Viewing Personal Data over Time

Sonja Knoll, Aaron Hoff, Danyel Fisher, Susan Dumais, Ed Cutrell
Microsoft Research
1 Microsoft Way, Redmond, Washington. USA 98052
{sonjak, aaron.hoff, danyelf, sdumais, cutrell}@microsoft.com

ABSTRACT
Desktop search applications are changing the way we interact with personal information: we have a choice of whether to view files within their traditional siloed hierarchies, or brought together in search results. In this position paper, we discuss the advantages of temporal ordering of desktop search results, and present an interface that aggregates search results over time.

Author Keywords
Desktop search, personal information management

INTRODUCTION
Desktop search applications are changing the way we interact with personal information. Traditionally, personal computer systems kept information siloed and hierarchical: email lived only within email clients, and could only be read from within them; files lived within folders, and could only be accessed from desktop applications; etc. While limited search capabilities were available, they were inconvenient, difficult and inconsistent.

The advent of desktop indexes changed that. With a background process generating and maintaining a desktop index, and a series of indexing plugins that can extract words from all files, the file system is conceptually flattened: keywords can be used to collect documents from anywhere. A user can search for content through email, word-processing documents, notes, spreadsheets, and other sources all at once. Indexes can also store metadata attributes; thus it is possible to search for emails that contain attachments, sent before a particular date, in HTML format, or from a particular person—all information that is stored in metadata, rather than content.

This is far from news. The Stuff I've Seen [3] and Phlat [2] projects have discussed the importance of exposing both content and rich metadata, and have presented tools that allow users to work with their personal information.

Most consumer applications—such as Google's Desktop Search and Microsoft's Windows Desktop Search—present a ranked list of hits or a rich list in response to search queries. The ordering is sometimes determined by relevance score and other times, it is sorted by a time attribute. Relevance rankings are traditionally timeless and based on text analysis; something I worked on several years ago is ranked in the same way as something from today. Even ordering by time is complex: as we have previously noted [3], there are many notions of "time" on a computer system and for users; files can be meaningfully labeled by the time they were created, last modified, or last accessed. We also know from research in cognitive psychology that people remember information not in terms of exact time, but rather in episodes and relative to other events in time [3]. It is critical to get all this right in order for the user to find the information they are looking for.

In this position paper, we discuss temporal aspects of desktop search results from two perspectives: mining novelty and information visualization. In each, we discuss several related projects and then present our own contribution.

Indexing and the Desktop
In order to easily access temporal patterns in desktop information, the classic desktop search index needs to be enhanced. Using current versions of desktop search, it is difficult to directly ask questions like "has there been a recent burst of activity around a particular topic or person?" Instead, every document that mentions the term would need to be retrieved, analyzed, and its metadata examined.

We have modified Windows Desktop Search to calculate a broader set of corpus statistics, to add new properties to the index, and to generate a forward index that can be used to later compute additional metadata. Both of the applications we discuss in this section are based on these modifications to Windows Desktop Search.

PERSONAL VISUALIZATION AND TEMPORALITY
Showing richer temporal information can be useful, especially for personal content. A number of research projects have addressed this question.

Research Approaches
LifeLines [10] is a visual representation of a user over time. Collected from specific sources, LifeLines can then show personal histories. LifeLines is designed for cases when rich metadata is available with extensive labeling that can divide events into categories and subcategories with beginnings and endings: health care records are used in their example.

LifeStreams [5] puts all documents in sequential order, allowing users to scroll through their documents through time. Integrated search and metadata selection are filters on the data.

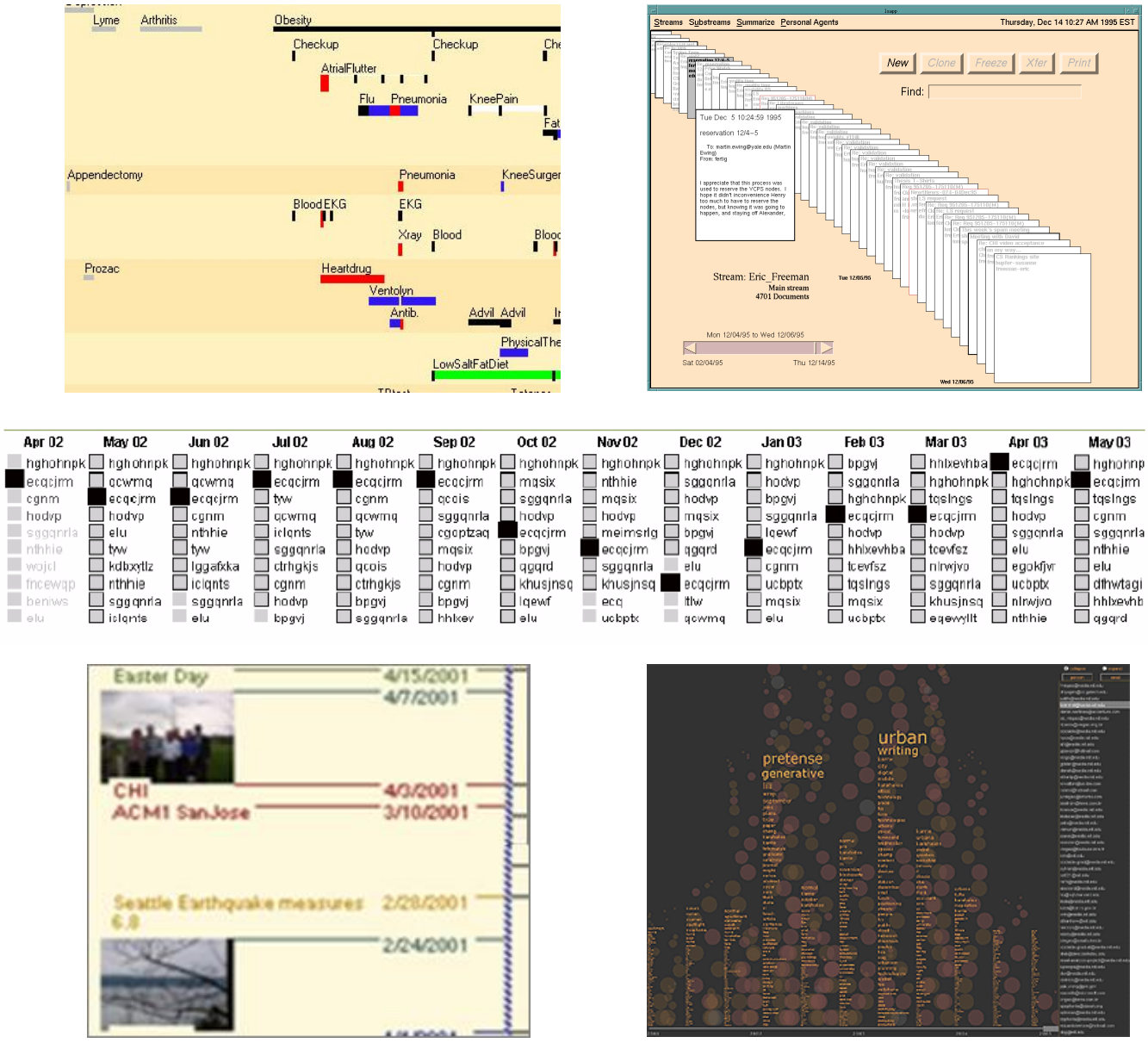

Figure 1: Five interfaces for personal information over time: LifeLines, LifeStreams, Soylent, Milestones, and Themail.

"Milestones" [11] adds to the standard desktop search list by augmenting the search results with a timeline and associated metadata. Times on the result list are annotated by entries from the user's own files (important calendar meetings and personal photographs), as well as collective information (general calendar and news events).

Limited to the email domain, Soylent [6] includes a temporal visualization that shows the top ten email correspondents in each month. The changes in that list over time can suggest different projects, priorities, and social circumstances.

Similarly, Themail [12] shows the most common words appearing in email over time. It allows users to see the general content of their email changing over time, and to look for major changes in their email at a glance.

Showing Personal Narratives

Personal Narratives is a visual tool that also shows information flows over time. A screenshot of Personal Narratives is shown in Figure 2. Like other temporal PIM visualization projects, Personal Narratives looks at changes in personal information flow over time. Unlike these other approaches, however, it does not attempt provide an overall view of the dataset, but rather shows results for a small number of terms at a time.

Personal Narratives is a revision of *Narratives* [7]. Narratives visualizes change in the number of blog entries

posted on news topics over time. In Personal Narratives we instead visualize the number of references to query terms over time.

Personal Narratives is driven by our modified version of Windows Desktop Search. We extract a subset of 200 words from each document, and create a time-labeled forward index based on them. This data is imported into a database which supports visualizing terms over time, showing the relative number of hits.

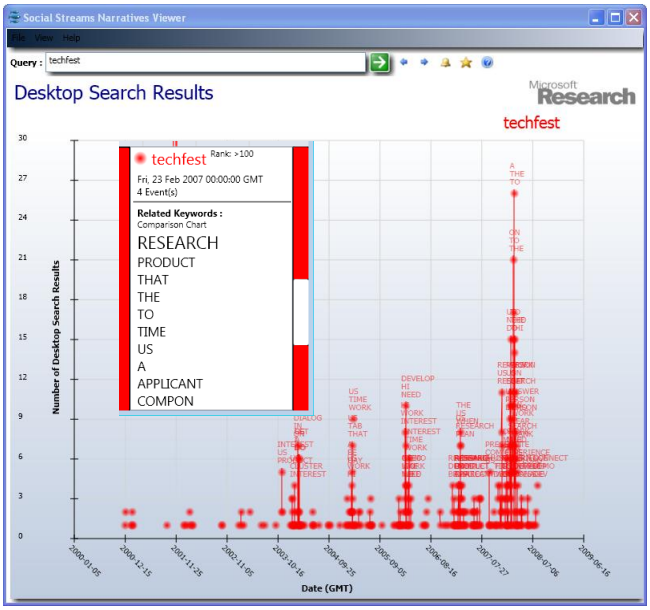

Figure 2: Personal Narratives, searching for the term "techfest". Inset, a tooltip from Feb 23, 2007.

In Figure 2, we show the interface for Personal Narratives. The screenshot covers the personal archive for one user over eight years. The timeline at the bottom runs from 2001 through 2008. Counts are shown for the term "techfest", which is an annual internal technology event at Microsoft, Thus, for this term, the red spikes occur once a year, building up in the email bursts coordinating techfest, then falling off after the event is over.

Narratives also identifies words that closely co-occur with the source term. In Figure 2, we see that "research" and "product" are the most closely associated words with "techfest" on one day of interest, Feb 23, 2007. (Other terms in the list are less significant for this discussion.)

CHANGES IN PERSONAL INFORMATION OVER TIME

The visualization from Narratives inspired us to consider ways that desktop information changes over time. The general research area of topic detection and tracking examines ways to handle text that changes over time.

Kleinberg [9] has looked at temporal clustering, extracting meaningful temporal clusters of topics, including a variety of domains including email. His work reports that personal email seems to cluster well into topics. Kleinberg extracts topics from ACM article titles (finding temporal trends in terms like "web crawl" and "relational database") and from email (where it is useful in identifying deadlines and other bursts of email activity).

Adar et al. [1] looked at how users revisit web pages, and how websites change over time. They found that web pages change frequently and that user behavior around web pages was characteristic both of the users and of the pages. Some pages were revisited often, while users revisited other pages very infrequently. By analogy, we might expect some PIM topics to have characteristic frequent, regular, or irregular patterns.

NewsJunkie [8] attempts to provide a personalized view of incoming news stories. Users can select a series of topics they are interested in, and NewsJunkie identifies novel stories on that general topic, given what the reader already knows.

Personal Search Gadget

The Personal Search Gadget is similar to NewsJunkie, in that it identifies novel material over time. It monitors a user's desktop index, looking for words and people that are important relative to a background model. The comparisons can be relative to the entire index, or to smaller subsets of time such as the previous week.

In Figure 3, we show a screenshot from the Personal Search Gadget. The application is packaged as a small desktop "gadget" for Windows Sidebar, and so is peripherally visible as the user continues with other tasks. In this example, it shows the top four words that are most different today relative to the entire desktop index. In this particular case, the user had been reading about weed killers, and thus both "glyphosate" and "roundup" have come to the top.

The user can dive in deeper to view recent and historical hits for any terms that the system surfaces; a click on the gadget generates a search to find all results for these terms.

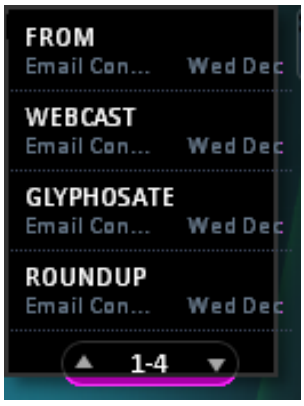

Figure 3: Desktop Search Gadget

CONCLUSIONS

A temporal perspective can be usefully applied to personal information. Indeed, we believe that, a temporal ordering is one of the most meaningful ways to approach personal information.

In this paper, we have reviewed several different directions that PIM research has followed for understanding how personal information evolves over time. We have discussed both a visualization and an adaptive desktop gadget that attempt to surface aspects of this temporal information. Personal Narratives attempts to visually explore the prevalence of topics over time, while the Personal Search Gadget looks for statistically unusual terms that have recently emerged. These, as well as other projects based around our enhanced Windows Desktop Search, will allow us to provide users with tools that allow them to better interact with and understand their personal histories.

REFERENCES

[1] Adar, E., Teevan, J. and Dumais, S. (2008). Large Scale Analysis of Web Revisitation Patterns. In *Proceedings of CHI 2008*, pp. 1197-1206.

[2] Cutrell, E., Robbins, D. C., Dumais, S. T. and Sarin, R. (2006). Fast, Flexible Filtering with Phlat - Personal Search and Organization Made Easy. In *Proceedings of CHI 2006*, pp. 261-270.

[3] Davies, G. and Thomson, D., Eds. (1988). *Memory in Context: Context in Memory*. Wiley: Chichester, England.

[4] Dumais, S. T., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R. and Robbins, D. C. (2003). Stuff I've Seen: A System for Personal Information Retrieval and Re-use. In *Proceedings of SIGIR 2003*, pp. 72-79.

[5] Fertig, S., Freeman, S. and Gelernter, D. (1996). Lifestreams: An Alternative to the Desktop Metaphor. In *Proceedings of CHI 1996*, pp. 410-411.

[6] Fisher, D. and Dourish, P. (2004). Social and Temporal Structures in Everyday Collaboration. In *Proceedings of CHI 2004*, 551-558.

[7] Fisher, D., Hoff, A., Robertson, G. and Hurst, M. (2008). Narratives: A Visualization to Track Narrative Events as They Develop. In *Proceedings of VAST 2008*, 115-122.

[8] E. Gabrilovich, S. Dumais and E. Horvitz (2004). Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty. In *Proceedings of WWW 2004*, pp. 482-490.

[9] Kleinberg, J. (2002). Bursty and Hierarchical Structure in Streams. In *Proceedings of KDD 2002*, 91-101.

[10] Plaisant, C., Milash, B., Rose, A., Widoff, S. and Shneiderman, B. (1996). LifeLines: Visualizing Personal Histories. In *Proceedings of CHI 1996*, pp. 221-227.

[11] Ringel (Morris), M., Cutrell, E., Dumais, S. and Horvitz, E. (2003). Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores. In *Proceedings of Interact 2003*, pp. 184-191.

[12] Viegas, F., Golder, S. and Donath, J. (2006). Visualizing Email Content: Portraying Relationships from Conversational Histories. In *Proceedings of CHI 2006*, pp. 979-988.