unique email addresses	3.999
Total recipients	16.692
Total direct recipients	14.480
Total copy recipients	2.212
Avg. recipients/message	1,67
Messages with in-reply-to	5.016
Total messages	**10.218**

Table I
Statistical overview of the email communication data used in this work. The data is from one single user, spanning over several email accounts.

As stated, the data used in our work is extracted from several email accounts of a user. It spans over a period of 4 years, from 2006 to 2010. It contains a total of over 10.000 messages, and of almost 4.000 different contacts. In

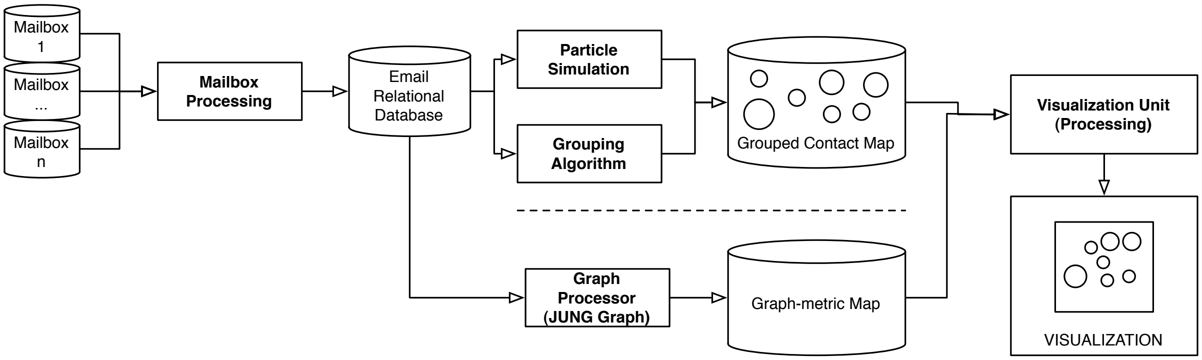

Figure 1. Overall system architecture. In a first step, the spatial layout and the grouping information is derived using the physical force relaxation simulation and a grouping algorithm. In a second step, this spatial layout is used in combination with a graph-based analysis of email communication to obtain a view focusing on a specific aspect of the communication.

order to efficiently use the email data, all the email header information (without the message bodies) was retrieved once using the IMAP4 protocol and then stored in a database. While the total size of all email messages including message bodies is approximately 6 gigabytes, the header information itself contains around 50 megabytes and can be processed quickly. Table I provides a statistical overview of the email communications used for the visualizations.

III. Spatial Email Layout

All communications extracted from the email accounts are processed and brought into a unified form, the resulting set of all messages is called $\mathbb{M}$. The spatial layout is generated by running a physical force relaxation simulation. Every contact (i.e. email address, with the special case that several email addresses belonging to the same person are collapsed into one single contact) in $\mathbb{M}$ is represented by a particle in a particle system. In analogy to the physical world, each particle is assigned a *mass*. The mass of the particle depends on the number of times the contact occurs in a communication (either as a sender or as a recipient). All particles in the simulation are repulsive towards one another (i.e. they have negative attraction). Messages define relations between contacts, and these relations are represented as elastic springs in the simulation. For every direct relation between two contacts (i.e. for every message where contact A is the sender, and B is the or a recipient, and v.v.), a weight update is performed for the involved particles: a spring is created between the two particles representing the contacts, and the weight of the particles is adapted. Both the particle mass and the strength of the springs between them depend on the frequency and nature of the communication they originate from.

$$\Delta m = \frac{1}{n_R} \cdot \frac{\gamma_i}{m} \tag{1}$$

The mass update of the particles, Δm, is calculated according to the following formula, where n_R is the number of recipients of a message, and γ_i is a communication specific growth constant (e.g. direct messages are rated higher than messages received as a carbon copy), and m is the current mass (see Equation 1). The strength update of the spring is calculated in an analogous manner, but with a different growth constant. The weight and spring update is illustrated in Figure 2. Note that for obtaining better clarity, the weight and strength updates are calculated using only the term $\frac{1}{n_R}$.

Once the particles with their respective weights and the springs between them have been created, the particles are initially placed at a random location in the space of the simulation. Then, the simulation is started, and we wait until the effect of the default mutual repulsion and the attraction through the springs have lead to a stable state after a certain relaxation time. In this state, the particles have a stable position, and the spatial map can be generated accordingly.

Through the simulation procedure, we gain a spatial model of the email communication network. As will be shown in the results section, the spatial layout already provides the user with an overview of her email communication that shows important clusters of contacts and their interrelation. Note that the absolute position of a node is not relevant (since it is arbitrary, for in every simulation run, it may be different). However, the relative position reveals important information, such as distance and closeness to neighbors and other clusters of nodes.

In addition to the clustering implicit in the relaxation simulation, an algorithm for grouping the contacts is employed during the simulation. Nodes that communicate with each other are assumed to influence one another, and for every communication between a sender and one or several recipients, the sender exerts a certain amount of influence on the recipients. After all messages have been processed, every

ISBN: 978-1-61208-129-8

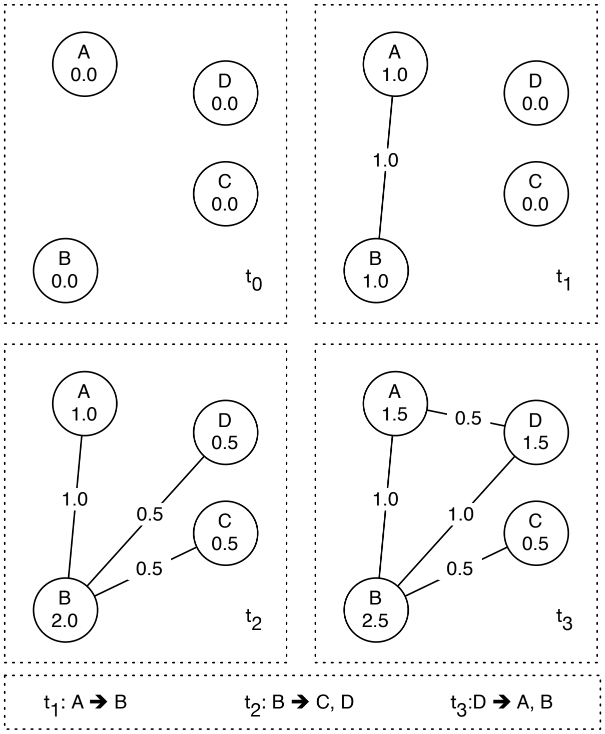

Figure 2. Illustration of the particle simulation using unit weights. The three communications 'A → B', 'B → C, D', and 'D → A, B' occur one after the other at times t_1, t_2, t_3. As they occur, the particle weights and the springs between them are updated accordingly.

contact node is influenced by one or several other nodes. If for any node, the relation between its own mass and the largest amount of influence exerted upon is below a certain threshold, the node is said to belong to the node with the largest influence on it. This relation is applied recursively, and results in the creation of groups within the contact network. The grouping is used to apply different colors to different groups, which results in an additional simplification of the interface. The spatially clustered and colored email contact network layout is the basis for the next step, the graph-based analysis of the communication.

IV. Graph-based Analysis

Similar to the case of the particle system, the entire email communication can be represented as a directed, weighted graph. It is constructed in a similar manner as the particle system used to derive the spatial layout: every contact is represented as a vertex, and every communication relation (sender to recipient) is represented as a directed edge. In a default case, the weight of the vertices and the weight of the edges can be derived as in the case of the particle simulation. A more interesting approach is to find various metrics according to which the vertices and edges are weighted. Depending on the metric chosen, different aspects of the email communication can be shown. While typical social network metrics such as *betweenness centrality* are more interesting for social networks that comprise several ego-networks, for our case, where we look at one single ego-network, other metrics have proven to be of more relevance. We have used the information contained in the email header information, especially: (a) whether a message is a direct reply to another message, and to which, (b) how deep threads run (a back and forth of replies and replies to replies), and (c) who forwards information. The results of these metrics will be shown in the results section. Note that since we can look at email communication as a graph, we can apply any metric we wish.

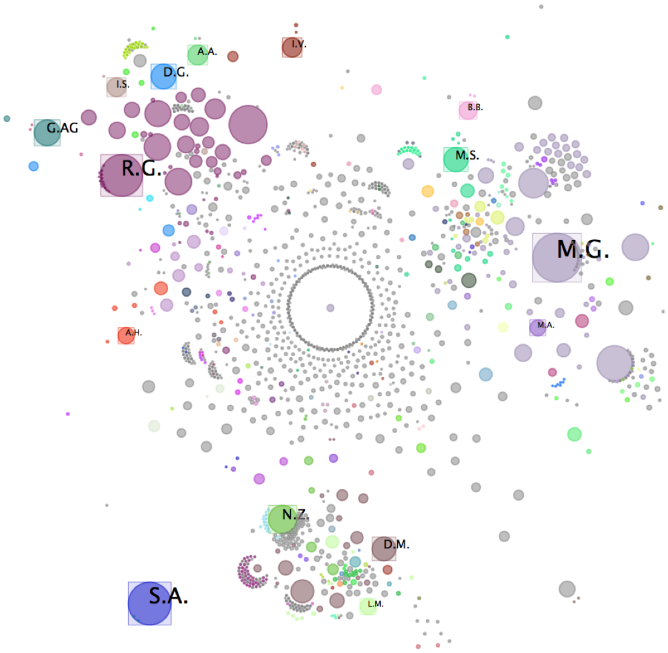

Figure 3. Spatial layout and grouping as resulting from relaxation simulation and grouping algorithm. The weights of the nodes reflect the number of overall occurrences in communications.

V. Results

Figure 3 shows the base layout derived from the relaxation simulation. The owner of the mailboxes is located in the center. As can be seen, three major clusters have been formed: one in the south, one in the northwest, and one to the east. Spatial proximity suggests knowledge of one another and communication with one another. The colors show additional structural information for clusters. The ring of small nodes around the owner of the mailbox are contacts that have only occurred few times in communication and that are not networked (i.e. they have never occurred in messages that were destined to multiple recipients). The size of each node is directly dependent on the number of times a contact has occurred in any communication.

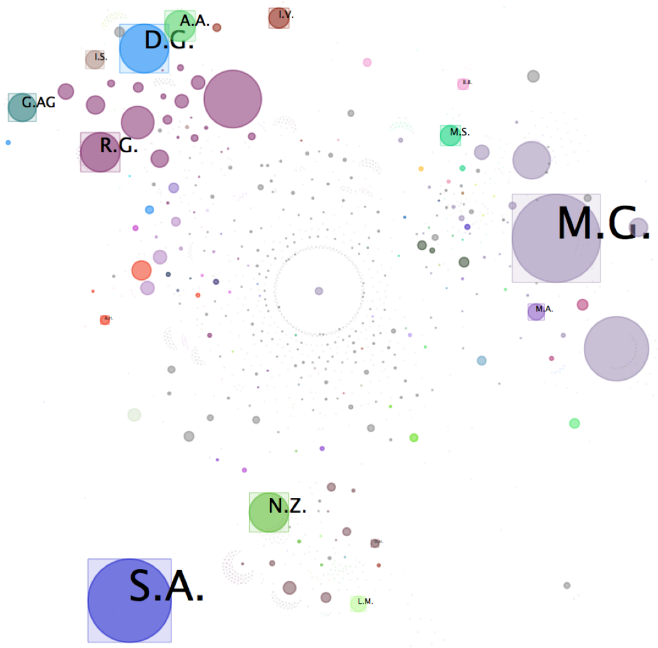

Figure 4. Weighting based on counting frequency and depth of replies between contacts. Three contact clusters can be clearly identified. The internal structure of each cluster can also be assumed: in two cases, one main communicator is identifiable, while in the case of the cluster to the northwest, there are several equal contributors.

Figure 4 shows a layout derived from the base layout, where the weighting of the graph depends on the reciprocity of communication. The header fields 'Message-in-reply-to' and 'References' are used to determine if a message was sent in reply to another message. A high number of replies (of various depths) in communication between two contacts suggests that they cooperate more closely than others and that the flow of information between them is two-sided. Within the three clusters and compared to the base layout, we can see how contacts with whom the owner of the mailbox communicates reciprocally are prioritized. Finally, in Figure 5, the layout is derived from information forwarding activity. The subject as well as forwarded headers are analyzed to determine whether a message is a forwarded message, and the nodes are weighted according to the number of messages they have forwarded. In addition, the edges between the nodes are weighted according to the number of messages forwarded between the two nodes. As can be seen, other nodes are prioritized than in the reciprocity example. The forwarding of information suggests organizational hierarchy, which may be formal or informal. Nodes that are weighted high are likely to be important organizers and distribute information in the network.

VI. Conclusion

We have shown how a spatial layout for email communication data can be derived based on a physical force relaxation simulation. The spatial layout is the basis for a two-dimensional interface for email that is targeted at e-memory applications. The base layout functions as a summarization of information in two manners: first, it generates clusters of related nodes and shows these clusters in relation to other clusters. Second, through the use of an grouping algorithm that calculates mutual influence of nodes, these can be further grouped by applying different colors.

Since the email communication data is represented as a graph, several metrics can be applied to it. This allows the analysis and visualization of various aspects of the communication, such as cooperation, information flow, and hierarchy. It was initially stated that our aim is to develop an e-memory application focusing on retrieving information and, especially, reflecting on it. If the application of various metrics on the graph-represented communication is seen as reflecting on that communication, our system provides not only reflecting capabilities, but also an interface to visualize them. Using a graphical representation of the results obtained allows users to gain insights into their communication patterns and networks. Apart from this user experience-centric view, further possibilities arise.

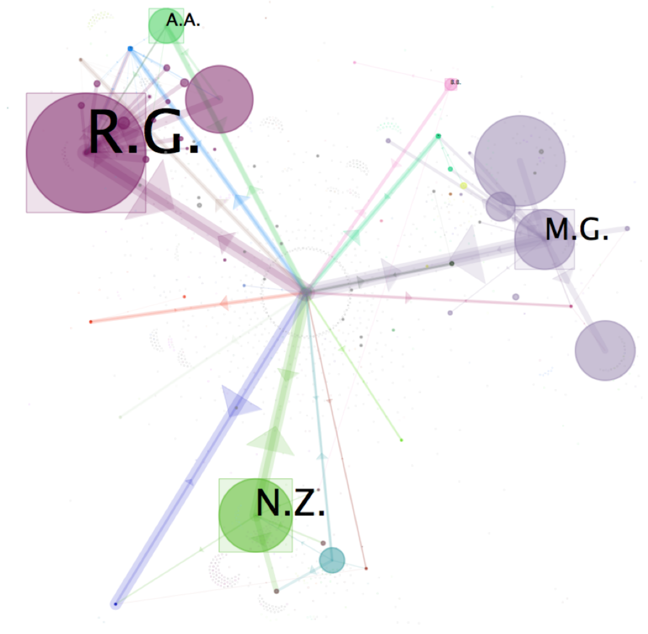

Figure 5. Weighting based on number of forwarded emails, with the direction of the forwarding indicated by arrows. This visualization shows who forwards information, which allows the assumption that this person has an important organizational role within the network.

The structural information derived from the communication network can be used to communicate more efficiently.

ISBN: 978-1-61208-129-8

For example, features such as the recently introduced *priority inbox* from Google (see [13]) can be based on our system. It allows the classification of message priority relative to the importance or function of the sender in an email communication network. We are currently also evaluating our system using multiple ego-networks of email communication. In such applications, the inferred knowledge is not limited to a single user, but is situated at an organizational level. In such a context, the scope of possible applications is even wider.

References

[1] Bush. V.: As We May Think. In: Atlantic Monthly, vol. 176, 1, pp. 101–108 (1945)

[2] Mann, S.: Wearable computing: A first step foward personal imaging. In: ACM Computer, vol. 30, 2, pp. 25–32 (1997)

[3] Gemmell, J., Bell, G., and Lueder, R.:MyLifeBits: a personal database for everything. In: Communications of the ACM, vol. 49, 1, pp. 88–95 (2006)

[4] Gemmell, J., Bell, G., Drucker, S., and Wong, C.: MyLifeBits: fullfilling the Memex vision. In: Proc. ACM International Conference on Multimedia, pp. 235–238 (2002)

[5] Sellen, A.J. and Whittaker, S.: Beyond total capture: a constructive critique of lifelogging. In: Communications of the ACM, vol. 53, 5, pp. 70–77 (2010)

[6] Frau, S., Roberts, J.C., and Boukhelifa, N.:Dynamic Coordinated Email Visualization. In: Proc. WSCG, pp. 187–193 (2005)

[7] Viegas, F.B., Golder, S., and Donath, J.: Visualizing Email Content: Portraying Relationships from Conversational Histories. Proc. CHI 2006, pp. 979–988 (2006)

[8] Viegas, F. B., Boyd, D., Nguyen, D.H., Potter, J., and Donath, J.: Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments. In: Proc. System Sciences, pp.10–19 (2004)

[9] Hansen, D., Shneiderman, B., and Smith, M.: Visualizing Threaded Conversation Networks: Mining Message Boards and Email Lists for Actionable Insights. Active Media Technology, pp. 47–62 (2010)

[10] Freire, M., Plaisant, C., Shneiderman, B., and Golbeck, J.: ManyNets: An Interface for Multiple Network Analysis and Visualization. In: Proc. CHI, pp. 213–222 (2010)

[11] JUNG Graph API. http://jung.sourceforge.net, last retrieved 16 December 2010.

[12] Processing. http://processing.org, last retrieved 16 December 2010.

[13] Gmail Priority Inbox. http://mail.google.com/mail/help/priority-inbox.html, last retrieved 16 December 2010.