Using Personal Information Visualization for Document Retrieval

Paulo Gomes
Instituto Superior Técnico
Av. Rovisco Pais, 49
1050-001 Lisboa, Portugal
+351-213100289
paulo.gomes@ist.utl.pt

Sandra Gama
Instituto Superior Técnico
Av. Rovisco Pais, 49
1050-001 Lisboa, Portugal
+351-213100289
sandra.gama@ist.utl.pt

Daniel Gonçalves
Instituto Superior Técnico
Av. Rovisco Pais, 49
1050-001 Lisboa, Portugal
+351-213100289
daniel.goncalves@inesc-id.pt

ABSTRACT

During our constant interaction with computers, we generate large amounts of personal information. However, it is often hard to find a certain item we are looking for, since our data is spread throughout several places and applications. Nevertheless, a meaningful visualization technique may be the solution to this problem. We present VisMe, an interactive integrated visualization system for personal information that allows users to meaningfully navigate and retrieve their data. Relevant concepts (people, subjects and documents) are uniformly displayed in interconnected timelines. Each of these items can be progressively expanded into new timelines, allowing relations between them to be explored in a simple, straightforward way. Several avenues can be simultaneously explored in context, thus giving users insights into their digital selves that current tools have a hard time providing. VisMe goes beyond traditional desktop search solutions by allowing not only keywords, but also the relations between different kinds of personal information to be used to retrieve personally relevant data.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces (GUI), Interaction styles, User-centered design*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

General Terms

Design, Human Factors

Keywords

Personal Information Retrieval, Personal Information Management, Information Visualization, User-Centered Design

SIGIR'10, Workshop on Desktop Search, July 23, 2010, Geneva, Switzerland.

1. INTRODUCTION

An increasing number of devices, such as laptop computers and smartphones, have pervaded our daily lives, providing us with the means to generate and store large amounts of personal information, from documents to emails. Either directly or indirectly, this data can help us understand who we are, what we do, and what we are interested in. ”What was I doing in January 2005? Where can I find the article John sent me two months ago?” These are questions that an effective analysis of our personal information should be able to answer. This not straightforward. In fact, despite considerable growth in storing capacity over the last years, methods and applications for managing and retrieving personal information have not suffered substantial improvements.

Hierarchical organization is one of the prevalent organizational systems, but it has several shortcomings. Its main disadvantage has to do with the effort requested from users to consistently classify every piece of data. Furthermore, different kinds of personal information are managed by different applications, with little or no links between themselves. A user's personal data is, thus, scattered and difficult to find.

A variety of systems have been developed to visualize different kinds of collections and histories, from emails to medical records, but while some of these systems can successfully reveal relevant patterns in the information, they are limited to particular sources. Those that are not, fail to provide a unified representation of information from different sources. We still lack an effective global visualization system for all of our personal information.

We propose a solution, VisMe, in which all relevant personal information is indexed as a whole. Links lost by applications, such as between a document in the filesystem and the email it was attached to, are recreated. Based on that index we developed a visualization centered on the most relevant autobiographic clues: time, people, and subject. VisMe allows the exploration of personal information in an efficient and understandable way. It abstracts from the different data sources to present semantically relevant information instead, and it allows several avenues to be explored simultaneously in context. As such, VisMe provides the means for finding information and retrieving interrelated items. Furthermore, by providing a synergistic visualization tool, VisMe allows users to efficiently navigate their data and helps them find patterns that are personally relevant.

2. RELATED WORK

Multiple applications have been developed to manage and retrieve personal information. Stuff I've Seen [3] accesses information independently of its initial form and the search is done using a word associated to the data or one of the many types of properties or metadata. The interface is halfway between browsing and keyword search. It also combines keyword and faceted search, making it easier to search for whatever criteria users do remember. MyLifeBits [4] stores and accesses virtually all data of a lifetime, inspired by Bush's Memex [1]. Given the quantity and diversity of information, MyLifeBits stores information and metadata.

Other solutions use information visualization as a way to search and explore personal data. Several applications have been created, usually focusing on a single information source, (email, instant messaging logs, or text documents). Themail [9] stands out from other email visualizations, by its simple and attractive interface and by its ability to display patterns in email content. ChrystalChat [8], an instant messaging visualization, displays a conversation space in an interesting three dimensional structure, even though its content representation, besides the textual display of the actual messages on demand, is limited to a peripheral mood indicator. There are also systems that allow searching and browsing to visualize information from multiple sources. Milestones in Time [6] takes a familiar list display and couples it with a timeline filled with landmarks to provide a simple and appealing interface for multimedia history search and browsing. FacetMap [7], with its facet bubbles, also avoids simple lists and manages to join a visual representation with the underlying searching mechanism in a simple and relatively effective way. Feldspar [2] allows users to interactively and incrementally construct association queries. Focusing more on the connections between entities and not on the entities themselves, it is possible to find things about the individuals that wouldn't be found if searching the items separately.

Evidently, none of these visualizations manages to provide a unified content overview of a heterogeneous collection of documents. This is the void VisMe attempts to fill: an interactive visualization of personal information taken from multiple sources which can help search for information and find relevant patterns while still allowing the micro-data (individual documents, emails, etc.) to be retrieved in context.

3. PROPOSED SOLUTION

To explore the personal document collection of a user, we first need a way to gather and index that information. Scribe [5], an automatic indexing application which is not the focus of this paper, is used for that purpose. It indexes and interconnects emails, documents, web pages, etc. Above this indexed data, a layer was developed to facilitate integration and to provide efficient access to the personal information. We then developed an interactive graphical user interface which handles personal information representation. We kept three goals in mind when designing this interface. First we wanted it to be simple to understand and manipulate. Second, we wanted to allow the data to be explored in context, as an inter-related whole, rather than seeing the results of individual queries one at a time. Finally, we wanted to treat information from different sources and natures in a uniform way.

The fundamental idea behind our solution is that every element in the visualization, namely keywords (most significant words extracted from each document according to their tf-idf weight), contacts (authors, senders, or receivers of information), and documents (files, individual emails, instant messaging logs, etc.), can be expanded to display the keywords, contacts, and documents which in turn are related to them (all documents from an author, all keywords in a document, all keywords in messages from a person, etc.). Each element in the visualization is represented by a word and three buttons from which three timelines can emerge, one for each facet (Figure 1).

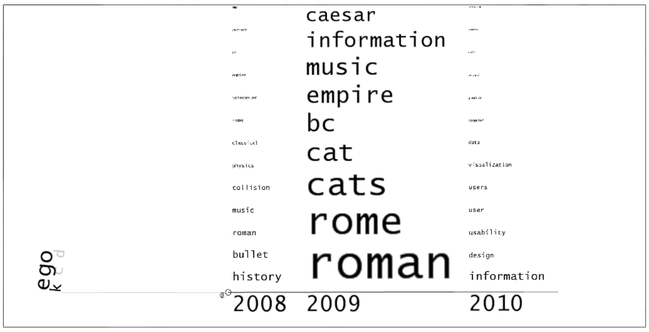

Figure 1: Expanded keywords

The visualization starts with an element representing the user ("ego"). Placing the cursor over it prompts the display of three buttons representing keywords, contacts, and documents ("k", "c", and "d"). Clicking on one of them creates a zoomable timeline on which the respective elements are displayed. The most representative elements appear larger and closer to the bottom and there is a size threshold under which elements are no longer shown. Timelines can be zoomed in progressively to display years, months, or days, by clicking on the desired period. Besides expanding a timeline perpendicularly with a simple click, users can drag the timeline out of the icon to whatever position and orientation they want. Following an expansion, any element in the timeline can be subjected to the same process as the initial element – and there is potentially no limit to this. The user is allowed a progressive exploration of their information on simple and comprehensive timelines.

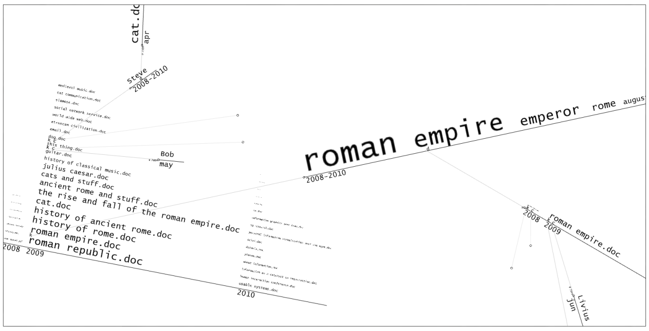

Figure 2: Timelines expanded from several elements

VisMe can maintain several expanded elements on the same canvas (Figure 2), which users can then translate, rotate, and zoom in and out of using the mouse buttons and scroll wheel, so that they can observe and interconnect the various results of their ongoing research. Besides controlling the view manually, users can also double click on a timeline to bring it into focus automatically.

When users move the mouse over an element, all its instances become highlighted. This makes it easy to follow the evolution of an element over time. The color depends on the position of the mouse over the element, as there is a gradient from red (left) to blue (right). Left clicking fixes the color, another left click resets it to black.

To use VisMe as a Desktop Search tool, we provide text search capabilities. Keeping a minimalistic design, simply pressing a character on the keyboard prompts the display of an input box on the top left corner of the screen containing the text as it is written and icons ("k", "c", and "d") to define the search. As each character in written, the resulting string is searched and a possible result is shown to the user as grey text next to the string, together with an indication of the number of results for the currently selected facet (Figure 3). Pressing the tab key will complete the string to match the currently visible result, pressing the up and down keys will cycle through results.

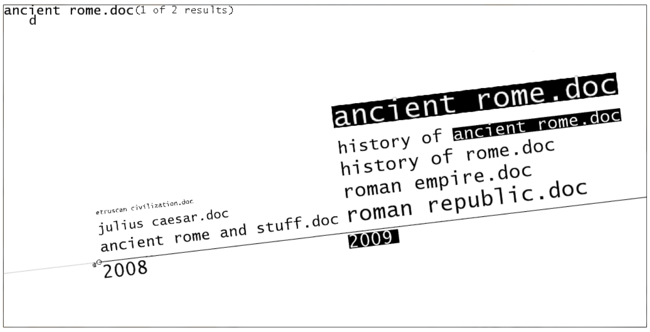
ancient rome.doc(1 of 2 results)
d
ancient rome.doc
history of ancient rome.doc
history of rome.doc
roman empire.doc
roman republic.doc
2009
etruscan civilization.doc
julius caesar.doc
ancient rome and stuff.doc
2008

Figure 3: Searching with VisMe

If the current string appears anywhere on the expanded timelines, the respective elements will be highlighted. If there is an element that matches the search string but is not representative enough to be displayed on the timeline, it will appear on top the respective column. It will be significantly larger than the element before it, showing that the element does not follow the same size convention as the rest of the timeline. Also, whether or not there is an exact match in any time period, the respective time will be highlighted at the bottom of the timeline.

Evidently, there is a limit to how much information can be displayed. To deal with cluttering, we implemented several measures. When the mouse is not above a timeline, only the most representative elements at the bottom are shown. Also, timelines initially appear in a horizontal state, with the most representative elements from all time periods. As such, only relatively thin lines of important elements are seen most of the time. We also let users reposition or hide existing timelines they may consider less relevant. By moving the view and reorganizing information, the necessary relevant information can be kept in view.

4. USER TESTS

Volunteers aged 17 to 29 ($\overline{x} = 23.7$, $\sigma = 2.7$) and with self reported high level of experience with computers ($\overline{x} = 3.7$, $\sigma = 0.47$, scale of 1 to 4) were asked to perform a series of tasks using the prototype application over a set of 1004 text documents authored by 102 people. This data set was crafted based on our insights on personal information spaces to be representative of a real collection of documents, with realistic trends and patterns for each tested combination of facets and enough documents and authors to make it hard to just stumble upon them without significant help from the interface. There were eight document recovery tasks, which consisted of recovering of a document with the knowledge of one or of a combination of its facets (time, most representative keyword, and author, as well as a single task in which the actual file name was given). For instance, "Find a document written by Bob about the Internet that you read in May of 2009" or just "Find a document about guitars". Before being handed the tasks, users were given a five minute demonstration of the prototype and all its features. Users were then given another five minutes to experiment with the interface freely. The order in which tasks were performed was random, in order to prevent result biases. These tasks were timed and recorded for later analysis. Users were asked to grade the difficulty of each task in a scale of 1 (very difficult) to 4 (very easy) upon completion. The time limit for each task was 150s, after which users would be told to move on to the next task. There was also a questionnaire to be answered at the end of the session which focused on the satisfaction with several general and particular aspects of the interface using a four point scale.

4.1 Results

Most tasks were completed successfully, (with failures per task in all twenty sessions $\overline{x} = 1.9$, $\sigma = 2.4$). On average, successful tasks were completed in about 52s ($\overline{x} = 52$, $\sigma = 27$). Users also considered the tasks to be easy ($\overline{x} = 3.4$, $\sigma = 0.899$). This provides some evidence that VisMe can be an effective retrieval tool for documents based on time, keywords, and authors.

There are, however, two exceptions. Both the tasks in which users were asked to locate a document based on a keyword-author combination (and, in one task, also time) have the greatest number of failures (6 and 5 in 20 sessions), the highest average completion times ($\overline{x} = 84.5$, $\sigma = 27.5$ and $\overline{x} = 88.7$, $\sigma = 27.6$) and were considered to be the most difficult tasks ($\overline{x} = 2.6$, $\sigma = 1.143$ and $\overline{x} = 2.55$, $\sigma = 0.999$). It is possible to complete these tasks, as the majority of users did, but evidently it is harder to find documents based on a combination of two facets other than time.

Completing these tasks requires either expanding documents from both facets and cross checking the results, or expanding documents from one facet and expanding the other facet out of each document one by one. Most importantly, with our data set, the number of documents that matched one of the facets (or one of the facets plus time) ranged from only two to a dozen, making it feasible to actually inspect each individual item or cross check the results of two separate timelines. More extensive result sets could have made these tasks impractical. Some users expressed some frustration with the fact that timelines were not filtered according to their hierarchy, as expanding a documents of a keyword expanded from a contact did not yield a list of documents written only by that author about that keyword.

The questionnaire shows that users were generally satisfied with the system ($\overline{x} = 3.35$, $\sigma = 0.49$). They did not find it difficult to use for the most part ($\overline{x} = 3.20$, $\sigma = 0.69$), but they did find it somewhat difficult to learn ($\overline{x} = 2.75$, $\sigma = 0.85$). They also felt at times that the system did not offer sufficient functionalities ($\overline{x} = 2.85$, $\sigma = 0.59$). Although users did not generally consider the control over the viewing area by rotating, scaling, and translating difficult ($\overline{x} = 3.15$, $\sigma = 0.81$) we did observe that many users were uncomfortable these actions. This may be explained by the short experience with an unfamiliar and unconventional interface, but perhaps there is room for improving and smoothing out the controls, even if only for the sake of novice users.

Finally, although the retrieval times were apparently higher than those expected from keyword-based desktop search tools, such tools lack the support for certain, more complex, tasks, such as the ones required from VisMe's users. For instance, to find a document based on the combination of author, date and subject users have to perform multiple keyword queries and to interrelate information on their own, greatly increasing the overall time. VisMe provides explicit support for those tasks. Although, as described, those are the tasks where VisMe still performs comparatively worse, task completion was high, proving that it was helpful, having been designed precisely to show multiple avenues of exploration at once and to leverage the user's memory of the context surrounding each document. Future improvements to the interface (see below) will further improve its performance for these complex tasks.

5. FUTURE WORK

The tests have shown there is a problem with the retrieval of a document based on the combination of two facets. One can expect the problem to be even worse if users were to attempt a combination of several authors and keywords. A possible solution to this problem lies in filtering. We have developed a working, although still untested, solution. In the current prototype, users can simply drag any keyword, contact, and file name, into any timeline, as many times and in any combination they want, to filter it. For instance, clicking and dragging the mouse from a keyword in one timeline to the space occupied by a second timeline will add the keyword as a filter to it. Active filters appear to the left of the timeline and a simple click will remove them. This has been extended to the search string that appears on the corner of the screen, which can also be dragged into any timeline, making it easy to filter a timeline according to any existing facet that users find through textual search. We are also planning on modifying the prototype so that filters can be passed through hierarchies of timelines.

Although it was not made specifically evident in the user tests (the necessary information on screen for each task was relatively small), we have also been working on additional solutions to cluttering. The current idea being worked on is collision detection with smooth separation of timelines.

6. CONCLUSION

An efficient, integrated, visualization of personal information could allow us to search and retrieve files or discover interesting patterns. We presented our solution, VisMe, an interactive personal information visualization system. The main idea behind VisMe is to progressively expand and lay out over time the information related to one of three facets: keywords, people, and title. It provides a unified and coherent representation of heterogeneous information and it can display an overview of the entire content of a document collection in a way that allows for the interrelation and interconnection of several of its elements. Usability tests validated VisMe's capabilities as an in-context document search and retrieval tool but also revealed complications in combining many facets in the same search, something we have attempted to solve by developing a filtering mechanism that can leverage the text search and the presence of multiple timelines on screen. Because the tests we conducted were not performed with the users' own information, they do not fully validate the solution. Ideally, VisMe takes advantage of a person's memory of their own information; the context around each document, the history of each contact, etc. The artificial data also made it impractical to compare document search and retrieval performance with traditional approaches, such as browsing document folders or searching. Still, the goal of this evaluation was to obtain an early validation of the initial progresses with our prototype, which we believe we did, and we are planning on performing more conclusive tests using actual personal information of users in the near future.

7. REFERENCES

[1] V. Bush and J. Wang. As we may think. *Atlantic Monthly*, 176:101–108, 1945.

[2] D. H. Chau, B. Myers, and A. Faulring. Feldspar: A system for finding information by association. In *ACM SIGCHI PIM2008, the Third International Workshop on Personal Information Management, Florence, Italy*, 2008.

[3] E. Cutrell, S. Dumais, and R. Sarin. New directions in personal search ui. In *SIGIR 2006 Workshop, Seattle, Washington*, 2006.

[4] J. Gemmell, G. Bell, and R. Lueder. Mylifebits: a personal database for everything. *Commun. ACM*, 49(1):88–95, 2006.

[5] D. Gonçalves and J. Jorge. In search of personal information: narrative-based interfaces. In *IUI 2008, New York, NY, USA*, pages 179–188, 2008.

[6] M. Ringel, E. Cutrell, S. T. Dumais, and E. Horvitz. Milestones in time: The value of landmarks in retrieving information from personal stores. In *INTERACT*, 2003.

[7] 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.

[8] A. Tat and S. Carpendale. Crystalchat: Visualizing personal chat history. *Hawaii International Conference on System Sciences*, 3:58c, 2006.

[9] F. Viégas, 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.